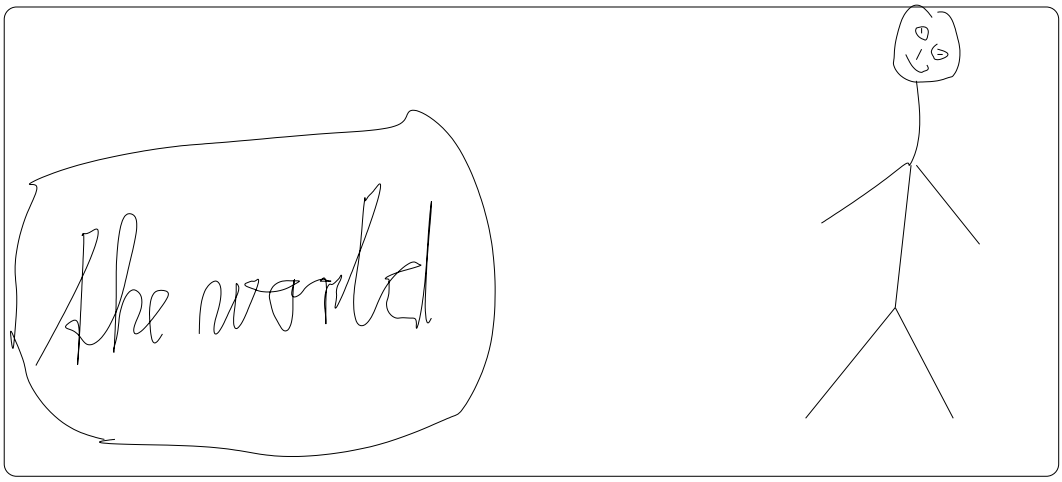
Between Truth and Description

Flemming Topsøe
University of Copenhagen
Department of Mathematics
topsoe@math.ku.dk

Presentation at Indian Institute of Technology
New Delhi, November 15th, 2006 *
(overheads will be posted on
www.math.ku.dk/~topsoe)

*presentation based on joint work with Peter Harremoës

Overview



Natures side Observers side (you!)

the game: basic ideas, philosophy
a reminder: Codes
Thm 0: Getting started: entropy, divergence
Thm A: MaxEnt principle derived from a game
Thm B: Observer “always” has an optimal strategy
Thm C: There is “almost always” equilibrium
Thm D: “Pythagorean” inequalities under equilibrium
Thm E: Possible Identification of optimal strategies
Thm F: Entropy loss and heavy tails
Zipf’s law: Stability and flexibility under entropy loss

Modelling the two sides: *Nature* **and** *Observer*

Viewed as a *game of conflict* with *Strategies*:
Nature chooses a *world* P among the *set of possible worlds*. Observer chooses a *code* κ.
(Code reflects observers *efforts* or *energy*).

The *cost function* c: $(P, \kappa) \curvearrowright c(P, \kappa)$ is a measure of *observation time*, *difficulty* or *complexity* seen from the point of view of Observer.

Two assumptions, a general and a special:
• Observer is a *minimizer*, Nature a *maximizer*.
Leads to *2-person zero-sum game*.
• A *Duality* $P \leftrightarrow \kappa$ provides a connection:
Code adapted to P and *world matching* κ (actually matching Observers expectations).
Technically: $\forall P^* \exists \kappa^* \, : \, c(P^*, \kappa^*) = \min_\kappa c(P^*, \kappa)$
and κ^* is unique (if cost is finite).

Key **example** involves *probability distributions* as worlds and *idealized codes* (tools of *description, representation* or *observation strategies*) as codes.

Philosophy

Modelling *asymmetric* 2-person zero sum games, one player having a mind, the other not. Observer has an optimal strategy, not necessarily so for Nature.

Why opposing goals?
Goal of Observer is clear, what about Nature? Will Nature react to choices of Observer, a mere human?

One view: No, Nature has no mind and has once and for all fixed the *laws of nature*. Nature is an absolute and we seek the *absolute truth*.

... But it is *you*, Observer, who model the world. Therefore, modelling of Nature and what *appears* as actions of Nature is not modelling the absolute but rather something which reflects
your knowledge about the world.

Nature then is another side of yourself. So you, Observer, are in conflict with another side of yourself, the side expressing knowledge – or absence of knowledge – about the world.

A reminder: Codes

alphabet $\mathbb{A}$	code-word	code-word length (κ)
a	11	2
e	00	2
i	01	2
o	100	3
u	1010	4
y	1011	4

To ensure unambiguous identification, code must be *prefix-free*: no code-word in the code-book must be the beginning of another. Denoting code-word lengths by κ_i ; $i \in \mathbb{A}$, with $\mathbb{A}$ denoting the *alphabet*, we realize that *Kraft's inequality*

$$\boxed{\sum_{i \in \mathbb{A}} 2^{-\kappa_i} \leq 1}$$

must hold – indeed, the binary subintervals of $[0, 1]$ which correspond, via successive bisections, to the various code-words must be pairwise disjoint, hence have total length at most 1. A converse is also true.

Coding letters in “A tale of two cities”

Letter	frequency		fixed length word	fixed length length	Huffman code word	Huffman code length	ideal length
a	47064	8.07 %	00000	5	1110	4	3.63
b	8140	1.40 %	00001	5	101111	6	6.16
c	13224	2.27 %	00010	5	01111	5	5.46
d	27485	4.71 %	00011	5	0110	4	4.41
e	72883	12.49 %	00100	5	000	3	3.00
f	13155	2.25 %	00101	5	111100	6	5.47
g	12120	2.08 %	00110	5	111101	6	5.59
h	38360	6.57 %	00111	5	1000	4	3.93
i	39786	6.82 %	01000	5	1010	4	3.87
j	622	0.11 %	01001	5	1111111110	10	9.87
k	4635	0.79 %	01010	5	11111110	8	6.98
l	21523	3.69 %	01011	5	10110	5	4.76
m	14923	2.56 %	01100	5	00111	5	5.29
n	41310	7.08 %	01101	5	1101	4	3.82
o	45118	7.73 %	01110	5	1100	4	3.69
p	9453	1.62 %	01111	5	101110	6	5.95
q	655	0.11 %	10000	5	1111111100	10	9.80
r	35956	6.16 %	10001	5	0010	4	4.02
s	36772	6.30 %	10010	5	1001	4	3.99
t	52396	8.98 %	10011	5	010	3	3.48
u	16218	2.78 %	10100	5	00110	5	5.17
v	5065	0.87 %	10101	5	1111110	7	6.85
w	13835	2.37 %	10110	5	01110	5	5.40
x	666	0.11 %	10111	5	1111111101	10	9.77
y	11849	2.03 %	11000	5	111110	6	5.62
z	213	0.04 %	11001	5	1111111111	10	11.42
total = 583.426		**100 %**	**mean = 5.00**		**mean = 4.19**		**H = 4.16**

Huffman $\approx$ *combinatorial entropy* (4.19 bits). Idealizing $\approx$ *entropy*. (4.16 bits). Theoretical units (nits rather than bits) corresponds to a change from base 2 to base e. Example also illustrates *redundancy*.

Prob., codes, entropy, redundancy and divergence

$(\mathbb{A}, M_+^1(\mathbb{A}))$: *alphabet*, with set of *prob. distributions.*
$(\mathbb{A}, K(\mathbb{A}))$: the set of *(idealized) codes* κ over $\mathbb{A}$, i.e. $\kappa : \mathbb{A} \to [0, \infty]$ satisfies $\sum_{i \in \mathbb{A}} \exp(-\kappa_i) = 1$

$P \leftrightarrow \kappa$ given by $\boxed{\kappa_i = -\ln p_i}$ or $\boxed{p_i = \exp(-\kappa_i).}$

$\langle \kappa, P \rangle = \sum_{i \in \mathbb{A}} \kappa_i p_i$: *average code length*

Definitions: $H(P) = \min_\kappa \langle \kappa, P \rangle$, the *entropy* of P,
$D(P\|\kappa) = \langle \kappa, P \rangle - H(P)$: *Red. (div)* btw. P and κ,
$D(P\|Q) = D(P\|\kappa)$ with $Q \leftrightarrow \kappa$: *Div.* btw. P and Q

$\langle \kappa, P \rangle = H(P) + D(P\|\kappa)$: *linking identity* or:
$\langle \kappa, P \rangle = H(P) + D(P\|Q$ with $Q \leftrightarrow \kappa$:

Theorem 0

$$H(P) = -\sum p_i \ln p_i \,,\; D(P\|Q) = \sum p_i \ln \frac{p_i}{q_i}$$

Code length game, MaxEnt, Optimal codes

$\mathcal{P} \subseteq M_+^1(\mathbb{A})$: the *preparation* (set of possible worlds).
Distributions in $\mathcal{P}$: *consistent worlds* (or distributions).

$\gamma(\mathcal{P})$: the *code length game*. Nature chooses $P \in \mathcal{P}$, Observer chooses κ, cost is average code length :
$c(P, \kappa) = \langle \kappa, P \rangle = \sum \kappa_i p_i$.

Optimal strategies:

- for Nature: consistent P^* with

$$\inf_{\kappa \in K(\mathbb{A})} \langle \kappa, P^* \rangle = \sup_{P \in \mathcal{P}} \inf_{\kappa \in K(\mathbb{A})} \langle \kappa, P \rangle$$

- for Observer: a *minimum risk code* κ^*:

$$\sup_{P \in \mathcal{P}} \langle \kappa^*, P \rangle = \inf_{\kappa \in K(\mathbb{A})} \sup_{P \in \mathcal{P}} \langle \kappa, P \rangle = \inf_{\kappa \in K(\mathbb{A})} R(\kappa | \mathcal{P})$$

Theorem A P^* optimal $\Leftrightarrow$ P^* *MaxEnt distribution*, the distribution in $\mathcal{P}$ with $H(P^*) = H_{max}(\mathcal{P})$.
Thus: *game leads to the Maximum Entropy principle*

Existence of optimal codes, minimax ineq.

Theorem B For every preparation $\mathcal{P}$, Observer has a unique optimal strategy, κ^* (given $R_{min}(\mathcal{P}) < \infty$).

Proof. Let $\overline{K}(\mathbb{A})$ be set of $\kappa : \mathbb{A} \to [0, \infty]$ with $\sum_{i \in \mathbb{A}} \exp(-\kappa_i) \leq 1$ in the topology of pointwise convergence. Extend previous definitions to $\overline{K}(\mathbb{A})$. Then $R(\cdot|\mathcal{P})$ is lower semi-continuous on the compact set $\overline{K}(\mathbb{A})$, hence assumes its minimal value. Clearly, minimum must be assumed for a $\kappa^* \in K(\mathbb{A})$. For uniqueness, apply geometric-arithmetic mean inequality to a mixture of two postulated optimal codes. □

By the general *minimax-inequality*,

$$\sup_{P \in \mathcal{P}} \inf_{\kappa \in K(\mathbb{A})} \langle \kappa, P \rangle \leq \inf_{\kappa \in K(\mathbb{A})} \sup_{P \in \mathcal{P}} \langle \kappa, P \rangle$$

or

$$H_{max} \leq R_{min} \, .$$

Equilibrium

If $H_{max} = R_{min}$: *equilibrium!* ("$<$" is possible)

Theorem C Assume $H_{max}(\mathcal{P}) < \infty$. Then
$\gamma(\mathcal{P})$ in equilibrium $\Leftrightarrow H_{max}(co\mathcal{P}) = H_{max}(\mathcal{P})$.

Proof. "$\Rightarrow$":
$H_{max}(co\mathcal{P}) \leq R_{min}(co\mathcal{P}) = R_{min}(\mathcal{P}) = H_{max}(\mathcal{P})$.
"$\Leftarrow$": To prove: $\gamma(co\mathcal{P})$ in equilibrium. Assume $\mathcal{P}$ convex. $(\kappa, P) \curvearrowright \langle \kappa, P \rangle$ on $\overline{K}(\mathbb{A}) \times M_+^1(\mathbb{A})$ is affine in each variable, lower semi-continuous in the first variable. By *Kneser's minimax theorem*,

$$\sup_{P \in \mathcal{P}} \min_{\kappa \in \overline{K}(\mathbb{A})} \langle \kappa, P \rangle = \min_{\kappa \in \overline{K}(\mathbb{A})} \sup_{P \in M_+^1(\mathbb{A})} \langle \kappa, P \rangle$$

As in proof of **B**, $\overline{K}(\mathcal{P})$ may be replaced by $K(\mathcal{P})$. Thus, $\gamma(\mathcal{P})$ is in equilibrium. $\square$

Properties under equilibrium

$(P_n)_{n\geq 1}$ *asymptotically optimal* if $(P_n)_{n\geq 1} \subseteq \mathcal{P}$ and $H(P_n) \to H_{max}(\mathcal{P})$. P^* (not necessarily in $\mathcal{P}$!) is *attractor* if $D(P_n\|P^*) \to 0$ for (P_n) asymptotically optimal. Need not exist. When it does, it is unique and $P_n \to P^*$ in total variation (by Pinsker's inequality).

Theorem D Any game $\gamma(\mathcal{P})$ in equilibrium has an attractor: Let κ^* be the optimal code. Then the attractor is the dist. P^* matching κ^*. Furthermore:

a: $H(P) + D(P\|P^*) \leq H_{max}(\mathcal{P})$ for all $P \in \mathcal{P}$,

b: $R_{min} + D(P^*\|\kappa) \leq R(\kappa|\mathcal{P})$ for all κ .

Proof. **a:** We have $\langle \kappa^*, P\rangle \leq R_{min}(\mathcal{P}) = H_{max}(\mathcal{P})$ for any $P \in \mathcal{P}$. Then use linking.
b: For $\kappa \in K(\mathbb{A})$ and (P_n) asymptotically optimal,

$$R(\kappa|\mathcal{P}) \geq \langle \kappa, P_n\rangle = H(P_n) + D(P_n\|\kappa)\,.$$

Then consider the limit as $n \to \infty$! □

Criteria enabling Identification

κ^* *robust* : $\langle \kappa^*, P \rangle$ finite and independent of $P \in \mathcal{P}$.
P^* *ess. consistent* : $\exists (P_n) \subseteq \mathcal{P} \ : \ D(P_n \| P^*) \to 0$.
P^* *ess. MaxEnt dist.*: attractor s.t. $H(P^*) = H_{max}$.
(κ^*, P^*) *opt. pair*: κ^* opt. code, P^* ess. MaxEnt dist.

Theorem E $(\kappa^* \, , \, P^*)$ given with $\kappa^* \leftrightarrow P^*$.
a: $R(\kappa^* | \mathcal{P}) \leq H(P^*) < \infty$, P^* ess. consistent $\Rightarrow \gamma(\mathcal{P})$ in equilibrium with optimal pair (κ^*, P^*).
b: κ^* robust, P^* consistent $\Rightarrow \gamma(\mathcal{P})$ in equilibrium with optimal pair (κ^*, P^*), P^* is even the unique MaxEnt distribution.

Proof. b follows from *a*. To prove *a:* Choose $(P_n) \subseteq \mathcal{P}$ such that $D(P_n \| P^*) \to 0$. By assumption,

$$R_{min}(\mathcal{P}) \leq R(\kappa^* | \mathcal{P}) \leq \limsup_{n \to \infty} \langle \kappa^*, P_n \rangle$$
$$= \limsup_{n \to \infty} H(P_n) + D(P_n \| P^*) = H_{max}(\mathcal{P}) \, . \square$$

Standard example: $\mathcal{P} = \{P | \langle E, P \rangle = \overline{E}\}$ (or "$\leq$") with E the *energy function*, e.g. on $\mathbb{A} = \mathbb{N}$.

A dialogue

S: Can you help me to identify the distribution behind some interesting data I am studying?
IT: OK, let me try. What do you know?
S: All observed values are non-negative integers.
IT: What else?
S: Well, I have reasons to believe that the mean value is 2.3.
IT: What more?
S: Nothing more.
IT: Are you sure?
S: I am!
IT: This then indicates the geometric distribution.

S: What! You are pulling my leg! This is a very special distribution and there are many, many other distributions which are consistent with my observations.

IT: Of course. But I am serious. In fact, any other distribution would mean that *you would have known something more.*

S: Hmmm. So the geometric distribution is the true distribution.

IT: I did not say that. The true distribution we cannot know about.

S: But what then did you say – or mean to say?

IT: Well, in more detail, certainty comes from observation. Based on your information, the best descriptor for you, until further observations are made, is the one adapted to the geometric distribution. In case you use any other descriptor there is a risk of a higher cost.

S: This takes the focus away from the phenomenon I am studying. Instead, you make statements about my behavior.

IT: Quite right. “Truth” and “reality” are human imaginations. All you can do is to make careful observations and reflect on what you see as best you can.

S: Hmmmm. You are moving the focus. Instead of all your philosophical talk I would like to think more pragmatically that the geometric distribution is indeed the true one. Then the variance should be about 7.6. I will go and check that.

IT: Good idea.

S: But what now if my data indicate a different variance?

IT: Well, then you would know something more, would you not? And I will change my opinion and point you to a better descriptor and tell you about the associated distribution in case you care to know.

S: But this could go on and on with revisions of opinion ever so often.

IT: Yes, but perhaps you should also consider what you are willing to know. Possibly I should direct you to a friend of mine, expert in complexity theory.

S: Good heavens no. Another expert! You have confused me sufficiently. But thanks for your time, anyhow. Goodbye!

Entropy loss

For a game $\gamma(\mathcal{P})$ in equilibrium with optimal code κ^*, matching distribution P^* we have $H(P^*) \leq H_{max}(\mathcal{P})$. If the inequality is strict, we have *entropy loss* or *collapse of the entropy function*. If equality holds, P^* is the essential MaxEnt distribution.

P^* has *potential entropy loss* if P^* is attractor for some model $\mathcal{P}$ in equilibrium with entropy loss. Requires ultra heavy tails! For convenience, take $\mathbb{A} = \mathbb{N}$.

P^* is *power-dominated* if, for some $a > 1$, $p_n^* \leq n^{-a}$, eventually. If P^* is *not* power-dominated, P^* is *hyperbolic*.

Example P with $p_n \approx \frac{1}{n(\ln n)^K}$ hyperbolic with finite entropy iff $K > 2$.

Theorem F P^* has potential entropy loss iff P^* is hyperbolic and $H(P^*) < \infty$.

Re Theorem **F**: Points to a potential for "generation" of entropy, almost contradicting the law of energy preservation. If a phenomenon is governed by a hyperbolic distribution P^*, this requires only finite "energy" , $H(P^*) = \langle \kappa^*, P^* \rangle$, but does lead to preparations $\mathcal{P}_h = \{\langle \kappa^*, P \rangle = h\}$ which operate at as high an "energy level" $H_{max} = h$ we wish.

This applies to *Zipf's law* and explains why a stable, yet flexible language is possible with a potential for unlimited expressive power.

Phenomena modelled by hyperbolic or other heavy-tailed distributions all seem to require high energies for their emergence. (creation of a language, of the internet, of large economies, of the universe (!) etc.).

Statistical handling of such phenomena is difficult, cf. Embrechts, Klüppelberg and Mikosch: "Modelling Extremal Events". Could it be that in some sense it is impossible to handle statistically data generated by a hyperbolic distribution?

Predicting the future!

The mixed game- and information theoretical ideas will be integrated in central parts of probability and statistics, thereby leading to a change of paradigm for these areas of science.

A main difficulty: Natural information theoretical proofs of basic limit theorems of probability theory and statistics must be in place for this development to take place. Though many results of this nature can be established, the central limit theorem still poses problems!