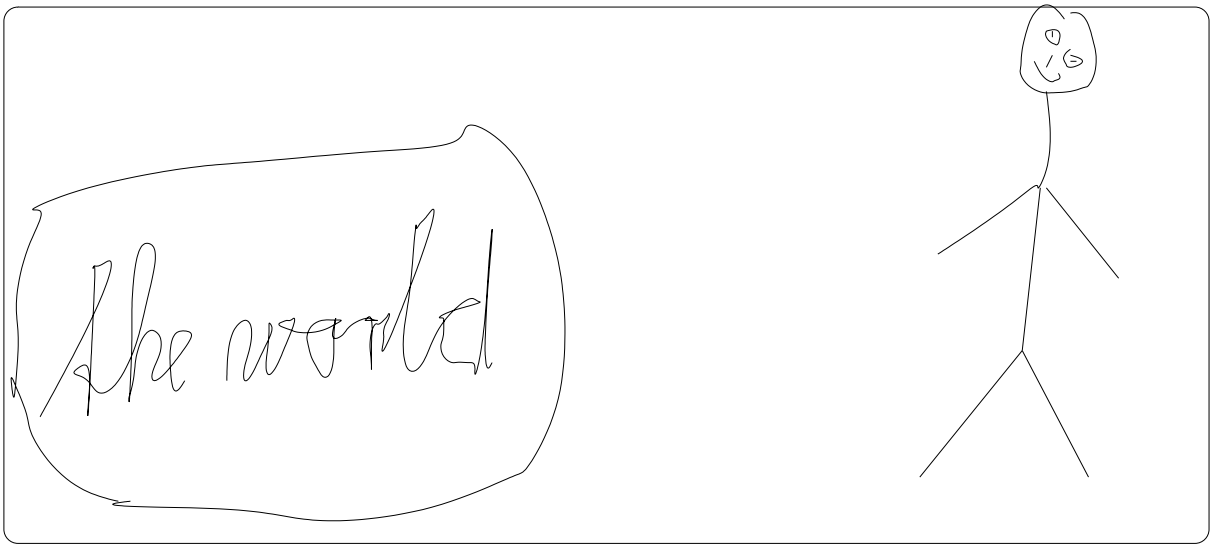

Between Truth and Description

Flemming Topsøe, University of Copenhagen
Department of Mathematics, topsoe@math.ku.dk
thanks to Peter Harremoës
- presentation based in part on
PH&FT:”maximum Entropy Fundamentals” and
PH&FT’s contribution to:
Handbook on the Philosophy of Information
support: Danish Natural Science Research Council
(overheads used are posted on
www.math.ku.dk/˜topsoe)

Overview

Nature and Observer

0: Getting started: entropy, divergence

A: MaxEnt principle derived from a game

B: Observer “always” has an optimal strategy

C: There is “almost always” equilibrium

D: “Pythagorean” inequalities under equilibrium

E: Possible Identification of optimal strategies

F: Entropy loss and heavy tails

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.

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 attempts to minimize, Nature to maxim. c.
Leads to *2-person zero-sum game*.
• A *Duality* $P \leftrightarrow \kappa$ provides a connection:
Code adapted to P and *world matching* κ
(P not necessarily ideal for Observers choice κ but matches Observers expectations as expressed by κ).
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 perhaps rather absence of knowledge about the world.

Prob., codes, entropy, redundancy and divergence

$(\mathbb{A}, M_+^1(\mathbb{A}))$: countable *alphabet*, set of pr. dist.

$(\mathbb{A}, K(\mathbb{A}))$: the set of *(idealized) codes* over $\mathbb{A}$, i.e. $\kappa : \mathbb{A} \to [0, \infty]$ such that $\sum_{i \in \mathbb{A}} \exp(-\kappa_i) = 1$

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

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

$\min_\kappa \langle \kappa, P \rangle$: 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 *Assume* $\gamma(\mathcal{P})$ *in equilibrium. Let* κ^* *be the optimal code. Then* P^* *matching* κ^* *is the attractor. Further*

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 : if $\langle \kappa^*, P \rangle$ 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 *(κ^* , 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 $H(P) < \infty \Leftrightarrow K > 2$.*

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

Proof. Assume first P^* is power-dominated. Clearly then, $H(P^*) < \infty$.

Consider $\gamma(\mathcal{P})$ in equilibrium with P^* as attractor. For the purpose of an indirect proof, assume that $H(P^*) < H_{max}(\mathcal{P})$.

Let $\kappa^* \leftrightarrow P^*$ and consider the *Dirichlet series*

$$Z(x) = \sum_{n \in \mathbb{N}} \exp(-x\kappa_n^*) \,.$$

For $Z(x) < \infty$, let Q_x be the distribution with point probabilities
$\exp(-x\kappa_n^*)/Z(x)$.

Choose $\beta < 1$ such that $H(P^*) < h < H_{max}(\mathcal{P})$ with $h = \langle \kappa^*, Q_\beta \rangle$. (possible since P^* is power-dominated, ...)

Consider preparation

$$\mathcal{P}_h = \{P \in M_+^1(\mathbb{N}) | \langle \kappa^*, P \rangle = h\} \,.$$

By construction, $\mathcal{P}_h \neq \emptyset$. A key part of the proof is to show that $H_{max} \geq h$

[We shall show that $\mathcal{P}_h$ is in equilibrium with P^* as attractor. To this end, consider an asymptotically optimal sequence $(P_k)_{k\geq 1}$ such that $H(P_k) > h$ for all $k \geq 1$. Then, as $\langle \kappa^*, P^* \rangle$ is less than and $\langle \kappa^*, P_k \rangle$ is larger than h, we can, for each $k \geq 1$, find a convex combination, say $Q_k = \alpha_k P^* + \beta_k P_k$, such that $Q_k \in \mathcal{P}_h$. By convexity of $D(\cdot \| P^*)$ and as $D(P_k \| P^*) \to 0$, we conclude from the linking identity applied to $\langle \kappa^*, Q_k \rangle$ that $\lim_{k\to\infty} H(Q_k) = h$, hence $H_{max}(\mathcal{P}_h) \geq h$.]

On the other hand, by the definition of $\mathcal{P}_h$, $R(\kappa^* | \mathcal{P}_h) = h$. It follows that $\gamma(\mathcal{P}_h)$ is in equilibrium with κ^* as optimal code, hence with P^* as attractor. However, by **b**, Theorem **E**, also Q_β is attractor for this preparation. As $Q_\beta \neq P^*$, we have a contradiction!

Now assume P^* is hyperbolic and $H(P^*) < \infty$. Fix $h > H(P^*)$ and consider $\mathcal{P}_h$. This works! $\square$

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* $\mathcal{P}_h$ *which operate at as high an "energy level"* H_{max} *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?

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.