# m-sophistication

Bruno Bauwens *

Department of Electrical Energy, Systems and Automation,
Ghent University,
Technologiepark 913, B-9052, Zwijnaarde, Belgium,
Bruno.Bauwens@ugent.be.

**Abstract.** The $m$-sophistication of a finite binary string $x$ is introduced as a generalization of some parameter in the proof that complexity of complexity is rare. A probabilistic near sufficient statistic of $x$ is given whose length is upper bounded by the $m$-sophistication of $x$ within small additive terms. This shows that $m$-sophistication is lower bounded by coarse sophistication and upper bounded by sophistication within small additive terms. It is also shown that $m$-sophistication and coarse sophistication can not be approximated by an upper or lower semicomputable function, not even within very large error. Furthermore, connections with important problems in the field of computability and statistics are discussed.

**Key words:** m-sophistication – sophistication – coarse sophistication – Halting probability – Busy Beaver function – sufficient statistic

## Introduction

The Kolmogorov complexity of a finite binary sequence is a measure for the amount of structure in the sequence. Computational and logical depth [1, 2, 4, 13, 22] expresses how much computation time is needed to optimally encode the structure in the sequence. Sophistication [3, 24] informally corresponds to the minimal description length of a function that is able to encode such structure. Here $m$-sophistication is introduced as a variant of logical depth [13]. It is shown that it is lower-bounded by sophistication and upper bounded by coarse sophistication within small bounds. This justifies the name $m$-sophistication. It is also discussed that $m$-sophistication is related to three important questions in the field of statistics and computability.

– If the Kolmogorov complexity $K(x)$ is low for some binary finite sequence $x$, then $x$ can be interpreted as "deterministically" generated, and "non-deterministically" generated otherwise. The structure function [23, 26, 31]

* Supported by a Ph.D grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). The work was carried out while the author was connected to the interdisciplinary Guislain research group at Ugent.

for some $x$, maps each natural number $k$ to the logarithm of the minimal cardinality of a set containing $x$ that has descriptional complexity below $k$. If for some $x$, the structure function decreases for low $k$ to the value $K(x) - k$, these sequences are called "positively random". Positive randomness is satisfied with high probability if $x$ is "stochastically" generated. Such $x$ allow a useful definition of frequentistic probabilities satisfying the Kolmogorov probability axioms. The requirement of positive randomness leads to a set variant of the definition of sophistication. The question rises whether for any $x$ that is negative random but not positive random, $x$ contains much information on the Halting problem.

- A sumtest for a computable semimeasure is an abstraction of a statistical significance test for a simple hypothesis [26]. It can be argued that for many composite hypotheses, a theoretical ideal statistical test is given by a sumtest for a lower semicomputable semimeasure [7, 9]. The question rises whether in some computability class, there exists an unbounded sumtest for some lower semicomputable semimeasure. It turns out that for the hypotheses of independence on the finite binary strings $x$, there are no unbounded computable and lower semicomputable sumtests, but there are upper semicomputable sumtests of maximal magnitude $l(x)$ [9, 12]. There are also no computable or lower semicomputable sumtests for a lower semicomputable universal semimeasure, but there are upper semicomputable sumtests of magnitude $\log l(x) - O(\log \log l(x))$ [5, 9]. The proof relies on the observation that the introduced $m$-sophistication for a universal lower semicomputable semimeasure $m$, is within logarithmic terms a sumtest for $m$.
- The coding theorem justifies the approximation of the logarithm of a lower semicomputable universal semimeasure by data-compression heuristics [16, 17, 29]. The hypothesis of a timeseries $x$ being influence-free of another timeseries $y$ corresponds to a universal lower semicomputable online semimeasure [7, 9, 15]. Also the approximation of such a semimeasure is related to online Kolmogorov complexity [7, 15]. The error in such a coding result is given by logarithmic terms in the $m$-sophistication [6, 9].

*Overview and results.* The paper uses definitions and observations from [14] and basically runs through the proof of the theorem that high complexity of complexity is rare as in [19], see also [18, 20, 26]. $m$-sophistication is a generalization of a parameter used in this proof. It allows some simple observations related to the questions above. Let $k$ be the $m$-sophistication of a finite sequence $x$. It is shown that the amount $K(x)$ of information in $x$ can be decomposed as $k$ bits of Halting information and $K(x) - k$ bits of additional information, within $2 \log k$ error terms. The first $k$ bits of the Halting probability compute an approximate sufficient statistic for $x$. It is shown that within $O(\log k)$ terms $m$-sophistication is larger than coarse sophistication, and smaller than sophistication. Finally it is shown that $m$-sophistication and coarse sophistication define within logarithmic terms a sumtest relative to the universal semimeasure, and that they have no lower and upper semicomputable approximation, not even within large error.

*Definitions and notation.* For an introduction to Kolmogorov complexity and computability we refer to [20, 26, 30], and for extensive specialized background to [18, 27, 28]. Let $\omega$ be the set of natural numbers. The binary strings $2^{<\omega}$ of finite length can be associated with $\omega$: $\varepsilon \leftrightarrow 0; [0] \leftrightarrow 1; [1] \leftrightarrow 2; \ldots$. For $x \in 2^{<\omega}$, let $l(x)$ denote the length of $x$. For $x \in \omega$, let $l(x)$ denote the length of $x$ in its binary expansion. Let $2^n$ and $2^{<n}$ be the sets of $x \in 2^{<\omega}$ with $l(x) = n$, and $l(x) < n$. Let $\omega^{<\omega}$ be the set of finite sequences in $\omega$. The real numbers in $[0,1]$ are associated with Cantor space[1]. For $\alpha \in 2^\omega$, $\alpha^k$ denotes $\alpha_1\alpha_2...\alpha_k$. For $x \in 2^{<\omega}$, $x^k$ denotes $x_1x_2...x_k$. For $\alpha \in [0,1]$, $\alpha^k$ denotes $0.\alpha_0 \ldots \alpha_k$. Notice that the finite binary sequences can be associated with the dyadic numbers, it is: the rational numbers $p/q$ such that $q = 2^k$ for some $k$. In this way one can observe for any $\alpha \in 2^\omega$ or $\alpha \in [0,1]$ that

$$\alpha^k \leqslant \alpha \leqslant \alpha + 2^{-k}.$$

A *semimeasure* $P$ is a positive real function that satisfies $\sum\{P(x) : x \in \omega\} \leqslant 1$. A *measure* is semimeasure that satisfies $\sum\{P(x) : x \in \omega\} = 1$. A semimeasure $P$ (multiplicatively) *dominates* a semimeasure $Q$, notation: $P \geqslant^* Q$, if a constant $c$ exists such that for all $x$: $cP(x) \geqslant Q(x)$. The notation $P \leqslant^* Q$, is used for $Q \geqslant^* P$, and $P =^* Q$, means that both $P \leqslant^* Q$ and $Q \leqslant^* P$. A set $S$ of semimeasures has a *universal element* $m$ if $m \in S$ and $m$ dominates all semimeasures in $S$. Let $f, g$ be functions depending on parameters $t$, $x$, and $n$. $f$ *dominates* $g$, iff there is a constant $c$ which satisfies for all $t, x$ and $n$: $f(t,x,n) + c \geqslant g(t,x,n)$, notation: $f \geqslant^+ g$. $c$ may depend on any parameter except $t, x, n$. $f \leqslant^+ g$ iff $g \geqslant^+ f$. $f =^+ g$ iff both $f \geqslant^+ g$ and $f \leqslant^+ g$.

If for some function $f$, and for some argument $x$ the function value $f(x)$ is defined, then it is written: $f(x)\downarrow$, otherwise it is written $f(x)\uparrow$. A *prefix-free Turing machine* is a Turing machine that defines a function

$$\Phi : \omega \times 2^{<\omega} \times \omega^{<\omega} \to \omega^{<\omega} : (t, p, x) \to \Phi_t(p|x),$$

such that for all $p, w \in 2^{<\omega}$ such that $w \neq \varepsilon$, and for all $x \in \omega^{<\omega}$, iff $\Phi(p|x)\downarrow$ then $\Phi(pw|x)\uparrow$. For $t \in \omega$, $p \in 2^{<\omega}$, and $x \in \omega^{<\omega}$: $\Phi_t(p|x) \downarrow = y$ means that $\Phi$ on input $p, x$ outputs $y$, and halts in less than $t$ computation steps. *We repeat that in this definition, as well as in the whole paper, $t$ is not a function, but $t \in \omega$.* A prefix-free Turing machine is *optimal universal*, iff for any prefix-free Turing machine $\Psi$, there exists a $w \in 2^{<\omega}$ such that for all $p, x$: $\Psi(p|x)\downarrow$ implies $\Phi(wp|x)$. From now on, a fixed optimal universal prefix-free Turing machine $\Phi$ is assumed.

A real function $f : \omega \to [0,1]$ is *computable* if there is a $p \in 2^{<\omega}$ such that for all $k, x$: $\Phi(p|x,k) \downarrow = f(x)^k$. An *enumeration* of a real function $f(x)$ is a computable real function $g(x,t)$ such that for all $t$: $g(u,t) \leqslant g(u,t+1)$ and such that $\lim_{t,k} g(u,t) = f(u)$. A *lower semicomputable* function $f$ is a function $f$ that has an enumeration. A function $f$ is *upper semicomputable* if $-f$ is lower semicomputable. With abuse of notation, an enumeration of $f$ is denoted as $f_t$.

[1] This association is not bijective since the real $0.a0111...$ equals the real $0.a1000...$ for any $a \in 2^{<\omega}$, however, this omission does not cause problems.

For $x, y \in \omega^{<\omega}$, let the *prefix-free Kolmogorov complexity* be

$$K_t(x) = \min\{l(p) : \Phi_t(p|y) \downarrow= x\}$$
$$K(x) = \lim_{t\to\infty} K_t(x).$$

For all $n \in \omega \subset \omega^{<\omega}$: $K(n) \leqslant^+ \log n + 2\log\log n$ and for all $x \in 2^n \subset \omega^{<\omega}$: $K(x) \leqslant^+ n + 2\log n$. Let $x^*$ represent the lexicographic first program that produces $x$. Prefix-free Kolmogorov complexity satisfies the additivity property

$$K(x, y) =^+ K(y) + K(x|y^*) =^+ K(y) + K(x|y, K(y)).$$

A Halting program can output itself, and thus also its own length, therefore $K(x) =^+ K(x, K(x))$. The coding theorem shows that

$$Q_{p,t}(x) = \sum\{2^{-l(p)} : \Phi_t(p) \downarrow= x\} \tag{1}$$
$$Q_{K,t}(x) = 2^{-K_t(x)} \tag{2}$$

define enumerations of lower semicomputable universal semimeasures. This implies that for any universal lower semicomputable semimeasure $m$, one has $-\log m(x) =^+ K(x)$.

## 1 Halting probability and a variant of a Busy Beaver function

This section introduces some technical results that will be used later. Furthermore, it also raises some questions, related to the dependency of $m$-sophistication on the choice of $m$.

In computability theory, the number $\Omega$ is typically defined as the prior probability that some universal prefix-free Turing machine halts [14, 19]. Here generalized version is studied: the probability that a universal lower semicomputable semimeasure is defined.

**Definition 1.** *Let $m_t$ be some enumeration of some universal lower semicomputable semimeasure over $\omega$.*

$$\Omega_{m,t} = \sum_{l(x)<t} m_t(x)$$
$$\Omega_m = \lim_{t\to\infty} \Omega_t$$

The original definition in [14, 19] is obtained by choosing $m_t = Q_{p,t}$, as in Equation (1). $\Omega_{Q_p}$ satisfies the following well known theorem.

**Theorem 1.** *[14, 19] For all $n$: $K(\Omega^n_{Q_p}) \geqslant^+ n$. There is a constant $c$ such that for all $n$, the Halting of any program $p \in 2^{<n}$ can be decided by $\Omega^{n+c}$.*

These properties of $\Omega_{Q_p}$ remain for general $\Omega_m$ by the same argument (see further). For $a, b$ elements or tuples of elements in $2^{<\omega}$, $\omega$, and/or $\omega^{<\omega}$, that possibly depend on parameters $t, x$ and $n$, the notation: $a \longrightarrow b$ or with explicit dependencies, $a(t, x, n) \longrightarrow b(t, x, n)$, means that $K(b(t, x, n)|a(t, x, n)) \leqslant^+ 0$. For $\alpha, \beta \in 2^\omega$, the relation $\alpha^n \longrightarrow \beta^n$ defines a partial order on $2^\omega$, which is equivalent with the 'domination' relation in [25]. $\Omega_{Q_p}$ is stable with respect to the choice of universal machine $\Phi$. Let $\Phi$ and $\Phi'$ be two optimal universal prefix-free Turing machines and let $Q_p$, and $Q'_p$ be defined as in equation (1), according to $\Phi$ and $\Phi'$. It is easily observed that

$$\Omega^n_{Q_p} \longleftrightarrow \Omega^n_{Q'_p}.$$

An other example of such a relation is

$$\Omega^n_{Q_p} \longrightarrow \Omega^n_{Q_K},$$

where $Q_K$ is defined in Equation (2). It is an interesting question whether the opposite direction also holds.

Following the proof that high $K(K(x)|x)$ is rare in [20], the times $t_n$ are defined. Fix some universal lower semicomputable semimeasure $m$, and let for each $n$:

$$t_n = \min\{t : \Omega^n_m \leqslant \Omega_{m,t}\}.$$

**Lemma 1.**

$$\Omega^n_m \longleftrightarrow n, t_n$$

*Proof.* By definition. □

The prefix-free Busy Beaver function is defined by:

$$PBB(n) = \max\{0\} \cup \{\Phi(p) : l(p) \leqslant n \wedge \Phi(p) \downarrow \in \omega\}.$$

Lemma 2 shows that $t_n$ is a very fast growing function that oscillates between $PBB(n - O(1))$ and $PBB(n + 2\log n + O(1))$.

**Lemma 2.** *For all* $n$

$$n \leqslant_+ K(t_n) \leqslant_+ n + 2\log n.$$

*There exists a constant* $c$ *such that:*

$$PBB(n - c) \leqslant t_n < PBB(n + 2\log n + c).$$

A proof is given in [11, 9]. The dependence of $t_n$ on the choice of $m$ is given by the subsequent corollary.

**Corollary 1.** *For all universal semimeasures $m$, and $m'$, there is some constant $c$ such that*

$$t_n < t'_{n+2\log n+c},$$

*with $t_n$ and $t'_n$ defined by $m$ and $m'$.*

*Proof.*

$$t_n \leqslant PBB(n+2\log n+c) < t'_{n+2\log n+2c}$$

□

A real number $\alpha \in 2^\omega$ is random if for any $n$: $K(\alpha^n) \geqslant^+ n$. It follows by Lemma 2 that

**Corollary 2.** *$\Omega_m$ is random.*

*Proof.* Since $n \leqslant^+ K(t_n) \leqslant^+ K(\Omega_m^n)$. □

By Corollary 1 it follows that

**Lemma 3.** *for $m, m'$ universal lower semicomputable semimeasures*

$$\Omega_m^n \longrightarrow \Omega_{m'}^{n-2\log n}.$$

*Proof.*

$$\Omega_m^n \longrightarrow n, t_n \longrightarrow n, t'_{n-2\log n} \longrightarrow \Omega_{m'}^{n-2\log n}$$

□

The question rises whether the set of all $\Omega_m$ for some universal lower semicomputable semimeasures has a maximal element relative to the $\longrightarrow$ order. Notice that it is shown in [25] that the set of all $\Omega_m$ with $m$ universal lower semicomputable semimeasures, corresponds to all computable enumerable random real numbers. Such an optimal $\Omega_m$ would give rise to an $m$-sophistication that is minimal within a constant.

## 2 $m$-sophistication and complexity of complexity

The logical depth of an $x \in 2^{<\omega}$ [13], is defined as the minimal computation time needed by a program of length close to $K(x)$ that produces $x$ [13]. Intuitively, this is the computation time needed to encode all structure in $x$. $m$-sophistication is now defined by taking the inverse function of $t : n \rightarrow t_n$ of the logical depth.

**Definition 2.** *For some $c \in \omega$, the $m$-sophistication of $x \in 2^{<\omega}$ is given by:*

$$k_c(x) = \min\{k : K_{t_k}(x) \leqslant K(x) + c\}.$$

For fixed $c$, the function $k_c(x)$ is not computable, nor lower semicomputable nor upper semicomputable by Proposition 1, however, from the definition it follows that it is limit-computable. From Corollary 1 it is observed that $k_c$ is relatively stable with respect to changes of universal lower semicomputable semimeasure $m$.

**Corollary 3.** *Let $m, m'$ be two universal lower semicomputable semimeasures and let $k$ and $k'$ be the corresponding $m$-sophistication and $m'$-sophistication, for any $c$*

$$k_c \leqslant^+ k'_c + 2\log k'_c.$$

As for sophistication (see further), also $m$-sophistication is unstable with respect to the parameter $c$.

**Lemma 4.** *For all $c$, there is a $c'$ such that for infinitely many $x$:*

$$k_c(x) - k_{c+c'}(x) \geqslant_+ l(x) - 4\log l(x).$$

Informally, one chooses an $x$ that is only a little compressible, by some constant $c+c'$, for $c'$ large enough, thus, $k_{c+c'}(x) = 0$, and such that this little compression is only possible within a time $t_{n-O(\log n)}$. Therefore, $k_c(x)$ is much larger. A formal detailed proof needs some care, and is given in [11, 9]. Lemma's 5 and 6 lead to Corollaries 4 and 5. Let $k'(x)$ be the $Q_K$-sophistication $k_0(x)$. High $Q_K$-sophistication is rare.

**Lemma 5.** *For any $i$ and $S_i = \{x : k'(x) \geqslant i\}$:*

$$m(S_i) \leqslant 2^{-i+1}.$$

*Proof.*

$$\frac{1}{2}m(S_i) \leqslant m(S_i) - m_{t_i}(S_i) \leqslant \Omega - \Omega_{t_i} \leqslant 2^{-i}.$$

□

**Lemma 6.** *For all $x$*

$$K(K(x)|x) \leqslant^+ k'(x) + 2\log k'(x).$$

*Proof.* Notice that $t_{k'(x)}, x \longrightarrow K(x)$, thus

$$K(K(x)|x) \leqslant^+ K(t_{k'(x)}) \leqslant^+ K(\Omega^{k'(x)}) \leqslant^+ k'(x) + 2\log k'(x),$$

□

**Corollary 4.** *[19, 20] There exists a constant $c > 0$ such that*

$$m(\{K(K(x)|x) \geqslant k\}) \leqslant c2^{-k-2\log k}.$$

**Lemma 7.** *For any $c$ large enough: $k' \geqslant k_c$.*

*Proof.* By some time-bounded version of the coding theorem [9]:

$$K_{t_{k'(x)+c}}(x) \leqslant^+ -\log m_{t_{k'(x)}}(x) =^+ -\log m(x) =^+ K(x).$$

□

A sumtest $d$ for a semimeasure $P$ is a function $d : 2^{<\omega} \to \mathbb{Z}$ such that

$$\sum_{x \in 2^{<\omega}} P(x) 2^{d(x)} \leqslant 1.$$

**Corollary 5.** *For $k = k'$ and for $k = k_c$ with $c$ large enough, $k - 2\log k$ defines a sumtest for $m$.*

*Proof.*

$$\sum_{x \in 2^{<\omega}} m(x) 2^{k'(x) - 2\log k'(x) - 2} \leqslant \sum_{k \in \omega} m(S_k) 2^{k - 2\log k - 2} \leqslant \sum_{k \in \omega} 2^{-2\log k - 1} \leqslant 1$$

□

$k_c$ and $k'$ are not computable, and not even a logarithmic lower bound can be computed.

**Proposition 1.** *For $k = k'$ and for $k = k_c$ with $c$ large enough, $k$ can not be approximated by a lower or upper semicomputable function within $k - 2\log k + O(1)$ error.*

A proof is given in [11, 9].

## 3 Sophistication and coarse sophistication

Typically a computable function is a partial computable function that is total in its domain. For prefix-free functions this definition can not longer be applied, since the domain of such a function can only be a strict subset of $2^{<\omega}$. Therefore a prefix-free function $f$ on $2^{<\omega}$ is defined to be computable iff the set

$$U_f = \{\alpha \in 2^\omega : \forall n \in \omega \big[ f(\alpha^n) \uparrow \big] \}$$

has measure zero.

**Definition 3.** *Let $f$ be a computable function. A function $f$-sufficient statistic for $x \in 2^{<\omega}$ is a computable prefix-free function $g$ such that there exists a $d \in g^{-1}(x)$ with*

$$K(g) + l(d) \leqslant K(x) + f(l(x)).$$

*The* sophistication *[24] of $x \in 2^{<\omega}$ is given by:*

$$k_c^{\mathrm{soph}}(x) = \min\{K(f) : f \text{ is a } c\text{-sufficient statistic of } x\}.$$

Notice that there is a slight deviation from [24, 32] since it is also required that $f$ is prefix-free. This is necessary to interpret sophistication as the length of a minimal sufficient statistic [21], which is defined there using prefix-free functions. Also remark that now Lemma 8 is true. Let $pbb(x)$ be the inverse of the Busy Beaver function, it is $pbb(x) = \min\{k : x \leqslant PBB(k)\}$. It is a very slow growing function, dominated by any unbounded non-decreasing function [12].

**Proposition 2.** *There exists a $c'$ such that for all $c, x$:*

$$k_{c+c'}(x) \leqslant^+ k_c^{\text{soph}}(x) + pbb(x).$$

A proof is given in [11, 9].

**Definition 4.** *A* probabilistic $f$-sufficient statistic *of $x \in 2^{<\omega}$, is a computable measure*[2] *$P$ such that*

$$K(P) - \log P(x) \leqslant K(x) + f(l(x)).$$

Since prefix-free functions are used here, probabilistic and function sufficient statistics are equivalent.

**Lemma 8.** *There is a constant $c$ such that every probabilistic $f$-sufficient statistic $P$ defines a function $(f+c)$-sufficient statistic $g$ with $abs(K(P)-K(g)) \leqslant c$, and every function $f$-sufficient statistic $g$ defines a probabilistic $(f+c)$-sufficient statistic $P$ with $abs(K(P) - K(g)) \leqslant c$.*

Let

$$P_k(x) = N 2^{-k}(m_{t_k}(x) - m_{t_{k-1}}(x)),$$

Where $N$ is a normalization constant such that $P_k$ defines a computable probability distribution. Notice that $2 \leqslant N < 4$. Also remark that this can be considered as the probabilistic equivalent of the "explicit minimal near sufficient set statistic" described in [21]. By the following Lemma, it follows that strings with high $m$-sophistication, contain a lot of information on the Halting problem.

**Lemma 9.** *For all $x$*

$$K(x|\Omega^{k'(x)}) \leqslant^+ K(x) - k'(x).$$

A proof is given in [11, 9]. To relate $P_k$ to sophistication, it is shown that it defines some $f$-sufficient statistic.

**Proposition 3.** *There exists a $c$ such that $P_{k'(x)}$ is a probabilistic $(2\log k'(x) + c)$-sufficient statistic for $x$. There exists a $c$ such that for any $c'$, there is a $k \leqslant^+ k_{c'}(x)$ such that $P_k$ is a $(3\log k_c(x) + c + c')$-sufficient statistic for $x$.*

A proof is given in [11, 9].

The online coding theorem [15] relates the logarithm of a universal lower semicomputable online semimeasures (causal semimeasure) to online Kolmogorov complexity, which is a variant of conditional Kolmogorov complexity $K(x|y)$ for $x, y \in 2^{<\omega}$, where $y_i$ is only known to the Turing machine $\Phi$ after it has computed $x^{i-1}$. The online coding theorem has an error term, which is improved for the length-conditional case in [6, 10]. In the proof of the improved online

[2] Remind that a *measure* is a semimeasure with $\sum_{x\in\omega} P(x) = 1$

coding theorem, an online computable semimeasure is associated with $P_{k'(x)}$. It is shown that the value of the logarithm of the universal lower semicomputable online semimeasure and the associated semimeasure for $x$ equals within a $O(\log k'(x))$ term. Since the associated semimeasure is computable, a variant of Shannon-Fano code can be applied.

In [8] it is shown that the result of Proposition can not be further improved to eliminate the logarithmic terms in order to consider $P_k$ as a probabilistic $c$-sufficient statistic. It is shown that any minimal sufficient statistic of some finite binary sequence can contain a substantial amount of non-Halting information. From the proof it can be conjectured that that in contrast with $m$-sophistication, sophistication does not define a sumtest. However, it is shown in [8] that $P_k$ defines a minimal typical model as defined in [33].

Sophistication is unstable with respect to the parameter $c$, therefore in [3] coarse sophistication is defined. The prefix-free variant is given by

$$k^{\text{csoph}}(x) = \min_c \{k_c(x) + c\}.$$

As a corollary of Proposition 3 it follows that:

**Corollary 6.** *For $k = k'(x)$ and $k = k_c(x)$ for $c$ large enough*

$$k^{\text{csoph}}(x) \leqslant^+ k(x) + 2\log k(x).$$

Together with Proposition 3, this shows that within small terms, $m$-sophistication is lower bound by sophistication and upper bounded by coarse sophistication. This also justifies the choice of the term $m$-sophistication.

**Proposition 4.** *$k^{\text{csoph}}(x) - 4\log k^{\text{csoph}}(x)$ defines a sumtest for $m$. $k^{\text{csoph}}$ can not be approximated by a lower or upper semicomputable function within $k - 2\log k + O(1)$ error.*

*Proof.* This follows from Corollary 6 and the same proof as 1. □

**Acknowledgments** The author is grateful for the anonymous referees for extensive comments, and to N.Vereshchagin for extensive comments on the manuscript of [9], from which this text is extracted. The author is also grateful to M.Li and P.Vitanyi, for the book [26], without such a good introductory and reference book this work would never have appeared.

## References

1. L. Antunes. *Useful information.* PhD thesis, Universidade do Porto, 2002.
2. L. Antunes, A. Costa, A. Souto, and P. Vitanyi. Computation and logic in the real world. In *Computability in Europe*, 2007.
3. L. Antunes and L. Fortnow. Sophistication revisited. *Theor. Comp. Sys.*, 45(1):150–161, 2009.

4. L. Antunes, A. Matos, A. Souto, and P. Vitanyi. Depth as randomness deficiency. *Theor. of Comput. Sys.*, 45(4):724–739, 2009.
5. B. Bauwens. Co-enumerable sumtests for the universal distribution. Submitted, 2008.
6. B. Bauwens. Ideal hypothesis testing and algorithmic information transfer, june 2009. Talk in *Conference on Logic, computability and randomness*, `www.lif.univ-mrs.fr/lce/bauwens.pdf`.
7. B. Bauwens. Influence tests I: ideal hypothesis tests and causal semimeasures. *ArXiv e-prints*, december 2009. `http://arxiv.org/abs/0912.2688`.
8. B. Bauwens. On the equivalence between minimal sufficient statistics, minimal typical models and initial segments of the Halting sequence. *ArXiv e-prints*, November 2009. `http://arxiv.org/abs/0911.4521`.
9. B. Bauwens. *Computability in statistical hypotheses testing, and characterizations of independence and directed influences in time series using Kolmogorov complexity*. PhD thesis, Ugent, may 2010.
10. B. Bauwens. Influence tests II: on-line coding results. In preparation, 2010.
11. B. Bauwens. Proofs of paper: m-sophistication. Technical report, Ugent, 2010. http://www.autoctrl.ugent.be/bruno/papers/msophProofs.pdf.
12. B. Bauwens and S. Terwijn. Notes on sum-tests and independence tests. Accepted for publication in Theor. Comput. Sys., open access, 2009.
13. C. H. Bennett. *Logical depth and physical complexity*, pages 227–257. Oxford University Press, Inc., New York, NY, USA, 1988.
14. G.J. Chaitin. A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, 22(3):329–340, 1975.
15. A. Chernov, S. Alexander, N. Vereshchagin, and V.Vovk. On-line probability, complexity and randomness. In *ALT '08: Proceedings of the 19th international conference on Algorithmic Learning Theory*, pages 138–153, Berlin, Heidelberg, 2008. Springer-Verlag.
16. R. Cilibrasi and P.M.B. Vitányi. Clustering by compression. *Trans. on Inform. Theory*, 51(4):1523–1545, 2005.
17. S. de Rooij and P.M.B. Vitányi. Approximating rate-distortion graphs of individual data: experiments in lossy compression and denoising. Submitted.
18. R. Downey and D. Hirschfeldt. Algorithmic randomness and complexity. To appear.
19. P. Gács. On the symmetry of algorithmic information. *Soviet Math. Dokl.*, 15:1477–1480, 1974.
20. P. Gács. Lecture notes on descriptional complexity and randomness. Technical report, Comput. Sci. Dept., Boston, 1988-2010. Technical report, `http://www.cs.bu.edu/faculty/gacs/papers/ait-notes.pdf`.
21. P. Gács, J. Tromp, and P.M.B. Vitányi. Algorithmic statistics. *IEEE Trans. Inform. Theory*, 47(6):2443–2463, 2001.
22. D.W. Juedes, J.I. Lathrop, and J.H. Lutz. Computational depth and reducibility. *Theor. Comput. Sci.*, 132(1-2):37–70, 1994.
23. A.N. Kolmogorov. Complexity of algorithms and objective definition of randomness. *Uspekhi Mat. Nauk*, 29(4), 1974. Abstract of a talk at Moscow Math. Soc. meeting 4/16/1974, translation in M.Li and P.M.B. Vitányi 2008, page 438.
24. M. Koppel. *The Universal Turing Machine: A Half-Century Survey*, chapter Structure, pages 435–452. R. Herken, Oxford University Press, 1988.
25. A. Kucera and T. Slaman. Randomness and recursive enumerability. *SIAM J. Comput.*, 31(1):199–211, 2001.

26. M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 2008.
27. A. Nies. *Computability and Randomness*. Oxford University Press, Inc., New York, 2009.
28. P. Odifreddy. *Classical Recursion Theory, Volume I*. North-Holland, 1989.
29. B. Ryabko, J. Astola, and A. Gammerman. Application of kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theoretical Computer Science*, 359(1-3):440–448, august 2006.
30. S. A. Terwijn. Syllabus computability theory, 2008.
31. N.K. Vereshchagin and P.M.B. Vitányi. Kolmogorov's structure functions and model selection. *IEEE Trans. Infor. Theory*, 50(12):3265–3290, 2004.
32. P.M.B. Vitányi. Meaningful information. *IEEE Trans. Inform. Theory*, 52(10):4617–4626, 2006.
33. P.M.B Vitányi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. Infor. Theory*, 46(2):446–464, 2000.