Information Theory in Intelligent Decision Making

Daniel Polani

Adaptive Systems and Algorithms Research Groups
School of Computer Science
University of Hertfordshire, United Kingdom

June 7, 2015

Information Theory in Intelligent Decision Making

Information Theory

Daniel Polani

Adaptive Systems and Algorithms Research Groups
School of Computer Science
University of Hertfordshire, United Kingdom

June 7, 2015

Motivation

Artificial Intelligence

- modelling cognition in humans
- realizing human-level "intelligent" behaviour in machines
- jumble of various ideas to get above points working

Question

Is there a joint way of understanding cognition?

Probability

- we have probability theory for a theory of uncertainty
- we have information theory for endowing probability with a sense of "metrics"

Motivation

Artificial Intelligence

- modelling cognition in humans
- realizing human-level "intelligent" behaviour in machines (just performance: not necessarily imitating biological substrate)
- jumble of various ideas to get above points working

Question

Is there a joint way of understanding cognition?

Probability

- we have probability theory for a theory of uncertainty
- we have information theory for endowing probability with a sense of "metrics"

Random Variables

Def.: Event Space

Consider an *event space* $\Omega = \{\omega_1, \omega_2, \dots\}$, finite or countably infinite with a (probability) measure $P_\Omega : \Omega \to [0,1]$ s.t. $\sum_\omega P_\Omega(\omega) = 1$. The ω are called *events*.

Random Variable

A *random variable* X is a map $X : \Omega \to \mathcal{X}$ with some outcome space $\mathcal{X} = \{x_1, x_2, \dots\}$ and induced probability measure $P_X(x) = P_\Omega(X^{-1}(x))$.
We also write instead

$$P_X(x) \equiv P(X = x) \equiv p(x) \ .$$

Neyman-Pearson Lemma I

Lemma

- Consider observations $x_1, x_2, \ldots, x_n$ of a random variable X and two potential hypotheses (distributions) p_1 and p_2 they could have been based upon.
- Consider the test for hypothesis p_1 to be given as $(x_1, x_2, \ldots, x_n) \in \mathcal{A}$ where $\mathcal{A} = \left\{ \mathbf{x} = (x'_1, x'_2, \ldots, x'_n) \,\middle|\, \frac{p_1(x'_1, x'_2, \ldots, x'_n)}{p_2(x'_1, x'_2, \ldots, x'_n)} \geq C \right\}$ with some $C \in \mathbb{R}^+$.
- Assuming the rate α of *false negatives* $p_1(\bar{\mathcal{A}})$ to be given.
  Generated by p_1, but not in $\mathcal{A}$
- If β is the rate of *false positives* $p_2(\mathcal{A})$

  Then: any test with false negative rate $\alpha' \leq \alpha$ has false positive rate $\beta' \geq \beta$.

(Cover and Thomas, 2006)

Neyman-Pearson Lemma II

Proof
(Cover and Thomas, 2006)

Let $\mathcal{A}$ as above and $\mathcal{B}$ some other acceptance region; $\chi_{\mathcal{A}}$ and $\chi_{\mathcal{B}}$ be the indicator functions. Then for all $\mathbf{x}$:

$$[\chi_{\mathcal{A}}(\mathbf{x}) - \chi_{\mathcal{B}}(\mathbf{x})]\,[p_1(\mathbf{x}) - Cp_2(\mathbf{x})] \geq 0\,.$$

Multiplying out & integrating:

$$\begin{aligned} 0 &\leq \sum_{\mathcal{A}}(p_1 - Cp_2) - \sum_{\mathcal{B}}(p_1 - Cp_2) \\ &= (1-\alpha) - C\beta - (1-\alpha') + C\beta' \\ &= C(\beta' - \beta) - (\alpha - \alpha') \qquad \blacksquare \end{aligned}$$

Neyman-Pearson Lemma III

Consideration

- assume events x i.i.d.
- test becomes:

$$\prod_i \frac{p_1(x_i)}{p_2(x_i)} \geq C$$

- logarithmize:

$$\sum_i \log \frac{p_1(x_i)}{p_2(x_i)} \geq \kappa \qquad (:= \log C)$$

Neyman-Pearson Lemma IV

Consideration

- assume events x i.i.d.
- test becomes:

$$\prod_i \frac{p_1(x_i)}{p_2(x_i)} \geq C$$

- logarithmize:

$$\sum_i \log \frac{p_1(x_i)}{p_2(x_i)} \geq \kappa \qquad (:= \log C)$$

Note

Average "evidence" growth per sample

$$\mathbf{E}\left[\log \frac{p_1(X)}{p_2(X)}\right] = \sum_{x \in \mathcal{X}} p(x) \log \frac{p_1(x)}{p_2(x)}$$

Neyman-Pearson Lemma V

Consideration

- assume events x i.i.d.
- test becomes:

$$\prod_i \frac{p_1(x_i)}{p_2(x_i)} \geq C$$

- logarithmize:

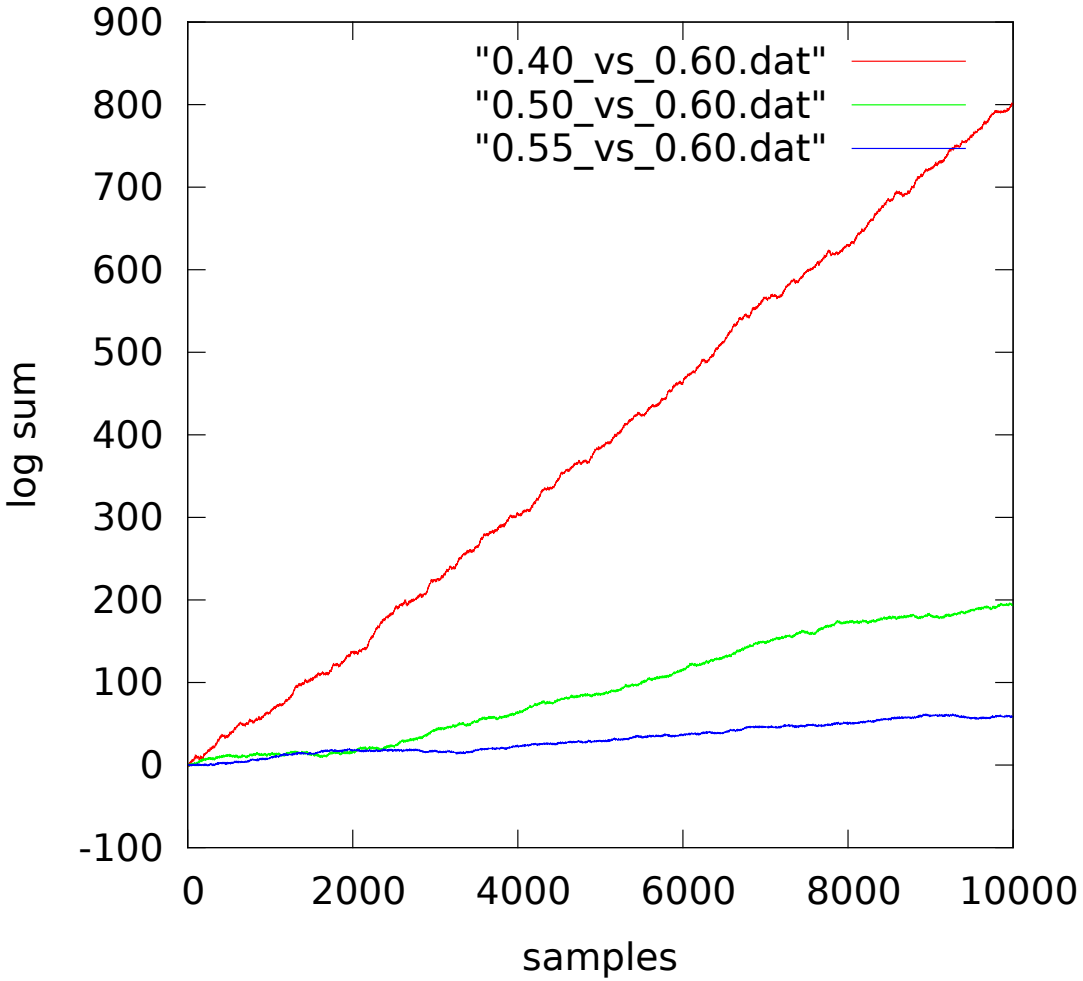
$$\sum_i \log \frac{p_1(x_i)}{p_2(x_i)} \geq \kappa \qquad (:= \log C)$$

Note: Kullback-Leibler Divergence

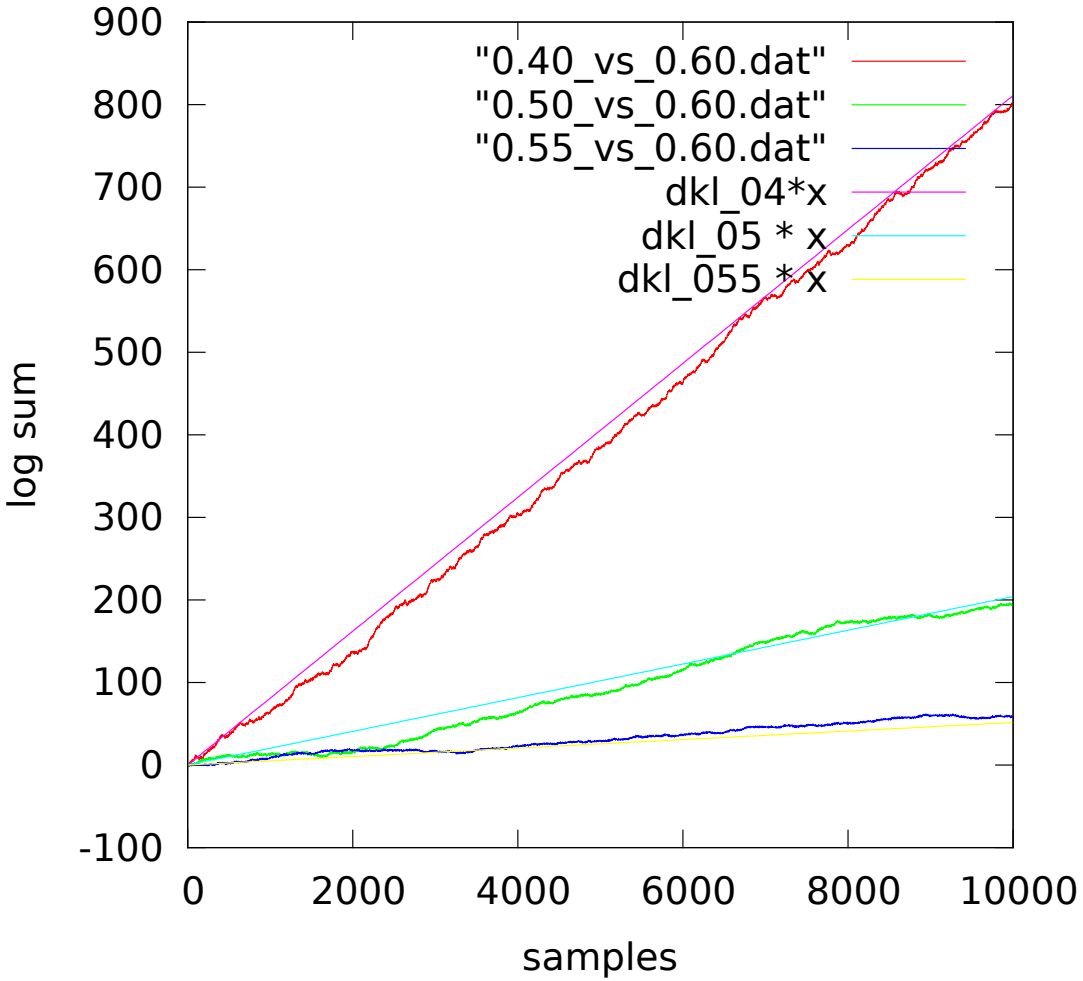
Average "evidence" growth per sample

$$D_{\mathrm{KL}}(p_1||p_2) = \mathbf{E}_{p_1}\left[\log \frac{p_1(X)}{p_2(X)}\right] = \sum_{x\in\mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)}$$

Neyman-Pearson Lemma VI



Neyman-Pearson Lemma VII



Part I

Information Theory

Structural Motivation

Intrinsic Pathways to Information Theory

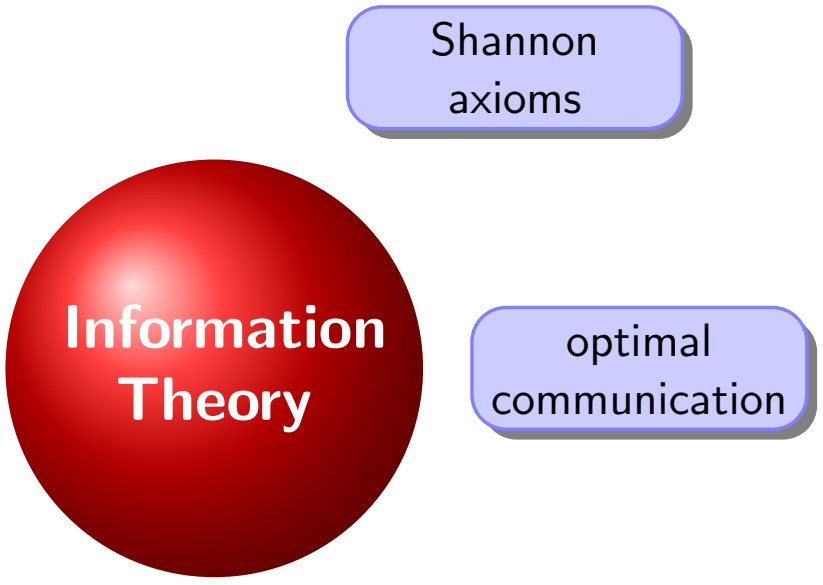
Structural Motivation

Intrinsic Pathways to Information Theory

optimal
communication

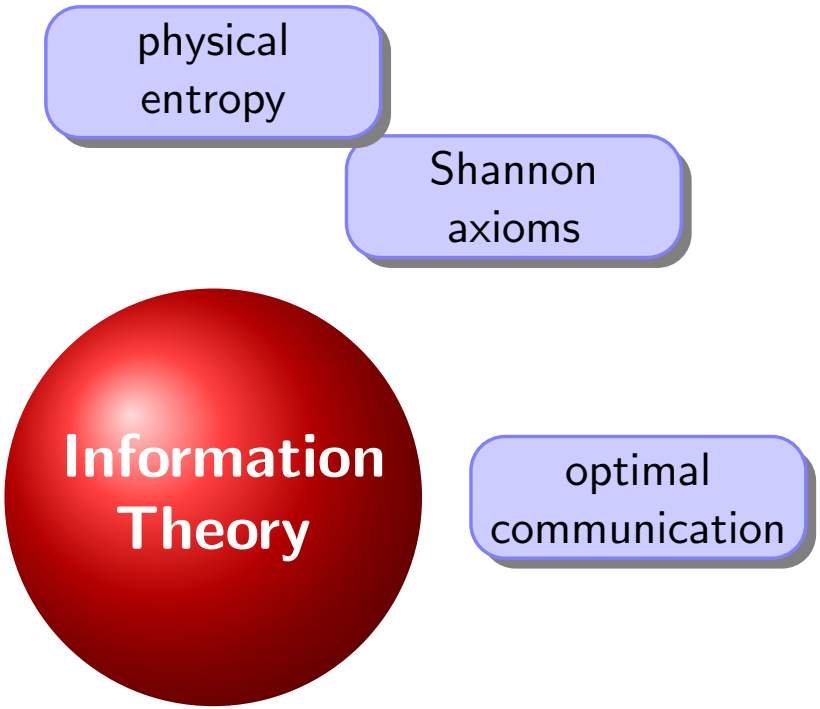
Structural Motivation

Intrinsic Pathways to Information Theory



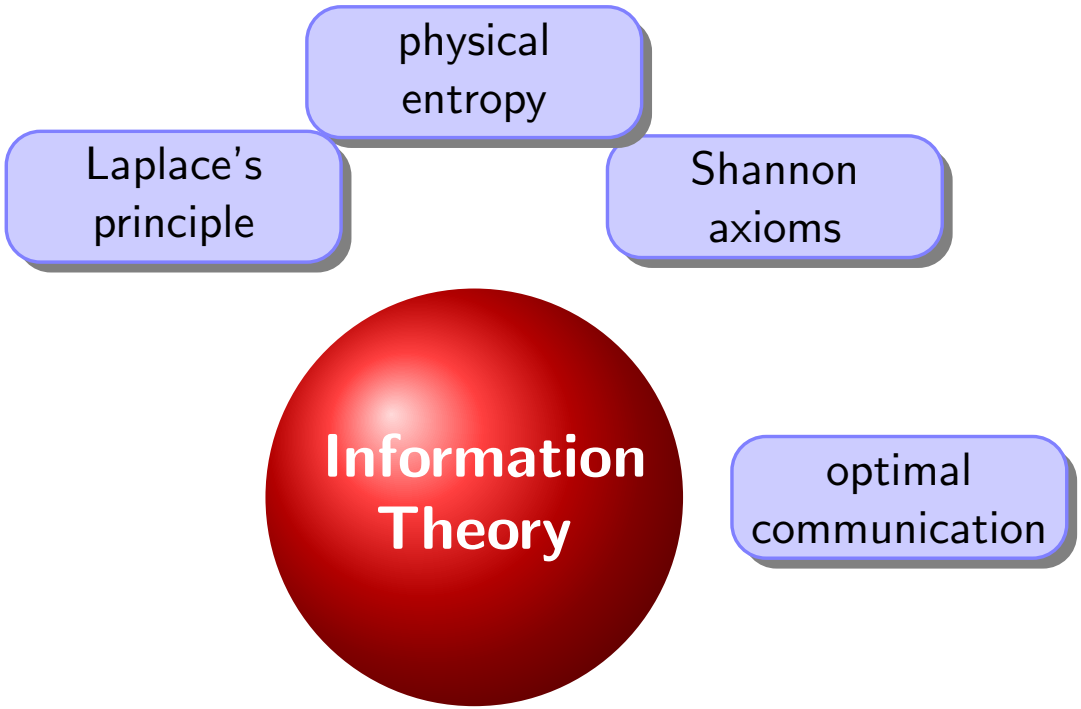
Structural Motivation

Intrinsic Pathways to Information Theory



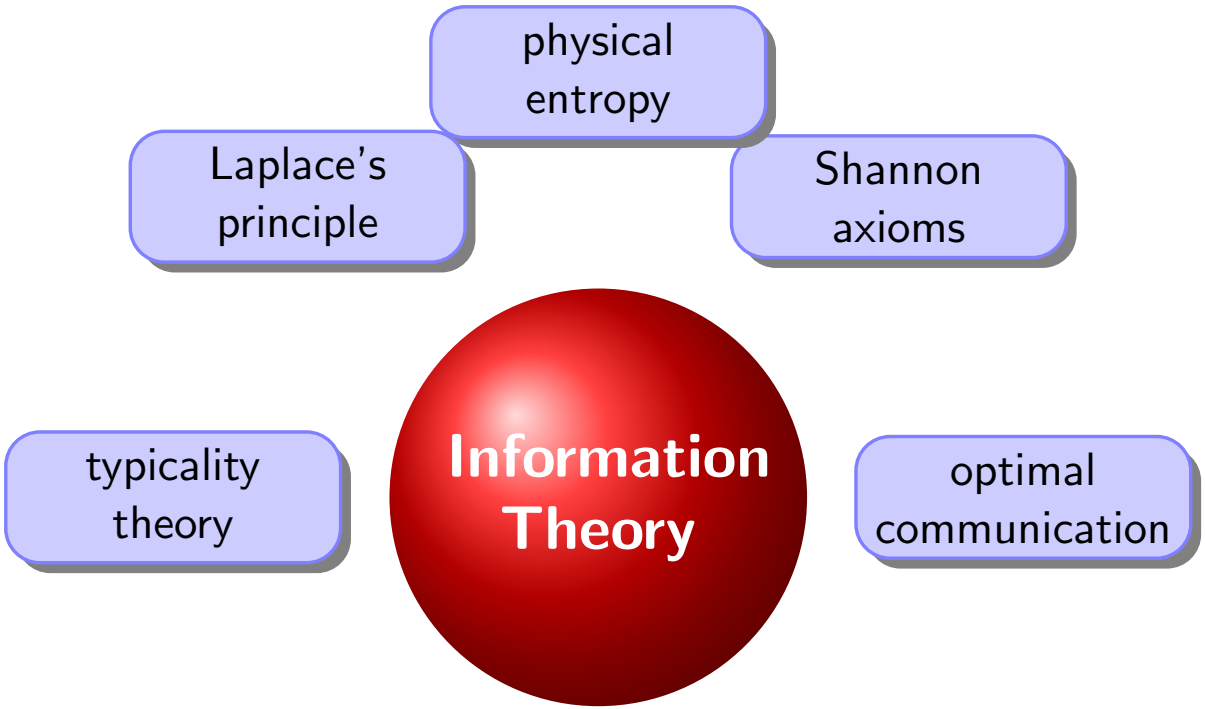
Structural Motivation

Intrinsic Pathways to Information Theory



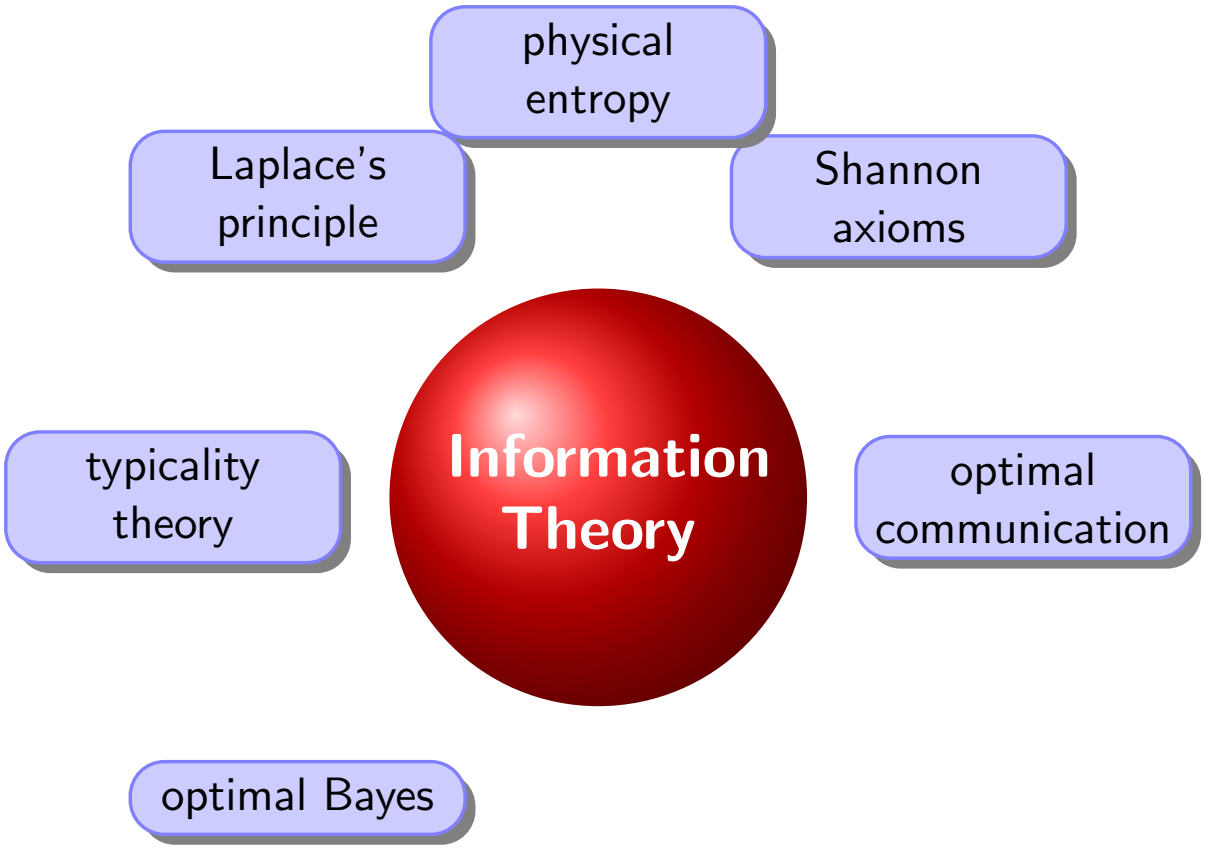
Structural Motivation

Intrinsic Pathways to Information Theory



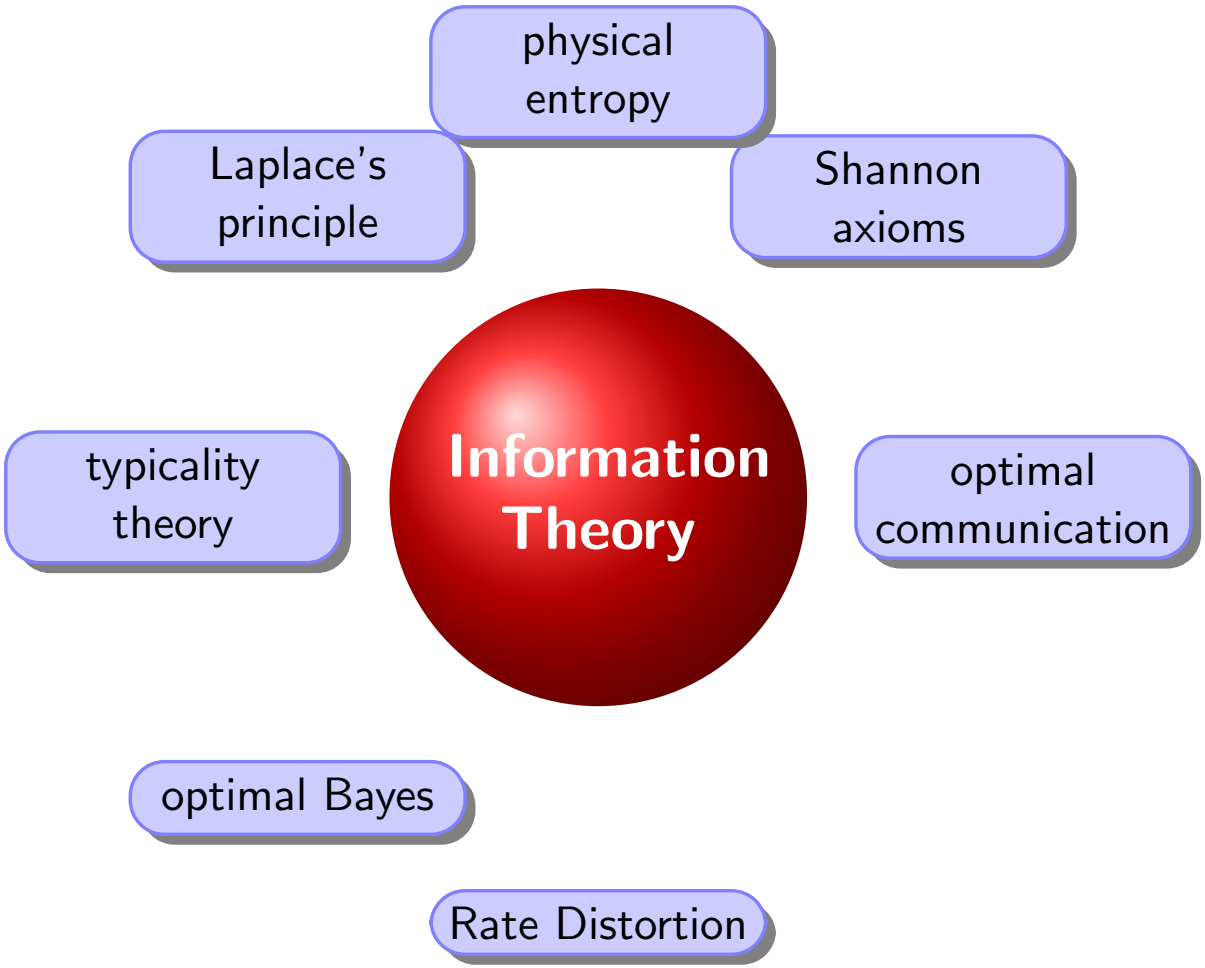
Structural Motivation

Intrinsic Pathways to Information Theory



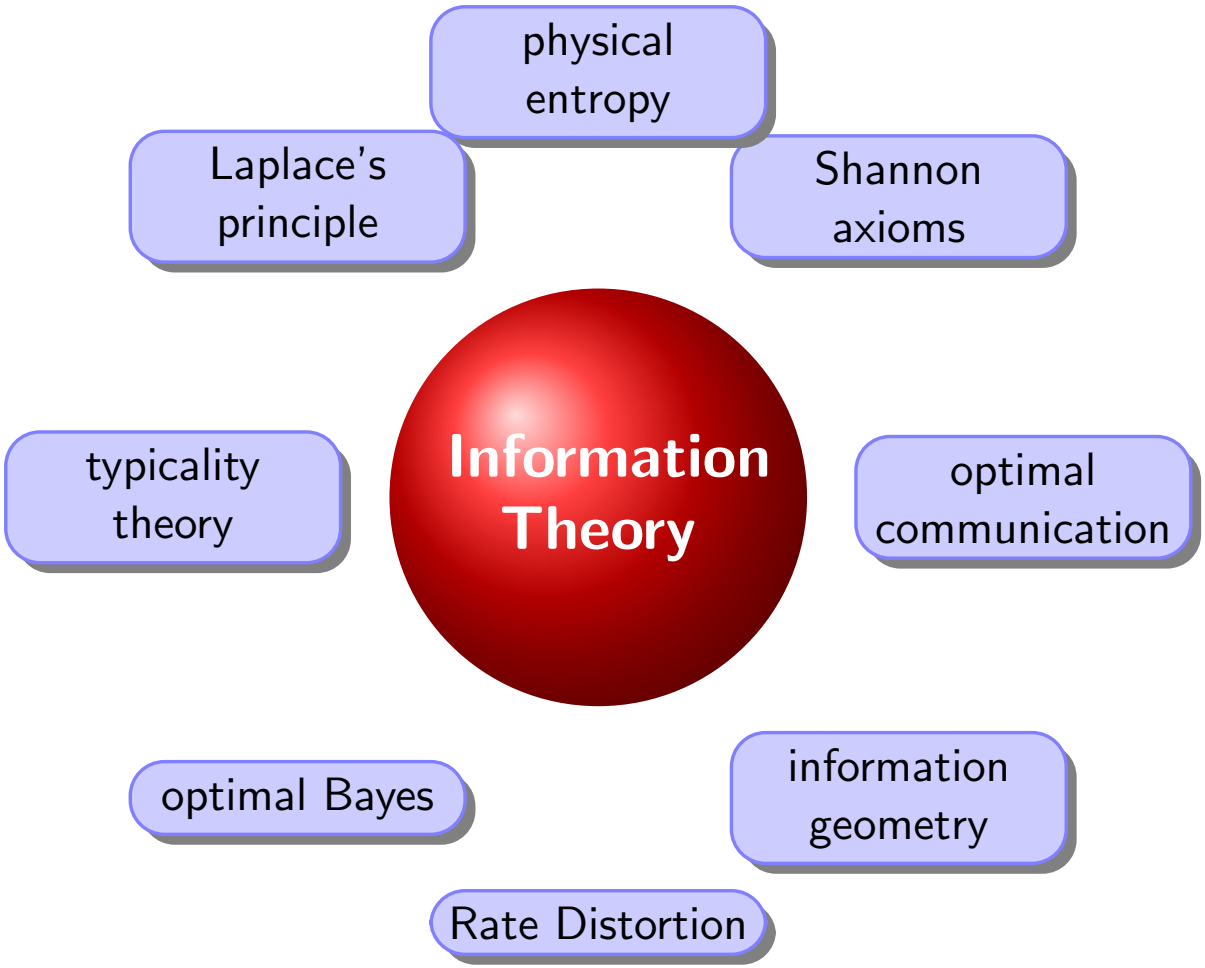
Structural Motivation

Intrinsic Pathways to Information Theory



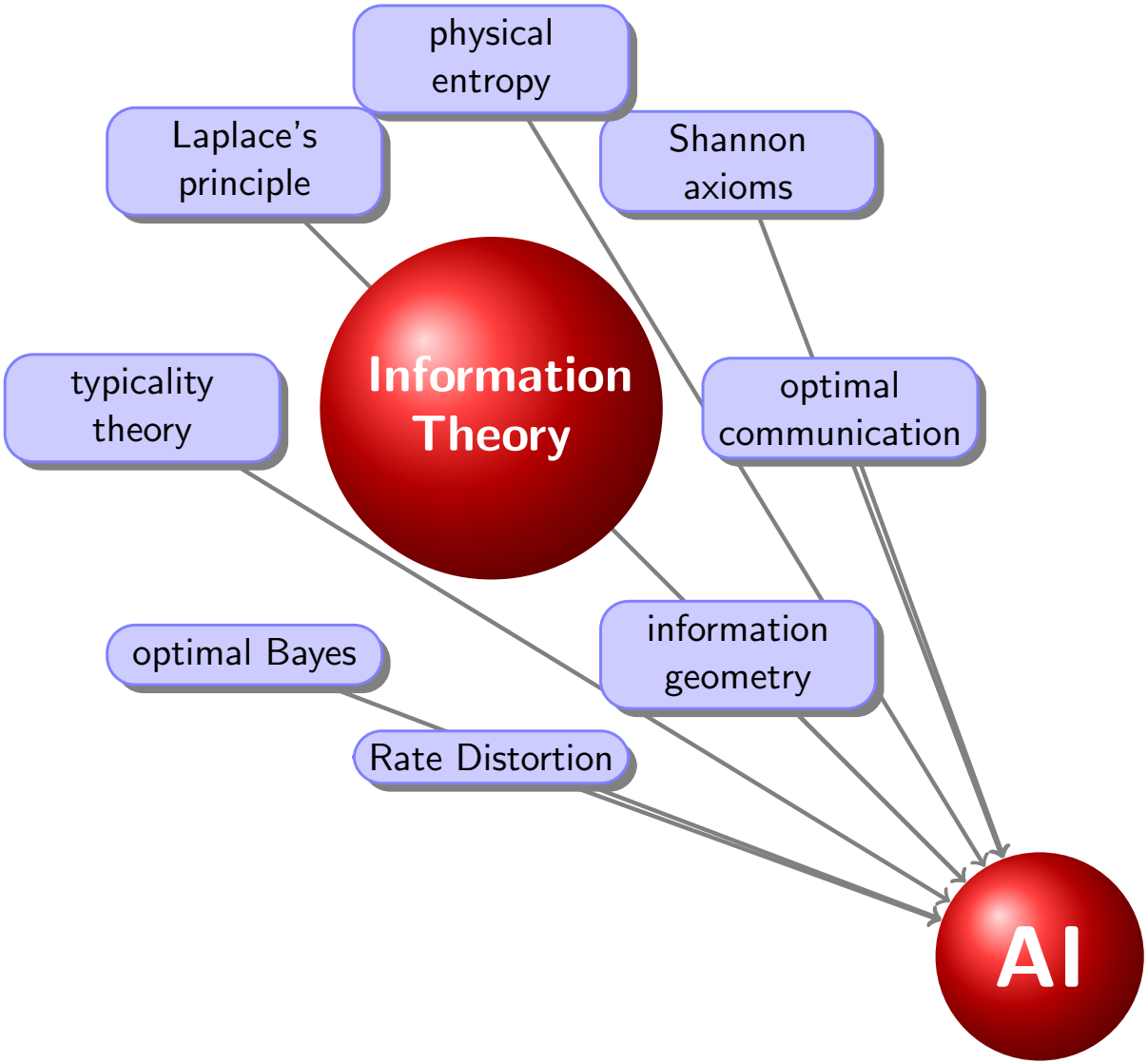
Structural Motivation

Intrinsic Pathways to Information Theory



Structural Motivation

Intrinsic Pathways to Information Theory



Optimal Communication

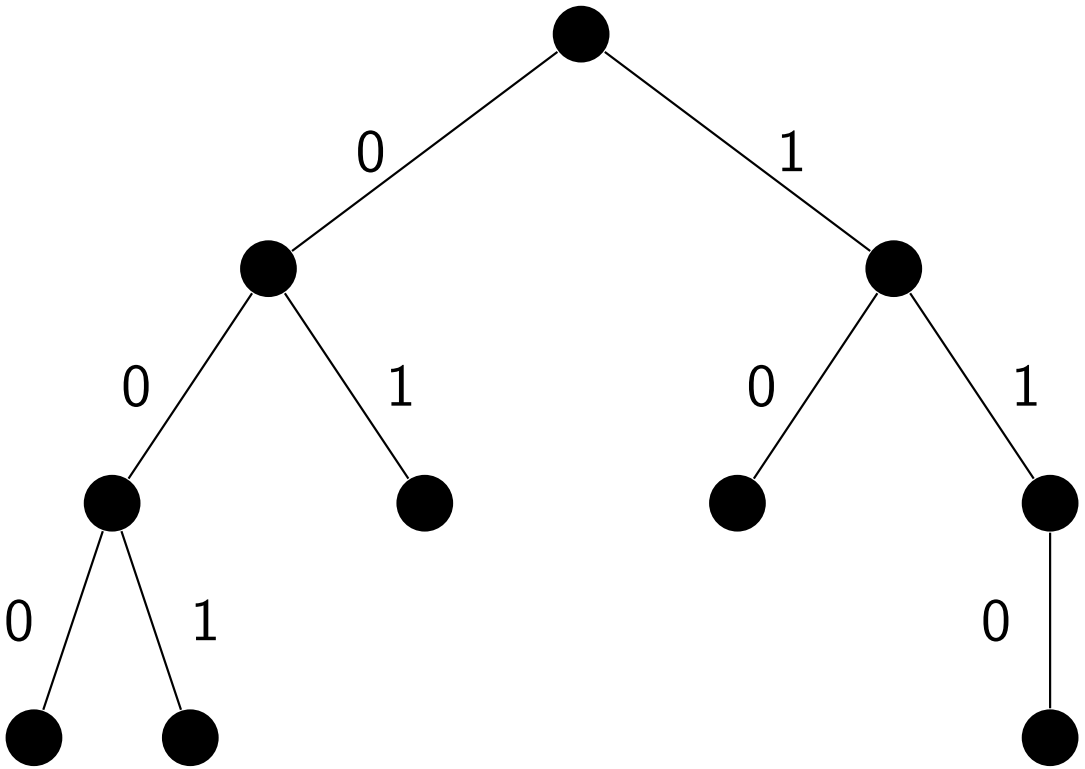
Codes

- task: send messages (disambiguate states) from sender to receiver
- consider self-delimiting codes (without extra delimiting character)
- simple example: **prefix codes**

Def.: Prefix Codes

codes where none is a prefix of another code

Prefix Codes



Kraft Inequality

Theorem

Assume events $x \in \mathcal{X} = \{x_1, x_2, \ldots x_k\}$ are coded using prefix codewords based on alphabet size $b = |\mathcal{B}|$, with lengths $l_1, l_2, \ldots, l_k$ for the respective events, then one has

$$\sum_{i=1}^{k} b^{l_i} \leq 1 \,.$$

Proof Sketch

(Cover and Thomas, 2006)

Let $l_{\max}$ be the length of the longest codeword. Expand tree fully to level $l_{\max}$. Fully expanded leaves are either: 1. codewords; 2. descendants of codewords; 3. neither.
An l_i codeword has $b^{l_{\max}-l_i}$ full-tree descendants, which must be different for the different codewords and there cannot be more than $b^{l_{\max}}$ in total. Hence

$$\sum b^{l_{\max}-l_i} \leq b^{l_{\max}} \qquad \blacksquare$$

Remark

The converse also holds.

Considerations — Most compact code

Assume

Want to code stream of events $x \in \mathcal{X}$ appearing with probability $p(x)$.

Minimize

Average code length: $\mathbf{E}[L] = \sum_i p(x_i)\, l_i$ under constraint $\sum_i b^{-l_i} \stackrel{!}{=} 1$

Note

1. try to make l_i as small as possible
2. make b^{-l_i} as large as possible
3. limited by Kraft inequality; ideally becoming equality

$$\sum_i b^{-l_i} = 1$$

as l_i are integers, that's typically not exact

Result

Differentiating Lagrangian

$$\sum_i p(x_i)\, l_i + \lambda \sum_i b^{-l_i}$$

w.r.t. l gives codeword lengths for "shortest" code:

$$l_i = -\log_b p(x_i)$$

Considerations — Most compact code

Assume

Want to code stream of events $x \in \mathcal{X}$ appearing with probability $p(x)$.

Minimize

Average code length: $\mathbf{E}[L] = \sum_i p(x_i)\, l_i$ under constraint $\sum_i b^{-l_i} \stackrel{!}{=} 1$

Note

1. try to make l_i as small as possible
2. make b^{-l_i} as large as possible
3. limited by Kraft inequality; ideally becoming equality

$$\sum_i b^{-l_i} = 1$$

as l_i are integers, that's typically not exact

Result

Differentiating Lagrangian

$$\sum_i p(x_i)\, l_i + \lambda \sum_i b^{-l_i}$$

w.r.t. l gives codeword lengths for "shortest" code:

$$l_i = -\log_b p(x_i)$$

Average Codeword Length

$$= \sum_i p(x_i) \cdot l_i = -\sum_x p(x) \log p(x)$$

In the following, assume binary log.

Entropy

Def.: Entropy

Consider the random variable X. Then the *entropy* $H(X)$ of X is defined as

$$H(X) := -\sum_x p(x) \log p(x)$$

with convention $0 \log 0 \equiv 0$

Entropy

Def.: Entropy

Consider the random variable X. Then the *entropy* $H(X)$ of X is defined as

$$H(X) := -\sum_x p(x) \log p(x)$$

with convention $0 \log 0 \equiv 0$

Interpretations

- average optimal codeword length
- uncertainty (about next sample of X)
- physical entropy
- much more ...

Quote

"Why don't you call it entropy. In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage."

John von Neumann

Entropy

Def.: Entropy

Consider the random variable X. Then the *entropy* $H(X)$ of X is defined as

$$H(X)[\equiv H(p)] := -\sum_x p(x) \log p(x)$$

with convention $0 \log 0 \equiv 0$

Interpretations

- average optimal codeword length
- uncertainty (about next sample of X)
- physical entropy
- much more ...

Quote

"Why don't you call it entropy. In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage."

John von Neumann

Probability/Code Mismatch

Consider events x following a probability $p(x)$, but modeler assuming mistakenly probability $q(x)$, with optimal code lengths $-\log q(x)$. Then "*code length waste per symbol*" given by

$$
\begin{aligned}
&-\sum_x p(x)\log q(x) + \sum_x p(x)\log p(x) \\
&= \sum_x p(x)\log\frac{p(x)}{q(x)} \\
&= D_{\mathsf{KL}}(p||q)
\end{aligned}
$$

A Tip of Types

(Cover and Thomas, 2006)

Method of Types: Motivation

- consider sequences with same empirical distribution
- how many of these with a particular distribution
- probability of such a sequence

Sketch of the Method

- consider binary event set $\mathcal{X} = \{0, 1\}$
  w.l.o.g.
- consider sample $x^{(n)} = (x_1, \ldots, x_n) \in \mathcal{X}^n$
- the *type* $\mathfrak{p}_x^{(n)}$ is the empirical distribution of symbols $y \in \mathcal{X}$ in sample $x^{(n)}$. I.e. $\mathfrak{p}_{x^{(n)}}(y)$ counts how often symbol y appears in $x^{(n)}$. Let $\mathcal{P}_n$ be set of types with denominator n.
  or dividing n
- for $\mathfrak{p} \in \mathcal{P}_n$, call the set of all sequences $x^{(n)} \in \mathcal{X}^n$ with type $\mathfrak{p}$ the *type class* $C(\mathfrak{p}) = \{x^{(n)} | \mathfrak{p}_{x^{(n)}} = \mathfrak{p}\}$.

Type Theorem

Type Count

If $|\mathcal{X}| = 2$, one has $|\mathcal{P}_n| = n + 1$ different types for sequences of length n.

easy to generalize

Important

$\mathcal{P}_n$ grows only polynomially, but $\mathcal{X}^n$ grows exponentially with n. It follows that (at least one) type must contain exponentially many sequences. This corresponds to the "macrostate" in physics.

Theorem

(Cover and Thomas, 2006)

If $x_1, x_2, \ldots, x_n$ is an i.i.d. drawn sample sequence drawn from q, then the probability of $x^{(n)}$ depends only on its type and is given by

$$2^{-n[H(\mathfrak{p}_{x^{(n)}}) + D_{\mathsf{KL}}(\mathfrak{p}_{x^{(n)}} || q)]}$$

Corollary

If $x^{(n)}$ has type q, then its probability is given by

$$2^{-nH(q)}$$

A large value of $H(q)$ indicates many possible candidates $x^{(n)}$ and high uncertainty, a small value few candidates and low uncertainty.

here, we interpret probability q as type

Laplace's Principle of Insufficient Reason I

Scenario

Consider $\mathcal{X}$. A probability distribution is assumed on $\mathcal{X}$, but it is unknown.
Laplace's principle of insufficient reason states that, in absence of any reason to assume that the outcomes are inequivalent, the probability distribution on $\mathcal{X}$ is assumed as equidistribution.

Question

How to generalize when something is known?

Answer: Types

Dominant Sample Sequence

Remember: sequence probability of sequences in type class $C(q)$

$$2^{-nH(q)}$$

A priori, a probability q maximizing $H(q)$ will generate dominating sequence types dominating all others.

Maximum Entropy Principle

Maximize: $H(q)$ with respect to q

Result: equidistribution $q(x) = \frac{1}{|\mathcal{X}|}$

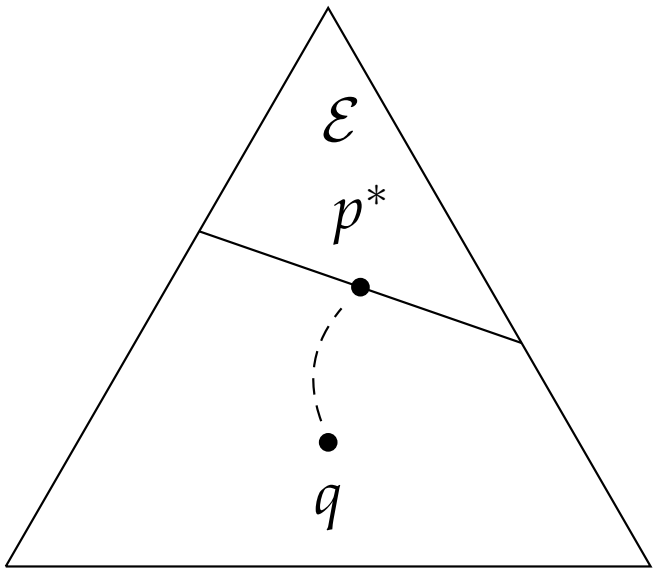
Sanov's Theorem I

Theorem

Consider i.i.d. sequence $X_1, X_2, \ldots, X_n$ of random variables, distributed according to $q(X)$. Let further $\mathcal{E}$ be a set of probability distributions.

Then (amongst other), if $\mathcal{E}$ is closed and with $p^* = \arg\min_{p \in E} D(p||q)$, one has

$$\frac{1}{n} \log q^{(n)}(\mathcal{E}) \longrightarrow -D(p^*||q)$$



Sanov's Theorem II

Interpretation

p is unknown, but one knows constraints for p (e.g. some condition, such as some mean value $\bar{U} \stackrel{!}{=} \sum_x p(x)U(x)$ must be attained, i.e. the set $\mathcal{E}$ is given), then the dominating types are those close to p^*.

Special Case

if prior q is equidistribution (indifference), then minimizing $D(p||q)$ under constraints $\mathcal{E}$ is equivalent to maximizing $H(p)$ under these constraints.

Jaynes' Maximum Entropy Principle

Sanov's Theorem III

Jaynes' Principle

- generalization of Laplace's Principle
- maximally uncommitted distribution

Maximum Entropy Distributions I

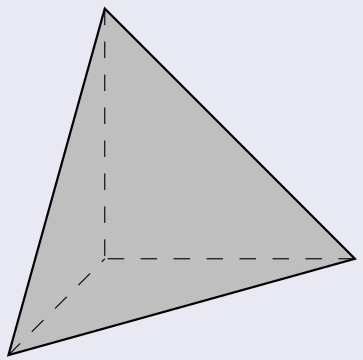
No constraints

We are interested in maximizing

$$H(X) = -\sum_x p(x) \log p(x)$$

over all probabilities p. The probability p lives in the simplex

$$\Delta = \{q \in \mathbb{R}^{|\mathcal{X}|} | \sum_i q_i = 1, q_i \geq 0\}$$



The maximization requires to respect constraints, of which we now consider only $\sum_x p(x) \stackrel{!}{=} 1$.

The edge constraints happen not to be invoked here.

Maximum Entropy Distributions II

No constraints

Unconstrained maximization via Lagrange:

$$\max_p [-\sum_x p(x) \log p(x) + \lambda \sum_x p(x)]$$

Taking derivative $\nabla_{p(x)}$ gives

$$-\log p(x) - 1 + \lambda \stackrel{!}{=} 0$$

. Thus $p(x) = e^{\lambda - 1} \equiv 1/|\mathcal{X}|$ — equidistribution

Maximum Entropy Distributions

Linear Constraints

Constraints are now

$$\sum_x p(x) \stackrel{!}{=} 1$$
$$\sum_x p(x) f(x) \stackrel{!}{=} \bar{f} \,.$$

Derive Lagrangian

$$0 = \quad -\sum_x p(x) \log p(x) + \lambda \sum_x p(x) + \mu \sum_x p(x) f(x)$$
$$-\log p(x) - 1 + \lambda + \mu f(x) = 0$$

so that one has

Boltzmann/Gibbs Distribution

$$p(x) = e^{\lambda - 1 + \mu f(x)}$$
$$= \frac{1}{Z} e^{\mu f(x)}$$

Maximum Entropy Distributions

Linear Constraints

Constraints are now

$$\sum_x p(x) \stackrel{!}{=} 1$$
$$\sum_x p(x) f(x) \stackrel{!}{=} \bar{f} \, .$$

Derive Lagrangian

$$0 = \nabla_P [- \sum_x p(x) \log p(x) + \lambda \sum_x p(x) + \mu \sum_x p(x) f(x)]$$
$$- \log p(x) - 1 + \lambda + \mu f(x) = 0$$

so that one has

Boltzmann/Gibbs Distribution

$$p(x) = e^{\lambda - 1 + \mu f(x)}$$
$$= \frac{1}{Z} e^{\mu f(x)}$$

Conditional Kullback-Leibler

D_{KL} can be conditional

$$D_{\mathsf{KL}}\left[p(Y|X)||q(Y||X)\right] = \sum_x p(x) D_{\mathsf{KL}}\left[p(Y|x)||q(Y|x)\right]$$

$$D_{\mathsf{KL}}\left[p(Y|x)||q(Y|x)\right]$$

Kullback-Leibler and Bayes

(Biehl, 2013)

Want to estimate $p(x|\theta)$, where θ is the parameter. Observe y. Seek "best" $q(x|y)$ for this y in the following sense:

1. minimize D_{KL} of true distribution to model distribution q

$$\min_{q} \qquad D_{\mathsf{KL}}[p(x|\theta)||q(x|y)]$$

Kullback-Leibler and Bayes

(Biehl, 2013)

Want to estimate $p(x|\theta)$, where θ is the parameter. Observe y. Seek "best" $q(x|y)$ for this y in the following sense:

1. minimize D_{KL} of true distribution to model distribution q
2. averaged over possible observations y

$$\min_q \qquad \sum_y p(y|\theta)\; D_{\mathsf{KL}}[p(x|\theta)||q(x|y)]$$

Kullback-Leibler and Bayes

(Biehl, 2013)

Want to estimate $p(x|\theta)$, where θ is the parameter. Observe y. Seek "best" $q(x|y)$ for this y in the following sense:

1. minimize D_{KL} of true distribution to model distribution q
2. averaged over possible observations y
3. averaged over θ

$$\min_q \int d\theta \; p(\theta) \; \sum_y p(y|\theta) \; D_{\mathsf{KL}}[p(x|\theta)||q(x|y)]$$

Kullback-Leibler and Bayes

(Biehl, 2013)

Want to estimate $p(x|\theta)$, where θ is the parameter. Observe y. Seek "best" $q(x|y)$ for this y in the following sense:

1. minimize D_{KL} of true distribution to model distribution q
2. averaged over possible observations y
3. averaged over θ

$$\min_q \int d\theta \; p(\theta) \sum_y p(y|\theta) \; D_{\mathsf{KL}}[p(x|\theta)||q(x|y)]$$

Result

$q(x|y)$ is the Bayesian inference obtained from $p(y|x)$ and $p(x)$

Conditional Entropies

Special Case: Conditional Entropy

$$H(Y|X=x) := -\sum_y p(y|x)\log p(y|x)$$

$$H(Y|X) := -\sum_x p(x)\sum_y p(y|x)\log p(y|x)$$

Information

Reduction of entropy (uncertainty) by knowing another variable

$$\begin{aligned} I(X;Y) &:= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= D_{\mathsf{KL}}[p(x,y)||p(x)p(y)] \end{aligned}$$

Reminder

Information is about sending messages. We considered most compact codes over a given noiseless channel. Now consider the situation where either:

1. channel is not noiseless but has noisy characteristics $p(\hat{x}|x)$ or
2. we cannot afford to spend average of $H(X)$ bits per symbol to transmit

Question

What happens? **Total collapse of transmission**

"Compromise"

- don't longer insist on perfect transmission
- accept compromise, measure **distortion** $d(x, \hat{x})$ between original x and transmitted $\hat{x}$
- small distortion good, large distortion "baaad"

Theorem: Rate Distortion Function

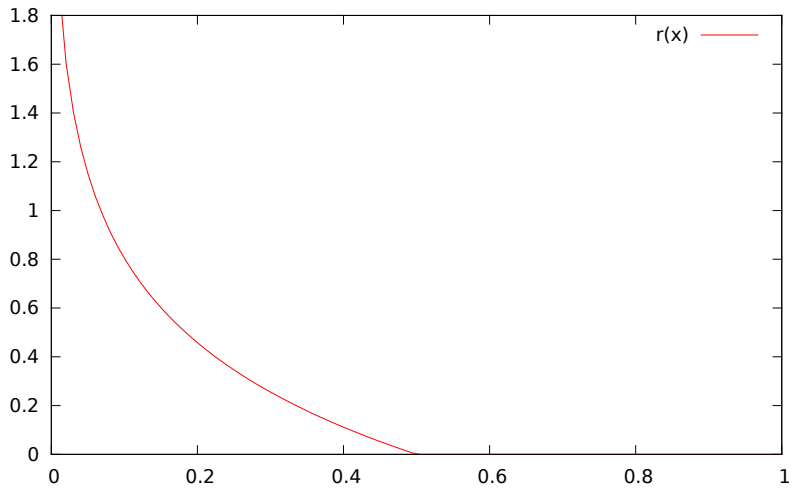
Given $p(x)$ for generation of symbols X,

$$R(D) := \min_{\substack{p(\hat{x}|x) \\ \mathbf{E}[d(X,\hat{X})]=D}} I(X;\hat{X})$$

where the mean is over $p(x, \hat{x}) = p(\hat{x}|x)p(x)$.

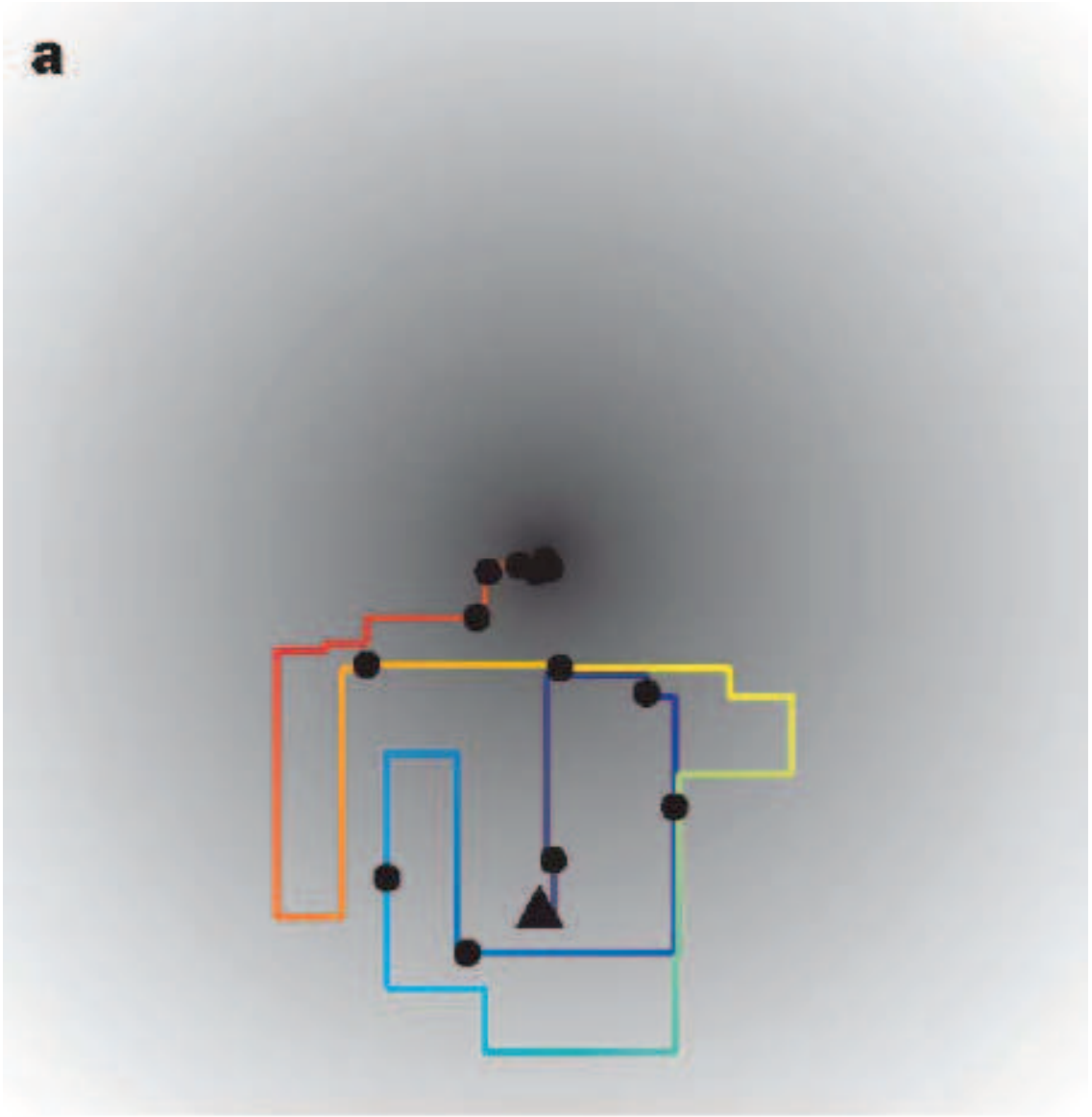
Rate/Distortion Theory II

Distortion



First Example: Infotaxis

(Vergassola et al., 2007)



Part II

References

Biehl, M. (2013). Kullback-leibler and bayes. Internal Memo.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, 2nd edition.

Vergassola, M., Villermaux, E., and Shraiman, B. I. (2007). 'infotaxis' as a strategy for searching without gradients. *Nature*, 445:406–409.