Information and Neural Coding:
Can we Quantify Relevance?

Naftali Tishby

Center for Neural Computation
The Hebrew University, Jerusalem, Israel
http://www.cs.huji.ac.il/~tishby
ITP-UCSB, Sep. 2001

Many thanks to:

Bill Bialek
Fernando Pereira

Noam Slonim
Gal Chechik
Amir Globerson
Nir Friedman
Motivation:

• Can we quantify relevant information?

• What is encoded by a neural code?

• How to analyze the response of complex cortical cells?

• A general design principle?
Information Theory and Neural Coding: Can We Quantify Relevant Information?

**Analysis of neural codes.**
(With Rob de Ruyter, W. Bialek, E. Schniedman, N. Brenner)

**Prelude:**
Organization of the auditory system

**Visual system:**
- Photoreceptors
- Retinal ganglion cells
- LGN
- V1
- IT
- Face cells

**Auditory system:**
- Hair cells
- Auditory nerve fibres
- Cochlear nucleus
- Superior Olive
- Medial Geniculate Body
- Auditory cortex
- Inferior Colliculus
- Species-specific calls?
- Auditory scene analysis?

**Stimulus**

**Spike trains**

**Word distribution** @ 1

**Total word distribution**

**Frequency**

**Localization and binaural detection**
Motivating case study:
Feature detectors of complex sounds

(Eli Nelken)

Separating a natural sound into its components
Information Theory and Neural Coding: Can We Quantify Relevant Information?

Neural responses in: Auditory Cortex (B), MGB (C), and IC (D)
to the same four acoustic stimuli (A).

**Mutual information**

*How much X is telling about Y?*

$I(X; Y)$: function of the joint probability distribution $p(x, y)$ —
minimal number of yes/no questions (bits) needed to ask about $x$ in
order to learn all we can about $y$.

Uncertainty removed about $X$ when we know $Y$:

$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
**Relevant Coding**

- What are the questions that we need to ask about $X$ in order to learn about $Y$?
- Need to partition $X$ into relevant domains, or clusters, between which we really need to distinguish...

**Bottlenecks and Neural Nets**

- Auto association: forcing compact representations
- $\hat{X}$ is a relevant code of $X$ w.r.t. $Y$
### A Simple Example...

<table>
<thead>
<tr>
<th></th>
<th>Israel</th>
<th>Health</th>
<th>www</th>
<th>Drug</th>
<th>Jewish</th>
<th>Dos</th>
<th>Doctor</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>Doc3</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Doc4</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Doc5</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc6</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Doc7</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Doc8</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc9</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Doc10</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

...  ...  ...  ...  ...  ...  ...  ...

### Simple Example

<table>
<thead>
<tr>
<th></th>
<th>Israel</th>
<th>Jewish</th>
<th>Health</th>
<th>Drug</th>
<th>Doctor</th>
<th>www</th>
<th>Dos</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc3</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc4</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc5</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc6</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc7</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>16</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Doc9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Doc10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>

...  ...  ...  ...  ...  ...  ...  ...
A new compact representation

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>36</th>
<th>25</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2</td>
<td>1</td>
<td>1</td>
<td>42</td>
<td>39</td>
<td>45</td>
<td>4</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>26</td>
<td>33</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The document clusters preserve the relevant information between the documents and words
• Q: How many bits are needed to determine the relevant representation?
  – need to index the max number of non-overlapping green blobs inside the blue blob:
    (mutual information!)

\[
2^{H(X | \hat{X})} \leq 2^{H(X)} p(\hat{x} | x) \hat{X}
\]

\[
2^{H(X)} / 2^{H(X | \hat{X})} \equiv 2^I(X, \hat{X})
\]

Information Bottleneck

• The distortion function determines the relevant part of the pattern…
• but what if we don’t know the distortion function but rather a relevance variable?

Examples:
  Speech vs its transcription
  Images vs objects-names
  Faces vs expressions
  Stimuli vs spike-trains (neural codes)
  Protein-sequences vs structure/function
  Documents vs text-categories
  Input vs responses
  etc…
• The idea: find a compressed signal $\hat{X}$ that needs short encoding (small $I(X, \hat{X})$) while preserving as much as possible the information on the relevant signal ($I(\hat{X}, Y)$).

A Variational Principle
We want a short representation of $X$ that keeps the information about another variable, $Y$, if possible.
The **Self Consistent Equations**

- **Marginal:** 
  \[ p(\hat{x}) = \sum_x p(\hat{x} \mid x) p(x) \]

- **Markov condition:** 
  \[ p(y \mid \hat{x}) = \sum_x p(y \mid x) p(x \mid \hat{x}) \]

- **Bayes’ rule:** 
  \[ p(x \mid \hat{x}) = \frac{p(x)}{p(\hat{x})} p(\hat{x} \mid x) \]

\[
\frac{\delta L[p(\hat{x}|x)]}{\delta p(\hat{x}|x)} = 0 \quad \Rightarrow \quad p(\hat{x}|x) = \frac{p(\hat{x})}{Z(x, \beta)} \exp(-\beta D_{KL}[x, \hat{x}])
\]

---

The emerged **effective distortion** measure:

\[
D_{KL}(x, \hat{x}) = D_{KL}[p(y \mid x) \mid p(y \mid \hat{x})]
\]

\[
= \sum_y p(y \mid x) \log \frac{p(y \mid x)}{p(y \mid \hat{x})}
\]

- Regular if \( p(y \mid \hat{x}) \) is absolutely continuous w.r.t. \( p(y \mid x) \)
- Small if \( \hat{x} \) predicts \( y \) as well as \( x \):

```
x -------- p(y|x) ----> y
      ↓                   ↓
p(\hat{x} \mid x)       p(y|\hat{x})
\hat{x} -------- p(y|\hat{x}) ----> y
```
The iterative algorithm: (Generalized Blahut-Arimoto)

The **Information Bottleneck** Algorithm

\[
\min_p \min_{\hat{x}} \min_{\hat{y}} \langle \log Z(x, \beta) \rangle = \text{“free energy”}
\]

\[
\min_p I(X, \hat{X}) + \beta \left\langle D_{KL}(x, \hat{x}) \right\rangle
\]

\[
\begin{align*}
    p_{t+1}(\hat{x} \mid x) &= \frac{p_t(\hat{x})}{Z_t(x, \beta)} \exp\left(-\beta D'_{KL}(x, \hat{x})\right) \\
p_t(\hat{x}) &= \sum_x p(x) p_t(\hat{x} \mid x) \\
p_t(y \mid \hat{x}) &= \sum_x p(y \mid x) p_t(x \mid \hat{x})
\end{align*}
\]
Information Theory and Neural Coding: Can We Quantify Relevant Information?

- The *information - plane*, the optimal $I(\hat{X}, Y)$ for a given $I(\hat{X}, X)$ is a concave function:

\[ \frac{\delta I(Y, \hat{X})}{\delta I(X, \hat{X})} = \beta^{-1} \]

![Diagram showing the information plane and possible phases](image)

Document classification - information curves

![Graph showing document classification information curves](image)
Information Theory and Neural Coding: Can We Quantify Relevant Information?

---

**Analysis of neural codes.**
(w Rob de Ruyter, W Bialek, E Schniedman, N Brenner)

---

**Stimulus**

**Spike trains**

**Trial**

**Total word distribution**

---

**Graphs**

- **Probability**
  - Binary Words (ordered by # of spikes)
  - Information (clusters responses)
  - # of clusters

---

Dr. Naftali Tishby, Hebrew University (ITP 9/18/01)
Neural responses in: Auditory Cortex (B), MGB (C), and IC (D) to the same four acoustic stimuli (A).
Naive approach: Spike-triggered averaging (reverse correlation)

- Taking windows of the spectrogram ("frames"), triggered on spikes (or other events).

Figure 3: A-C examples of frames. D. MI single spikes provide about short acoustic segments as a function of number of neurons. (IC-blue, MGB-magenta, A1-red). Analysis was repeated for several groups of cells of each set size. Error bars designate the standard error of the mean MI over these sets. E. Data rescaled with regard to maximal MI level. For each region, the level of MI obtained when linearly adding single cells was calculated (black line). Data from all regions is then rescaled with respect to this single-cell baseline.
Clustering was performed by normalizing each frame to a sum of 1, and applying IB clustering. Each figure represents the center (average) of a cluster of frames. (unit6712)

The joint probability matrix

- Before IB compression
- After IB compression
An independent matrix was created from our S&R marginal probabilities, with a true MI of zero., the bias is estimated both numerically and by first two terms of Panzeri expansion, yielding two-fold difference.
Bimanual motor task

Test trials

Mutual information between stimulus velocity and neural HMM States

-500 -400 -300 -200 -100 0 100 200 300 400 500 600 Shift (ms)

-0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 (bit/s)

Dr. Naftali Tishby, Hebrew University (ITP 9/18/01)
Information Theory and Neural Coding: Can We Quantify Relevant Information?

Figure 3: A. Original conditional distribution $p(N|M=m)$. Each row is a conditional distribution for a different value of $M$. B. A reordering of the lines in A, so that rows between two black lines fall into the same cluster, in the $F$ mapping.

Figure 4: Upper row: Clusters of movements generated by the AIB algorithm. Each
Regression as relevant encoding

- Extracting relevant information is fundamental for many problems in learning:
  \[
  y_i = f(x_i, \theta) + \xi_i
  \]

  \[
  X = \{x^{(n)} | (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}
  \]

  \[
  Y = \emptyset = \{\theta\}
  \]

  \[
  p(x^{(n)}, \theta) = p(\theta) \prod_k p(x_k) \exp \left[-\frac{1}{2\sigma^2} \sum_k \left| y_k - f(x_k, \theta) \right|^2 \right]
  \]

- Knowing the parametric class we can calculate \( p(X, Y) \), without sampling!

Manifold of relevance

The self consistent equations:

\[
\log p(x | \hat{x}) = \log p(x) - \log Z(x, \beta) - \beta D(x | \hat{x})
\]

\[
\log p(y | \hat{x}) = \log \sum_x p(y | x) p(x | \hat{x})
\]

Assuming a continuous manifold for \( \hat{x} \)

\[
\begin{align*}
\frac{\delta \log p(x | \hat{x})}{\delta \hat{x}} &= \beta M_x[\hat{x}] \frac{\delta \log p(y | \hat{x})}{\delta \hat{x}} \\
\frac{\delta \log p(y | \hat{x})}{\delta \hat{x}} &= M_y[\hat{x}] \frac{\delta \log p(x | \hat{x})}{\delta \hat{x}}
\end{align*}
\]

Coupled (local in \( \hat{x} \)) eigenfunction equations, with \( \beta \) as an eigenvalue.
**Generalization** as relevant encoding

The *two sample* problem:
the probability that two samples come from one
source \( y_i = f(x_i, \theta) + \xi_i \)

\[
X = \{ s_1^{(n)} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \}
\]

\[
Y = \{ s_2^{(n)} = \{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2}), ..., (x_{2n}, y_{2n})\} \}
\]

\[
p(s_1^{(n)}, s_2^{(n)}) = \int d\theta p(\theta) \prod_{k=1}^{2n} p(x_k) \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{2n} \left| y_n - f(x_n, \theta) \right|^2 \right]
\]

Knowing the function class we can estimate \( p(X, Y) \).  
Convergence depends on the class complexity.

---

**Multivariate Information Bottleneck**

- Complex relationship between many variables
- Multiple unrelated dimensionality reduction schemes
- Trade between known and desired dependencies
- Express IB in the language of Graphical Models
- Multivariate extension of Rate-Distortion Theory
Information Theory and Neural Coding: Can We Quantify Relevant Information?

Generalized Bottleneck: Extending the dependency graph (Nir Friedman)

\[ L[p_1(\bar{T} | \bar{x}), p_2(\bar{x} | \bar{T}), p(\bar{T})] = \tilde{I}^{G_1}(\bar{X}, \bar{T}) - \beta \tilde{I}^{G_2}(\bar{T}, \bar{Y}) \]

\[ \tilde{I}(X_1, X_2, \ldots, X_n) = \sum_i p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{\prod_i p(x_i)} \]

(Multi-information)

Figure 2: A DAG representing multiple independent variables that summarize the connection between \(X\) and \(Y\).

Multi-information

Multi-information

\[ I(X_1, \ldots, X_n) = E[\log \frac{p(X_1, \ldots, X_n)}{p(X_1) \ldots p(X_n)}] \]

- Information random variables jointly contain about each other
- Generalizes mutual information
Conclusions

There may be a common principle behind...

- Noise filtering
- Time series prediction
- Categorization and classification
- Feature extraction
- Supervised and unsupervised learning
- Visual and auditory segmentation
- Clustering
- Self organized representation
- ...

Summary

- We present a general information theoretic approach for extracting relevant information.
- It is a natural generalization of Rate-Distortion theory with similar converging and optimality proofs.
- Unifies learning, feature extraction, filtering, and prediction...
- Applications (so far) include:
  - Word sense disambiguation
  - Document classification and categorization
  - Spectral analysis
  - Neural codes
  - Bioinformatics, ...
  - Data clustering based on multi-distance distributions