2012 IEEE 53rd Annual Symposium on Foundations of Computer Science

Rarity for Semimeasures.

Leonid A. Levin
Boston University
http://www.cs.bu.edu/fac/Lnd

Abstract—**The notion of Kolmogorov-Martin-Löf Random sequences is extended from computable to enumerable distributions. This allows definitions of various other properties, such as mutual information in infinite sequences. Enumerable distributions (as well as distributions faced in some finite multi-party settings) are semimeasures; handling those requires care.**

I. Introduction.

This is just an extended abstract conforming to IEEE policies. I suggest reading [L 12] first. [Solomonoff 64], [Kolmogorov 65] noted that many characteristics of finite objects, such as their complexity (the shortest description length) can be defined invariantly: their dependence on the programming language is limited to an additive constant. This led to the development of very robust concepts of randomness, information, etc. intrinsic to objects themselves, not to the mechanism that supposedly generated them.

These concepts are easy to define for for integers; the case of emerging objects, such as prefixes x of other (possibly infinite) sequences α is more subtle. While x can be encoded as integers, the code carries more information than x themselves. The information in x is a part of information in α, *i.e.,* is nondecreasing in extensions. The code of x has an extra information about the (arbitrary) cut-off point, not intrinsic to the α, and thus distortive.

Per Martin-Löf extended the concept of randomness and its deficiency (***rarity***) to prefixes of infinite sequences, assuming their probability distribution is computable. Yet, many important distributions are only lower-enumerable (r.e.). For instance, universal probability $\mathbf{M}$ is the largest within a constant factor *r.e.* distribution. While all sequences are random with respect to it, it has derivative distributions with more informative properties. In particular, Mutual Information in two sequences is their ***dependence***, *i.e.,* rarity with respect to the distribution generating them independently with universal probability each.

The purpose of this article is to extend the concept of sequence rarity to *r.e.* distributions. The definition proposed respects the randomness conservation laws and is the strongest (*i.e.,* largest) possible among such definitions. Among applications of this concept is the definition of mutual information in infinite sequences and their prefixes.

Enumerable distributions are of necessity semimeasures: infimums of sets of measures. They are essential for handling algorithms that have no time limit and so can diverge. However the benefits of semimeasures are not limited to this use. They make a good description of widespread situations where the specific probability distribution is unknown (*e.g.,* due to interaction with a party that cannot be modeled).

II. Conventions and Background.

Let $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{N}$, $B = \{0,1\}$, $S = B^*$, $\Omega = B^{\mathbb{N}}$ be, respectively, the sets of reals, rationals, integers, bits, finite, and infinite binary sequences; $x_{[n]}$ is the n-bit prefix and $\|x\|$ is the bit-length of $x{\in}S$. A real function f and its values are ***enumerable*** or *r.e.* ($-f$ is co-r.e.) if its subgraph $\{(x,q) : f(x) > q \in \mathbb{Q}\}$ is. X^+ means $X \cap \{x \geq 0\}$. ***Elementary*** ($f{\in}\mathcal{E}$) are functions $f : \Omega \to \mathbb{Q}$ depending on a finite number of digits; $\mathbf{1}{\in}\mathcal{E}$ is their unity: $\mathbf{1}(\omega) = 1$. $\overline{\mathcal{E}}$ is the set of all lower semicontinuous functions $\Omega \to \mathbb{R}$. When unambiguous, I identify objects in clear correspondence: *e.g.,* prefixes with their codes or their sets of extensions, sets with their characteristic functions, etc. ***Majorant*** is an *r.e.* function largest, up to a constant factor, among *r.e.* functions in its class.

A. Integers: Complexity, Randomness, Rarity.

Let us define Kolmogorov ***complexity*** $\mathbf{K}(x)$ as $\lceil -\log \mathbf{m}(x) \rceil$ where $\mathbf{m} : \mathbb{N} \to \mathbb{R}$ is the ***uni-***

0272-5428/12 $26.00 © 2012 IEEE
DOI 10.1109/FOCS.2012.50

versal measure, *i.e.,* a majorant *r.e.* function with $\sum_x \mathbf{m}(x) \leq 1$. It was introduced in [ZL 70], and noted in [L 73], [L 74], [Gacs 74] to be a modification (restriction to self-delimiting codes) of the least length of binary programs for x defined in [Kolmogorov 65]. While technically different, $\mathbf{m}$ relies on intuition similar to that of [Solomonoff 64]. The proof of the existence of the largest function was a straightforward modification of proofs in [Solomonoff 64], [Kolmogorov 65] which have been a keystone of the informational complexity theory.

For $x{\in}\mathbb{N}, y{\in}\mathbb{N}$ or $y{\in}\Omega$, similarly, $\mathbf{m}(\cdot|\cdot)$ is the largest *r.e.* real function with $\sum_x \mathbf{m}(x|y) \leq 1$; $\mathbf{K}(x|y) \overset{\text{df}}{=} \lceil -\log \mathbf{m}(x|y) \rceil$ (and is the least length of self-delimiting programs transforming y into x).

[Kolmogorov 65] considers ***rarity*** $\mathbf{d}(x) \overset{\text{df}}{=} \|x\| - \mathbf{K}(x)$ of uniformly distributed $x \in B^n$. Our modified $\mathbf{K}$ allows extending this to other measures μ on $\mathbb{N}$. A μ-test is $f : \mathbb{N} \to \mathbb{R}$ with mean $\mu(f) \leq 1$ (and, thus, small values $f(x)$ on randomly chosen x). For computable μ, a majorant *r.e.* test is $\mathbf{m}(x)/\mu(x)$. This suggests defining $\mathbf{d}(x|\mu)$ as $|\log \mu(x)| - \mathbf{K}(x) = \lfloor \log(\mathbf{m}(x)/\mu(x)) \rfloor \pm O(1)$.

B. Integers: Information.

In particular, $x = (a, b)$ distributed with $\mu = \mathbf{m} \otimes \mathbf{m}$, is a pair of two independent, but otherwise completely generic, finite objects. Then, $\mathbf{I}(a : b) \overset{\text{df}}{=} \mathbf{d}((a, b)|\mathbf{m} \otimes \mathbf{m}) = \mathbf{K}(a) + \mathbf{K}(b) - \mathbf{K}(a, b)$ measures their ***dependence*** or ***mutual information***. It was shown (see [ZL 70]) by Kolmogorov and Levin to be close (within $\pm O(\log \mathbf{K}(a, b))$) to the expression $\mathbf{K}(a) - \mathbf{K}(a|b)$ of [Kolmogorov 65]. Unlike this earlier expression (see [Gacs 74]), our $\mathbf{I}$ is symmetric and monotone: $\mathbf{I}(a : b) \leq \mathbf{I}((a, a') : b) + O(1)$ (which will allow extending $\mathbf{I}$ to Ω); it equals $\mathbf{K}(a) - \mathbf{K}(a|(b, \mathbf{K}(b))) \pm O(1)$ and satisfies the following Independence Conservation Inequalities [L 74], [L 84]: For any computable transformation A and measure μ, and some family $t_{a,b}$ of μ-tests

$$\mathbf{I}(A(a) : b) \leq \mathbf{I}(a : b) + O(1),$$

$$\mathbf{I}((a, w) : b) \leq \mathbf{I}(a : b) + \log t_{a,b}(w) + O(1).$$

(The $O(1)$ error terms reflect the constant complexities of A, μ.) So, independence of a from b is preserved in random processes, in deterministic computations, their combinations, etc. These inequalities are not obvious (and false for the original 1965 expression $\mathbf{I}(a : b) = \mathbf{K}(a) - \mathbf{K}(a/b)$) even with A, say, simply cutting off half of a. An unexpected aspect of $\mathbf{I}$ is that x contains all information about $k = \mathbf{K}(x)$, $\mathbf{I}(x : k) = \mathbf{K}(k) \pm O(1)$, despite $\mathbf{K}(k|x)$ being $\sim\|k\|$ or $\sim \log \|x\|$, in the worst case [Gacs 74]. One can view this as an "Occam Razor" effect: with no initial information about it, x is as hard to obtain as its simplest (k-bit) description.

All the above works as well for the $\mathbf{I}_z$ variation of $\mathbf{I}$ allowing all algorithms access to oracle z.

C. Reals: Measures and Rarity.

A measure on Ω is a function $\mu(x) = \mu(x0) + \mu(x1)$, for $x{\in}S$. Its mean $\mu(f)$ is a functional on $\mathcal{E}$, linear: $\mu(cf + g) = c\mu(f) + \mu(g)$ and ***normal:*** $\mu(\mathbf{1}) \leq 1$, $\mu(\mathcal{E}^+) {\subset} \mathbb{R}^+$. It extends to other functions, as usual. μ-tests are functions $f \in \overline{\mathcal{E}}$, $\mu(f) \leq 1$; computable μ have ***universal*** (*i.e.,* majorant *r.e.*) Martin-Löf tests $\mathbf{T}_\mu(\alpha) = \sum_i \mathbf{m}(\alpha_{[i]})/\mu(\alpha_{[i]})$. ***Random*** are α of rarity $\mathbf{d}(\alpha|\mu) \overset{\text{df}}{=} \lfloor \log(1 + \mathbf{T}_\mu(\alpha)) \rfloor < \infty$.

Continuous transformations $A : \Omega {\to} \Omega$ induce normal linear operators $A^* : f {\mapsto} g$ over $\mathcal{E}$, where $g(\omega) = f(A(\omega))$. So obtained, A^* are ***deterministic***: $A(\min\{f, g\}) = \min\{A(f), A(g)\}$. Operators that are not, correspond to probabilistic transformations (their inclusion is the benefit of the dual representation), and $g(\omega)$ is then the expected value of $f(A(\omega))$. Such A also induce A^{**} transforming input distributions μ to output distributions $\varphi = A^{**}(\mu) : \varphi(f) = \mu(A^*(f))$.

To avoid congestion, I often omit the *, identifying A with A^*, A^{**}, and $\omega{\in}\Omega$ in their inputs with measures $\mu : f \mapsto f(\omega)$. Same for partial transformations below and their concave duals.

D. Partial Operators, Semimeasures, Complexity of Prefixes.

Algorithms are not always total: focusing output to a single sequence may go slowly and fail.

Definition 1: 1) *Partial* continuous transformations (PCT) are compact subsets $A \subset \Omega {\times} \Omega$ with $A(\alpha) = \{\beta : (\alpha, \beta){\in}A\} \neq \emptyset$. If $A(\alpha)$ is singleton $\{\omega\}$, I identify it with $\omega{\in}\Omega$.

2) Dual of PCT A is the operator A^* mapping $f{\in}\mathcal{E}$ to $g{\in}\overline{\mathcal{E}}$, where $g(\alpha) = \min_{\beta \in A(\alpha)} f(\beta)$.

PCT turn input measures φ into ***semimeasures*** that map $f{\in}\mathcal{E}$ on outputs of A to their mean:

Definition 2: 1) A semimeasure μ is a functional that is normal: $\mu(-\mathbf{1}) \geq -1$, $\mu(\mathcal{E}^+) \subset \mathbb{R}^+$, and concave: $\mu(cf{+}g) \geq c\mu(f){+}\mu(g)$, $c \in \mathbb{Q}^+$ (*e.g.,* $\mu(x) \geq \mu(x0){+}\mu(x1)$, for $x{\in}S$). μ extends beyond $\mathcal{E}$ as is usual for internal measures. μ is ***deterministic*** if $\mu(\min\{f,g\}) = \min\{\mu(f),\mu(g)\}$, and ***binary*** if $\mu(f^3) = (\mu(f))^3$, $\mu(\mathbf{1}) = 1$.

2) Concave normal operators $A : \mathcal{E}^+{\to}\overline{\mathcal{E}}^+$ transform input points ω and input distributions (measures or semimeasures) φ into their output distributions $\mu = A(\varphi)$, where $\mu(f) = \varphi(A(f))$. Operators A are deterministic or binary if semimeasures $A(\omega)$ are.

Proposition 1: Operators A^* dual of PCT are concave, normal, deterministic, and binary. Each such A^* is a dual of a PCT.

Proposition 2: There exists a ***universal*** *i.e.,* majorant (on $\mathcal{E}^+$) r.e., semimeasure $\mathbf{M}$.

[ZL 70] used a this $\mathbf{M}$ to define ***complexity*** $\mathbf{KM}(x)$ of prefixes x of $\alpha{\in}\Omega$ as $\lceil -\log \mathbf{M}(x)\rceil$.

III. Rarity

Coarse Graining. We use $\lambda(x) = 2^{-\|x\|}$ as a typical continuous computable measure, though any of them could be used instead. Some considerations require reducing semimeasures to smaller linear functionals, *i.e.,* measures. Thus, restricting inputs ω of a PCT A to those with a singleton output $A(\omega){\in}\Omega$, results in a maximal measure $\mu_1 \leq \mu = A(\lambda)$. However much information is lost this way, *e.g.,* some computable A have no recursive in $1/\mu(x)$ bound on $1/\mu_1(x), x{\in}S$. To preserve information about finite prefixes of $\omega{\in}\Omega$, we will require linearity of μ_1 only on a subspace of $\mathcal{E}$. Thus, restricting inputs just to those that result in at least n-bit output produces a distribution μ_1 that is linear only on a subspace of all functions $f(\alpha)$ in $\mathcal{E}$ that depend only on $\alpha_{[n]}$. Such subspaces $\hat{E}$ must be ***lattices*** (*i.e.,* closed under $\min\{f,g\}$) for the greatest μ_1 to exist.

E-measures are semimeasures linear on the lattice vector subspace $\hat{E}$ generated by $E \subset \mathcal{E}$. From [Choquet, Meyer 63] one can derive:

Lemma 1: Each semimeasure μ, for each E, has the largest (on $\hat{E}^+$) E-measure $\mu_E \leq \mu$.

For convenience we will consider only $\hat{E}$ including constants and represent them as $\{f(A(\omega))\}$ for some total continuous linear transformation A and all $f{\in}\,\mathcal{E}$. An example of E is the space of all functions in $\mathcal{E}$ dependent only on the n-bit prefix of $\omega{\in}\Omega$ (with $A(\omega) = \omega_{[n]}000\ldots$).

Now, I will extend the concept of rarity $\mathbf{T}(\cdot|\mu)$, $\mathbf{d} = \lfloor\log(1{+}\mathbf{T})\rfloor$ from computable measures μ to r.e. semimeasures. The idea is for $\mathbf{d}(\alpha|\mu)$ to be bounded by $\mathbf{d}(\omega|\lambda)$ if $\alpha = A(\omega)$, $\mu \geq A(\lambda)$. Coarse graining on a lattice $\hat{E}$, rougher than the whole $\mathcal{E}$, allows to define rarity not only for $\alpha{\in}\Omega$ but also for its prefixes. For semimeasures, rarity of extensions do not determine rarity of a prefix.

$\mathbf{T}(\cdot|\mu)$ for a measure μ is a single *r.e.* function $\Omega \to \mathbb{R}^+$ with ≤ 1 mean. It is obtained by averaging an *r.e.* family of such functions. This fails if μ is a semimeasure: its mean of sum can exceed the sum of means. So, $\mathbf{T}(\cdot|\mu)$ will be an expression $\vee.F$ with $F{\subset}\mathcal{E}$.

Definition 3: $\vee_E F$ for an $E{\subset}\mathcal{E}$ and a closed down $F{\subset}\mathcal{E}^+$ (*i.e.,* $0 \leq f \leq g{\in}F \Rightarrow f{\in}F$), denotes $\sup(F \cap E)$. t^A_E for an operator A is $\vee_E F$ where $F = \{f : A(f) \leq \mathbf{T}(\cdot|\lambda)\}$. ***Regular*** semimeasures are $\mu = A(\lambda)$ for a deterministic normal concave *r.e.* A.

Not every *r.e.* μ is regular but each has a regular *r.e.* $\mu_1 \leq \mu$ such that $\mu(x) = \mu_1(x)$ for $x{\in}S$.

Proposition 3: Each *r.e.* μ, among all deterministic normal concave *r.e.* A such that $A(\lambda) \leq \mu$, has a universal one $A = U_\mu$ *i.e.,* such that $t^{U_\mu}_E = O(t^A_E)$ for each such A. $\mu \leq 2U_\mu(\lambda)$ for regular μ.

Definition 4: $\mathbf{T}_E(\varphi|\mu)$ for semimeasures φ, μ, is the mean: $\varphi_E(t^{U_\mu}_E)$ for U_μ defined above. Indexes E are dropped if $E = \mathcal{E}$; $\mu' = U_\mu(\lambda)$; $\mathbf{d} = \lfloor\log(1{+}\mathbf{T})\rfloor$.

From the results of [Gacs 86] one can derive:

Lemma 2: $\mathbf{d}(\cdot|\mathbf{M}) = O(1)$ for the universal regular semimeasure $\mathbf{M}$.

Let f_1 for $f : \Omega^2 \to \mathbb{R}$ be $\beta \mapsto f(\alpha,\beta)$. Let $\nu = \mu{\otimes}\varphi$ be a semimeasure on Ω^2 such that $\nu(f) = \mu(\nu(f_1))$, $A(E)$ be $\{f : A(f){\in}\overline{E}{\subset}\overline{\mathcal{E}}\}$, $E{\otimes}\mathcal{E}$ contain functions $h(\alpha,\beta) = f(\alpha)g(\beta), g{\in}E, f{\in}\mathcal{E}$.

Theorem 1: For each deterministic *r.e.* A, all φ, lattice subspaces $E{\subset}\mathcal{E}$, *r.e.* μ, the test $\mathbf{T}$ satisfies the following Conservation Inequalities:

1) $\mathbf{d}_{A(E)}(A(\varphi)|A(\mu)) \leq \mathbf{d}_E(\varphi|\mu) + O(1)$.

2) $\mathbf{d}_{E\otimes\mathcal{E}}(\varphi \otimes \lambda|\mu \otimes \lambda) \leq \mathbf{d}_E(\varphi|\mu) + O(1)$.

While $\mu(\vee_{\mathcal{E}} F)$ can exceed 1, $\mathbf{T}$ shares the following property with Martin-Löf tests:

Corollary 1: $\mathbf{d}_E(\phi'|\phi') = 0$ for any E, *r.e.* ϕ (thus $\mathbf{d}_E(\phi|\phi) \leq 1$ if ϕ is regular).

These tests are the strongest (largest) extensions of Martin-Löf tests for computable μ. We formalize this for the case of $\omega{\in}\Omega$. Covering other φ is straightforward but more cumbersome.

Proposition 4: $\mathbf{d}(\omega|\mu)$ is the largest up to $+O(1)$ semicontinuous on ω nonincreasing on μ extension of Martin-Löf tests.

Now, like for the integer case, mutual information $\mathbf{I}(\alpha:\beta)$ can be defined as the deficiency of independence, *i.e.,* rarity for the distribution where α,β are assumed each universally distributed (a vacuous assumption, see *e.g.,* Lemma 2) but independent of each other:

$$\mathbf{I}(\alpha:\beta)\overset{\mathrm{df}}{=}\mathbf{d}((\alpha,\beta)|\mathbf{M}\otimes\mathbf{M}).$$

Its conservation inequalities are just special cases of Theorem 1.

Acknowledgment.

This work was partially supported by NSF grant CCF-1049505.

References

[DAN] *Doklady* AN SSSR (= Soviet Math. Doclady).

[Choquet, Meyer 63] Gustave Choquet, Paul-Andre Meyer. Existence et unicite des representations integrales dans les convexes compacts quelconques. Ann. Inst. Fourier, 13/1:139-154, 1963.

[Gacs 74] Peter Gács. On the Symmetry of algorithmic information. [DAN] 15:1477, 1974.

[Gacs 86] Peter Gács. Every sequence is reducible to random one. *Information and Control,* 70/2-3:186-192, 1986.

[Kolmogorov 65] Andrei N. Kolmogorov. Three approaches to the concept of the amount of information. *Probl. Inf. Transm.*, 1(1):1-7, 1965.

[L 12] Leonid A. Levin. Enumerable distributions, randomness, dependence. 2012, http://arxiv.org/abs/1208.2955

[L 73] Leonid A. Levin. On the concept of a random sequence. [DAN] 14(5):1413-1416, 1973.

[L 74] Leonid A. Levin. Laws of information conservation (nongrowth) and aspects of the foundations of probability theory. *Probl Pered. Inf.= Probl. Inf. Transm.* 10(3):206-210, 1974.

[L 84] Leonid A. Levin. Randomness conservation inequalities. *Inf.& Control* 61(1):15-37, 1984.

[Solomonoff 64] R.J. Solomonoff. A formal theory of inductive inference. *Inf.&Cntr* 7(1), 1964.

[ZL 70] Alexander Zvonkin, Leonid A. Levin. The complexity of finite objects and the algorithmic concepts of information and randomness. *UMN = Russian Math. Surveys* 25(6):83-124, 1970.