# Randomness, complexity and information in metric spaces

Péter Gács

Computer Science Department
Boston University

Fall 07

# Martin-Löf's theory of randomness

(As presented by Levin). Let $X$ be the space $\Sigma^*$ of finite strings, or the space $\Sigma^\omega$ of infinite strings. Let $\mu$ be a probability measure over $X$. A test

$$t_\mu(x)$$

quantifies the nonrandomness of outcome $x \in X$ with respect to $\mu$. In Martin-Löf's theory, measure $\mu$ is assumed to be "computable" and fixed. Required:

- $\int t_\mu(x)\mu(dx) \leqslant 1$. (The measure of "non-random" objects is small.)
- $t$ is lower semicomputable in $x$. (Sooner or later we will recognize non-randomness.)

Test $u$ is universal if $\forall t\ \exists c > 0\ \forall x\ t_\mu(x) > c \cdot u_\mu(x)$.

**Theorem 1**

*There is a universal test $\tilde{\mathbf{t}}_\mu(x)$.*

# Test in terms of complexity

I assume familiarity with description (Kolmogorov) complexity. Let $X = \Sigma^*$. For $x \in X$, denote the complexity (the prefix version) of $x$ by

$$H(x)$$

(same as $K(x)$ in Li-Vitányi). Let $\tilde{\mathbf{d}}_\mu(x) = \log \tilde{\mathbf{t}}_\mu(x)$, called the deficiency of randomness of $x$ with respect to $\mu$.

## Theorem 2

*Over the set of finite strings,*

$$\tilde{\mathbf{d}}_\mu(x) \stackrel{+}{=} -\log \mu(x) - H(x).$$

*Over the set of infinite strings,*

$$\tilde{\mathbf{d}}_\mu(x) \stackrel{+}{=} \sup_n \; -\log \mu(x_{\leqslant n}) - H(x_{\leqslant n}).$$

Constants in $\stackrel{+}{=}$ depend on $\mu$.

# Conservation of randomness

For a computable function $f : X \to Y$, and probability measure $\mu$ over $X$, define the output distribution $f^*\mu$ over $Y$ by

$$(f^*\mu)(y) = \mu(f^{-1}(x)).$$

If $\mu$ is computable then it can be seen that $f^*\mu$ is also computable. The following theorem implies that if $x$ is random with respect to $\mu$ then $f(x)$ is random with respect to $f^*\mu$:

## Proposition 3

$$\tilde{\mathbf{d}}_{f^*\mu}(f(x)) \stackrel{+}{<} \tilde{\mathbf{d}}_{\mu}(x).$$

# "Apriori probability"

If $\mathbf{m}(x) = 2^{-H(x)}$ is treated as a measure, then

$$\tilde{\mathbf{d}}_{\mathbf{m}}(x) \stackrel{+}{=} -\log \mathbf{m}(x) - H(x) = 0$$

shows that all strings are random with respect to $\mathbf{m}$.
But: $\mathbf{m}$ is not a probability measure (only a "semimeasure"), and is not computable (only lower semicomputable).
Still: this idea (over infinite sequences) is used in inductive inference (Solomonoff).

# Arbitrary measures: uniform tests

Restriction to computable measures $\mu$ is unnatural (it is particularly baffling to probabilists). How to extend the definition to arbitrary measures? Idea: just use (over $X = \Sigma^*$):

$$-\log \mu(x) - H(x).$$

Alas, this test does not conserve randomness (easy counterexample). New idea (following early work of Levin): test $t_\mu(x)$:

1. $\int t_\mu(x)\mu(dx) \leqslant 1$.
2. $t$ is lower semicomputable in $(\mu, x)$.

To make sense of 2 equip the space of measures with a computability structure. Levin has done this for some compact spaces (like infinite binary sequences).

# Levin's uniform tests

With appropriately defined notion of test, claims:

- Universal uniform test $\mathbf{t}_\mu(x)$; let $\mathbf{d}_\mu(x) = \log \mathbf{t}_\mu(x)$.
- Randomness conservation
- Neutral measure $M$: for all every $x$ we have $\mathbf{t}_M(x) \leqslant 1$ ("apriori probability").
- Lower semicomputable neutral semimeasure Semimeasures (semi-additive measures) are introduced; there is a lower semicomputable semi-measure $M$ that is neutral (and universal).
- Information $I(x:y)$ (appropriately defined) is essentially equal to $\mathbf{d}_{M\times M}(x,y)$: "defect of independence". This allows information conservation to be proved as special case of randomness conservation.

## My goals

- Natural, general definition of test (achieved). ?
- Non-compact spaces, too (achieved).
- Expressing the test via complexity (partial success).
- See which of Levin's results survive:
  - Universal uniform test yes.
  - Randomness conservation yes.
  - Neutral measure yes.
  - Neutral l.sc. semimeasure no, not even in the compact case or for finite strings.
  - Information conservation from randomness conservation: ?.

# Constructive topology

Computability extended: instead of only about random strings, to speak of random real numbers, even about a random path of the Brownian motion (non-compact space). (For the special case of Brownian motion the concept has been worked out already by Asarin.) I assume familiarity with topological spaces. For the constructive version, I (essentially) follow Weihrauch et al.

Constructive topological space:

$$\mathbf{X} = (X, \beta, \nu),$$

where $X$ is the underlying set, $\beta$ is a basis of open neighborhoods, $\nu$ is an *enumeration* of $\beta$: $\beta = \{\nu(1), \nu(2), \dots\}$.

Open set: a union of basis elements. R.e. open set: a union of a r.e. set of basis elements.

# Computable functions

Let $f : X \to Y$.
Continuous: $f^{-1}(V)$ is open for all basis elements $V \subseteq Y$.
Computable: $f^{-1}(V)$ is r.e. open, uniformly in the enumerated basis elements $V$.
Lower semicomputable: a constructive version of "lower semicontinuity": the set

$$\{ (x, r) : f(x) > r \}$$

is a r.e. open subset of $X \times \mathbb{Q}$.
Point $x \in \mathbf{X}$ is computable if the constant function $0 \mapsto x$ is.
Conditional description complexity $H(x \mid y)$ can be generalized to the case where $y$ is coming from a computable topological space. The interpreter function used in the definition must be computable in $y$.

# Computable metric space

$\mathbf{X} = (X, d, D, \alpha)$.
$d$ is a distance function over $X$.
$D \subseteq X$ is countable, dense (so, $\mathbf{X}$ is separable).
$\alpha$ is an enumeration of $D$.
Condition: $d(x, y)$ is computable for $x, y \in D$.
A computable metric space is automatically a constructive topological space. Basic balls: balls with center in $D$ and rational radius.
The space $X$ is effectively compact if for every $k$ one can compute a covering of $X$ by basic balls.

# Topology of measures

I assume familiarity with measures, defined on the Borel sets of a topological space $\mathbf{X}$. We will always require $\mathbf{X}$ to be a complete computable metric space.

Weak convergence: $\mu_i \to \mu$ if $\mu_i f \to \mu f$ for all bounded continuous functions $f$.

Example (Dirac delta): $\delta_{x_i} \to \delta_x$ if $x_i \to x$.

Prokhorov distance: $p(\mu, \nu) = \inf\{\, \varepsilon : \forall$ Borel $A,\ \nu A^{\varepsilon} < \mu A + \varepsilon \,\}$.

Wasserstein distance: $W(\mu, \nu) = \inf_{f \in \mathrm{Lip}} |\mu f - \nu f|$.

Dense set of measures: finite rational combinations of measures of form $\delta_x$ for $x \in D$.

This turns the set of probability measures into a computable metric space $\mathbf{M}(\mathbf{X})$.

### Theorem 4

*If $f : X \to \mathbb{R}$ is computable then $\mu \mapsto \mu f$ is computable.*

But, for example, for any ball $B = B(x, r)$, $(x \in D, r \in \mathbb{Q})$, the function $\mu \mapsto \mu(B)$ is not computable. Let $B_i$ be an enumeration of all basic balls.

### Theorem 5 (Hoyrup, Rojas)

*Measure $\mu$ is computable if and only if the function*

$$\langle i_1, \dots, i_k \rangle \mapsto \mu(B_{i_1} \cup \dots \cup B_{i_k})$$

*is lower semicomputable.*

# Uniform tests

- $\int t_\mu(x)\mu(dx) \leqslant 1$.
- $t$ is lower semicomputable in $(\mu, x)$.

**Theorem 6 (Hoyrup, Rojas)**

*There is a universal uniform test* $\mathbf{t}_\mu(x)$.

(I had this theorem only under a certain condition on the space.)

Randomness with respect to computable measures has certain—intuitively meaningful—monotonicity and convexity:

- $\mu \leqslant c\nu$ implies $\mathbf{t}_\nu(x) \stackrel{*}{<} c\mathbf{t}_\mu(x)$.
- $\mu = \frac{1}{n}\sum_{i=1}^n \mu_i$ implies $\mathbf{t}_\mu \stackrel{*}{>} \min_i \mathbf{t}_{\mu_i}$.

These properties do not survive for the uniform test: let $\mu_0$ be uniform over $[0, 1]$, and $\mu_1$ uniform over $[0, 1/2]$, $\mu_2$ uniform over $[1/2, 1]$. Let $p < 1/2$ be random with respect to $\mu_0$, let $\nu_1 = p\mu_1 + (1-p)\mu_2$, and $\nu_2 = (1-p)\mu_1 + p\mu_2$. Then $p$ is not random with respect to either $\nu_1$ or $\nu_2$, but Then $\mu_0 \leqslant p^{-1}\nu_1$ and also $\mu_0 = (\nu_1 + \nu_2)/2$.

# Randomness conservation

**Theorem 7**

*Let $f : X \to Y$ be computable. Then*

$$\mathbf{d}_{f^*\mu}(f(x)) \stackrel{+}{<} \mathbf{d}_\mu(x).$$

There is a more general theorem, for computable random transitions.

# Relation to complexity
## Nicest cases

### Theorem 8

*If* $\mathbf{X}$ *is discrete,*

$$\mathbf{d}_\mu(x) \stackrel{+}{=} -\log \mu(x) - H(x \mid \mu).$$

For other spaces, we do not have a nice characterization for the uniform tests, so we assume that $\mu$ is computable.

### Theorem 9

*On the space of infinite sequences, for computable measure* $\mu$ *we have*

$$\mathbf{d}_\mu(x) \stackrel{+}{=} \sup_n \; -\log \mu(x_{\leqslant n}) - H(x_{\leqslant n}).$$

# Cells

Sometimes other spaces can be mapped to the space of infinite sequences.
For a computable sequence of functions $b_1, b_2, \dots$, with $b_i : X \to \mathbb{R}$, let

$$\Phi_{i,0} = \{\, x \in X : b_i(x) < 0 \,\},$$
$$\Phi_{i,1} = \{\, x \in X : b_i(x) > 0 \,\}.$$

We say that the sequence $\{b_i\}$ is separating if $x_1 \neq x_2$ implies $\exists j\ b_j(x_1) \cdot b_j(x_2) < 0$.
It is isolating if the nonempty finite intersections of the sets $\Phi_{i,j}$ form an enumerated basis computationally equivalent to the canonical one.
An isolating sequence is always separating.

**Theorem 10**

*If the space is effectively compact then a separating sequence is also isolating.*

Fix a separating sequence $\{b_i\}$, let

$$X^0 = \{\, x \in X : b_j(x) \neq 0,\ j = 1, 2, \dots \,\}.$$

For $x \in X^0$ let

$$\sigma_i(x) = j \text{ if } x \in \Phi_{i,j},$$
$$\sigma_{[n]}(x) = (\sigma_1(x), \dots, \sigma_n(x)).$$

For a binary string $s_1 \cdots s_n = s$, we define the $n$-cell

$$\Gamma(s) = \Gamma_n(x) = \{\, x : \sigma_{[n]}(x) = s \,\}.$$

If $\{b_i\}$ is isolating then the nonempty sets $\Gamma(s)$ form an enumerated basis over the subspace $X^0$.
On the set $X^0$, the cells behave somewhat like binary subintervals: they divide $X^0$ in half, then each half again in half, etc.

A measure $\mu$ is regular for the sequence $\{b_i\}$ if $\mu(X^0) = 1$.

## Theorem 11

*If the sequence $\{b_i\}$ is isolating, the measure $\mu$ is computable and regular for $\{b_i\}$ then for $x \in X_0$ we have*

$$\mathbf{d}_\mu(x) \stackrel{+}{=} \sup_n \; -\log \mu(\Gamma_n(x))) - H(\Gamma_n(x)).$$

*For $x \in X \setminus X^0$ we have $\mathbf{d}_\mu(x) = \infty$.*

## Question 1

*Find a nice characterization for the general uniform test in terms of complexity.*

# Neutral measure

## Theorem 12 (Levin)

*If $\mathbf{X}$ is compact then there is a measure $M$ with the property that for all $x$, $\mathbf{t}_M(x) \leqslant 1$.*

(Proof using Sperner's Lemma.)

Noncompact spaces? No. The discrete space $\mathbf{X} = \mathbb{N}$ has no neutral measure. But, we can compactify $\mathbb{N}$. A neutral measure $M$ over $\overline{\mathbb{N}} = \mathbb{N} \cup \infty$ is only a semimeasure over $\mathbb{N}$.

Is there a neutral measure with some nice computability property?

## Theorem 13

*No neutral measure over $\overline{\mathbb{N}}$ is lower semicomputable or upper semicomputable over $\mathbb{N}$.*

# Information

## Relative algorithmic entropy

$$H_\nu(x) = -\mathbf{d}_\nu(x)$$

is a generalization of complexity (algorithmic entropy). Indeed, generalizing to non-probability measures $\nu$ (example: the counting measure #)

$$H_\#(x) \stackrel{+}{=} H(x).$$

This is in analogy to the definition of relative (information-theoretical) entropy of $\mu$ with respect to $\nu$,

$$\mathcal{H}_\nu(\mu) = -\int \log \frac{d\mu}{d\nu} d\mu,$$

(which is the negative of the so-called Kullback distance). Special cases: $\nu = \#$ gives ordinary entropy. For $\nu =$ Lebesgue measure gives $-\int f(x) \log f(x) dx$.

# Addition theorem

Let us generalize the well-known addition theorem

$$H(x,y) \stackrel{+}{=} H(x) + H(y \mid x,\ H(x)).$$

## Theorem 14 (General Addition)

$$H_{\mu\times\nu}(x,y) \stackrel{+}{=} H_{\mu}(x \mid \nu) + H_{\nu}(y \mid x,\ H_{\mu}(x \mid \nu),\ \mu).$$

The proof is somewhat subtle.

## Question 2

*Applications?*

# Information

Classical information between two random variables $X, Y$ with distribution $\mu_{X,Y}$ is

$$\mathcal{I}(X:Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y) \tag{1}$$
$$= -\mathcal{H}_{\mu_X \times \mu_Y}(\mu_{X,Y}). \tag{2}$$

(2) is the Kullback distance of $\mu_{X,Y}$ from the product $\mu_X \times \mu_Y$.
The analog of (1) is the algorithmic mutual information

$$I(x:y) = H(x) + H(y) - H(x,y).$$

If we defined deficiency of randomness as

$$\overline{\mathbf{d}}_\mu(x) = -\log \mu(x) - H(x),$$

then the analog of (2) is

$$I(x:y) = \overline{\mathbf{d}}_{\mathbf{m}\times\mathbf{m}}(x,y) = -\overline{H}_{\mathbf{m}\times\mathbf{m}}(x,y),$$

which can be seen as the deficiency of independence between $x$ and $y$.

Alas, we had to discard $\overline{\mathbf{d}}$.

## Question 3

*Is* $I(x : y) = \mathbf{d}_{M \times M}(x, y)$ *with some neutral measure* $M$ *over* $\overline{\mathbb{N}}$*?*

Levin's use of a similar formula allowed him to derive his information conservation inequality from randomness conservation.

## Question 4

*What is the natural definition of information (having the best properties) in the continuous case?*

A candidate for the case with cells is

$$I(x : y) = \sup_{m,n} I(\sigma_{[m]}(x) : \sigma_{[n]}(y)).$$

Other possibility, with underlying measures $\mu, \nu$:

$$I_{\mu,\nu}(x : y) = H_{\mu}(x \mid \nu) + H_{\nu}(y \mid \mu) - H_{\mu \times \nu}(x, y).$$

How much does this depend on $\mu, \nu$? What if we use a neutral measure here?

# Entropy in physics

## The model

Assume that our system is that of classical mechanics, a dynamical system with $X$ as a state space (phase space, configuration space), and a dynamic

$$x \mapsto U^t x,$$

where time $t$ is discrete or continuous. We have a measure $L$ invariant under $U^t$ (think of Liouville's theorem).
We assume an isolating set of functions $\{b_i\}$ and so will speak about cells. Asssume the functions $b_i$ arranged in decreasing order of interest. At the beginning are some "macroscopic" ones like temperature, pressure, then come pressure, concentrations in the different compartments of space, and so on.
When the data of interest have been specified we arrive at a cell $\Gamma_n(x)$, a coarse-grained description of the system.

## Example 15 (The baker's map)

$X$ = the set of doubly infinite binary sequences $x = \dots x_{-1}x_0x_1x_2\dots$ with the shift transformation $(U^t x)_i = x_{i+t}$ over discrete time. Let $x^n = x_{-\lfloor n/2 \rfloor} \cdots x_{\lceil n/2 \rceil - 1}$. The $n$-cells are of the form $\Gamma_n(x) = \Gamma(x^n)$, with volume

$$L(\Gamma(x^n)) = 2^{-n}.$$

# Boltzmann entropy

Boltzmann defined entropy as

$$\log L(\Gamma_n(x)).$$

This definition is, of course, dependent on the choice of the functions $b_i$ and the fineness $n$ of the partition. In practice these variations are negligible compared to $\log L(\Gamma_n(x))$ (of the order of $10^{23}$).
One way of expressing the second law of thermodynamics is to say that in typical systems of physics, entropy

$$\log L(\Gamma_n(U^t x))$$

increases over time, until it reaches its maximum near $\log L(X)$. This can only be true in a statistical sense, requires some strong mixing properties of the map $U^t$.

## "Physical" entropy

For the baker's map (otherwise a very nicely mixing system), all $n$-cells have the same measure no matter what fixed precision we choose, so the volume $L(\Gamma_n(U^t z))$ is constant in $t$. This suggests difficulties with Boltzmann's definition.
Using some more interesting considerations, Zurek recommended a quantity similar to

$$H^n(x) = \log L(\Gamma_n(x)) + H(\Gamma_n(x)),$$

calling it "physical entropy"; we call it coarse-grained algorithmic Boltzmann entropy. So, we add the complexity of the cell to the logarithm of its size.
For typical applications in classical physics, the correction term is negligible.
In the baker's map, it can be shown that $H^n(x)$ increases fast to its maximum (which is 0) for almost all sequences.

The function $H^n(x)$ depends only moderately on the choice of the functions $b_i$ and on $n$. Indeed, let

$$H_L(x) = -\mathbf{d}_L(x)$$

be the relative algorithmic entropy of $x$ with respect to $L$ (a finite measure now). We will call it fine-grained entropy in the physical context.
The theorem characterizing the randomness defect in terms of complexity translates to

$$H_L(x) = \inf_n L(\Gamma_n(x)) + H(\Gamma_n(x)) = \inf_n H^n(x).$$

So $H^n(x)$ can be viewed as the $n$th approximation of the fine-grained entropy $H_L(x)$. On the other hand, the latter is essentially invariant with respect to the choice of $b_i$ and $n$.

## Question 5

*Prove the increase of $H^n(x)$ for some interesting maps $U^t$. Maybe some hyperbolicity properties of the map suffice.*