Bits of Meaning

or

Efficient Data Representations

that Preserve Information

Naftali Tishby

School of Computer Science and Engineering

and Center for Neural Computation

The Hebrew University, Jerusalem, Israel

http://www.cs.huji.ac.il/~tishby
ALT & DS Conference October 2003
Sapporo, Japan

Thanks to the organizers

http://www.cs.huji.ac.il/labs/learning/Papers/IBM_list.html
◆ 23 Faculty members (from 7 departments; 3 campuses)

◆ 60 Ph.D. Students! (Largest Ph.D. Program in Neural Computation World-Wide)

◆ Wide, deep and highly demanding curriculum
Outline:

The fundamental dilemma: Simplicity/Complexity versus Accuracy

- What can be learnt – extracting relevant information?
- Can we quantify it?
- The “right level” of description

◆ A Variational Principle
  - Lessons from Statistical Physics and Information Theory
  - [Mutual] Information as the Fundamental Function
  - Information Bottleneck (IB) and Efficient Representations

◆ Applications and Extensions
  - Words, documents and meaning…
  - Learning from the Fly …
  - IB and Graphical Models
  - Sufficient Dimensionality Reduction (SDR)

◆ Conclusions: The importance of being principled…
Learning Theory: The fundamental dilemma...

Good models should enable prediction of new data...

Tradeoff between accuracy and simplicity
Or in unsupervised learning ...
Lessons from Statistical Physics

Physical systems with many degrees of freedom:
What is the right level of description?

- “Microscopic Variables” – dynamical variables
  (positions, momenta, …)
- “Variational Functions” – Thermodynamic potentials
  (Energy, Entropy, Free-energy, …)
- Competition between order (energy) and disorder (entropy) …

Emergent “relevant” description ➞ Order Parameters
  (magnetization, …)
Statistical Models in Learning and AI

Statistical models of complex systems:
What is the right level of description?

- "Microscopic Variables" – probability distributions
  (over the observed and hidden variables)
- "Variational Functions" – Log-likelihood, Entropy,…?
  (Multi-Information, “Free-energy”,…)

- Competition between accuracy and complexity

Emerged “relevant” description → Features
(clusters, sufficient statistics,…)
The Fundamental Dilemma (of science): Model Complexity vs Prediction Accuracy
Can we quantify it...?

When there is a *(relevant)* prediction or *distortion measure*

**Accuracy** ⇔ *small average distortion* *(good predictions)*

**Complexity** ⇔ *short description* *(high compression)*

A general tradeoff between distortion and compression:

Shannon’s Information Theory
Shannon’s Information Theory

- Source
- Channel
- Receiver
- Sent messages
- Received messages
- Symbols
- Source Entropy
- Channel Capacity
- Decoding
- Error Correction
- Decompression
- Compression
- Rate vs Distortion
- Capacity vs Efficiency
Shannon's axioms for Entropy ("Missing Information")

- **S1**: For a random variable $X$, its “missing information”, $S[X]$, is a continuous functional of its probability distribution $p(x)$.

- **S2**: when the variable has $K$ equally probable values $p(x)=1/K$, then $S[X]$ is a monotonic function of $K$.

- **S3**: If $X$ is partitioned into two subsets, $X_1$ and $X_2$, then the amount of shared information is additive in the following sense:
  \[ S[X] = S[(X_1,X_2)] + p(x_1)S[X_1] + p(x_2)S[X_2] \]

**Theorem (Shannon 48):**

\[
S[X] = - \sum_{x \in X} p(x) \log p(x)
\]
**Axioms for “shared network information”**

- **M1**: For two nodes, X and Y, the shared information, $I[X;Y]$, is a continuous function of $p(x,y)$.

- **M2**: when one of the variables defines the other uniquely (1-1 mapping from X to Y) and both $p(x)=p(y)=1/K$, then $I[X;Y]$ is a monotonic function of K.

- **M3**: If X is partitioned into two subsets, X1 and X2, then the amount of shared information is additive in the following sense:

  $$I[X;Y] = I[(X_1,X_2);Y] + p(x_1)I[X;Y|X_1] + p(x_2)I[X;Y|X_2]$$

- **M4**: Shared information is symmetric: $I[X;Y] = I[Y;X]$.

- **M5**: $I[X_1;X_2;...;X_{N+1}] = I[X_1;X_2;...;X_N] + I[(X_1,X_2,...,X_N);X_{N+1}]$

---

**Theorem:**

$$I[X_1;X_2;...;X_N] = \sum_{x_1,x_2,...,x_N} p(x_1,x_2,...,x_N) \log \frac{p(x_1,x_2,...,x_N)}{p(x_1)p(x_2)\cdots p(x_N)}$$
We need to index the max number of non-overlapping green blobs inside the blue blob:

(mutual information!)

\[
2^{nH(X)} \quad / \quad 2^{nH(X|\hat{X})} \quad \equiv \quad 2^{nI(X,\hat{X})}
\]
The Dual Variational problems of IT: 
Rate vs Distortion and Capacity vs Efficiency

Q: What is the simplest representation – fewer bits/sec (Rate) for a given expected distortion (accuracy)?

A: \((RDT, \text{Shannon 1960})\) solve:

\[
R(D) = \min_{\langle d \rangle \leq D} I(X, \hat{X})
\]

Q: What is the maximum number of bits/sec that can be sent reliably (prediction rate) for a given efficiency of the channel (power, cost)?

A: \((\text{Capacity-power tradeoff, Shannon 1956})\) solve:

\[
C(E) = \max_{\langle e \rangle \leq E} I(\hat{X}, Y)
\]
Rate-Distortion theory

The tradeoff between expected representation size (Rate) and the expected distortion is expressed by the rate-distortion function:

\[ R(D) = \min_{\langle d \rangle \leq D} I(X, \hat{X}) \]

By introducing a Lagrange multiplier, \( \beta \), we have a variational principle:

\[ L[p(\hat{x} | x)] = I(X, \hat{X}) + \beta \langle d(x, \hat{x}) \rangle_{p(x, \hat{x})} \]

with:

\[ \langle d(x, \hat{x}) \rangle_{p(x, \hat{x})} = \sum_{x, \hat{x}} p(x)p(\hat{x} | x)d(x, \hat{x}) \]
The *Rate-Distortion function*, \( R(D) \), is the *optimal* rate for a given distortion and is a convex function.
The Capacity - Efficiency Tradeoff

The graph illustrates the relationship between capacity and efficiency. The x-axis represents efficiency (power) and the y-axis represents capacity at E (bits/erg). The achievable region is highlighted on the graph.
Double matching of source to channel eliminates the need for coding (?!)

\[ R(D) = \frac{1}{2} \log \frac{\sigma^2}{D} \]

\[ C(E) = \frac{1}{2} \log \left( 1 + \frac{E}{\sigma^2} \right) \]

In the case of a Gaussian channel and square distortion, double matching means:

\[ D = \sigma^2 \left/ \left( 1 + \frac{E}{\sigma^2} \right) \right. \]

And No coding required!
“There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel ... if we consider channels in which there is a "cost" associated with the different input letters...

The solution to this problem amounts, mathematically, to maximizing a mutual information under linear inequality constraint ... which leads to a capacity-cost function $C(E)$ for the channel...

In a somewhat dual way, evaluating the rate-distortion function $R(D)$ for source amounts, mathematically, to minimizing a mutual Information ... again under a linear inequality constraint.

This duality can be pursued further and is related to the duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.”

C. E. Shannon
Bottlenecks and Neural Nets

- **Auto association**: forcing compact representations
- $\hat{X}$ provide a “relevant code” of $X$ w.r.t. $Y$

![Diagram](image)
Bottlenecks in Nature...
(w E Schniedman, Rob de Ruyter, W Bialek, N Brenner)
The idea: find a compressed variable $\hat{X}$ that enables short encoding of $X$ (small $I(X,\hat{X})$) while preserving as much as possible the information on the relevant signal $Y$ (large $I(\hat{X},Y)$)
A Variational Principle

We want a short representation of $X$ that keeps the information about another variable, $Y$, if possible.

\[ L[p(\hat{x} | x)] = I(X, \hat{X}) - \beta I(\hat{X}, Y) \]
Self Consistent Equations

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

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

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

\[
\frac{\delta L[p(\hat{x} | x)]}{\delta p(\hat{x} | x)} = 0 \quad \Rightarrow \quad p(\hat{x} | x) = \frac{p(\hat{x})}{Z(x, \beta)} \exp(-\beta D_{KL}(x, \hat{x}))
\]
The emerged **effective distortion** measure:

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

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

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

\[ x \xrightarrow{p(y|x)} y \]

\[ \downarrow \quad p(\hat{x} | x) \]

\[ \hat{x} \xrightarrow{p(y|\hat{x})} y \]
I-Projections on a set of distributions (Csiszar 75,84)

- The \textit{i-projection} of a distribution $q(x)$ on a convex set of distributions $L$:

$$\text{IPR}(q, L) = \arg \min_{\tilde{p} \in L} D_{KL} [\tilde{p} \mid q]$$
The Blahut-Arimoto Algorithm

\[
\min_{p(\hat{x})} \min_{p(\hat{x}|x)} \langle \log Z (x, \beta) \rangle = \min_{p(\hat{x}), p(\hat{x}|x)} I(X, \hat{X}) - \beta \langle d \rangle
\]

\[
\begin{align*}
p_{t+1}(\hat{x} | x) &= \frac{p_t(\hat{x})}{Z_t(x, \beta)} \exp \left( - \beta d(x, \hat{x}) \right) \\
p_t(\hat{x}) &= \sum_x p(x) p_t(\hat{x} | x)
\end{align*}
\]
The **Information Bottleneck** Algorithm

\[
\begin{aligned}
\min_{p(y|\hat{x})} \min_{p(\hat{x})} \min_{p(\hat{x}|x)} & \left\langle \log Z(x, \beta) \right\rangle = \\
\min_{p(y|\hat{x}), p(\hat{x}), p(\hat{x}|x)} & I(X, \hat{X}) - \beta \left\langle D_{KL}(x, \hat{x}) \right\rangle
\end{aligned}
\]

\[
\begin{aligned}
p_{t+1}(\hat{x} | x) &= \frac{p_t(\hat{x})}{Z_t(x, \beta)} \exp \left( - \beta D^t_{KL}(x, \hat{x}) \right) \\
p_t(\hat{x}) &= \sum_x p(x)p_t(\hat{x} | x) \\
p_t(y | \hat{x}) &= \sum_x p(y | x)p_t(x | \hat{x})
\end{aligned}
\]
The IB emergent effective distortion measure:

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

**IB Coding Theorem** *(Wyner 1975, Bachrach, Navot, Tishby 2003):*

The IB variational problem provides a tight convex lower bound on the expected size of the achievable representations of $X, \hat{X}$, that maintain at least $I(\hat{X}, Y)$ of mutual information on the variable $Y$. The bound can be achieved by: $|\hat{X}| \leq |X| + 2$.

Formally equivalent to “Channel coding with Side-Information” (introduced in a very different context by Wyner, 75).

The same bound is true for the dual channel coding problem and the optimal representation constitutes a “perfectly matched source-channel” and requires “no further coding”.

**IB Algorithm Convergence** *(Tishby 99, Friedman, Slonim, Tishby 01):*

The Generalized Arimoto-Blauht IB algorithm converges to a local optimum of the IB functional, (including its generalized multivariate case).
| Source \([ p(x) ] \) | Channel: \([ q(y|x) ] \) |
|---|---|
| **Entropy:** \( H = \min_q I(p, q) \) \((d) = 0\) | **Capacity:** \( C = \max_p I(p, q) \) |
| **Rate-Distortion function:** \( R(D) = \min_q I(p, q) \) \((d) \leq D\) | **Capacity-Expense function:** \( C(E) = \max_p I(p, q) \) \((e(p)) \leq E\) |
| **Constrained Reliability-Rate function**  
\( F(R, D) = \max_q \min_{\hat{p}} D_{KL}(\hat{p} \mid p) \)  
\((I(\hat{p}, q) \geq R) \cap (d) \leq D\) | **Constrained Efficiency-Rate function**  
\( F(R, E) = \max_p \min_{\hat{q}} D_{KL}(\hat{q} \mid q) \)  
\((I(p, \hat{q}) \leq R) \cap (e(p)) \leq E\) |
| **Information Bottleneck (source)**  
\( L[I_X, p(x, y)] = \min_{\hat{p}} I(X; \hat{X}) \) \((I(\hat{p}, q) \geq I_Y)\) | **Information Bottleneck (channel)**  
\( L(I_Y, p(x, y)) = \max_q I(Y; \hat{X}) \) \((I(\hat{p}, q) \leq I_X)\) |
The Information – curve, the optimal $I(\hat{X}, X)$ for a given $I(\hat{X}, Y)$ is a concave function:

$$\frac{\partial I(Y, \hat{X})}{\partial I(X, \hat{X})} = \beta^{-1}$$
Applications and Extensions

- Document classification and categorization
- Extracting relevant features for human faces
- Adding Side Information
- IB for Graphical models
- Nonlinear Sufficient Dimensionality Reduction (SDR)
## A Simple Example...

| Audio | Health | www | Drug | noise | Dos | Doctor |...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></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>
</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>
</tr>
<tr>
<td>Doc3</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>20</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>
</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>
</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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
## Simple Example

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>Noise</th>
<th>Health</th>
<th>Drug</th>
<th>Doctor</th>
<th>www</th>
<th>Dos</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>
</tr>
<tr>
<td>Doc4</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc8</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Doc3</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>20</td>
<td>1</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>
</tr>
<tr>
<td>Doc9</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>16</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc5</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Doc7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>12</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>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
A new compact representation

| Audio | Noise | Health | Drug | Doctor | www | Dos | ...
|-------|-------|--------|------|--------|-----|-----|------
| Cluster1 | 36 | 25 | 1 | 1 | 1 | 1 | 0 | ...
| Cluster2 | 1 | 1 | 42 | 39 | 45 | 4 | 2 | ...
| Cluster3 | 1 | 4 | 3 | 0 | 4 | 26 | 33 | ...
| ... | ... | ... | ... | ... | ... | ... | ... |

The document clusters preserve the relevant information between the documents and words
Scaling in Natural Language

Any subset of the language has the same exponent!

\[
\frac{\delta I_Y}{\delta I_X} \propto \frac{1 - I_Y}{1 - I_X}
\]

\[I_Y = \frac{I(\hat{X}, Y)}{I(X, Y)}\]

\[I_X = I(\hat{X}, X)/H(X)\]

\[y = 1.92x - 0.866\]

\[1 - I_Y \cong c(1 - I_X)^{1.92}\]
Learning from the Fly...
(w E Schniedman, Rob de Ruyter, W Bialek, N Brenner)
Analysis of fly H1 neural codes.
(w E Schniedman, Rob de Ruyter, W Bialek,, N Brenner)
The fly's Information curves
[Schneidman, Slonim, Tishby, de Ruyter van Steveninck, Bialek]

- The information curves for compressing the neural code of different flies:

\[
\frac{I(T_1; X_2)}{I(X_1; X_2)}
\]

Similar firing rate

Higher firing rate
Cluster specialized filters
Extensions
The Importance of being Principled

◆ Discriminative IB –
  Information bottleneck with Side Information
  (w Chechik and Globerson, NIPS 02, UAI 03, NIPS 03)

◆ Multivariate Information Bottleneck
  Using graphical models to express the compression and prediction trade-off in a multivariate setting
  (with Nir Friedman and Noam Slonim, UAI-01, NIPS-02)

◆ Sufficient Dimensionality Reduction (SDR, SDR-SI)
  Finding continuous data representation that preserves information
  (with Globerson, ICML-02, with Chechik & Globerson UAI-03)
Discriminative Information-Bottleneck (IB with side (irrelevance) information)

\[
\min I(X;Y^-) \quad \max I(X;Y^+) 
\]
Relevant face features

Men – $p^+(x,y)$

Women – $p^-(x,y)$
Result of SDR with d=1
Illumination feature
Adding side information

$\lambda \ 0:0.1:2.3$
Multivariate IB
[Friedman, Mosenzon, Slonim, Tishby, UAI-2001, NIPS 2002]

Input variables

$X_1 \ X_2 \ X_3 \ X_4 \ldots \ X_n$

Parameters

$P(T_j | P_{\text{gin}})$

Gin dependencies
(minimize)

T1 T2 T3 \ldots \ Tk

Compression (Bottleneck) variables

Gout dependencies
(maximize)

Input variables

$X_1 \ X_2 \ X_3 \ X_4 \ldots \ X_n$

-Two Graphical models, $G_{\text{in}}$ and $G_{\text{out}}$, define the dependencies and the (multi)-information tradeoff
Graphical models
(AKA Bayesian Networks)

- A Graphical model over \((X_1, ..., X_n)\) is a DAG \(G\) in which vertices correspond to the random variables.

- \(P(X_1, ..., X_n)\) is consistent with \(G\) iff each \(X_i\) is independent of all the other (non-descendant) variables, given its parents \(Pa_i\).

\[
P(X_1, ..., X_n) = \prod_i P(X_i | Pa_i^c)
\]
Application to gene-expression data

Data: Gene expression of 500 “informative” genes Vs. 72 Leukemia samples (Golub et al, 1999)
Example: AIB compression of gene expression samples

Data after symmetric AIB compression:

10 Gene clusters

8 Sample clusters

P(Ts|Tg)
Multi-category document classification

- Consider a document collection with different topics, and different writing styles:
Multi-category document classification (cont.)

- One possible “legitimate” partition is by the topic:

Politics  Sport  Internet  Science
Multi-category document classification (cont.)

- And another possible “legitimate” partition is by the writing style:

There might be more than one “legitimate” partition...
Summary

- Shannon’s Information Theory suggests a compelling framework for quantifying the “Complexity-Accuracy” dilemma, by unifying source and channel coding as a Min Max of mutual information.

- The IB method turns this unified coding principle into algorithms for extracting “informative” relevant structures from empirical joint probability distributions $P(X_1, X_2)$...

- The Multivariate IB extends this framework to extract “informative” structures from multivariate distributions, $P(X_1, ..., X_n)$, via Min Max $I(X_1, ..., X_n)$, and Graphical models.

- IB can be further extended for Sufficient Dimensionality Reduction, a method for finding informative continuous low dimensional representations via Min Max $I(X_1, X_2)$.
See: www.cs.huji.ac.il/~tishby

- Slonim and Tishby, “Agglomerative information Bottleneck” (NIPS 1999)
- Slonim and Tishby, “Document classification from word-clusters, via the Information Bottleneck Method”, (SIGIR 2000)
- Tishby and Slonim, “Data clustering by Markovian Relaxation”, (NIPS 2000)
- Bialek and Tishby, “Extarting relevant information” (NECI TR, 1999)
- Friedman, Mosenzon, Slonim, Tishby, “ Multivariate information bottleneck”, (UAI 2001)
- Slonim, Friedman, Tishby, “Agglomerative information bottleneck”, (NIPS
Many thanks to:

Fernando Pereira
William Bialek
Nir Friedman
Noam Slonim
Gal Chechik
Amir Globerson
Ran Bachrach
Amir Navot
Ilya Nemenman