Extracting Relevant Information from Samples

Naftali Tishby

School of Computer Science and Engineering
Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem, Israel

ISAiM 2008
Outline

1. Mathematics of relevance
   - Motivating examples
   - Sufficient Statistics
   - Relevance and Information

2. The Information Bottleneck Method
   - Relations to learning theory
   - Finite sample bounds
   - Consistency and optimality

3. Further work and Conclusions
   - The Perception Action Cycle
   - Temporary conclusions
Examples: Co-occurrence data
(words-topics, genes-tissues, etc.)
Example: Objects and pixels
Example: Neural codes (e.g. de-Ruyter and Bialek)
Neural codes (Fly H1 cell recording, with Rob de-Ruyter and Bill Bialek)
Sufficient statistics

What captures the *relevant properties* in a sample about a parameter?

- Given an i.i.d. sample $x^{(n)} \sim p(x|\theta)$

**Definition (Sufficient statistic)**

A sufficient statistic: $T(x^{(n)})$ is a function of the sample such that

$$p(x^{(n)}|T(x^{(n)}) = t, \theta) = p(x^{(n)}|T(x^{(n)}) = t).$$

**Theorem (Fisher Neyman factorization)**

$T(x^{(n)})$ is sufficient for $\theta$ in $p(x|\theta) \iff$ there exist $h(x^{(n)})$ and $g(T, \theta)$ such that

$$p(x^{(n)}|\theta) = h(x^{(n)})g(T(x^{(n))}, \theta).$$
Sufficient statistics

What captures the *relevant properties* in a sample about a parameter?

- Given an i.i.d. sample $x^{(n)} \sim p(x|\theta)$

**Definition (Sufficient statistic)**

A sufficient statistic: $T(x^{(n)})$ is a function of the sample such that

$$p(x^{(n)}|T(x^{(n)}) = t, \theta) = p(x^{(n)}|T(x^{(n)}) = t).$$

**Theorem (Fisher Neyman factorization)**

$T(x^{(n)})$ is sufficient for $\theta$ in $p(x|\theta) \iff$ there exist $h(x^{(n)})$ and $g(T, \theta)$ such that

$$p(x^{(n)}|\theta) = h(x^{(n)})g(T(x^{(n))), \theta).$$
Sufficient statistics

What captures the *relevant properties* in a sample about a parameter?

- Given an i.i.d. sample \( x^{(n)} \sim p(x|\theta) \)

**Definition (Sufficient statistic)**

A sufficient statistic: \( T(x^{(n)}) \) is a function of the sample such that

\[
p(x^{(n)}|T(x^{(n)}) = t, \theta) = p(x^{(n)}|T(x^{(n)}) = t).
\]

**Theorem (Fisher Neyman factorization)**

\( T(x^{(n)}) \) is sufficient for \( \theta \) in \( p(x|\theta) \) \( \iff \) there exist \( h(x^{(n)}) \) and \( g(T, \theta) \) such that

\[
p(x^{(n)}|\theta) = h(x^{(n)})g(T(x^{(n)}), \theta).
\]
Minimal sufficient statistics

- There are always trivial (complex) sufficient statistics - e.g. the sample itself.

**Definition (Minimal sufficient statistic)**

\( S(x^{(n)}) \) is a *minimal sufficient statistic* for \( \theta \) in \( p(x|\theta) \) if it is a function of any other sufficient statistics \( T(x^{(n)}) \).

- \( S(X^n) \) gives the coarser *sufficient partition* of the \( n \)-sample space.
- \( S \) is unique (up to 1-1 map).
There are always trivial (complex) sufficient statistics - e.g. the sample itself.

**Definition (Minimal sufficient statistic)**

\( S(x^{(n)}) \) is a *minimal sufficient statistic* for \( \theta \) in \( p(x|\theta) \) if it is a function of any other sufficient statistics \( T(x^{(n)}) \).

- \( S(X^n) \) gives the coarser *sufficient partition* of the \( n \)-sample space.
- \( S \) is unique (up to 1-1 map).
Minimal sufficient statistics

- There are always trivial (complex) sufficient statistics - e.g. the sample itself.

**Definition (Minimal sufficient statistic)**

\[ S(x^{(n)}) \] is a *minimal sufficient statistic* for \( \theta \) in \( p(x|\theta) \) if it is a function of any other sufficient statistics \( T(x^{(n)}) \).

- \( S(X^n) \) gives the coarser *sufficient partition* of the \( n \)-sample space.

- \( S \) is unique (up to 1-1 map).
There are always trivial (complex) sufficient statistics - e.g. the sample itself.

**Definition (Minimal sufficient statistic)**

$S(x^{(n)})$ is a *minimal sufficient statistic* for $\theta$ in $p(x|\theta)$ if it is a function of any other sufficient statistics $T(x^{(n)})$.

- $S(X^n)$ gives the coarser *sufficient partition* of the $n$-sample space.
- $S$ is unique (up to 1-1 map).
What distributions have sufficient statistics?

Theorem (Pitman, Koopman, Darmois.)

Among families of parametric distributions whose domain does not vary with the parameter, only in \textit{exponential families},

\[ p(x|\theta) = h(x) \exp \left( \sum_r \eta_r(\theta) A_r(x) - A_0(\theta) \right), \]

there are sufficient statistics for $\theta$ with bounded dimensionality: $T_r(x^{(n)}) = \sum_{k=1}^{n} A_r(x_k)$, (additive for i.i.d. samples).
What distributions have sufficient statistics?

**Theorem (Pitman, Koopman, Darmois.)**

*Among families of parametric distributions whose domain does not vary with the parameter, only in *exponential families*,

\[
p(x|\theta) = h(x) \exp \left( \sum_r \eta_r(\theta) A_r(x) - A_0(\theta) \right),
\]

*there are sufficient statistics for \(\theta\) with bounded dimensionality: \(T_r(x^{(n)}) = \sum_{k=1}^n A_r(x_k)\), (additive for i.i.d. samples).*
Definition (Mutual Information)

For any two random variables $X$ and $Y$ with joint pdf $P(X = x, Y = y) = p(x, y)$, Shannon’s mutual information $I(X; Y)$ is defined as

$$I(X; Y) = \mathbb{E}_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)}.$$ 

- $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \geq 0$
- $I(X; Y) = D_{KL}[p(x, y)||p(x)p(y)]$, maximal number (on average) of independent bits on $Y$ that can be revealed from measurements on $X$. 

Naftali Tishby  
Extracting Relevant Information from Samples
Definition (Mutual Information)

For any two random variables $X$ and $Y$ with joint pdf $P(X = x, Y = y) = p(x, y)$, Shannon’s mutual information $I(X; Y)$ is defined as

$$I(X; Y) = \mathbb{E}_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)}.$$

- $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \geq 0$
- $I(X; Y) = D_{KL}[p(x, y)||p(x)p(y)]$, maximal number (on average) of independent bits on $Y$ that can be revealed from measurements on $X$. 

Sufficiency and Information

Motivating examples
The Information Bottleneck Method
Further work and Conclusions

Sufficient Statistics
Relevance and Information

Mathematics of relevance

Naftali Tishby
Extracting Relevant Information from Samples
## Definition (Mutual Information)

For any two random variables $X$ and $Y$ with joint pdf $P(X = x, Y = y) = p(x, y)$, Shannon’s mutual information $I(X; Y)$ is defined as

\[ I(X; Y) = \mathbb{E}_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} . \]

- $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \geq 0$
- $I(X; Y) = D_{KL}[p(x, y)|p(x)p(y)]$, maximal number (on average) of independent bits on $Y$ that can be revealed from measurements on $X$. 

**Naftali Tishby**

*Extracting Relevant Information from Samples*
Properties of Mutual Information

- Key properties of mutual information:

Theorem (Data-processing inequality)
When $X \rightarrow Y \rightarrow Z$ form a Markov chain, then

$$I(X; Z) \leq I(X; Y)$$

- data processing can’t increase (mutual) information.

Theorem (Joint typicality)
The probability of a typical sequence $y^{(n)}$ to be jointly typical with an independent typical sequence $x^{(n)}$ is

$$P(y^{(n)}|x^{(n)}) \propto \exp(-nI(X; Y)).$$
Properties of Mutual Information

- Key properties of mutual information:

**Theorem (Data-processing inequality)**

When $X \rightarrow Y \rightarrow Z$ form a Markov chain, then

$$I(X; Z) \leq I(X; Y)$$

- data processing can’t increase (mutual) information.

**Theorem (Joint typicality)**

The probability of a typical sequence $y^{(n)}$ to be jointly typical with an independent typical sequence $x^{(n)}$ is

$$P(y^{(n)}|x^{(n)}) \propto \exp(-nI(X; Y)).$$
Properties of Mutual Information

**Key properties of mutual information:**

**Theorem (Data-processing inequality)**

When $X \rightarrow Y \rightarrow Z$ form a Markov chain, then

$$I(X; Z) \leq I(X; Y)$$

- *Data processing can't increase (mutual) information.*

**Theorem (Joint typicality)**

The probability of a typical sequence $y^{(n)}$ to be jointly typical with an independent typical sequence $x^{(n)}$ is

$$P(y^{(n)}|x^{(n)}) \propto \exp(-nI(X; Y)).$$
When the parameter $\theta$ is a random variable (we are Bayesian), we can characterize sufficiency and minimality using mutual information:

**Theorem (Sufficiency and Information)**

- $T$ is sufficient statistics for $\theta$ in $p(x|\theta)$ if and only if
  \[ I(T(X^n); \theta) = I(X^n; \theta). \]

- If $S$ is minimal sufficient statistics for $\theta$ in $p(x|\theta)$, then:
  \[ I(S(X^n); X^n) \leq I(T(X^n); X^n). \]

That is, among all sufficient statistics, minimal maintain the least mutual information on the sample $X^n$. 
Sufficiency and Information

When the parameter $\theta$ is a random variable (we are Bayesian), we can characterize sufficiency and minimality using mutual information:

**Theorem (Sufficiency and Information)**

- $T$ is sufficient statistics for $\theta$ in $p(x|\theta)$ if
  \[ I(T(X^n); \theta) = I(X^n; \theta). \]

- If $S$ is minimal sufficient statistics for $\theta$ in $p(x|\theta)$, then:
  \[ I(S(X^n); X^n) \leq I(T(X^n); X^n). \]

That is, among all sufficient statistics, minimal maintain the least mutual information on the sample $X^n$. 

---

Motivating examples
The Information Bottleneck Method
Further work and Conclusions
Sufficient Statistics
Relevance and Information
Naftali Tishby
Extracting Relevant Information from Samples
When the parameter $\theta$ is a random variable (we are Bayesian), we can characterize sufficiency and minimality using mutual information:

**Theorem (Sufficiency and Information)**

- $T$ is sufficient statistics for $\theta$ in $p(x|\theta)$ if
  \[ I(T(X^n); \theta) = I(X^n; \theta). \]

- If $S$ is minimal sufficient statistics for $\theta$ in $p(x|\theta)$, then:
  \[ I(S(X^n); X^n) \leq I(T(X^n); X^n). \]

That is, among all sufficient statistics, minimal maintain the least mutual information on the sample $X^n$. 

Naftali Tishby

Extracting Relevant Information from Samples
When the parameter $\theta$ is a random variable (we are Bayesian), we can characterize sufficiency and minimality using mutual information:

**Theorem (Sufficiency and Information)**

- $T$ is sufficient statistics for $\theta$ in $p(x|\theta)$ $\iff$
  \[ I(T(X^n); \theta) = I(X^n; \theta). \]

- If $S$ is minimal sufficient statistics for $\theta$ in $p(x|\theta)$, then:
  \[ I(S(X^n); X^n) \leq I(T(X^n); X^n). \]

That is, among all sufficient statistics, minimal maintain the least mutual information on the sample $X^n$. 

Naftali Tishby

Extracting Relevant Information from Samples
Given \((X, Y) \sim p(x, y)\), the above theorem suggests a definition for the relevant part of \(X\) with respect to \(Y\). Find a random variable \(T\) such that:

- \(T \leftrightarrow X \leftrightarrow Y\) form a Markov chain
- \(I(T; X)\) is minimized (minimality, complexity term)
  while \(I(T; Y)\) is maximized (sufficiency, accuracy term).

Equivalent to the minimization of the following Lagrangian:

\[
\mathcal{L}[p(t|x)] = I(X; T) - \beta I(Y; T)
\]

subject to the Markov conditions. Varying the Lagrange multiplier \(\beta\) yields an information tradeoff curve, similar to RDT.

\(T\) is called the Information Bottleneck between \(X\) and \(Y\).
The Information Bottleneck: Approximate Minimal Sufficient Statistics

Given \((X, Y) \sim p(x, y)\), the above theorem suggests a definition for the relevant part of \(X\) with respect to \(Y\). Find a random variable \(T\) such that:

- \(T \leftrightarrow X \leftrightarrow Y\) form a Markov chain
- \(I(T; X)\) is minimized (minimality, complexity term)
- while \(I(T; Y)\) is maximized (sufficiency, accuracy term).

Equivalent to the minimization of the following Lagrangian:

\[
\mathcal{L}[p(t|x)] = I(X; T) - \beta I(Y; T)
\]

subject to the Markov conditions. Varying the Lagrange multiplier \(\beta\) yields an information tradeoff curve, similar to RDT.

\(T\) is called the Information Bottleneck between \(X\) and \(Y\).
Given \((X, Y) \sim p(x, y)\), the above theorem suggests a definition for \textit{the relevant part} of \(X\) with respect to \(Y\).

Find a random variable \(T\) such that:

- \(T \leftrightarrow X \leftrightarrow Y\) form a Markov chain
- \(I(T; X)\) is minimized (minimality, \textit{complexity} term)
  while \(I(T; Y)\) is maximized (sufficiency, \textit{accuracy} term).

Equivalent to the minimization of the following Lagrangian:

\[
L[p(t|x)] = I(X; T) - \beta I(Y; T)
\]

subject to the Markov conditions. Varying the Lagrange multiplier \(\beta\) yields an \textit{information tradeoff curve}, similar to RDT.

\(T\) is called the \textit{Information Bottleneck} between \(X\) and \(Y\).
The Information Bottleneck: Approximate Minimal Sufficient Statistics

- Given \((X, Y) \sim p(x, y)\), the above theorem suggests a definition for *the relevant part* of \(X\) with respect to \(Y\). Find a random variable \(T\) such that:
  - \(T \leftrightarrow X \leftrightarrow Y\) form a Markov chain
  - \(I(T; X)\) is minimized (minimality, complexity term) while \(I(T; Y)\) is maximized (sufficiency, accuracy term).

- Equivalent to the minimization of the following Lagrangian:
  \[
  \mathcal{L}[p(t|x)] = I(X; T) - \beta I(Y; T)
  \]

subject to the Markov conditions. Varying the Lagrange multiplier \(\beta\) yields an *information tradeoff curve*, similar to RDT.

- \(T\) is called the *Information Bottleneck* between \(X\) and \(Y\).
The Information Bottleneck: Approximate Minimal Sufficient Statistics

Given \((X, Y) \sim p(x, y)\), the above theorem suggests a definition for *the relevant part* of \(X\) with respect to \(Y\). Find a random variable \(T\) such that:

- \(T \leftrightarrow X \leftrightarrow Y\) form a Markov chain
- \(I(T; X)\) is minimized (minimality, *complexity* term) while \(I(T; Y)\) is maximized (sufficiency, *accuracy* term).

Equivalent to the minimization of the following Lagrangian:

\[
\mathcal{L}[p(t|x)] = I(X; T) - \beta I(Y; T)
\]

subject to the Markov conditions. Varying the Lagrange multiplier \(\beta\) yields an *information tradeoff curve*, similar to RDT.

\(T\) is called the *Information Bottleneck* between \(X\) and \(Y\).
The **Information-Curve** for Multivariate Gaussian variables (GGTW 2005).

\[ \beta^{-1} = 1 - \lambda_1 \]
Outline

1. Mathematics of relevance
   - Motivating examples
   - Sufficient Statistics
   - Relevance and Information

2. The Information Bottleneck Method
   - Relations to learning theory
   - Finite sample bounds
   - Consistency and optimality

3. Further work and Conclusions
   - The Perception Action Cycle
   - Temporary conclusions
The IB Algorithm I (Tishby, Periera, Bialek 1999)

How is the Information Bottleneck problem solved?

- \[
    \frac{\delta L}{\delta p(t|x)} = 0 + \text{the Markov and normalization constraints,}
    \]
  yields the (bottleneck) self-consistent equations:

The bottleneck equations

1. \[
    p(t|x) = \frac{p(t)}{Z(x, \beta)} \exp(-\beta D_{KL}[p(y|x)\|p(y|t)])
    \]
2. \[
    p(t) = \sum_x p(t|x)p(x)
    \]
3. \[
    p(y|t) = \sum_x p(y|x)p(x|t)
    \]

\[
    Z(x, \beta) = \sum_t p(t) \exp(-\beta D_{KL}[p(y|x)\|p(y|t)])
    \]
\[
    D_{KL}[p(y|x)\|p(y|t)] = \mathbb{E}_{p(y|x)} \log \frac{p(y|x)}{p(y|t)} = d_{IB}(x, t) - \text{an effective distortion measure on the } q(y) \text{ simplex.} 
    \]
How is the Information Bottleneck problem solved?

\[ \frac{\delta L}{\delta p(t|x)} = 0 + \text{the Markov and normalization constraints,} \]
\[ \text{yields the (bottleneck) self-consistent equations:} \]

The bottleneck equations

\[ p(t|x) = \frac{p(t)}{Z(x, \beta)} \exp(-\beta D_{KL}[p(y|x) \| p(y|t)]) \] (1)

\[ p(t) = \sum_x p(t|x)p(x) \] (2)

\[ p(y|t) = \sum_x p(y|x)p(x|t) , \] (3)

\[ Z(x, \beta) = \sum_t p(t) \exp(-\beta D_{KL}[p(y|x) \| p(y|t)]) \]

\[ D_{KL}[p(y|x) \| p(y|t)] = \mathbb{E}_{p(y|x)} \log \frac{p(y|x)}{p(y|t)} = d_{IB}(x, t) \] - an effective distortion measure on the q(y) simplex.
As showed in (Tishby, Periera, Bialek 1999) iterating these equations converges for any $\beta$ to a consistent solution:

**Algorithm:** randomly initiate; iterate for $k \geq 1$

\[ p_{k+1}(t|x) = \frac{p_k(t)}{Z(x, \beta)} \exp(-\beta D_{KL}[p(y|x) || p_k(y|t)]) \] (4)

\[ p_k(t) = \sum_x p_k(t|x)p(x) \] (5)

\[ p_k(y|t) = \sum_x p(y|x)p_k(x|t) \] (6)
Relation with learning theory

Issues often raised about IB:

- If you assume you know $p(x, y)$ - what else is left to be learned or modeled?
  
  **A:** Relevance, meaning, explanations...

- How is it different from statistical modeling (e.g. Maximum Likelihood)?
  
  **A:** It’s not about statistical modeling.

- Is it supervised or unsupervised learning? (Wrong question - none and both)

- What if you only have a finite sample? can it generalize?

- What’s the advantage of maximizing information about $Y$ (rather than other cost/loss)?

- Is there a "coding theorem" associated with this problem (what is good for)?
Issues often raised about IB:

- If you assume you know $p(x, y)$ - what else is left to be learned or modeled?
  A: Relevance, meaning, explanations...
- How is it different from statistical modeling (e.g. Maximum Likelihood)?
  A: It’s not about statistical modeling.
- Is it supervised or unsupervised learning? (wrong question - none and both)
- What if you only have a finite sample? can it generalize?
- What’s the advantage of maximizing information about $Y$ (rather than other cost/loss)?
- Is there a "coding theorem" associated with this problem (what is good for)?
Relation with learning theory

Issues often raised about IB:

- If you assume you know $p(x, y)$ - what else is left to be learned or modeled?
  A: Relevance, meaning, explanations...

- How is it different from statistical modeling (e.g. Maximum Likelihood)?
  A: it’s not about statistical modeling.

- Is it supervised or unsupervised learning?
  (wrong question - none and both)

  - What if you only have a finite sample? can it generalize?
  - What’s the advantage of maximizing information about $Y$ (rather than other cost/loss)?
  - Is there a "coding theorem" associated with this problem (what is good for)?
Relation with learning theory

Issues often raised about IB:

- If you assume you know $p(x, y)$ - what else is left to be learned or modeled?
  A: Relevance, meaning, explanations...

- How is it different from statistical modeling (e.g. Maximum Likelihood)?
  A: it’s not about statistical modeling.

- Is it supervised or unsupervised learning? (wrong question - none and both)

- What if you only have a finite sample? can it generalize?
  - What’s the advantage of maximizing information about $Y$ (rather than other cost/loss)?
  - Is there a "coding theorem" associated with this problem (what is good for)?
Relation with learning theory

Issues often raised about IB:

- If you assume you know $p(x, y)$ - what else is left to be learned or modeled?
  
  A: Relevance, meaning, explanations...

- How is it different from statistical modeling (e.g. Maximum Likelihood)?
  
  A: it’s not about statistical modeling.

- Is it supervised or unsupervised learning? (wrong question - none and both)

- What if you only have a finite sample? can it generalize?

- What’s the advantage of maximizing information about $Y$ (rather than other cost/loss)?

- Is there a "coding theorem" associated with this problem (what is good for)?
Relation with learning theory

Issues often raised about IB:

- If you assume you know $p(x, y)$ - what else is left to be learned or modeled?
  
  A: Relevance, meaning, explanations...

- How is it different from statistical modeling (e.g. Maximum Likelihood)?
  
  A: it’s not about statistical modeling.

- Is it supervised or unsupervised learning? (wrong question - none and both)

- What if you only have a finite sample? can it generalize?

- What’s the advantage of maximizing information about $Y$ (rather than other cost/loss)?

- Is there a "coding theorem" associated with this problem (what is good for)?
Mathematics of relevance
The Information Bottleneck Method
Further work and Conclusions

A Validation theorem
Notation: \( \hat{\cdot} \) denotes empirical quantities using an iid sample \( S \) of size \( m \).

**Theorem (Ohad Shamir & NT, 2007)**

For any fixed random variable \( T \) defined via \( p(t|x) \), and for any confidence parameter \( \delta > 0 \), it holds with probability of at least \( 1 - \delta \) over the sample \( S \) that \( |I(X; T) - \hat{I}(X; T)| \) is upper bounded by:

\[
(|T| \log(m) + \log |T|) \sqrt{\frac{\log(8/\delta)}{2m}} + \frac{|T| - 1}{m},
\]

and similarly \( |I(Y; T) - \hat{I}(Y; T)| \) is upper bounded by:

\[
(1 + \frac{3}{2} |T|) \log(m) \sqrt{\frac{2 \log(8/\delta)}{m}} + \frac{(|Y| + 1)(|T| + 1) - 4}{m}.
\]
Proof idea: We apply McDiarmid’s inequality to bound the sample variations of the empirical Entropies, and a recent bound by Liam Paninski on entropy estimation.

The bounds on the information curve are independent of the cardinality of \( X \) (normally the larger variable) and weakly on \(|Y|\). The bounds are larger for large \( T \), which increase with \( \beta \), as expected.

The information curve can be approximated from a sample of size \( m \sim O(|Y||T|) \), much smaller than needed to estimate \( p(x, y) \).

But how about the quality of the estimated variable \( T \) (defined by \( p(t|x) \) itself?)
Proof idea: We apply McDiarmid’s inequality to bound the sample variations of the empirical Entropies, and a recent bound by Liam Paninski on entropy estimation.

The bounds on the information curve are independent of the cardinality of $X$ (normally the larger variable) and weakly on $|Y|$. The bounds are larger for large $T$, which increase with $\beta$, as expected.

The information curve can be approximated from a sample of size $m \sim O(|Y| |T|)$, much smaller than needed to estimate $p(x, y)$!

But how about the quality of the estimated variable $T$ (defined by $p(t|x)$ itself?)
Proof idea: We apply McDiarmid’s inequality to bound the sample variations of the empirical Entropies, and a recent bound by Liam Paninski on entropy estimation.

The bounds on the information curve are independent of the cardinality of $X$ (normally the larger variable) and weakly on $|Y|$. The bounds are larger for large $T$, which increase with $\beta$, as expected.

The information curve can be approximated from a sample of size $m \sim O(|Y||T|)$, much smaller than needed to estimate $p(x, y)$!

But how about the quality of the estimated variable $T$ (defined by $p(t|x)$ itself?)
Proof idea: We apply McDiarmid’s inequality to bound the sample variations of the empirical Entropies, and a recent bound by Liam Paninski on entropy estimation.

The bounds on the information curve are independent of the cardinality of $X$ (normally the larger variable) and weakly on $|Y|$. The bounds are larger for large $T$, which increase with $\beta$, as expected.

The information curve can be approximated from a sample of size $m \sim O(|Y||T|)$, much smaller than needed to estimate $p(x, y)$!

But how about the quality of the estimated variable $T$ (defined by $p(t|x)$ itself?)
Generalization bounds

**Theorem (Shamir & NT 2007)**

For any confidence parameter $\delta \geq 0$, we have with probability of at least $1 - \delta$, for any $T$ defined via $p(t|x)$ and any constants $a, b_1, \ldots, b_{|T|}, c$ simultaneously:

$$
|I(X; T) - \hat{I}(X; T)| \leq \sum_t f\left(\frac{n(\delta)\|p(t|x) - b_t\|}{\sqrt{m}}\right)
+ \frac{n(\delta)\|H(T|x) - a\|}{\sqrt{m}},
$$

$$
|I(Y; T) - \hat{I}(Y; T)| \leq 2\sum_t f\left(\frac{n(\delta)\|p(t|x) - b_t\|}{\sqrt{m}}\right)
+ \frac{n(\delta)\|\hat{H}(T|y) - c\|}{\sqrt{m}}.
$$

where $n(\delta) = 2 + \sqrt{2 \log \left(\frac{|Y|+2}{\delta}\right)}$, and $f(x)$ is monotonically increasing and concave in $|x|$, defined as:

$$
f(x) = \begin{cases} 
|x| \log(1/|x|) & |x| \leq 1/e \\
1/e & |x| > 1/e
\end{cases}
$$
Corollary

*Under the conditions and notation of Thm. 10, we have that if:*

\[
m \geq e^2 |X| \left(1 + \sqrt{\frac{1}{2} \log \left(\frac{|Y| + 2}{\delta}\right)}\right)^2,
\]

*then with probability of at least \(1 - \delta\), \(|I(X; T) - \hat{I}(X; T)|\) is upper bounded by*

\[
n(\delta) \frac{\frac{1}{2} |T| \sqrt{|X|} \log \left(\frac{4m}{n^2(\delta)|X|}\right) + \sqrt{|X|} \log(|T|)}{2\sqrt{m}},
\]

*and \(|I(Y; T) - \hat{I}(Y; T)|\) is upper bounded by*

\[
n(\delta) \frac{|T| \sqrt{|X|} \log \left(\frac{4m}{n^2(\delta)|X|}\right) + \sqrt{|Y|} \log(|T|)}{2\sqrt{m}}.
\]
If \( m \sim |X||Y| \) and \( |T| \ll \sqrt{|Y|} \) the bound is tight. This is much less than needed to estimate \( p(x, y) \).

We also obtain a statistical consistency result:

**Theorem (IB is consistent (Shamir & NT 2007))**

For any given \( \beta \), let \( A \) be the set of IB optimal \( p(t|x) \). As \( m \to \infty \), the optimal \( p(t|x) \) with respect to the empirical \( \hat{p}(x,y) \), converges in total variation distance to \( A \) with probability 1 as \( m \to \infty \).

Finally, despite its apparent non-convexity, the IB solution is optimal and unique in a well defined sense (Harremoes & NT 2007, Shamir & NT 2007).
If \( m \sim |X| |Y| \) and \( |T| \ll \sqrt{|Y|} \) the bound is tight. This is much less than needed to estimate \( p(x, y) \).

We also obtain a statistical consistency result:

**Theorem (IB is consistent (Shamir & NT 2007))**

For any given \( \beta \), let \( A \) be the set of IB optimal \( p(t|x) \). As \( m \to \infty \), the optimal \( p(t|x) \) with respect to the empirical \( \hat{p}(x, y) \), converges in total variation distance to \( A \) with probability 1 as \( m \to \infty \).

Finally, despite its apparent non-convexity, the IB solution is optimal and unique in a well defined sense (Harremoes & NT 2007, Shamir & NT 2007).
Consistency and optimality

- If $m \sim |X||Y|$ and $|T| \ll |\sqrt{Y}|$ the bound is tight. This is much less than needed to estimate $p(x, y)$.
- We also obtain a statistical consistency result:

**Theorem (IB is consistent (Shamir & NT 2007))**

*For any given $\beta$, let $A$ be the set of IB optimal $p(t|x)$. As $m \to \infty$, the optimal $p(t|x)$ with respect to the empirical $\hat{p}(x, y)$, converges in total variation distance to $A$ with probability 1 as $m \to \infty$.*

- Finally, despite its apparent non-convexity, the IB solution is optimal and unique in a well defined sense (Harremoes & NT 2007, Shamir & NT 2007).
Consistency and optimality

- If \( m \sim |X| |Y| \) and \( |T| \ll \sqrt{|Y|} \) the bound is tight. This is much less than needed to estimate \( p(x, y) \).
- We also obtain a statistical consistency result:

**Theorem (IB is consistent (Shamir & NT 2007))**

*For any given \( \beta \), let \( A \) be the set of IB optimal \( p(t|x) \). As \( m \to \infty \), the optimal \( p(t|x) \) with respect to the empirical \( \hat{p}(x, y) \), converges in total variation distance to \( A \) with probability 1 as \( m \to \infty \).*

- Finally, despite its apparent non-convexity, the IB solution is optimal and unique in a well defined sense (Harremoes & NT 2007, Shamir & NT 2007).
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mathematics of relevance</td>
</tr>
<tr>
<td></td>
<td>- Motivating examples</td>
</tr>
<tr>
<td></td>
<td>- Sufficient Statistics</td>
</tr>
<tr>
<td></td>
<td>- Relevance and Information</td>
</tr>
<tr>
<td>2</td>
<td>The Information Bottleneck Method</td>
</tr>
<tr>
<td></td>
<td>- Relations to learning theory</td>
</tr>
<tr>
<td></td>
<td>- Finite sample bounds</td>
</tr>
<tr>
<td></td>
<td>- Consistency and optimality</td>
</tr>
<tr>
<td>3</td>
<td>Further work and Conclusions</td>
</tr>
<tr>
<td></td>
<td>- The Perception Action Cycle</td>
</tr>
<tr>
<td></td>
<td>- Temporary conclusions</td>
</tr>
</tbody>
</table>
An exciting new application of IB is for characterizing optimal steady-state interaction between an organism and its environment:

Summary

- Relevance can be identified with an extension of the classical notion of *minimal sufficient statistics*.
- Can be quantified using information theoretic notions, leading to the IB principle.
- Yielding practical algorithms for extracting relevant variables.
- Can be done efficiently and consistently from empirical data, but isn’t standard learning theory.
- Has many applications, most exciting so far in biology and cognitive science.
Summary

Relevance can be identified with an extension of the classical notion of \emph{minimal sufficient statistics}.

Can be quantified using information theoretic notions, leading to the IB principle.

Yielding practical algorithms for extracting relevant variables.

Can be done efficiently and consistently from empirical data, but isn’t standard learning theory.

Has many applications, most exciting so far in biology and cognitive science.
Summary

- Relevance can be identified with an extension of the classical notion of *minimal sufficient statistics*.
- Can be quantified using information theoretic notions, leading to the IB principle.
- Yielding practical algorithms for extracting relevant variables.
- Can be done efficiently and consistently from empirical data, but isn’t standard learning theory.
- Has many applications, most exciting so far in biology and cognitive science.
Summary

- Relevance can be identified with an extension of the classical notion of *minimal sufficient statistics*.
- Can be quantified using information theoretic notions, leading to the IB principle.
- Yielding practical algorithms for extracting relevant variables.
- Can be done efficiently and consistently from empirical data, but isn’t standard learning theory.
- Has many applications, most exciting so far in biology and cognitive science.
Summary

- Relevance can be identified with an extension of the classical notion of *minimal sufficient statistics*.
- Can be quantified using information theoretic notions, leading to the IB principle.
- Yielding practical algorithms for extracting relevant variables.
- Can be done efficiently and consistently from empirical data, but isn’t standard learning theory.
- Has many applications, most exciting so far in biology and cognitive science.
Thank You!