The Information Bottleneck Method

by Naftali Tishby, Fernando C. Pereira, and William Bialek

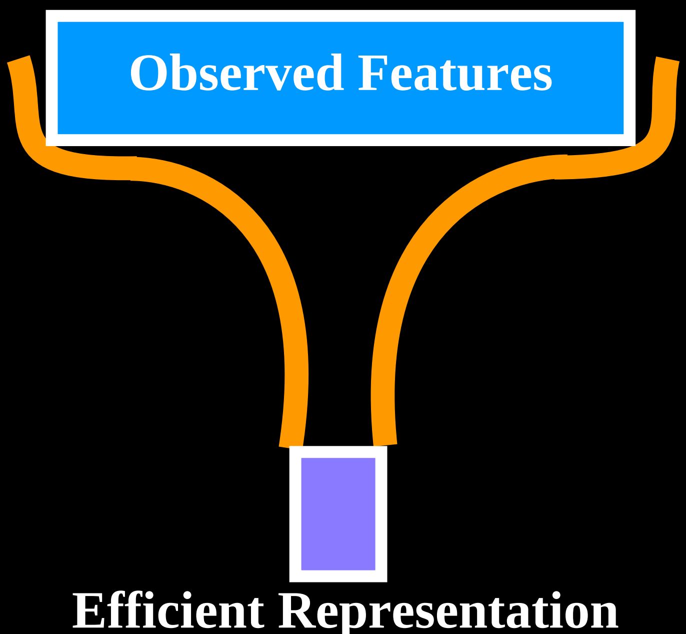
Presented by Kristin Branson

Cog Sci 260: Seminar on Machine Learning

February 11, 2003

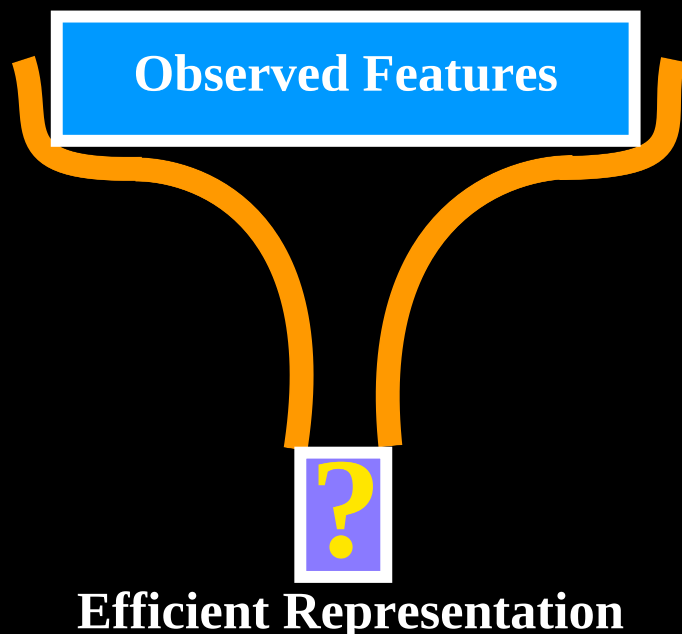
Problem

- How do we extract an efficient representation of the relevant information contained in a large set of features?



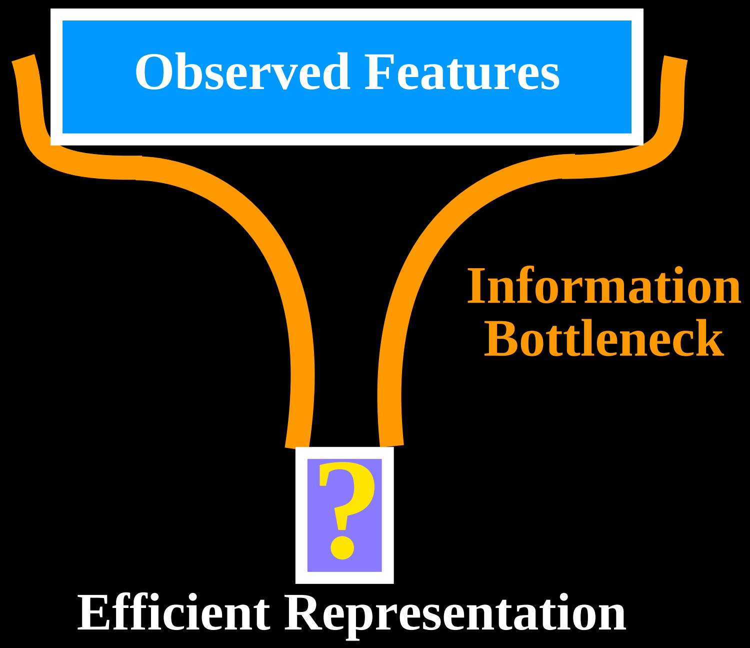
Problem

- How do we extract an efficient representation of the relevant information contained in a large set of features?
- What information is relevant?



Problem

- How do we extract an efficient representation of the relevant information contained in a large set of features?
- What information is relevant?
- The Information Bottleneck Method answers this.



Overview

- The Information Bottleneck Method extends elements of rate distortion theory to supervised information extraction.
- Relevant information is the information in a pattern X useful for predicting a label Y.
- Rate distortion theory is applied to maximize the amount of information about Y retained for a particular length description.

Outline

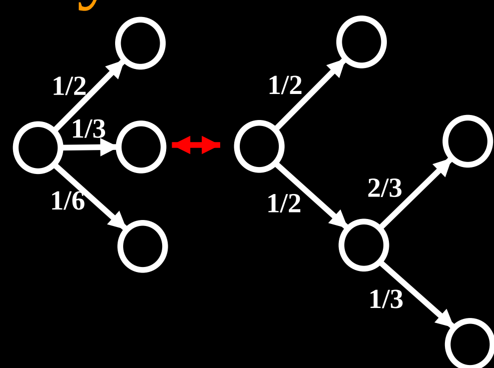
- Information Theory Definitions.
- Rate Distortion Theory Overview.
- Application of rate distortion theory in the Information Bottleneck.
- Practical uses of the information bottleneck.

Information Theory Definitions

- Entropy is the uncertainty of a random variable,

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = -\mathrm{E}_X[\log p(x)].$$

- Interpretation: number of bits needed to represent x.
- This definition falls out of three requirements:
  - $H(p, 1-p)$ is a continuous function of p.
  - $H(p_1, ..., p_n)$ is symmetric w.r.t. its arguments.
  - Grouping:

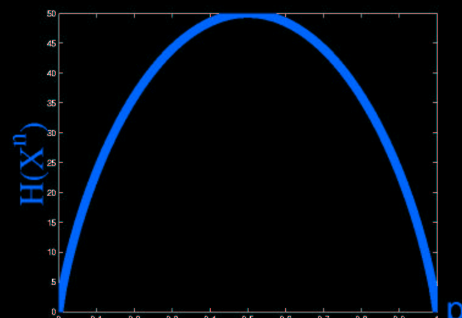


Information Theory Definitions

- Example: If X^n is the outcome of n coin tosses, $P(heads) = p$,

$$\begin{aligned} H(X^n) &= -E_{X_1,\ldots,X_n}[\log P(X_1,\ldots,X_n)] \\ &= -nE_X[\log P(X)] \\ &= -n\left[p\log p + (1-p)\log(1-p)\right] \end{aligned}$$

- If $p = 1/2$, then $H(X^n) = n$ is maximal.
- If $p = 0$, then $H(X^n) = 0$ is minimal.



Information Theory Definitions

- Joint entropy is the entropy of a pair of r.v.s.

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y).$$

- Conditional entropy is the uncertainty of one r.v. given knowledge of the other,

$$H(Y|X) = -\sum_{x\in\mathcal{X}} p(x) \overbrace{\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)}^{H(Y|X=x)}$$

- A good property: $H(X,Y) = H(X) + H(Y|X)$.

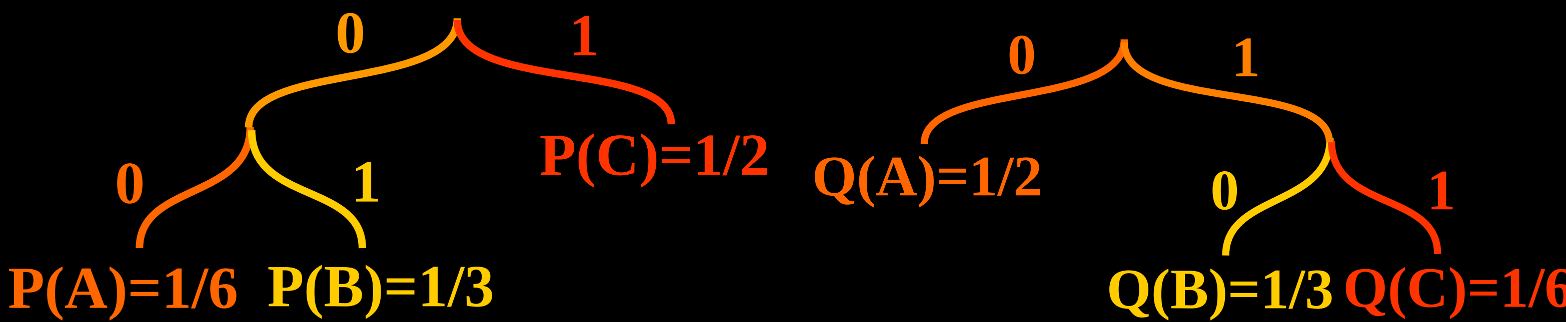
Information Theory Definitions

Relative Entropy (Kullback Liebler distance) is a measure of the inefficiency of assuming distribution q when the true distribution is p,

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Information Theory Definitions

Example: Huffman code of X with true distribution P.



- Coding X using P requires $2P(A) + 2P(B) + P(C) = 1.5$ bits on average.
- Coding X using Q requires $P(A) + 2P(B) + 2P(C) = 1\frac{5}{6}$ bits on average.

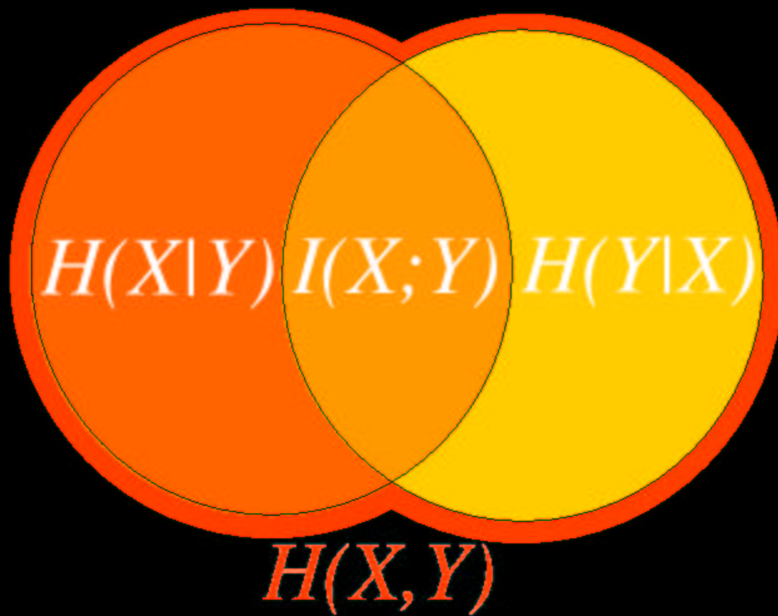
Information Theory Definitions

- Mutual Information is the relative entropy between the joint distribution and the product distribution,

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= H(X) - H(X|Y).
\end{aligned}
$$

- Interpretation: reduction in uncertainty of X due to knowledge of Y.

Information Theory Summary



- Each r.v. has its own uncertainty, $H(X)$ and $H(Y)$.
- The joint entropy is the total entropy of both r.v.s.
- The conditional entropy is that particular to a r.v.
- The shared entropy is the mutual information.

Rate Distortion Theory Idea

- Goal: Determine how well we can represent a r.v. X using a compressed representation $\tilde{X}$.
- "Goodness" is defined as both minimizing the description length of $\tilde{X}$ and a specified measure of distance between X and $\tilde{X}$.

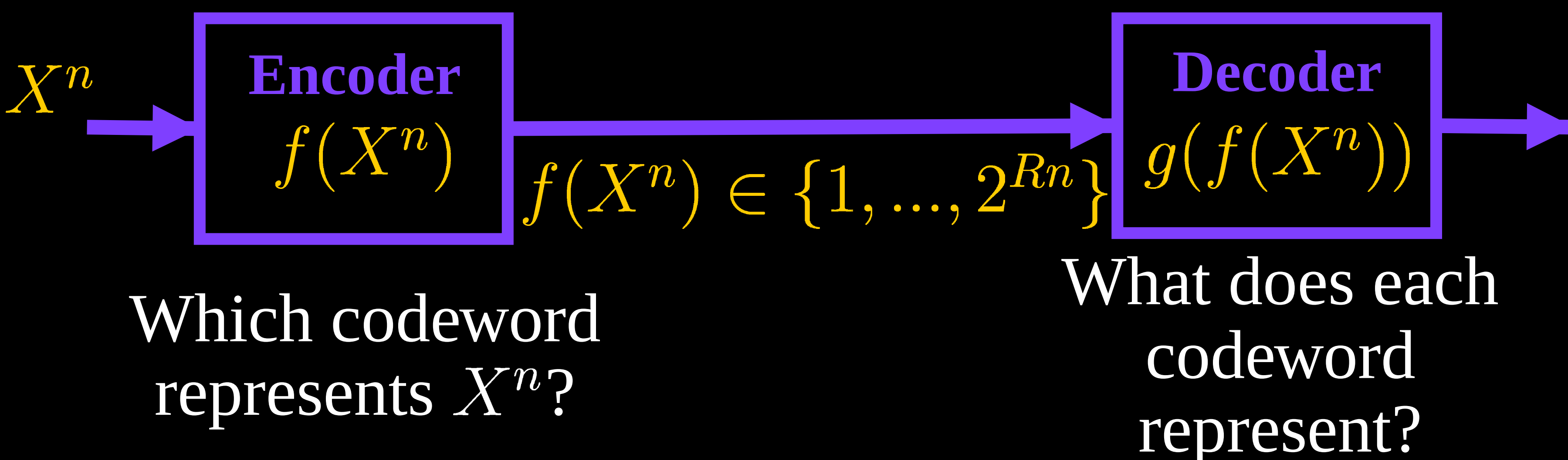
Rate

Distortion $d : \mathcal{X} \times \tilde{\mathcal{X}} \to R^+$

- In other words, we are trying to find a lossy compression of X.

Rate Distortion Theory Definitions

- A $(2^{nR}, n)$ rate distortion code consists of an encoding function, $f_n : \mathcal{X}^n \rightarrow \{1, ..., 2^{nR}\}$, and a decoding function $g_n : \{1, ..., 2^{nR}\} \rightarrow \tilde{\mathcal{X}}^n$.



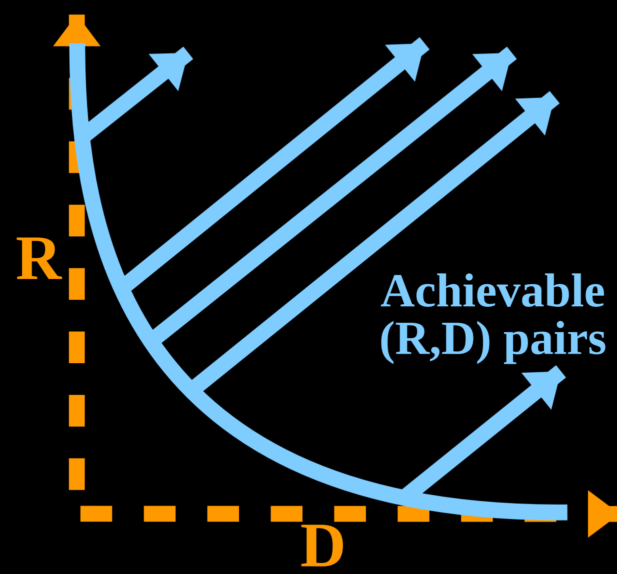
- The distortion of this code is

$$D = E[d(X^n, g_n(f_n(X^n)))].$$

Rate Distortion Theory Definitions

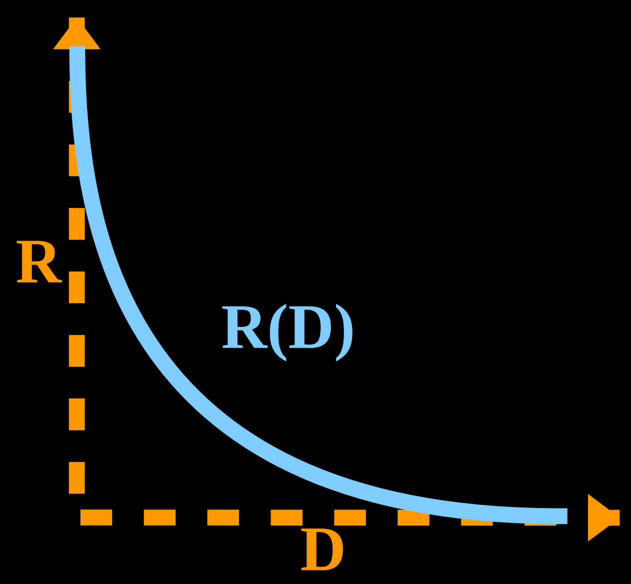
A rate distortion pair (R, D) is achievable if there exists a $(2^{nR}, n)$ rate distortion code (f_n, g_n) s.t.

$$\lim_{n\to\infty} E[d(X^n, g_n(f_n(X^n)))] \leq D.$$



Rate Distortion Theory Definitions

The rate distortion function $R(D)$ is the infimum of rates R s.t. (R, D) is achievable.

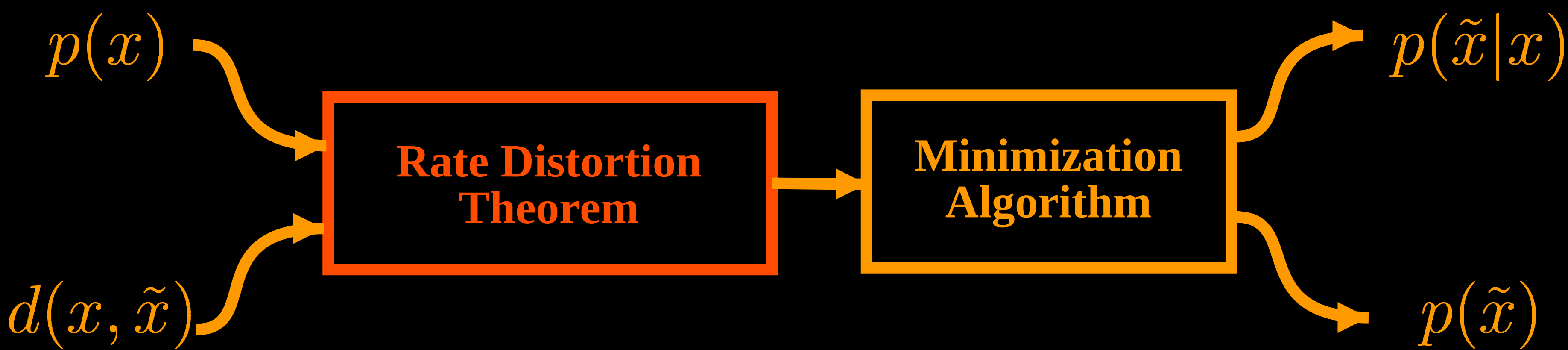


The Rate Distortion Theorem

The rate distortion function for an i.i.d. source X with distribution $p(x)$ and distortion function $d(x,\tilde{x})$ is

$$R(D) = \min_{p(\tilde{x}|x)\in S_D} I(X;\tilde{X}),$$

where $S_D = \{p(\tilde{x}|x) : E_{X,\tilde{X}}[d(X,\tilde{X})] \leq D\}$.



Minimization Algorithm

The distribution $p(\tilde{x}|x)$ that minimizes $I(X;\tilde{X})$ can be calculated using Lagrange multipliers by minimizing

$$\begin{aligned}\mathcal{F}[p(\tilde{x}|x)] &= I(X;\tilde{X}) + \beta E[d(X,\tilde{X})] + \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x) \\ &= \sum_{x,\tilde{x}} p(x,\tilde{x}) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} \\ &+ \beta \sum_{x,\tilde{x}} p(x,\tilde{x}) d(x,\tilde{x}) + \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x)\end{aligned}$$

β can be set to force the distortion below D. Varying β sweeps out the rate distortion function.

Minimization Algorithm

$$
\begin{aligned}
\frac{\partial \mathcal{F}}{\partial p(\tilde{x}|x)} &= \frac{\partial p(x,\tilde{x})}{\partial p(\tilde{x}|x)} \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + \frac{p(x,\tilde{x})}{p(\tilde{x}|x)} \\
&- \sum_{x'} p(x',\tilde{x}) \frac{\partial p(\tilde{x})}{\partial p(\tilde{x}|x)} \frac{1}{p(\tilde{x})} + \beta \frac{\partial p(x,\tilde{x})}{\partial p(\tilde{x}|x)} d(x,\tilde{x}) + \lambda(x) \\
&= p(x) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + \frac{p(x,\tilde{x})}{p(\tilde{x}|x)} - \sum_{x'} p(x',\tilde{x}) \frac{p(x)}{p(\tilde{x})} \\
&\quad + \beta p(x) d(x,\tilde{x}) + \lambda(x) \\
&= p(x) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + p(x) - \frac{p(x)}{p(\tilde{x})} \sum_{x'} p(x',\tilde{x}) \\
&\quad + \beta p(x) d(x,\tilde{x}) + \lambda(x) \\
&= p(x) \left[\log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + 1 - 1 + \beta d(x,\tilde{x}) + \frac{\lambda(x)}{p(x)} \right]
\end{aligned}
$$

Minimization Algorithm

$$
\begin{aligned}
0 &= p(x)\left[\log\frac{p(\tilde{x}|x)}{p(\tilde{x})} + \beta d(x,\tilde{x}) + \frac{\lambda(x)}{p(x)}\right] \\
\log\frac{p(\tilde{x}|x)}{p(\tilde{x})} &= -\beta d(x,\tilde{x}) - \frac{\lambda(x)}{p(x)} \\
p(\tilde{x}|x) &= p(\tilde{x})\exp\left[-\beta d(x,\tilde{x}) - \frac{\lambda(x)}{p(x)}\right] \\
p(\tilde{x}|x) &= p(\tilde{x})\exp\left[-\beta d(x,\tilde{x})\right]\exp\left[-\frac{\lambda(x)}{p(x)}\right] \\
p(\tilde{x}|x) &= \frac{p(\tilde{x})}{Z(x,\beta)}\exp\left[-\beta d(x,\tilde{x})\right]
\end{aligned}
$$

Unfortunately, the equation for $p(\tilde{x}|x)$ depends on the unknown $p(\tilde{x})$.

Minimization Algorithm

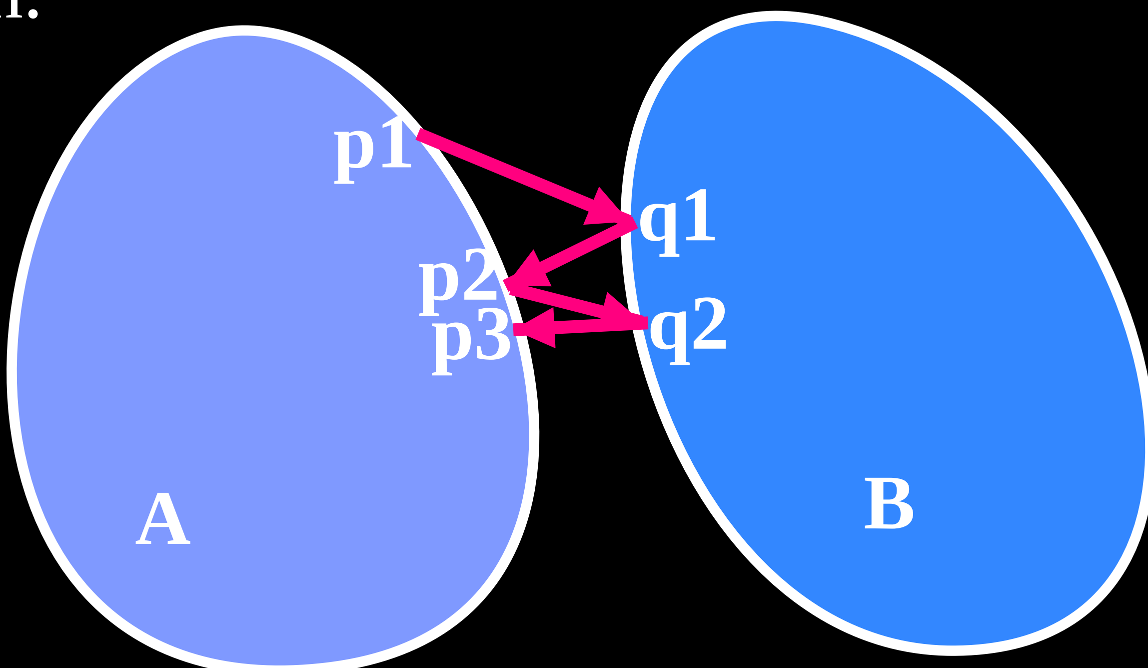
We can rewrite $R(D)$ as follows:

$$
\begin{aligned}
R(D) &= \min_{p(\tilde{x}|x)\in S_D, p(\tilde{x})\in A'_D} I(X;\tilde{X}) \\
&= \min_{p(\tilde{x}|x)\in S_D, p(\tilde{x})\in A'_D} D(p(x)p(\tilde{x}|x)||p(x)p(\tilde{x})) \\
&= \min_{p(x)p(\tilde{x}|x)\in B_D, p(x)p(\tilde{x})\in A_D} D(p(x)p(\tilde{x}|x)||p(x)p(\tilde{x}))
\end{aligned}
$$

This is the distance between two sets of probability distributions, A_D and B_D.

Minimization Algorithm

- We can iteratively estimate $p(\tilde{x}|x)$ and $p(\tilde{x})$ that minimize $D(p(x)q(\tilde{x}|x)||p(x)r(\tilde{x}))$ (Blahut-Arimoto algorithm).
- This is guaranteed to converge to the optimal solution.



Minimization Algorithm

- The distribution $p(\tilde{x})$ that minimizes $I(X;\tilde{X})$ is

$$r^*(\tilde{x}) = \sum_x p(x)p(\tilde{x}|x).$$

- This can be shown by proving

$$D(r^*||r) =$$

$$D(p(x,\tilde{x})||r(\tilde{x})p(x)) - D(p(x,\tilde{x})||r^*(\tilde{x})p(x))$$

and therefore is non-negative for all $r(\tilde{x})$.

Minimization Summary

- We cannot directly solve for $p(\tilde{x}|x)$ using Lagrange multipliers because our result depends on $p(\tilde{x})$.
- We convert the problem of minimizing $I(X;\tilde{X})$ to that of minimizing the KL distance between two sets of probability distributions.
- This problem can be solved by iteratively minimizing the distance w.r.t. $p(\tilde{x}|x)$ and w.r.t. $p(\tilde{x})$.
- There are closed form solutions to each of these minimization problems.
- Note: we have found the partitioning function $f(x) = p(\tilde{x}|x)$, not the compressed representation, $g(f(x))$.

Rate Distortion Theory Summary

- We have found a procedure for computing the optimal partitioning of X into codewords $\tilde{X}$, $p(\tilde{x}|x)$.
- The solution depends on the choice of the distortion measure, $d(x, \tilde{x})$.
- Directly choosing a distortion function is equivalent to specifying what features are relevant.
- The main idea of the Information Bottleneck Method is to use a supervised definition of relevance, the features of X relevant for predicting another variable Y.

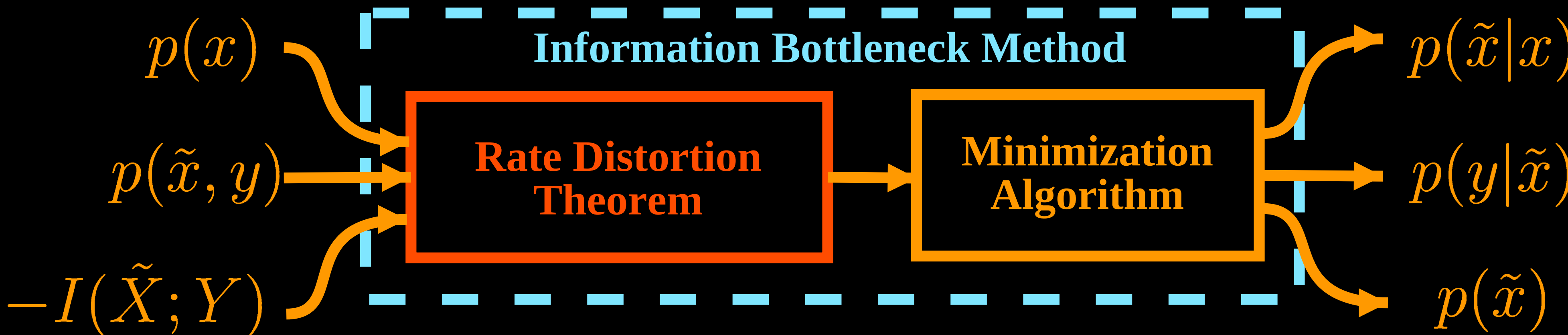
The Information Bottleneck Idea

- Goal: Determine how well we can represent a r.v. X using a compressed representation $\tilde{X}$.
- "Goodness" is defined as minimizing the rate and maximizing the information captured about the relevance variable, Y.
- The amount of information about Y in the compressed representation $\tilde{X}$ is the mutual information,

$$I(\tilde{X};Y) = \sum_{y,\tilde{x}} p(y,\tilde{x}) \log \frac{p(y,\tilde{x})}{p(y)p(x)}.$$

The Information Bottleneck Idea

The Information Bottleneck Method is an extension of the Rate Distortion Theorem using a supervised definition of relevance.



Minimization Algorithm

The distribution $p(\tilde{x}|x)$ that minimizes $I(X;\tilde{X})$ can be calculated using Lagrange multipliers by minimizing

$$\begin{aligned}\mathcal{F}[p(\tilde{x}|x)] &= I(X;\tilde{X}) - \beta I(\tilde{X},Y) - \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x) \\ &= \sum_{x,\tilde{x}} p(x,\tilde{x}) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} \\ &- \beta \sum_{\tilde{x},y} p(\tilde{x},y) \log \frac{p(\tilde{x},y)}{p(\tilde{x})p(y)} + \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x)\end{aligned}$$

Note: We specify $|\tilde{\mathcal{X}}|$ instead of a maximum distortion.

Minimization Algorithm

Computing the partial of $\mathcal{F}$ w.r.t. $p(\tilde{x}|x)$ yields

$$\frac{\partial \mathcal{F}}{\partial p(\tilde{x}|x)} = p(x)\left[\log\frac{p(\tilde{x}|x)}{p(\tilde{x})} - \beta\sum_y p(y|x)\log\frac{p(y|\tilde{x})}{p(y)} - \frac{\lambda(x)}{p(y)}\right]$$

Setting this equal to 0 and solving for $p(\tilde{x}|x)$ yields

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)}\exp(-\beta D(p(y|x)||p(y|\tilde{x}))).$$

As before, this solution depends on unknown distributions, $p(\tilde{x})$ and $p(y|\tilde{x})$.

Minimization Algorithm

- We can iteratively estimate $p(\tilde{x}|x)$, $p(\tilde{x})$, and $p(y|\tilde{x})$ that minimize $\mathcal{F}[p(\tilde{x}|x), p(\tilde{x}), p(y|\tilde{x})]$.
- This will converge to an optimal solution.
- The distribution $p(\tilde{x})$ that minimizes $\mathcal{F}$ is

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

- The distribution $p(y|\tilde{x})$ that minimizes $\mathcal{F}$ is

$$p(y|\tilde{x}) = \sum_y p(y|x)p(x|\tilde{x}).$$

Information Bottleneck Comparison

Compare the partitioning functions

$$p_{IB}(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp(-\beta D(p(y|x)||p(y|\tilde{x}))),$$

$$p_{RD}(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp(-\beta d(x,\tilde{x})).$$

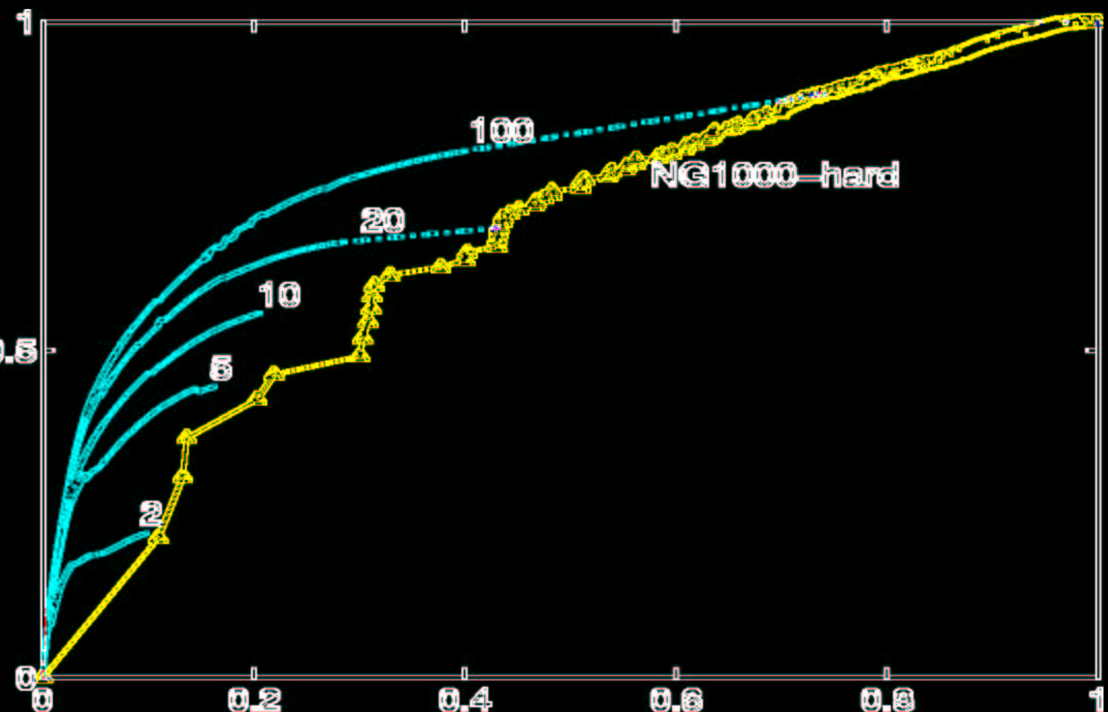
The KL distance emerged as the relevant effective distortion measure, even though it was not assumed.

Information Bottleneck Comparison

Not only have we found the encoding function $f(x) = p(\tilde{x}|x)$, but we have found the decoding function $g(\tilde{x}) = p(y|\tilde{x})$ (given $|\tilde{\mathcal{X}}|$).

The Information Plane

- How do we choose β and $|\tilde{\mathcal{X}}|$?
- For each value of $|\tilde{\mathcal{X}}|$, different values of β sweep out a curve in the Information Plane.



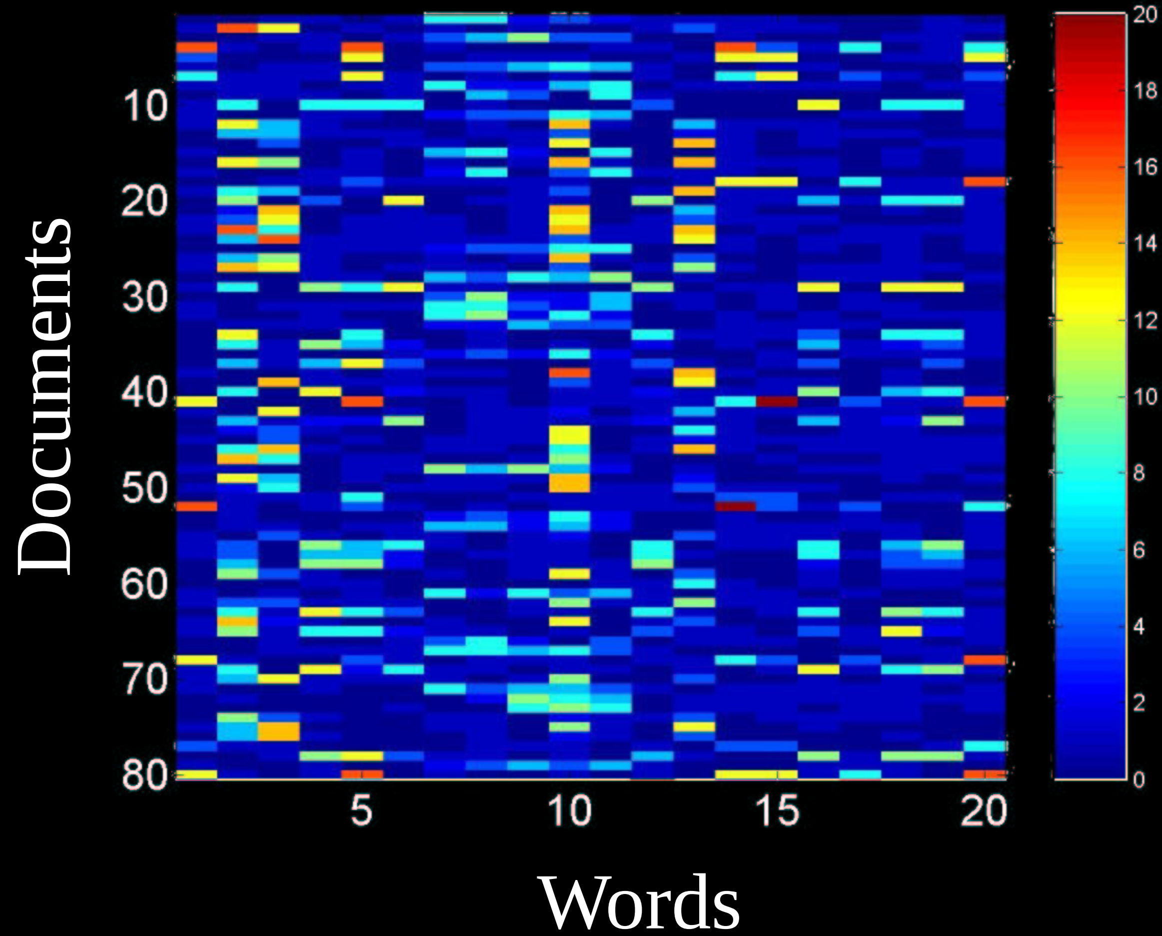
$I(\tilde{X};Y)/I(X;Y)$

$I(\tilde{X};X)/H(X)$

Practical Uses

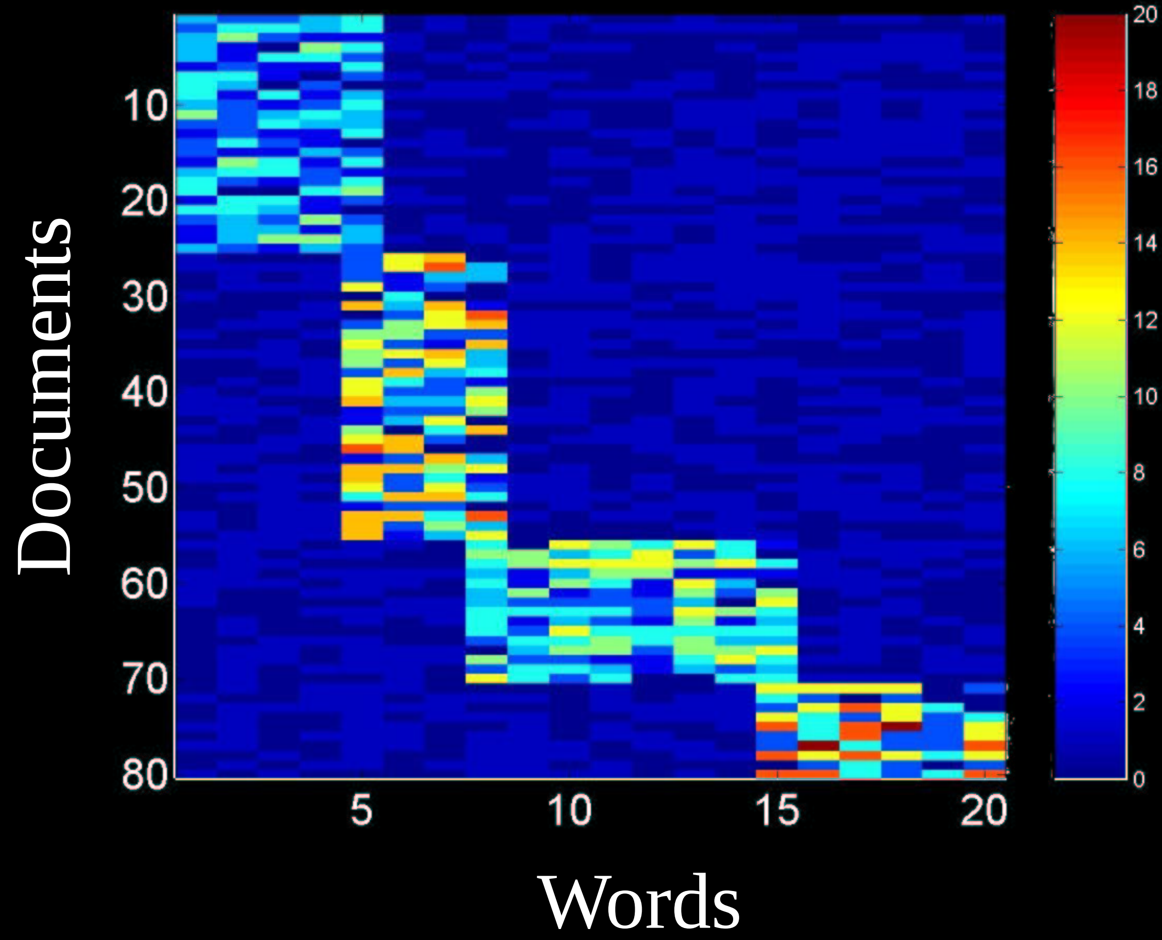
- We have a stochastic mapping of the observed patterns to an efficient representation.
- How is this useful?
- Deterministic Annealing can be used to cluster the features.
- The Agglomerative Information Bottleneck finds a greedy hierarchical clustering of the features for $\beta \to \infty$.

Document Clustering Application



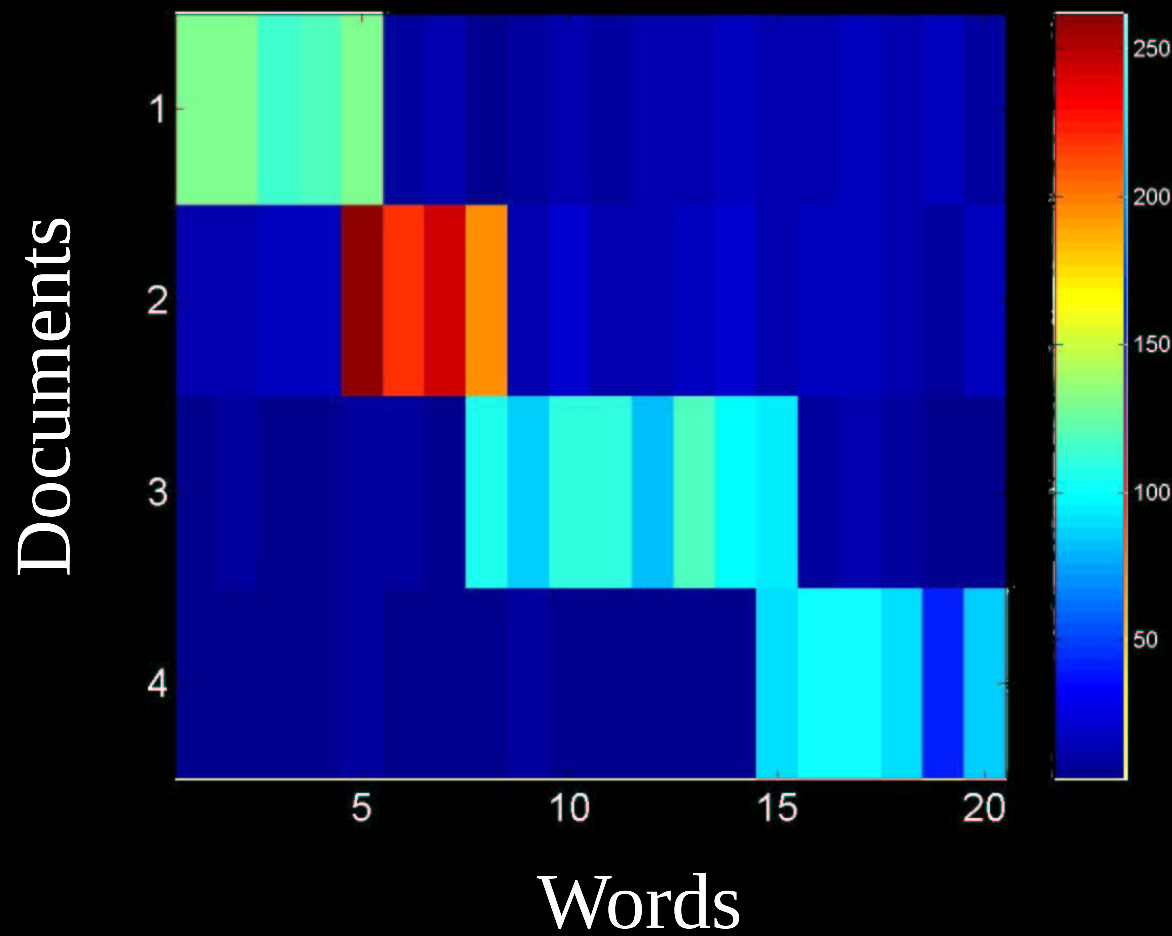
Documents-words counts matrix.

Document Clustering Application



Permuted documents-words counts matrix.

Document Clustering Application

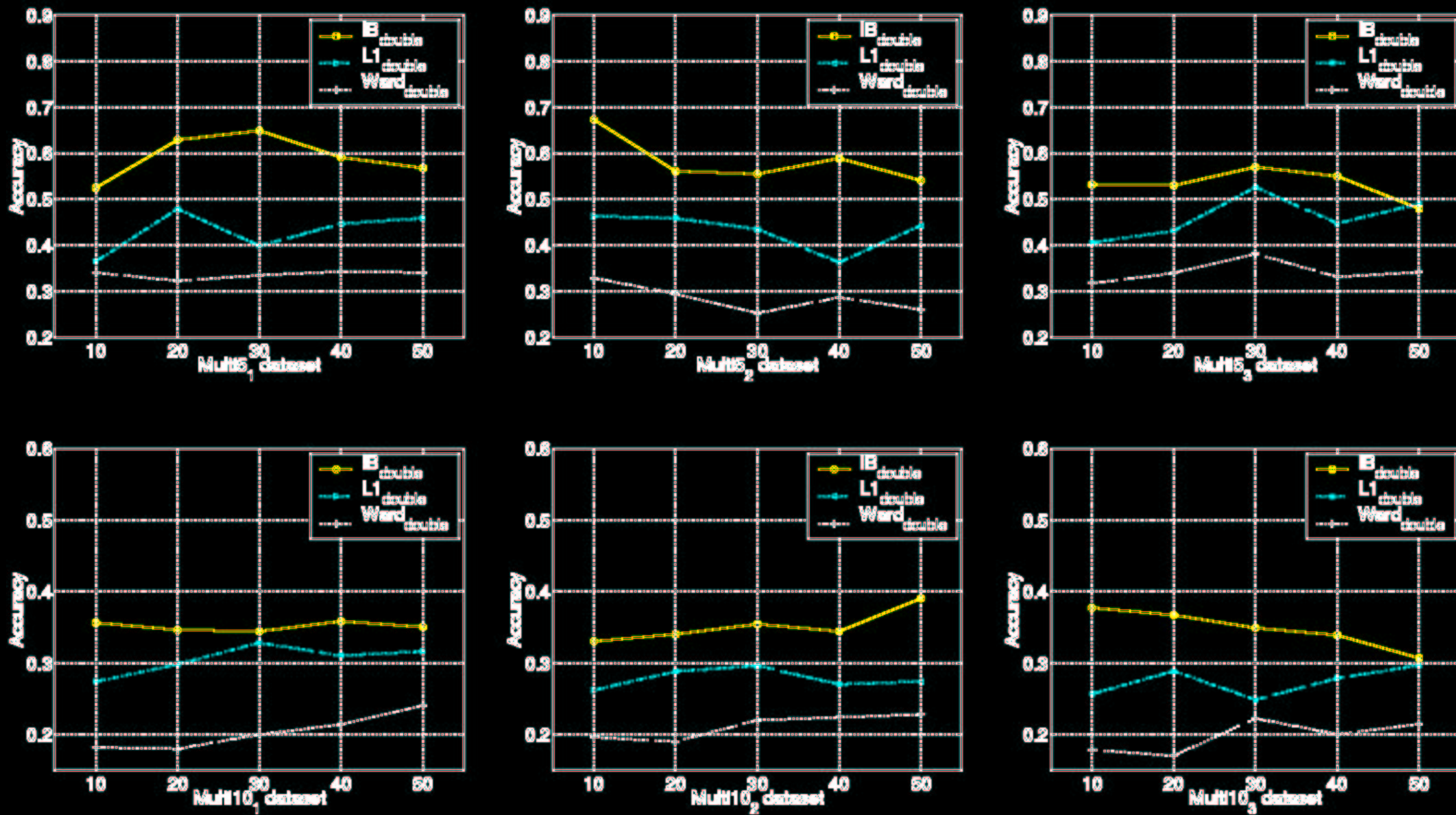


Document Clusters.

Document Clustering Application

- The Information Bottleneck Method was applied to document clustering.
- First, a partitioning of the words, $p(\tilde{w}|w)$, is found that preserves information about the documents, D.
- Second, the original document representation is replaced by a representation based on the word-clusters.
- Finally, a partitioning of the documents, $p(\tilde{d}|d)$, is found that preserves information about the words, W.

Document Clustering Results



Information Bottleneck Summary

- The Information Bottleneck Method extends elements of rate distortion theory to supervised information extraction.
- Relevant information is defined as the information in X useful for predicting Y.
- A guaranteed iterative minimization algorithm is applied to find the partitioning of X into $\tilde{X}$, $p(\tilde{x}|x)$.
- The solution is equivalent to using the KL distance $D(p(y|x)||p(y|\tilde{x}))$ for a distortion measure.

References

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley Interscience, New York.

Slonim, N. and Tishby, N. (1999). Agglomerative information bottleneck.

Slonim, N. and Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In *Research and Development in Information Retrieval*, pages 208–215.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.