Technical Report IDSIA-16-04

Sequential Predictions based on Algorithmic Complexity

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland[*]
marcus@idsia.ch http://www.idsia.ch/~marcus

Submitted: Oct. 2003 Published: Oct. 2005

Abstract

This paper studies sequence prediction based on the monotone Kolmogorov complexity $Km=-\log m$, i.e. based on universal deterministic/one-part MDL. m is extremely close to Solomonoff's universal prior M, the latter being an excellent predictor in deterministic as well as probabilistic environments, where performance is measured in terms of convergence of posteriors or losses. Despite this closeness to M, it is difficult to assess the prediction quality of m, since little is known about the closeness of their posteriors, which are the important quantities for prediction. We show that for deterministic computable environments, the "posterior" and losses of m converge, but rapid convergence could only be shown on-sequence; the off-sequence convergence can be slow. In probabilistic environments, neither the posterior nor the losses converge, in general.

Keyword

Sequence prediction; Algorithmic Information Theory; Solomonoff's prior; Monotone Kolmogorov Complexity; Minimal Description Length; Convergence; Self-Optimization.

[*] Part of this work appeared in the proceedings of the 2003 COLT conference [Hut03b].

Contents

1 Introduction

In this work we study the performance of Occam's razor based sequence predictors. Given a data sequence x_1, x_2, ..., x_{n-1} we want to predict (certain characteristics) of the next data item x_n. Every x_t is an element of some domain $\mathcal{X}$, for instance weather data or stock-market data at time t, or the t^{th} digit of π. Occam's razor [LV97], appropriately interpreted, tells us to search for the simplest explanation (model) of our data $x_1,...,x_{n-1}$ and to use this model for predicting x_n. Simplicity, or more precisely, effective complexity can be measured by the length of the shortest program computing sequence $x:=x_1...x_{n-1}$. This length is called the algorithmic information content of x, which we denote by $\tilde K(x)$. $\tilde K$ stands for one of the many variants of "Kolmogorov" complexity (plain, prefix, monotone, ...) or for $-\log\tilde k(x)$ of universal distributions/measures $\tilde k(x)$.

Algorithmic information theory mainly considers binary sequences. For finite alphabet $\mathcal{X}$ one could code each $x_t\in\mathcal{X}$ as a binary string of length $\lceil\log|\mathcal{X}|\rceil$, but this would not simplify the analysis in this work. The reason being that binary coding would *not* reduce the setting to bit by bit predictions, but to predict a block of bits before observing the true block of bits. The only difference in the analysis of general alphabet versus binary block-prediction is in the convention of how the length of a string is defined.

The most well-studied complexity regarding its predictive properties is $KM(x)= -\log M(x)$, where $M(x)$ is Solmonoff's [Sol64, Eq.(7)] universal prior. Solomonoff has shown that the posterior $M(x_t|x_1...x_{t-1})$ rapidly converges to the true data generating distribution [Sol78]. In [Hut01b, Hut03a] it has been shown that M is also an excellent predictor from a decision-theoretic point of view, where the goal is to minimize loss. In any case, for prediction, the posterior $M(x_t|x_1...x_{t-1})$, rather than the prior $M(x_1...x_t)$, is the more important quantity.

Most complexities $\tilde K$ coincide within an additive logarithmic term, which implies that their "priors" $\tilde k=2^{-\tilde K}$ are close within polynomial accuracy. Some of them are extremely close to each other. Many papers deal with the proximity of various complexity measures [Lev73a, Gác83, ...]. Closeness of two complexity measures is regarded as indication that the quality of their prediction is similarly good [LV97, p.334]. On the other hand, besides M, little is really known about the closeness of "posteriors", relevant for prediction.

Aim and conclusion. The main aim of this work is to study the predictive properties of complexity measures other than KM. The monotone complexity Km is, in a sense, closest to Solomonoff complexity KM. While KM is defined via a mixture of infinitely many programs, the conceptually simpler Km approximates KM by the contribution of the single shortest program. This is also closer to the spirit of Occam's razor. Km is a universal deterministic/one-part version of the popular Minimal Description Length (MDL) principle. We mainly concentrate on Km because it has a direct interpretation as a universal deterministic/one-part MDL predictor, and it is closest to the excellent performing KM, so we expect predictions based on other $\tilde{K}$ not to be better.

The main conclusion we will draw is that closeness of priors does neither necessarily imply closeness of posteriors, nor good performance from a decision-theoretic perspective. It is far from obvious, whether Km is a good predictor in general, and indeed we show that Km can fail (with probability strictly greater than zero) in the presence of noise, as opposed to KM. We do not suggest that Km fails for sequences occurring in practice. It is not implausible that (from a practical point of view) minor extra (apart from complexity) assumptions on the environment or loss function are sufficient to prove good performance of Km. Some complexity measures like the prefix complexity K, fail completely for prediction.

Contents. *Section 2* introduces notation and describes how prediction performance is measured in terms of convergence of posteriors or losses. *Section 3* summarizes known predictive properties of Solomonoff's prior M. *Section 4* introduces the monotone complexity Km and the prefix complexity K and describes how they and other complexity measures can be used for prediction. In *Section 5* we enumerate and relate eight important properties, which general predictive functions may posses or not: proximity to M, universality, monotonicity, being a semimeasure, the chain rule, enumerability, convergence, and self-optimization. Some later needed normalization issues are also discussed. Furthermore, convergence of non-semimeasures that are close to M is proven. *Section 6* contains our main results. Monotone complexity Km is analyzed quantitatively w.r.t. the eight predictive properties. Qualitatively, for deterministic, computable environments, the posterior converges and is self-optimizing, but rapid convergence could only be shown on-sequence; the (for prediction equally important) off-sequence convergence can be slow. In probabilistic environments, m neither converges, nor is it self-optimizing, in general. *Section 7* presents some further results: Poor predictive performance of the prefix complexity K is shown and a simpler MDL-inspired way of using Km for prediction is briefly discussed. *Section 8* contains an outlook and a list of open question, including the convergence speed of m, natural Turing machines, non-self-optimization for general Turing machines and losses, other complexity measures, two-part MDL, extra conditions on environments, and other generalizations.

2 Notation and Setup

Strings and natural numbers. We write $\mathcal{X}^*$ for the set of finite strings over finite alphabet $\mathcal{X}$, and $\mathcal{X}^\infty$ for the set of infinity sequences. We use letters i,t,n for natural numbers, x,y,z for finite strings, ϵ for the empty string, $\ell(x)$ for the length of string x, and $\omega = x_{1:\infty}$ for infinite sequences. We write xy for the concatenation of string x with y. For a string of length n we write $x_1x_2...x_n$ with $x_t\in\mathcal{X}$ and further abbreviate $x_{1:n} := x_1x_2...x_{n-1}x_n$ and $x_{<n} := x_1...x_{n-1}$. For a given sequence $x_{1:\infty}$ we say that x_t is on-sequence and $\bar{x}_t \neq x_t$ is off-sequence. x'_t may be on- or off-sequence.

Prefix sets/codes. String x is called a (proper) prefix of y if there is a $z(\neq\epsilon)$ such that $xz = y$. We write $x* = y$ in this case, where $*$ is a wildcard for a string, and similarly for infinite sequences. A set of strings is called prefix-free if no element is a proper prefix of another. A prefix-free set $\mathcal{P}$ is also called a prefix code. Prefix codes have the important property of satisfying Kraft's inequality $\sum_{x\in\mathcal{P}}|\mathcal{X}|^{-\ell(x)} \leq 1$.

Asymptotic notation. We abbreviate $\lim_{t\to\infty}[f(t)-g(t)] = 0$ by $f(t) \overset{t\to\infty}{\longrightarrow} g(t)$ and say f converges to g, without implying that $\lim_{t\to\infty}g(t)$ itself exists. The big O-notation $f(x) = O(g(x))$ means that there are constants c and $x_0 > 0$ such that $|f(x)| \leq c|g(x)|\,\forall x > x_0$. The small o-notation $f(x) = o(g(x))$ abbreviates $\lim_{x\to\infty}f(x)/g(x) = 0$. We write $f(x) \overset{\times}{\leq} g(x)$ for $f(x) = O(g(x))$ and $f(x) \overset{+}{\leq} g(x)$ for $f(x) \leq g(x) + O(1)$. Corresponding equalities can be defined similarly. They hold if the corresponding inequalities hold in both directions. $\sum_{t=1}^\infty a_t^2 < \infty$ implies $a_t \overset{t\to\infty}{\longrightarrow} 0$. We say that a_t converges fast or rapidly to zero if $\sum_{t=1}^\infty a_t^2 \leq c$, where c is a constant of reasonable size; $c = 100$ is reasonable, maybe even $c = 2^{30}$, but $c = 2^{500}$ is not.[1] The number of times for which a_t deviates from 0 by more than ε is finite and bounded by c/ε^2; no statement is possible for *which* t these deviations occur. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$ or $\#\mathcal{S}$. For properties $A(t)\in\{true,false\}$ we say

$A(t)$ is valid for ... t	almost all	most	many	finitely many
iff $\#\{t\leq n : A(t)\}$	$\overset{+}{=} n$	$= n - o(n)$	$\overset{\times}{=} n$	$\leq c \quad (\exists c)$

(Semi)measures. We call $\rho : \mathcal{X}^* \to [0,1]$ a (semi)measure *iff* $\sum_{x_n\in\mathcal{X}}\rho(x_{1:n}) \overset{(\leq)}{=} \rho(x_{<n})$ and $\rho(\epsilon) \overset{(\leq)}{=} 1$. $\rho(x)$ is interpreted as the ρ-probability of sampling a sequence which starts with x. In case of a semimeasure the gap $g_n = 1 - \sum_{x_{1:n}}\rho(x_{1:n}) \geq 0$ may be interpreted as the possibility/probability of *finite* sequences of length less than n [ZL70, Sch00], or as an evidence gap in Dempster-Shafer theory [Dem68, Sha76]. The conditional probability (posterior)

$$\rho(x_t|x_{<t}) := \frac{\rho(x_{1:t})}{\rho(x_{<t})} \qquad (1)$$

[1] Environments of interest have reasonable complexity K, but 2^K is not of reasonable size.

is the ρ-probability that a string $x_1...x_{t-1}$ is followed by (continued with) x_t. We call ρ deterministic if $\exists\omega:\rho(\omega_{1:n})=1\ \forall n$. In this case we identify ρ with ω.

Convergent predictors. We assume that μ is the "true"[2] sequence generating measure, also called environment. If we know the generating process μ, and given past data $x_{<t}$ we can predict the probability $\mu(x_t|x_{<t})$ of the next data item x_t. Usually we do not know μ, but estimate it from $x_{<t}$. Let $\rho(x_t|x_{<t})$ be an estimated probability[3] of x_t, given $x_{<t}$. Closeness of $\rho(x_t|x_{<t})$ to $\mu(x_t|x_{<t})$ is expected to lead to "good" predictions:

Consider, for instance, a weather data sequence $x_{1:n}$ with $x_t=1$ meaning rain and $x_t=0$ meaning sun at day t. Given $x_{<t}$ the probability of rain tomorrow is $\mu(1|x_{<t})$. A weather forecaster may announce the probability of rain to be $y_t:=\rho(1|x_{<t})$, which should be close to the true probability $\mu(1|x_{<t})$. To aim for

$$\rho(x_t'|x_{<t}) \overset{(fast)}{\longrightarrow} \mu(x_t'|x_{<t}) \quad \text{for} \quad t\to\infty \tag{2}$$

seems reasonable. A sequence of random variables $z_t=z_t(\omega)$ (like $z_t=\rho(x_t|x_{<t})-\mu(x_t|x_{<t})$) is said to converge to zero with μ-probability 1 (w.p.1) if the set $\{\omega: z_t(\omega)\overset{t\to\infty}{\longrightarrow}0\}$ has μ-measure 1. z_t is said to converge to zero in mean sum (i.m.s) if $\sum_{t=1}^\infty\mathbf{E}[z_t^2]\leq c<\infty$, where $\mathbf{E}$ denotes μ-expectation. Convergence i.m.s. implies convergence w.p.1 (rapid if c is of reasonable size).

Depending on the interpretation, a ρ satisfying (2) could be called consistent or self-tuning [KV86]. One problem with using (2) as performance measure is that closeness cannot be computed, since μ is unknown. Another disadvantage is that (2) does not take into account the value of correct predictions or the severity of wrong predictions.

Self-optimizing predictors. More practical and flexible is a decision-theoretic approach, where performance is measured w.r.t. the true outcome sequence $x_{1:n}$ by means of a loss function, for instance $\ell_{x_ty_t}:=(x_t-y_t)^2$, which does not involve μ. More generally, let $\ell_{x_ty_t}\in[0,1]\subset I\!R$ be the received loss when performing some prediction/decision/action $y_t\in\mathcal{Y}$ and $x_t\in\mathcal{X}$ is the t^{th} symbol of the sequence. Let $y_t^\Lambda\in\mathcal{Y}$ be the prediction of a (causal) prediction scheme Λ. The true probability of the next symbol being x_t, given $x_{<t}$, is $\mu(x_t|x_{<t})$. The μ-expected loss (given $x_{<t}$) when Λ predicts the t^{th} symbol is

$$l_t^\Lambda(x_{<t}) \;:=\; \sum_{x_t}\mu(x_t|x_{<t})\ell_{x_ty_t^\Lambda}.$$

The goal is to minimize the μ-expected loss. More generally, we define the Λ_ρ sequence prediction scheme

$$y_t^{\Lambda_\rho} := \arg\min_{y_t\in\mathcal{Y}}\sum_{x_t}\rho(x_t|x_{<t})\ell_{x_ty_t}, \tag{3}$$

[2] Also called *objective* or *aleatory* probability or *chance*.

[3] Also called *subjective* or *belief* or *epistemic* probability.

which minimizes the ρ-expected loss. If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss ($l_t^{\Lambda_\mu} \leq l_t^\Lambda$ for all Λ). An important special case is the error loss $\ell_{xy}=1-\delta_{xy}$ with $\mathcal{Y}=\mathcal{X}$. In this case Λ_ρ predicts the y_t which maximizes $\rho(y_t|x_{<t})$, and $\sum_t \mathbf{E}[l_t^{\Lambda_\rho}]$ is the expected number of prediction errors (where $y_t^{\Lambda_\rho} \neq x_t$). The natural decision-theoretic counterpart of (2) is to aim for

$$l_t^{\Lambda_\rho}(x_{<t}) \overset{(fast)}{\longrightarrow} l_t^{\Lambda_\mu}(x_{<t}) \quad \text{for} \quad t\to\infty \tag{4}$$

what is called (without the fast supplement) self-optimization in control-theory [KV86].

3 Predictive Properties of $M=2^{-KM}$

We define a prefix/monotone Turing machine T as a Turing machine with a binary unidirectional input tape, an unidirectional output tape with alphabet $\mathcal{X}$, and some bidirectional work tapes. We say T halts on input p with output x and write "$T(p)=x$ halts" if p is to the left of the input head and x is to the left of the output head after T halts. The set of p on which T halts forms a prefix code. We call such codes p *self-delimiting* programs. We write $T(p)=x*$ if T outputs a string starting with x; T need not to halt in this case. p is called *minimal* if $T(q)\neq x*$ for all proper prefixes of p. The set of all prefix/monotone Turing machines $\{T_1,T_2,...\}$ can be effectively enumerated. There exists a universal prefix/monotone Turing machine U which can simulate every T_i. A function is called computable if there is a Turing machine which computes it. A function is called enumerable if it can be approximated from below. Let $\mathcal{M}_{comp}^{msr}$ be the set of all computable measures, $\mathcal{M}_{enum}^{semi}$ the set of all enumerable semimeasures, and $\mathcal{M}_{det}$ be the set of all deterministic measures ($\widehat{=}\mathcal{X}^\infty$).[4]

Levin [ZL70, LV97] has shown the existence of an enumerable universal semimeasure M ($M \overset{\times}{\geq} \nu \; \forall\nu\in\mathcal{M}_{enum}^{semi}$). An explicit expression due to Solomonoff [Sol64, Eq.(7)] is

$$M(x) \;:=\; \sum_{p:U(p)=x*} 2^{-\ell(p)}, \qquad KM(x) := -\log M(x). \tag{5}$$

The sum is over all (possibly nonhalting) minimal programs p which output a string starting with x. This definition is equivalent to the probability that U outputs a string starting with x if provided with fair coin flips on the input tape. M can be used to characterize randomness of individual sequences: A sequence $x_{1:\infty}$ is (Martin-Löf) μ-random, *iff* $\exists c: M(x_{1:n}) \leq c\cdot\mu(x_{1:n})\forall n$. For later comparison, we summarize the (excellent) predictive properties of M [Sol78, Hut01a, Hut03a, Hut04] (the numbering will become clearer later):

[4] $\mathcal{M}_{enum}^{semi}$ is enumerable, but $\mathcal{M}_{comp}^{msr}$ is not, and $\mathcal{M}_{det}$ is uncountable.

Theorem 1 (Properties of $M=2^{-KM}$) *Solomonoff's prior M defined in (5) is a (i) universal, (v) enumerable, (ii) monotone, (iii) semimeasure, which (vi) converges to μ i.m.s., and (vii) is self-optimizing i.m.s. More quantitatively:*

- (*vi*) $\sum_{t=1}^{\infty}\mathbf{E}[\sum_{x'_t}(M(x'_t|x_{<t})-\mu(x'_t|x_{<t}))^2] \overset{+}{\leq} \ln2\cdot K(\mu)$, *which implies* $M(x'_t|x_{<t}) \overset{t\to\infty}{\longrightarrow} \mu(x'_t|x_{<t})$ *i.m.s. for* $\mu\in\mathcal{M}_{comp}^{msr}$.

- (*vii*) $\sum_{t=1}^{\infty}\mathbf{E}[(l_t^{\Lambda_M}-l_t^{\Lambda_\mu})^2] \overset{+}{\leq} 2\ln2\cdot K(\mu)$, *which implies* $l_t^{\Lambda_M} \overset{t\to\infty}{\longrightarrow} l_t^{\Lambda_\mu}$ *i.m.s. for* $\mu\in\mathcal{M}_{comp}^{msr}$,

where $K(\mu)$ is the length of the shortest program computing function μ.

4 Alternatives to Solomonoff's Prior M

The goal of this work is to investigate whether some other quantities that are closely related to M also lead to good predictors. The prefix Kolmogorov complexity K is closely related to KM ($K(x)=KM(x)+O(\log\ell(x))$). $K(x)$ is defined as the length of the shortest halting program on U with output x:

$$K(x) := \min\{\ell(p) : U(p) = x \text{ halts}\}, \qquad k(x) := 2^{-K(x)}. \tag{6}$$

In Section 7 we briefly discuss that K completely fails for predictive purposes. More promising is to approximate $M(x)=\sum_{p:U(p)=x*}2^{-\ell(p)}$ by the dominant contribution in the sum, which is given by

$$m(x) := 2^{-Km(x)} \quad \text{with} \quad Km(x) := \min_p\{\ell(p) : U(p) = x*\}. \tag{7}$$

Km is called *monotone complexity* and has been shown to be *very* close to KM [Lev73a, Gác83] (see Theorem 6(*o*)). It is natural to call a sequence $x_{1:\infty}$ *computable* if $Km(x_{1:\infty})<\infty$. KM, Km, and K are ordered in the following way:

$$0 \;\leq\; K(x|\ell(x)) \;\overset{+}{\leq}\; KM(x) \;\leq\; Km(x) \;\leq\; K(x) \;\overset{+}{\leq}\; \ell(x)\cdot\log|\mathcal{X}|+2\log\ell(x). \tag{8}$$

The second inequality follows from the fact that, given n and Kraft's inequality $\sum_{x\in\mathcal{X}^n}M(x)\leq 1$, there exists for $x\in\mathcal{X}^n$ a Shannon-Fano code of length $-\log M(x)$, which is effective since M is enumerable. The other inequalities are obvious from the definitions. There are many complexity measures (prefix, Solomonoff, monotone, plain, process, extension, ...) which we generically denote by $\tilde K\in\{K,KM,Km,...\}$ and their associated "predictive functions" $\tilde k(x):=2^{-\tilde K(x)}\in\{k,M,m,...\}$. This work is mainly devoted to the study of m.

Note that $\tilde k$ is generally not a semimeasure, so we have to clarify what it means to predict using $\tilde k$. One popular approach which is at the heart of the (one-part) MDL principle is to predict the y which minimizes $\tilde K(xy)$ (maximizes $\tilde k(xy)$), where x are past given data: $y_t^{MDL}:=\text{argmin}_{y_t}\tilde K(x_{<t}y_t)$.

For complexity measures $\tilde{K}$, the conditional version $\tilde{K}_|(x|y)$ is often defined[5] as $\tilde{K}(x)$, but where the underlying Turing machine U has additionally access to y. The definition $\tilde{k}_|(x|y):=2^{-\tilde{K}_|(x|y)}$ for the conditional predictive function $\tilde{k}$ seems natural, but has the disadvantage that the crucial chain rule (1) is violated. For $\tilde{K}=K$ and $\tilde{K}=Km$ and most other versions of $\tilde{K}$, the chain rule is still satisfied approximately (to logarithmic accuracy), but this is not sufficient to prove convergence (2) or self-optimization (4). Therefore, we define $\tilde{k}(x_t|x_{<t}):=\tilde{k}(x_{1:t})/\tilde{k}(x_{<t})$ in the following, analogously to semimeasures ρ (like M). A potential disadvantage of this definition is that $\tilde{k}(x_t|x_{<t})$ is not enumerable, whereas $\tilde{k}_|(x_t|x_{<t})$ and $\tilde{k}(x_{1:t})$ are.

We can now embed MDL predictions minimizing $\tilde{K}$ into our general framework: MDL coincides with the $\Lambda_{\tilde{k}}$ predictor for the error loss:

$$y_t^{\Lambda_{\tilde{k}}} \;=\; \arg\max_{y_t} \tilde{k}(y_t|x_{<t}) \;=\; \arg\max_{y_t} \tilde{k}(x_{<t}y_t) \;=\; \arg\min_{y_t} \tilde{K}(x_{<t}y_t) \;=\; y_t^{MDL} \qquad (9)$$

In the first equality we inserted $\ell_{xy}=1-\delta_{xy}$ into (3). In the second equality we used the chain rule (1). In both steps we dropped some in argmax ineffective additive/multiplicative terms independent of y_t. In the third equality we used $\tilde{k}=2^{-\tilde{K}}$. The last equality formalizes the one-part MDL principle: given $x_{<t}$ predict the $y_t\in\mathcal{X}$ which leads to the shortest code p. Hence, validity of (4) tells us something about the validity of the MDL principle. (2) and (4) address what (good) prediction *means.*

5 General Predictive Functions

We have seen that there are predictors (actually the major one studied in this work) Λ_ρ, but where $\rho(x_t|x_{<t})$ is not (immediately) a semimeasure. Nothing prevents us from replacing ρ in (3) by an arbitrary function $b_| : \mathcal{X}^* \to [0,\infty)$, written as $b_|(x_t|x_{<t})$. We also define general functions $b:\mathcal{X}^*\to[0,\infty)$, written as $b(x_{1:n})$ and $b(x_t|x_{<t}) := \frac{b(x_{1:t})}{b(x_{<t})}$, which may not coincide with $b_|(x_t|x_{<t})$. Most terminology for semimeasure ρ can and will be carried over to the case of general predictive functions b and $b_|$, but one has to be careful which properties and interpretations still hold:

Definition 2 (Properties of predictive functions) *We call functions b,$b_|$: $\mathcal{X}^*\to[0,\infty)$ (conditional) predictive functions. They may possess some of the following properties:*

o) Proximity: $b(x)$ *is "close" to the universal prior $M(x)$*

i) Universality: $b\overset{\times}{\geq}\mathcal{M}$, *i.e.* $\forall\nu\in\mathcal{M}\,\exists c>0 : b(x)\geq c\cdot\nu(x)\forall x.$

ii) Monotonicity: $b(x_{1:t})\leq b(x_{<t})\ \forall t,x_{1:t}$

iii) Semimeasure: $\sum_{x_t} b(x_{1:t})\leq b(x_{<t})$ *and* $b(\epsilon)\leq 1$

iv) Chain rule: $b(x_{1:t})=b.(x_t|x_{<t})b(x_{<t})$

[5] Usually written without index $|$.

$v)$ Enumerability: *b is lower semicomputable*

$vi)$ Convergence: $b.(x_t'|x_{<t})\overset{t\to\infty}{\longrightarrow}\mu(x_t'|x_{<t})\ \forall\mu\in\mathcal{M},x_t'\in\mathcal{X}$ *i.m.s. or w.p.1*

$vii)$ Self-optimization: $l_t^{\Lambda_{b.}}\overset{t\to\infty}{\longrightarrow}l_t^{\Lambda_\mu}$ *i.m.s. or w.p.1*

where $b.$ refers to b or $b_|$

The importance of the properties $(i)-(iv)$ stems from the fact that they together imply convergence (vi) and self-optimization (vii). Regarding proximity (o) we left open what we mean by "close". We also did not specify $\mathcal{M}$ but have in mind all computable measures $\mathcal{M}_{comp}^{msr}$ or enumerable semimeasures $\mathcal{M}_{enum}^{semi}$, possibly restricted to deterministic environments $\mathcal{M}_{det}$.

Theorem 3 (Predictive relations)

$a)$ $(iii)\Rightarrow(ii)$: *A semimeasure is monotone.*

$b)$ (i),(iii),(iv) $\Rightarrow$ (vi): *The posterior $b.$ as defined by the chain rule (iv) of a universal semimeasure b converges to μ i.m.s. for all $\mu\in\mathcal{M}$.*

$c)$ (i),(iii),$(v)\Rightarrow(o)$: *Every w.r.t. $\mathcal{M}_{enum}^{semi}$ universal enumerable semimeasure coincides with M within a multiplicative constant.*

$d)$ (vi) $\Rightarrow$ (vii): *Posterior convergence i.m.s./w.p.1 implies self-optimization i.m.s./w.p.1.*

Proof sketch. (a) follows trivially from dropping the sum in (iii), (b) is Solomonoff's major result [Sol78, LV97, Hut01a, Hut04], (c) is due to Levin [ZL70], (d) follows from $0\leq l_t^{\Lambda_{b.}}-l_t^{\Lambda_\mu}\leq\sum_{x_t'}|b.(x_t'|x_{<t})-\mu(x_t'|x_{<t})|$, since $\ell\in[0,1]$ [Hut03a, Thm.4(ii)]. □

We will see that (i),(iii),(iv) are crucial for proving (vi),(vii).

Normalization. Let us consider a scaled b version $b_{norm}(x_t|x_{<t}):=c(x_{<t})b(x_t|x_{<t})$, where $c>0$ is independent of x_t. Such a scaling does not affect the prediction scheme Λ_b (3), i.e. $y_t^{\Lambda_b}=y_t^{\Lambda_{b_{norm}}}$, which implies $l_t^{\Lambda_{b_{norm}}}=l_t^{\Lambda_b}$. Convergence $b(x_t'|x_{<t})\to\mu(x_t'|x_{<t})$ implies $\sum_{x_t'}b(x_t'|x_{<t})\to 1$ if μ is a measure, hence also $b_{norm}(x_t'|x_{<t})\to\mu(x_t'|x_{<t})$ for[6] $c(x_{<t}):=[\sum_{x_t'}b(x_t'|x_{<t})]^{-1}$. Speed of convergence may be affected by normalization, either positively or negatively. Assuming the chain rule (1) for b_{norm} we get

$$b_{norm}(x_{1:n})=\prod_{t=1}^n\frac{b(x_{1:t})}{\sum_{x_t}b(x_{1:t})}=d(x_{<n})b(x_{1:n}),\qquad d(x_{<n}):=\frac{1}{b(\epsilon)}\prod_{t=1}^n\frac{b(x_{<t})}{\sum_{x_t}b(x_{1:t})}$$

Whatever b we start with, b_{norm} is a measure, i.e. (iii) is satisfied with equality. Convergence and self-optimization proofs are now eligible for b_{norm}, provided universality (i) can be proven for b_{norm}. If b is a semimeasure, then $d\geq 1$, hence

[6] Arbitrarily we define $b_{norm}(x_t|x_{<t})=\frac{1}{|\mathcal{X}|}$ if $\sum_{x_t'}b(x_t'|x_{<t})=0$.

$M_{norm}\geq M\overset{\times}{\geq}\mathcal{M}_{enum}^{semi}$ is universal and converges (vi) with the same bound (Theorem 1(vi)) as for M. On the other hand, $d(x_{<n})$ may be unbounded for $b=k$ and $b=m$, so normalization does not help us in these cases for proving (vi). Normalization transforms a universal non-semimeasure into a measure, which may no longer be universal.

Universal Non-Semimeasures. If $b\overset{\times}{\geq}M$ is a universal semimeasure, then b is as good for prediction as M. The bounds are loosened by at most an additive constant. For b still dominating M, but no longer being a semimeasure, we believe that (vi) and (vii) can be violated. Bounds can be shown without any further assumptions on b on-sequence and if we demand a lower *and* upper bound on b, i.e. $b\overset{\times}{=}M$, then also off-sequence:

Theorem 4 (Convergence of Universal Non-Semimeasures) *For every predictive function b, and real numbers a and c it holds:*

$$\begin{array}{lrcl} a) & \sum_{t=1}^n 1-b(x_t|x_{<t}) & \leq & \ln 2\cdot KM(x_{1:n})+\ln a^{-1} \quad \textit{if} \quad aM(x)\leq b(x)\,\forall x, \\ b) & \sum_{t=1}^n\sum_{\bar x_t\neq x_t} b(\bar x_t|x_{<t}) & \leq & \frac{c}{a}\ln 2\cdot KM(x_{1:n}) \quad \textit{if} \quad aM(x)\leq b(x)\leq cM(x)\,\forall x. \end{array}$$

For computable $x_{1:\infty}$ this implies: $b(\bar x_t|x_{<t})\to 0$ and $b_{norm}(\bar x_t|x_{<t})\to 0$ for $\bar x_t\neq x_t$, and $b(x_t|x_{<t})\to 1$ if $b(x_t|x_{<t})\leq 1$ and $b_{norm}(x_t|x_{<t})\to 1$ for $t\to\infty$.

Remarks. If b additionally is a semimeasure, i.e. $\sum_{\bar x_t\neq x_t}b(\bar x_t|x_{<t})\leq 1-b(x_t|x_{<t})$ then (a) implies an improved off-sequence bound. Note that $b(\bar x_t|x_{<t})\to 0$ does not imply $b(x_t|x_{<t})\to 1$. Furthermore, although b_{norm} is a measure, convergence cannot be concluded similarly to (10), since b_{norm} may not be universal due to a possibly unbounded normalizer $d(x_{<t})$.

Proof.
(**a**)
$$\sum_{t=1}^n 1-b(x_t|x_{<t}) \;\leq\; \sum_{t=1}^n \ln b(x_t|x_{<t})^{-1} \;=\; \ln b(x_{1:n})^{-1}$$
$$\leq\; \ln[aM(x_{1:n})]^{-1} \;=\; \ln 2\cdot KM(x_{1:n})+\ln a^{-1}$$

(**b**)

$$b(\bar x_t|x_{<t}) \;\leq\; b(\bar x_t|x_{<t})\cdot\frac{b(x_{<t})}{aM(x_{<t})} = \frac{b(x_{<t}\bar x_t)}{aM(x_{<t})} \;\leq\; \frac{cM(x_{<t}\bar x_t)}{aM(x_{<t})} \;=\; \frac{c}{a}M(\bar x_t|x_{<t}).$$

For every semimeasure it holds:

$$\sum_{t=1}^n\sum_{\bar x_t\neq x_t}\rho(\bar x_t|x_{<t}) \;\leq\; \sum_{t=1}^n 1-\rho(x_t|x_{<t}) \;\leq\; -\sum_{t=1}^n \ln\rho(x_t|x_{<t}) \;=\; -\ln\rho(x_{1:n})$$

Combining both bounds and using that M is a semimeasure we get

$$\sum_{t=1}^n\sum_{\bar x_t\neq x_t} b(\bar x_t|x_{<t}) \;\leq\; \frac{c}{a}\sum_{t=1}^n\sum_{\bar x_t\neq x_t} M(\bar x_t|x_{<t}) \;\leq\; \frac{c}{a}\ln 2\cdot KM(x_{1:n}).$$

□

6 Predictive Properties of $m\!=\!2^{-Km}$

We can now state which predictive properties of m hold, and which not. We first summarize the qualitative predictive properties of m in Corollary 5, and subsequently present detailed quantitative results in Theorems $6(o)-(vii)$, followed by an item-by-item explanation, discussion and detailed proofs.

Corollary 5 (Properties of $\boldsymbol{m\!=\!2^{-Km}}$) *For $b=m=2^{-Km}$, where Km is the monotone Kolmogorov complexity (7), the following properties of Definition 2 are satisfied/violated: (o) For every $\mu\in\mathcal{M}_{comp}^{msr}$ and every μ-random sequence $x_{1:\infty}$, $m(x_{1:n})$ equals $M(x_{1:n})$ within a multiplicative constant. m is (i) universal (w.r.t. $\mathcal{M}\!=\!\mathcal{M}_{comp}^{msr}$), (ii) monotone, and (v) enumerable, but is $\neg(iii)$ not a semimeasure. m satisfies (iv) the chain rule by definition for $m.\!=\!m$, but for $m.\!=\!m_|$ the chain rule is only satisfied to logarithmic order. For $m.\!=\!m$, m (vi) converges and (vii) is self-optimizing for deterministic $\mu\!\in\!\mathcal{M}_{comp}^{msr}\!\cap\!\mathcal{M}_{det}$, but in general not for probabilistic $\mu\!\in\!\mathcal{M}_{comp}^{msr}\!\setminus\!\mathcal{M}_{det}$.*

The lesson to learn is that although m is very close to M in the sense of (o) and m dominates all computable measures μ, predictions based on m may nevertheless fail (cf. Theorem 1).

Some proof ideas. (o) [ZL70, Thm.3.4] and [Lev73a]. (i) [Lev73a]. (ii) from $Km(xy)\geq Km(x)$ (see definition of Km). $\neg(iii)$ follows from (i),(iv),$\neg(vi)$ and Theorem 3b with $m_|\!:=\!m$. (iv) follows within log from $Km\!=\!K\!+\!O(\log)$ and [LV97, Thm.3.9.1], $\neg(iv)$, since it does not even hold within an additive constant. (v) immediate from definition. (vi) similarly as for M. $\neg(vi)$ Use $m_|\!\in\!2^{-\mathbb{N}_0}$ and define a $\mu_|\!\notin\!2^{-\mathbb{N}_0}$. (vii) follows from (vi). $\neg(vii)$ For the monotone Turing machine U defined by $U(1x0)=x0$, the loss $\ell_{00}=\ell_{11}=0$, $\ell_{10}=1$, $\ell_{01}=\frac{2}{3}$ and a Bernoulli($\frac{1}{2}$) process $\mu(x_t|x_{<t})\!=\!\frac{1}{2}$ one can show $y_t^{\Lambda_m}\!=\!0\!\neq\!1\!=\!y_t^{\Lambda_\mu}$, which implies $l_t^{\Lambda_m}\!=\!\frac{1}{2}\!>\!\frac{1}{3}\!=\!l_t^{\Lambda_\mu}$. Extending U to a universal Turing machine by $U(0^{s+1}p)\!=\!U'(p)$ leaves this result intact with probability $\geq\!1\!-\!2^{-s}$, since random strings cannot be compressed (by U'). □

6.0 Proximity of $m\!=\!2^{-Km}$

The following closeness/separation results between Km and KM are known:

Theorem 6 (o) (Proximity of $\boldsymbol{m\!=\!2^{-Km}}$)

(1) $\forall\mu\!\in\!\mathcal{M}_{comp}^{msr}\,\forall\mu$-*random* $\omega\,\exists c_\omega : Km(\omega_{1:n})\!\leq\!KM(\omega_{1:n})\!+\!c_\omega\,\forall n,$ *[Lev73a]*

(2) $KM(x)\!\leq\!Km(x)\!\leq\!KM(x)\!+\!2\log Km(x)\!+\!O(1)\,\forall x.$ *[ZL70, Thm.3.4]*

¬(3) $\forall c : Km(x)\!-\!KM(x)\!\geq\!c$ *for infinitely many* $x.$ *[Gác83]*

Remarks. The first line (o_1) shows that m is close to M within a multiplicative constant for nearly all strings in a very strong sense. $\sup_n \frac{M(\omega_{1:n})}{m(\omega_{1:n})} \leq 2^{c_\omega}$ is finite for every ω which is random (in the sense of Martin-Löf) w.r.t. *any* computable μ, but note that the constant c_ω depends on ω. Levin falsely conjectured the result to be true for *all* ω, but could only prove it to hold within logarithmic accuracy (o_2). A later result by Gács $\neg(o_3)$, indeed, shows that $Km - KM$ is unbounded (for infinite alphabet it can even increase logarithmically).

Proof. The first two properties are due to Levin and are proven in [Lev73a] and [ZL70, Thm.3.4], respectively. The third property follows easily from Gács result [Gác83], which says that if g is some monotone co-enumerable function for which $Km(x) - KM(x) \leq g(\ell(x))$ holds for all x, then $g(n)$ must be $\overset{+}{\geq} K(n)$. Assume $Km(x) - KM(x) \geq \log \ell(x)$ only for finitely many x. Then there exists a c such that $Km(x) - KM(x) \leq \log \ell(x) + c$ for *all* x. Gács' theorem now implies $\log n + c \overset{+}{\geq} K(n)\, \forall n$, which is wrong due to Kraft's inequality $\sum_n 2^{-K(n)} \leq 1$. □

6.1 Universality of $m = 2^{-Km}$

Theorem 6 (i) (Universality of $m = 2^{-Km}$)

(1) $Km(x) \overset{+}{\leq} -\log \mu(x) + K(\mu)$ *if* $\mu \in \mathcal{M}_{comp}^{msr}$, *[LV97, Thm.4.5.4]*

(2) $m \overset{\times}{\geq} \mathcal{M}_{comp}^{msr}$, *but* $m \overset{\times}{\not\geq} \mathcal{M}_{enum}^{semi}$ *(unlike* $M \overset{\times}{\geq} \mathcal{M}_{enum}^{semi}$*).*

Remarks. The first line (i_1) can be interpreted as a "continuous" coding theorem for Km and recursive μ. It implies (by exponentiation) that m dominates all computable measures (i_2). Unlike M it does *not* dominate all enumerable semimeasures. Dominance is a key feature for good predictors. From a practical point of view the assumption that the true generating distribution μ is a proper measure and computable seems not to be restrictive. The problem will be that m is not a semimeasure.

Proof. The first line is proven in [LV97, Thm.4.5.4]. Exponentiating this result gives $m(x) \geq c_\mu \mu(x)\, \forall x, \mu \in \mathcal{M}_{comp}^{msr}$, i.e. $m \overset{\times}{\geq} \mathcal{M}_{comp}^{msr}$. Exponentiation of $\neg(o_3)$ implies $m(x) \overset{\times}{\not\geq} M(x) \in \mathcal{M}_{enum}^{semi}$, i.e. $m \overset{\times}{\not\geq} \mathcal{M}_{enum}^{semi}$. □

6.2 Monotonicity of $m = 2^{-Km}$

Monotonicity of Km is obvious from the definition of Km and is the origin of calling Km monotone complexity:

Theorem 6 (ii) (Monotonicity of $m = 2^{-Km}$)

$$Km(xy) \geq Km(x) \in I\!N_0, \quad 0 < m(xy) \leq m(x) \in 2^{-I\!N_0} \leq 1 = m(\epsilon).$$

6.3 Non-Semimeasure Property of $m = 2^{-Km}$

While m is monotone, it is not a semimeasure. The following theorem shows and quantifies how the crucial semimeasure property is violated for m in an essential way.

Theorem 6 (iii) (Non-Semimeasure property of $\boldsymbol{m = 2^{-Km}}$)

$\neg$(1) *If $x_{1:\infty}$ is computable, then $\sum_{x_t} m(x_{1:t}) \not\leq m(x_{<t})$ for almost all t,*

$\neg$(2) *If $Km(x_{1:t}) = o(t)$, then $\sum_{x_t} m(x_{1:t}) \not\leq m(x_{<t})$ for most t.*

Remark. On the other hand, at least for computable environments, multiplying Theorem $6(vi_{1\&3})$ by $m(x_{<t})$ shows that asymptotically the violation gets small, i.e. $\sum_{x_t} m(x_{1:t}) \xrightarrow{t\to\infty} m(x_{<t})$ for computable $x_{1:\infty}$.

Proof. Simple violation of the semimeasure property can be inferred indirectly from m possessing properties (i),(iv),$\neg(vi)$ (see Definition 2) and Theorem $3b$. To prove $\neg(\mathbf{iii_1})$ we first note that $Km(x) < \infty$ for all finite strings $x \in \mathcal{X}^*$, which implies $m(x_{1:n}) > 0$. Hence, whenever $Km(x_{1:n}) = Km(x_{<n})$, we have $\sum_{x_n} m(x_{1:n}) > m(x_{1:n}) = m(x_{<n})$, a violation of the semimeasure property. $\neg(\mathbf{iii_2})$ now follows from

$$\begin{aligned}
\#\{t \leq n : \sum_{x_t} m(x_{1:t}) \leq m(x_{<t})\} \;&\leq\; \#\{t \leq n : Km(x_{1:t}) \neq Km(x_{<t})\} \\
&\leq\; \sum_{t=1}^{n} [Km(x_{1:t}) - Km(x_{<t})] \;=\; Km(x_{1:n}),
\end{aligned}$$

where we exploited (ii) in the last inequality. □

6.4 Chain Rule for $m = 2^{-Km}$

Theorem 6 (iv) (Chain rule for $\boldsymbol{m = 2^{-Km}}$)

(1) $0 < m(x|y) := \frac{m(yx)}{m(y)} \leq 1$.

$\neg$(2) *If $m_|(x|y) := 2^{-\min_p\{\ell(p):U(p,y)=x*\}}$, then $\exists x,y : m(yx) \neq m_|(x|y) \cdot m(y)$.*

$\neg$(3) $Km(yx) = Km_|(x|y) + Km(y) \pm O(\log \ell(xy))$.

Remarks. Line 1 shows that the chain rule can be satisfied by definition. With such a definition, $m(x|y)$ is strictly positive like $M(x|y)$, but not necessarily strictly less than 1, unlike $M(x|y)$. Nevertheless it is bounded by 1 due to monotonicity of m, unlike for k (see Theorem 7). If a conditional monotone complexity $Km_| = -\log m_|$ is defined similarly to the conditional Kolmogorov complexity $K_|$, then the chain rule is only valid within logarithmic accuracy (lines 2 and 3).

Proof $(\mathbf{iv_1})$ is immediate from (ii). $\neg(\mathbf{iv_2})$ follows from the fact that equality does not even hold within an additive constant, i.e. $Km(yx) \overset{+}{\neq} Km(x|y) + Km(y)$. The proof of the latter is similar to the one for K (see [LV97]). $\neg(\mathbf{iv_3})$ follows within log from $Km = K + O(\log)$ and Theorem $7(iv)$. □

6.5 Enumerability of $m=2^{-Km}$

m shares the obvious enumerability property with M and Km shares the obvious co-enumerability property with K:

Theorem 6 (v) (Enumerability of $\boldsymbol{m=2^{-Km}}$)

(1) *m is enumerable, i.e. lower semicomputable.*

(2) *Km is co-enumerable, i.e. upper semicomputable.*

6.6 Convergence of $m=2^{-Km}$

Theorem 6 (vi) (Convergence of $\boldsymbol{m=2^{-Km}}$)

(1) $\sum_{t=1}^n|1-m(x_t|x_{<t})|\leq\frac{1}{2}Km(x_{1:n}),\quad m(x_t|x_{<t})\xrightarrow{fast}1$ *for comp.* $x_{1:\infty}$.

(2) *Indeed,* $m(x_t|x_{<t})\neq 1$ *at most* $Km(x_{1:\infty})$ *times.*

(3) $\sum_{t=1}^n\sum_{\bar x_t\neq x_t}m(\bar x_t|x_{<t})\leq 2^{Km(x_{1:n})},\qquad m(\bar x_t|x_{<t})\xrightarrow{slow?}0$ *for comp.* $x_{1:\infty}$.

(4) $\sum_{t=1}^n\sum_{\bar x_t\neq x_t}m(\bar x_t|x_{<t})\overset{\times}{\leq}[Km(x_{1:n})]^3,\quad m(\bar x_t|x_{<t})\xrightarrow{fast?}0$ *for comp.* $x_{1:\infty}$.

¬(5) $\forall s\,\exists\,U,x_{1:\infty}:Km(x_{1:\infty})=s$ *and* $\sum_{t=1}^\infty\sum_{\bar x_t\neq x_t}m(\bar x_t|x_{<t})\geq 2^s-2$.

¬(6) $\exists\mu\in\mathcal{M}_{comp}^{msr}\setminus\mathcal{M}_{det}:m_{(norm)}(x_t|x_{<t})\overset{t\to\infty}{\not\longrightarrow}\mu(x_t|x_{<t})\,\forall x_{1:\infty}$

Remarks. Line 1 shows that the on-sequence predictive properties of m for deterministic computable environments are excellent. The predicted m-probability[7] of x_t given $x_{<t}$ converges rapidly to 1 for reasonably simple $x_{1:\infty}$. A similar result holds for M.

The stronger result (second line), that $m(x_t|x_{<t})$ deviates from 1 at most $Km(x_{1:\infty})$ times, does not hold for M.

Note that without constraint on the predictive function b, perfect on-sequence prediction could trivially be achieved by defining $b_{\cdot}(x'_t|x_{<t})\equiv 1\ \forall x'_t$, which correctly predicts x_t with "probability" 1. But since we do not know the true outcome x_t in advance, we need to predict the probability of x'_t well for all $x'_t\in\mathcal{X}$. $m(|)$ also converges off-sequence for $\bar x_t\neq x_t$ (to zero as it should be), but the bound (third line) is much weaker than the on-sequence bound (first line), so rapid convergence cannot be concluded, unlike for M, where $M(x_t|x_{<t})\xrightarrow{fast}1$ implies $M(\bar x_t|x_{<t})\xrightarrow{fast}0$, since $\sum_{x'_t}M(x'_t|x_{<t})\leq 1$. Consider an environment $x_{1:\infty}$ describable in 500 bits, then bound (vi_3) does not exclude $m(\bar x_t|x_{<t})$ from being 1 (maximally wrong) for all $t=1..2^{500}$; with asymptotic convergence being of pure academic interest.

Line 4 presents a bound polynomial in Km, which is theoretically better than the exponential bound of line 3, but there is a pitfall due to the hidden multiplicative constant.

[7] We say "probability" just for convenience, not forgetting that $m(\cdot|x_{<t})$ is not a proper (semi)probability distribution.

Line 5 shows that for particular universal Turing machines this constant can be exponentially large. Note that this does not contradict the polynomial bound, since the multiplicative constant 2^{c_U} is allowed to depend on U. For a reasonable Turing machine, the compiler constant c_U is of reasonable size, but 2^{c_U} is unreasonably large. Let U' be a Turing machine which you regard as reasonable. Then, for e.g. $s=64=O(1)$, the U constructed in the proof is as reasonable as U' in the sense that a program of U' needs only to be prefixed by a short 64 bit word to run on U (the compiler constant between U and U' is small). In this sense, there are *reasonable* Turing machines U for which m makes the unreasonably large number of $2^{64}-2$ prediction errors on the trivial sequence $0_{1:\infty}$, as we will show.

Line 6 shows that the situation is provably worse in the probabilistic case. There are computable measures μ for which neither $m(x_t|x_{<t})$ nor $m_{norm}(x_t|x_{<t})$ converge to $\mu(x_t|x_{<t})$ for any $x_{1:\infty}$. So while [VL00, Thm.11] and [LV97, Thm.5.2.3] stating that $\mu(x_{t:t+l}|x_{<t})\overset{\times}{=}m(x_{t:t+l}|x_{<t})$ for μ-random $x_{1:\infty}$ and fixed l is correct, the conclusion [VL00, Cor.2] and [LV97, Cor.5.2.2] that (m is good for prediction in the sense that) maximizing $\mu(\cdot|x_{<t})$ is asymptotically equivalent to maximizing $m(\cdot|x_{<t})$, is wrong. For this to be true we would need convergence without multiplicative fudge, and which also holds off-sequence, i.e. $m_{(norm)}(x'_t|x_{<t})\to\mu(x'_t|x_{<t})$, but which $\neg(vi_6)$ just shows to fail (even on-sequence).

Proof (vi$_{\mathbf{1\&2}}$) $\quad \#\{t\le n: m(x_t|x_{<t})\neq 1\} \;\le\; \sum_{t=1}^n 2|1-m(x_t|x_{<t})| \;\le$

$$\le\; -\sum_{t=1}^n \log m(x_t|x_{<t}) \;=\; -\log m(x_{1:n}) \;=\; Km(x_{1:n}).$$

In the first inequality we used $m:=m(x_t|x_{<t})\in 2^{-\mathbb{N}_0}$, hence $1\le 2|1-m|$ for $m\neq 1$. In the second inequality we used $1-m\le -\frac{1}{2}\log m$, valid for $m\in[0,\frac{1}{2}]\cup\{1\}\supset 2^{-\mathbb{N}_0}$. In the first equality we used (the log of) the chain rule n times. For computable $x_{1:\infty}$ we have $\sum_{t=1}^\infty|1-m(x_t|x_{<t})|\le\frac{1}{2}Km(x_{1:\infty})<\infty$, which implies $m(x_t|x_{<t})\to 0$ (fast if $Km(x_{1:\infty})$ is of reasonable size). This shows the first two lines of (vi).

(vi$_\mathbf{3}$) Fix a sequence $x_{1:\infty}$ and define $\mathcal{Q}:=\{x_{<t}\bar x_t : t\in\mathbb{N}, \bar x_t\neq x_t\}$. $\mathcal{Q}$ is a prefix-free set of finite strings. For any such $\mathcal{Q}$ and any semimeasure ρ, one can show that $\sum_{x\in\mathcal{Q}}\rho(x)\le 1$.[8] Since M is a semimeasure lower-bounded by m we get

$$\sum_{t=1}^n\sum_{\bar x_t\neq x_t} m(x_{<t}\bar x_t) \;\le\; \sum_{t=1}^\infty\sum_{\bar x_t\neq x_t} m(x_{<t}\bar x_t) \;=\; \sum_{x\in\mathcal{Q}} m(x) \;\le\; \sum_{x\in\mathcal{Q}} M(x) \;\le\; 1.$$

With this, and using monotonicity of m we get

$$\sum_{t=1}^n\sum_{\bar x_t\neq x_t} m(\bar x_t|x_{<t}) = \sum_{t=1}^n\sum_{\bar x_t\neq x_t}\frac{m(x_{<t}\bar x_t)}{m(x_{<t})} \le \sum_{t=1}^n\sum_{\bar x_t\neq x_t}\frac{m(x_{<t}\bar x_t)}{m(x_{1:n})} \le \frac{1}{m(x_{1:n})} = 2^{Km(x_{1:n})}$$

Finally, for an infinite sum to be finite, its elements must converge to zero.

[8] This follows from $1\ge\rho(A\cup B)\ge\rho(A)+\rho(B)$ if $A\cap B=\{\}$, $\Gamma_x\cap\Gamma_y=\{\}$ if x not prefix of y and y not prefix of x, where $\Gamma_x:=\{\omega:\omega_{1:\ell(x)}=x\}$, hence $\sum_{x\in\mathcal{Q}}\rho(\Gamma_x)\le\rho(\bigcup_{x\in\mathcal{Q}}\Gamma_x)\le 1$, and noting that $\rho(x)$ is actually an abbreviation for $\rho(\Gamma_x)$.

(vi$_4$) For $t\leq n$ we can bound

$$m(\bar x_t|x_{<t}) \;\equiv\; \frac{m(x_{<t}\bar x_t)}{m(x_{<t})} \;\overset{\times}{\leq}\; Km^2(x_{<t})\frac{M(x_{<t}\bar x_t)}{M(x_{<t})} \;\leq\; Km^2(x_{1:n})M(\bar x_t|x_{<t})$$

In the first inequality we exploited Theorem 6(o_2) in the exponentiated form $M(x)/Km^2(x)\overset{\times}{\leq}m(x)\leq M(x)$. In the last inequality we used monotonicity of m. Using Theorem 4 with $a=c=1$ and $b=M$ and $KM\leq Km$ we get

$$\sum_{t=1}^n\sum_{\bar x_t\neq x_t} m(\bar x_t|x_{<t}) \;\overset{\times}{\leq}\; Km^2(x_{1:n})\sum_{t=1}^n\sum_{\bar x_t\neq x_t} M(\bar x_t|x_{<t}) \;\leq\; \ln 2\cdot Km^3(x_{1:n}).$$

Note that using (o_1) instead of (o_2) leads to a bound $2^{c_\omega}\ln 2\cdot Km(\omega)$, which for computable ω is also finite, but of unspecified magnitude due to the factor 2^{c_ω}.

$\neg$**(vi$_5$)** Fix $s\in I\!\!N$ and let $t\in T:=\{1,...,2^s-2\}$. We define a universal monotone Turing machine U by $U(0^s)=0^\infty$ and $U(q)=0^{t-1}1*$ for $q\in\{0,1\}^s\setminus\{0^s,1^s\}$, where $t\in T$ is the natural number represented by the s-bit string q (any coding will do). Only for the purpose of making U universal, we define $U(1^sp)=U'(p)$ for $p\in\{0,1\}^*$ and U' being some (other, e.g. your favorite) universal Turing machine. Obviously the length of the shortest programs on U for $0_{1:\infty}$, $0_{<t}1$ and $0_{<t}$ is s, i.e. $Km(0_{1:\infty})=Km(0_{<t})=Km(0_{<t}1)=s$, which implies $m(1|0_{<t})=1$. So for $x_{1:\infty}=0_{1:\infty}$, we have

$$\sum_{t=1}^\infty\sum_{\bar x_t\neq x_t} m(\bar x_t|x_{<t}) \;\geq\; \sum_{t=1}^{2^s-2} m(1|0_{<t}) \;=\; 2^s-2,$$

which proves $\neg(iv_5)$. Note that $m_{norm}(1|0_{<t})\geq\frac{1}{|\mathcal X|}$, i.e. save a factor of $|\mathcal X|$ the same lower bound holds for m_{norm}. Note also that on-sequence prediction is perfect, since $m(0|0_{<t})=1\;\forall t\in I\!\!N$.

Remark. It is instructive to see why $M(\bar x_t|x_{<t})$ converges fast to 0 for this U: The single program of size s for $0_{<t}1$ is outweighed by the 2^s-t programs of size s for $0_{<t}$. Ignoring the contributions from U', we have $M(1|0_{<t})\approx\frac{1\cdot 2^{-s}}{(2^s-t)\cdot 2^{-s}}=\frac{1}{2^s-t}$, hence $\sum_{t=1}^{2^s-2}M(1|0_{<t})\approx s\cdot\ln 2$.

$\neg$**(vi$_6$)** We show that the range of $m_{(norm)}$ is not dense in [0,1] and then choose a μ not in the closure of the range. For binary alphabet $\mathcal X=\{0,1\}$, the proof is particularly simple: We choose $\mu(1|x_{<t})=\frac38$, hence $\mu(0|x_{<t})=\frac58$. Since $m(x_t|x_{<t})\in 2^{-I\!\!N_0}=\{1,\frac12,\frac14,\frac18,...\}$, we have $|m(x_t|x_{<t})-\mu(x_t|x_{<t})|\geq\frac18\,\forall t,\forall x_{1:\infty}$. Similarly for

$$m_{norm}(x_t|x_{<t}) \;=\; \frac{m(x_t|x_{<t})}{m(0|x_{<t})+m(1|x_{<t})} \;\in\; \left\{\frac{2^{-n}}{2^{-n}+2^{-m}}:n,m\in I\!\!N_0\right\} \;=$$
$$=\;\left\{\frac{1}{1+2^z}:z\in Z\!\!\!Z\right\} \;=\; \frac{1}{1+2^{Z\!\!\!Z}} \;=\; \left\{...,\frac19,\frac15,\frac13,\frac12,\frac23,\frac45,\frac89,...\right\}$$

we choose $\mu(1|x_{<t})=1-\mu(0|x_{<t})=\frac{5}{12}$, which implies $|m_{norm}(x_t|x_{<t})-\mu(x_t|x_{<t})|\geq\frac{1}{12}$ $\forall t$, $\forall x_{1:\infty}$.

Consider now a general alphabet $\mathcal{X}=\{1,...,|\mathcal{X}|\}$, and the unnormalized m first. If $|\mathcal{X}|$ is not a power of 2 we define $\mu(x_t|x_{<t})=|\mathcal{X}|^{-1}$. If $|\mathcal{X}|$ is a power of 2 we define $\mu(x_t|x_{<t})=\frac{4}{3}|\mathcal{X}|^{-1}$ for even x_t and $\mu(x_t|x_{<t})=\frac{2}{3}|\mathcal{X}|^{-1}$ for odd x_t. μ is a measure, $0\neq\mu(x_t|x_{<t})\notin 2^{-I\!N_0}$, but $m(x_t|x_{<t})\in 2^{-I\!N_0}$. The only cluster[9] point of $2^{-I\!N_0}$ is 0, since $0\neq\mu\notin 2^{-I\!N_0}$ there exists $\gamma>0$ such that $(\mu-\gamma,\mu+\gamma)\cap 2^{-I\!N_0}=\{\}$, hence $|m(x_t|x_{<t})-\mu(x_t|x_{<t})|\geq\gamma\,\forall t,\forall x_{1:\infty}$ for some $\gamma>0$.

For m_{norm} we proceed as follows: With $z_i:=Km(1|x_{<t})-Km(i|x_{<t})\in\mathbb{Z}$, we have $m_{norm}(1|x_{<t})^{-1}=1+\sum_{i=2}^{|\mathcal{X}|}2^{z_i}$. We define $\mathcal{S}:=\{1+m_2+...+m_{|\mathcal{X}|}:m_i\in 2^{\mathbb{Z}}\cup\{0\}\forall i\}\not\ni 0$ and $\mathcal{I}:=\mathcal{S}^{-1}=\{x^{-1}:x\in\mathcal{S}\}$. By construction, $m_{norm}(1|x_{<t})\in\mathcal{I}$, and by symmetry also $m_{norm}(x_t|x_{<t})\in\mathcal{I}$. The cross product $\mathcal{I}^{|\mathcal{X}|}:=\mathcal{I}\times\overset{|\mathcal{X}|times}{\ldots\ldots\ldots}\times\mathcal{I}$ is a closed and countable set, since $2^{\mathbb{Z}}\cup\{0\}$ is closed and countable, and finite sums, inversions, and cross products of closed/countable sets, are closed/countable.[10] With $\Delta:=\{\mathbf{v}\in I\!R^{|\mathcal{X}|}:0<v_i<1,\,\sum_{i=1}^{|\mathcal{X}|}v_i=1\}$ being the open $|\mathcal{X}|-1$ dimensional simplex, we have $m_{norm}(\cdot|x_{<t})\in\mathcal{I}^{|\mathcal{X}|}\cap\Delta$ (e.g. $\mathcal{I}^2\cap\Delta=\{(\frac{1}{1+2^z},\frac{1}{1+2^{-z}}):z\in\mathbb{Z}\}$). Since $\Delta\backslash\mathcal{I}^{|\mathcal{X}|}$ is open and nonempty (due to countability of $\mathcal{I}^{|\mathcal{X}|}$), there exists $\mu(\cdot|x_{<t})\in\Delta\backslash\mathcal{I}^{|\mathcal{X}|}$ and a Box:$=\{\mathbf{v}:|v_i-\mu(i|x_{<t})|<\gamma\}$ of sufficiently small size $\gamma>0$ surrounding μ, such that Box$\cap\mathcal{I}^{|\mathcal{X}|}=\{\}$, which implies the desired result $|m(x_t|x_{<t})-\mu(x_t|x_{<t})|\geq\gamma$.

Remark. There is an easy proof for the weaker statement $m_{norm}(x'_t|x_{<t})\not\to\mu(x'_t|x_{<t})$, where x'_t may be off-sequence: For $\mu(0|x_{<t})=\frac{1}{4}=1-\mu(1|x_{<t})$ we have $\frac{\mu(1|x_{<t})}{\mu(0|x_{<t})}=3\notin 2^{\mathbb{Z}}$, while $\frac{m_{norm}(1|x_{<t})}{m_{norm}(0|x_{<t})}\in 2^{\mathbb{Z}}$. This implies that the posterior of m_{norm} cannot be too close to the posterior of μ for *all* x'_t, i.e. $\exists x'_t$ and $c>0$: $|m_{norm}(x'_t|x_{<t})-\mu(x'_t|x_{<t})|\geq c$ ($c=\frac{1}{20}$ possible). One advantage of this proof is that it also goes through for infinite alphabet $\mathcal{X}$. □

6.7 Self-optimization of $m=2^{-Km}$

Theorem 6 (vii) (Self-optimization of $\boldsymbol{m=2^{-Km}}$)

(1) $l_t^{\Lambda_m}(x_{<t})\overset{slow?}{\longrightarrow}l_t^{\Lambda_\omega}:=\text{argmin}_{y_t}\ell_{x_ty_t}$ *if* $\omega\equiv x_{1:\infty}$ *is computable.*

(2) $\Lambda_m=\Lambda_{m_{norm}}$, *i.e.* $y_t^{\Lambda_m}=y_t^{\Lambda_{m_{norm}}}$ *and* $l_t^{\Lambda_m}=l_t^{\Lambda_{m_{norm}}}$.

¬(3) $\forall|\mathcal{Y}|>2\,\exists\ell,\mu:l_t^{\Lambda_m}/l_t^{\Lambda_\mu}=c>1\,\forall t$ ($c=\frac{6}{5}-\varepsilon$ *possible*).

¬(4) $\exists\ell,\mu:l_t^{\Lambda_m}/l_t^{\Lambda_\mu}=c>1$ *for many t with μ-probability* $\geq\frac{1}{2}$ ($c=\sqrt{2}-\varepsilon$ *possible*).

¬(5) $\forall$ *non-degenerate*[11] $\ell\,\exists U,\mu:l_t^{\Lambda_m}/l_t^{\Lambda_\mu}\overset{t\to\infty}{\not\longrightarrow}1$ *with high probability.*

Remarks. Since (vi) implies (vii_1) by continuity, we have convergence of the instantaneous losses for computable environments $x_{1:\infty}$, but since convergence off-sequence is potentially slow, the convergence of the losses to optimum is potentially slow.

[9] A point $p\in I\!R^n$ is called a cluster point of a set $\mathcal{S}\subseteq I\!R^n$, if every open set of $I\!R^n$ which contains p, intersects $\mathcal{S}$.

[10] W.r.t. standard topology on $I\!R^n$.

[11] A formal definition of *non-degenerate* is given in the remarks after the theorem.

Non-convergence $\neg(vi_6)$ in probabilistic environments does not necessarily imply that Λ_m is not self-optimizing, since different predictive functions can lead to the same predictor Λ. But $\neg(vii_4)$ shows that Λ_m is not self-optimizing even in Bernoulli environments μ for particular losses ℓ with probability $\geq\frac{1}{2}$.

Interestingly, excluding binary action alphabets allows for a stronger for-sure statement $\neg(vii_3)$.

In $\neg(vii_5)$, non-self-optimization is shown for *any non-degenerate loss function* (especially for the error loss, cf. (9)), for specific choices of the universal Turing machine U. Loss ℓ is defined to be non-degenerate *iff* $\bigcap_{x\in\mathcal{X}}\{\tilde{y}:\ell_{x\tilde{y}}=\min_y\ell_{xy}\}=\{\}$. Assume the contrary that a *single* action $\tilde{y}$ is optimal for *every* outcome x, i.e. that (argmin$_y$ can be chosen such that) $\text{argmin}_y\ell_{xy}=\tilde{y}\,\forall x$. This implies $y_t^{\Lambda_\rho}=\tilde{y}\,\forall\rho$, which implies $l_t^{\Lambda_m}/l_t^{\Lambda_\mu}\equiv 1$. So the non-degeneracy assumption is necessary (and sufficient).

Proof (vii$_1$) follows from $(vi_{1\&3})$ and Theorem 3d.

(**vii$_2$**) That normalization does not affect the predictor, follows from the definition of $y_t^{\Lambda_\rho}$ (3) and the fact that argmin() is not affected by scaling its argument.

$\neg$(**vii$_3$**) Non-convergence of m does not necessarily imply non-convergence of the losses. For instance, for $\mathcal{X}=\mathcal{Y}=\{0,1\}$, and $\omega_t':=1/0$ for $\mu(1|x_{<t})\gtrless\gamma:=\frac{\ell_{01}-\ell_{00}}{\ell_{01}-\ell_{00}+\ell_{10}-\ell_{11}}$, one can show that $y_t^{\Lambda_\mu}=y_t^{\Lambda_{\omega'}}$, hence convergence of $m(x_t|x_{<t})$ to 0/1 and not to $\mu(x_t|x_{<t})$ could nevertheless lead to correct predictions.

Consider now $x\in\mathcal{X}=\{0,1\}$, $y\in\mathcal{Y}=\{0,1,2\}$. To prove $\neg(vii_3)$ we define a loss function such that $y_t^{\Lambda_\mu}\neq y_t^{\Lambda_\rho}$ for any ρ with same range as m_{norm} and for some μ. The loss function $\ell_{x0}=x$, $\ell_{x1}=\frac{3}{8}$, $\ell_{x2}=\frac{2}{3}(1-x)$, and $\mu:=\mu(1|x_{<t})=\frac{2}{5}$ will do. The ρ-expected loss under action y is $l_\rho^y:=\sum_{x_t=0}^1\rho(x_t|x_{<t})\ell_{x_ty}$; $l_\rho^0=\rho$, $l_\rho^1=\frac{3}{8}$, $l_\rho^2=\frac{2}{3}(1-\rho)$ with $\rho:=\rho(1|x_{<t})$ (see Figure 1). Since $l_\mu^0=l_\mu^2=\frac{2}{5}>\frac{3}{8}=l_\mu^1$, we have $y_t^{\Lambda_\mu}=1$ and $l_t^{\Lambda_\mu}=l_\mu^1=\frac{3}{8}$. For $\rho\leq\frac{1}{3}$, we have $l_\rho^0<l_\rho^1<l_\rho^2$, hence $y_t^{\Lambda_\rho}=0$ and $l_t^{\Lambda_\rho}=l_\mu^0=\frac{2}{5}$. For $\rho\geq\frac{1}{2}$, we have $l_\rho^2<l_\rho^1<l_\rho^0$, hence $y_t^{\Lambda_\rho}=2$ and $l_t^{\Lambda_\rho}=l_\mu^2=\frac{2}{5}$. Since $m_{norm}\not\in(\frac{1}{3},\frac{1}{2})$, $\Lambda_{m_{norm}}$ predicts 0 or 2, hence $l_t^{\Lambda_m}=l_\mu^{0/2}=\frac{2}{5}$. Since $\Lambda_{m_{norm}}=\Lambda_m$, this shows that $l_t^{\Lambda_m}/l_t^{\Lambda_\mu}=\frac{16}{15}>1$. The constant $\frac{16}{15}$ can be enlarged to $\frac{6}{5}-\varepsilon$ by setting $\ell_{x1}=\frac{1}{3}+\varepsilon$ instead of $\frac{3}{8}$.

For $\mathcal{Y}=\{0,...,|\mathcal{Y}|-1\}$, $|\mathcal{Y}|>3$, we extend the loss function by defining $\ell_{xy}=1$ $\forall y\geq 3$, ensuring that actions $y\geq 2$ are never favored. For $\mathcal{X}=\{0,...,|\mathcal{X}|-1\}$, $|\mathcal{X}|>2$, we extend μ and define $\mu(x_t|x_{<t})=0$ $\forall x_t\geq 2$. Furthermore, we define $\ell_{xy}=0$ for $x\geq 2$ and $y<3$. This ensures that the extra components of $m_{norm}(x_t|x_{<t})$ with $x_t\geq 2$ do not contribute to $l_{m_{norm}}^y$. Finally, and this is important, we define, solely for the purpose of this proof, $m_{norm}(x_t|x_{<t})=\frac{m(x_t|x_{<t})}{m(0|x_{<t})+m(1|x_{<t})}$, such that $m_{norm}(0|x_{<t})+m_{norm}(1|x_{<t})=1$ (rather than $\sum_{x_t=0}^{|\mathcal{X}|-1}m_{norm}(x_t|x_{<t})=1$) (Normalization influences the analysis, but not the result). With these extensions, the analysis of the $|\mathcal{X}|=2$, $|\mathcal{Y}|=3$ case applies, which finally shows $\neg(vii)$. In general, a non-dense range of $\rho(x_t|x_{<t})$ implies $l_t^{\Lambda_\rho}\not\to l_t^{\Lambda_\mu}$, provided $|\mathcal{Y}|\geq 3$.

$\neg$(**vii$_4$**) We consider binary $\mathcal{X}=\mathcal{Y}=\{0,1\}$ first. The proof idea and notation is similar to $\neg(vii_3)$. We choose a $\mu:=\mu(1|x_{<t})\not\in\frac{1}{1+2^{\mathbb{Z}}}$. Let $a,b\in\frac{1}{1+2^{\mathbb{Z}}}$ with $a<\mu<b$ be the nearest (to μ) possible values of $m_{norm}\in\frac{1}{1+2^{\mathbb{Z}}}$. For a fixed sequence $x_{1:\infty}$, we have either $m(1|x_{<t})\leq a$ for (infinitely) many t or $m(1|x_{<t})\geq b$ for (infinitely)

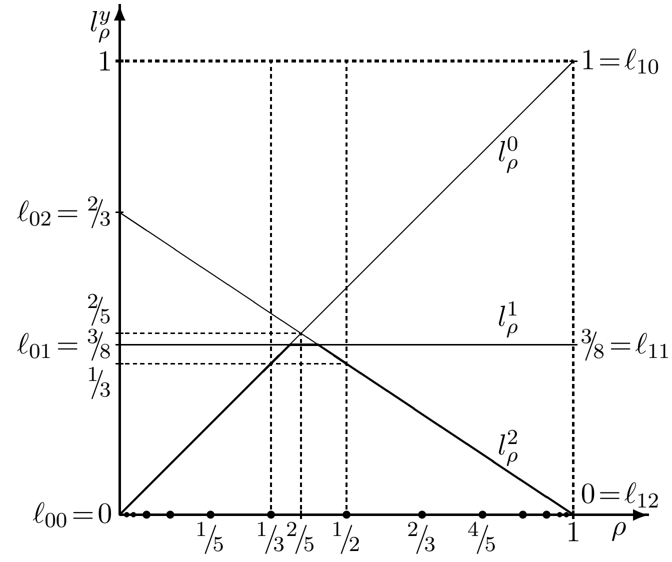

Figure 1 (Example loss used in proof of Theorem 6¬(*vii*)) *The ρ-expected expected losses l^y_ρ under actions $y\in\mathcal{Y}=\{0,1,2\}$ for $\mathcal{X}=\{0,1\}$ and loss function $\ell_{00}=\ell_{12}=00$, $\ell_{01}=\ell_{11}=\frac{3}{8}$, $\ell_{02}=\frac{2}{3}$, and $\ell_{10}=1$ are displayed as solid lines.*

many t (or both). Choosing $x_{1:\infty}$ at random, we have either $m(1|x_{<t})\leq a$ for many t with μ-probability $\geq\frac{1}{2}$ or $m(1|x_{<t})\geq b$ for many t with μ-probability $\geq\frac{1}{2}$ (or both). Assume the former; for the latter the proof is analogous. We consider a loss function such that $l^1_a>l^0_a$ and $l^1_\mu<l^0_\mu$. Then also $l^1_m>l^0_m$ whenever $m\leq a$, which is the case for many t by assumption. Hence $l_t^{\Lambda_m}/l_t^{\Lambda_\mu}=l^0_\mu/l^1_\mu=c>1$. For instance, choose $\mu=\sqrt{2}-1$ and $\ell_{00}=0$ and $\ell_{10}=1$ ($\Rightarrow l^0_\rho=\rho$). We get $c=\sqrt{2}-O(\varepsilon)$ by choosing $\ell_{01}=\frac{1}{2}+\varepsilon$ and $\ell_{11}=0$ ($\Rightarrow l^1_\rho=(\frac{1}{2}+\varepsilon)(1-\rho)$) in the former case with $a=\frac{1}{3}$ (and $\ell_{01}=1-\varepsilon$ and $\ell_{11}=0$ ($\Rightarrow l^1_\rho=(1-\varepsilon)(1-\rho)$) in the latter case with $b=\frac{1}{2}$ and $l^1_b<l^0_b$ and $l^1_\mu>l^0_\mu$). The generalization to general $\mathcal{X}$ and $\mathcal{Y}$ can be performed similarly to $\neg(vii_3)$.

$\neg(\mathbf{vii_5})$ We first present a simple proof for a particular loss function and $\mathcal{X}=\mathcal{Y}=\{0,1\}$, which contains the main idea also used to prove the general result. We define a monotone Turing machine U by $U(1x0)=x0$ for all $x\in\mathcal{X}^*$. More precisely, if the first bit of the input tape of U contains 1, U copies the half-infinite input tape (without the first 1) to the output tape, but always withholds the output until a 0 appears. We have $Km(x1)=Km(x10)=\ell(x)+2=Km(x0)+1$, which implies $m_{norm}(1|x)=\frac{1}{3}$ and $m_{norm}(0|x)=\frac{2}{3}$. For the loss function $\ell_{00}=\ell_{11}=0$, $\ell_{10}=1$, $\ell_{01}=\frac{2}{3}$ and a Bernoulli($\frac{1}{2}$) process $\mu(x_t|x_{<t})=\frac{1}{2}$ we get $l^1_\mu=\frac{1}{2}\cdot\frac{2}{3}<\frac{1}{2}=l^0_\mu$ and $l^1_{m_{norm}}=\frac{2}{3}\cdot\frac{2}{3}>\frac{1}{3}=l^0_{m_{norm}}$, hence $l_t^{\Lambda_m}/l_t^{\Lambda_\mu}=l^0_\mu/l^1_\mu=\frac{3}{2}>1$. U is not yet universal. We make U universal by additionally defining $U(0^{s+1}p)=U'(p)$ for some (large, but reasonable) $s\in\mathbb{N}$ and some (other) universal monotone TM U'. We have to check whether this can alter (lower) the monotone complexity. Fix n. Every x of length n has description $1x0$ of length $n+2$, so U' only matters if $U'(p)=x*$ for some p of length $<n-s+1$. Since there

are at most 2^{n-s} minimal programs of length $\leq n-s$, the fraction of problematic x is at most 2^{-s}. Since x is drawn at random, the loss ratio $l_t^{\Lambda_m}/l_t^{\Lambda_\mu}=\frac{3}{2}$, hence, holds with high probability ($\geq 1-2^{-s}$). A martingale argument (see below) shows that this implies $l_t^{\Lambda_m}/l_t^{\Lambda_\mu}\overset{t\to\infty}{\not\longrightarrow}1$ (w.h.p.).

We now consider the case of general loss and alphabets. In case where ambiguities in the choice of y in $\text{argmin}_y\ell_{xy}$ matter we consider the set of solutions $\{\text{argmin}_y\ell_{xy}\}:=\{\tilde y:\ell_{x\tilde y}=\min_y\ell_{xy}\}\neq\{\}$. By assumption, ℓ is non-degenerate, i.e. $\bigcap_{x\in\mathcal{X}}\{\text{argmin}_y\ell_{xy}\}=\{\}$. Let $\mathcal{X}_m$ be a minimal subset of $\mathcal{X}$ with $\bigcap_{x\in\mathcal{X}_m}\{\text{argmin}_y\ell_{xy}\}=\{\}$. Take any decomposition $\mathcal{X}_0\dot\cup\mathcal{X}_1=\mathcal{X}_m$ with $\mathcal{X}_0\neq\{\}\neq\mathcal{X}_1$, which is possible, since $|\mathcal{X}_m|\geq 2$. We have $\mathcal{Y}_i:=\bigcap_{x\in\mathcal{X}_i}\{\text{argmin}_y\ell_{xy}\}\neq\{\}$, since $\mathcal{X}_m$ is minimal. Further, $\mathcal{Y}_0\cap\mathcal{Y}_1=\mathcal{Y}_m=\{\}$. It is convenient to choose $|\mathcal{X}_1|=1$. W.l.g. we assume $\mathcal{X}_1=\{1\}$.

Define some $\mathcal{Q}\subset\{0,1\}^s$, $|\mathcal{Q}|=|\mathcal{X}_0|$, a bijection $b:\mathcal{Q}\to\mathcal{X}_0$, and a one-to-one (onto $\mathcal{A}$) decoding function $d:\{0,1\}^s\to\mathcal{A}$ with $\mathcal{A}=\mathcal{X}_0 1^s\cup 1\{0,1\}^s\setminus 1\mathcal{Q}\subset\mathcal{X}^{s+1}$ as $d(x)=b(x)1^s$ for $x\in\mathcal{Q}$ and $d(x)=1x$ for $x\in\{0,1\}^s\setminus\mathcal{Q}$ with a large $s\in\mathbb{N}$ to be determined later. We extend d to $d:(\{0,1\}^s)^*\to\mathcal{A}^*$ by defining $d(z_1...z_k)=d(z_1)...d(z_k)$ for $z_i\in\{0,1\}^s$ and define the inverse coding function $c:\mathcal{A}\to\{0,1\}^s$ and its extension $c:\mathcal{A}^*\to(\{0,1\}^s)^*$ by $c=d^{-1}$.

Roughly, U is defined as $U(1p_{1:sn}q)=d(p_{1:sn})b(q)1^s$ for $q\in\mathcal{Q}$. More precisely, if the first bit of the binary input tape of U contains 1, U decodes the successive blocks of size s, but always withholds the output until a block $q\in\mathcal{Q}$ appears. U is obviously monotone. Universality will be guaranteed by defining $U(0p)$ appropriately, but for the moment we set $U(0p)=\epsilon$. It is easy to see that for $x\in\mathcal{A}^*$ we have

$$\begin{array}{rcll} Km(xx_0)= & Km(xx_01^s) & = & \ell(c(x))+s+1 \quad \text{for } x_0\in\mathcal{X}_0, \\ Km(x1)= & Km(x1z0_{1:s+1}) & = & \ell(c(x))+2s+1 \quad \text{for any } z\in\{0,1\}^s\setminus\mathcal{Q}, \\ Km(xy)= & & = & \infty \quad \text{for any } y\in\mathcal{X}\setminus(\mathcal{X}_0\cup\{1\}). \end{array} \tag{10}$$

Hence, $m_{norm}(x_0|x)=[|\mathcal{X}_0|+2^{-s}]^{-1}\overset{s\to\infty}{\longrightarrow}1$ and $m_{norm}(1|x)=[2^s|\mathcal{X}_0|+1]^{-1}\overset{s\to\infty}{\longrightarrow}0$ and $m_{norm}(y|x)=0$. For $t-1\in(s+1)\mathbb{N}$ we get $l_m^{y_t}:=\sum_{x_t}m_{norm}(x_t|x_{<t})\ell_{x_ty_t}\overset{s\to\infty}{\longrightarrow}\frac{1}{|\mathcal{X}_0|}\sum_{x_t\in\mathcal{X}_0}\ell_{x_ty_t}$. This implies

$$y_t^{\Lambda_m}\in\{\arg\min_{y_t}l_m^{y_t}\}\subseteq\{\arg\min_y\frac{1}{|\mathcal{X}_0|}\sum_{x\in\mathcal{X}_0}\ell_{xy}\}=\bigcap_{x\in\mathcal{X}_0}\{\arg\min_y\ell_{xy}\}\equiv\mathcal{Y}_0. \tag{11}$$

Inclusion $\subseteq$ holds for sufficiently large finite s. Equality $=$ holds, since the set of points which are global maxima of a linear average of functions coincides with the set of points which simultaneously maximize all these functions, if the latter is nonempty.

We now define $\mu(z)=|\mathcal{A}|^{-1}=2^{-s}$ for $z\in\mathcal{A}$ and $\mu(z)=0$ for $z\in\mathcal{X}^{s+1}\setminus\mathcal{A}$, extend it to $\mu(z_1...z_k):=\mu(z_1)\cdot...\cdot\mu(z_k)$ for $z_i\in\mathcal{X}^{s+1}$, and finally extend it uniquely to a measure on $\mathcal{X}^*$ by $\mu(x_{<t}):=\sum_{x_{t:n}}\mu(x_{1:n})$ for $\mathbb{N}\ni t\leq n\in(s+1)\mathbb{N}$. For $x\in\mathcal{A}^*$ we have $\mu(x_0|x)=\mu(x_0)=\mu(x_01^s)=2^{-s}\overset{s\to\infty}{\longrightarrow}0$ and $\mu(1|x)=\mu(1)=\sum_{y\in\mathcal{X}^s}\mu(1y)=\sum_{y\in\{0,1\}^s\setminus\mathcal{Q}}\mu(1y)=(2^s-|Q|)\cdot 2^{-s}=1-|X_0|2^{-s}\overset{s\to\infty}{\longrightarrow}1$. For $t-1\in(s+1)\mathbb{N}$ we get $l_\mu^{y_t}:=\sum_{x_t}\mu(x_t|x_{<t})\ell_{x_ty_t}\overset{s\to\infty}{\longrightarrow}\ell_{1y_t}$. This implies

$$y_t^{\Lambda_\mu}\in\{\arg\min_{y_t}l_\mu^{y_t}\}\subseteq\{\arg\min_y\ell_{1y}\}\equiv\mathcal{Y}_1 \quad \text{for sufficiently large finite } s. \tag{12}$$

Since $\mathcal{Y}_0 \cap \mathcal{Y}_1 = \{\}$, (11) and (12) imply $y_t^{\Lambda_m} \neq y_t^{\Lambda_\mu}$, which implies $l_t^{\Lambda_m} \neq l_t^{\Lambda_\mu}$ (otherwise the choice $y_t^{\Lambda_m} = y_t^{\Lambda_\mu}$ would have been possible), which implies $l_t^{\Lambda_m}/l_t^{\Lambda_\mu} = c > 1$ for $t-1 \in (s+1)I\!N$, i.e. for (infinitely) many t.

What remains to do is to extend U to a universal Turing machine. We extend U by defining $U(0zp) = U'(p)$ for any $z \in \{0,1\}^{3s}$, where U' is some universal Turing machine. Clearly, U is now universal. We have to show that this extension does not spoil the preceding consideration, i.e. that the shortest code of x has sufficiently often the form $1p$ and sufficiently seldom the form $0p$. Above, μ has been chosen in such a way that $c(x)$ is a Shannon-Fano code for μ-distributed strings, i.e. $c(x)$ is with high μ-probability a shortest code of x. More precisely, $\ell(c(x)) \leq Km_T(x) + s$ with μ-probability at least $1 - 2^{-s}$, where Km_T is the monotone complexity w.r.t. any decoder T, especially $T = U'$. This implies $\min_p\{\ell(0p) : U(0p) = x*\} = 3s + 1 + Km_{U'}(x) \geq 3s + 1 + \ell(c(x)) - s > \ell(c(x)) + s + 1 \geq \min_p\{\ell(1p) : U(1p) = x*\}$, where the first $\geq$ holds with high probability $(1-2^{-s})$ and the last $\geq$ holds with μ-probability 1. This shows that the expressions (10) for Km are with high probability (w.h.p.) not affected by the extension of U. Altogether this shows $l_t^{\Lambda_m}/l_t^{\Lambda_\mu} = c > 1$ w.h.p.

A martingale argument can strengthen this result to yield non-selfoptimizingness. For $z_t := \frac{M(\omega_{1:t})}{\mu(\omega_{1:t})}$ we have $z_0 = 1$, $\mathbf{E}[z_t] \leq 1$, and $\mathbf{E}[z_t|\omega_{<t}] \leq z_{t-1}$, hence $-z_t$ is a non-positive semi-martingale. [Doo53, Thm.4.1s,p324] now implies that $z_\infty := \lim_{t\to\infty} z_t$ exists w.p.1 and $\mathbf{E}[z_\infty] \leq \lim_{t\to\infty}\mathbf{E}[z_t] \leq 1$. The Markov inequality now yields

$$\mathbf{P}[\lim_{t\to\infty}(KM(\omega_{1:t}) + \log\mu(\omega_{1:t})) \leq -s] = \mathbf{P}[z_\infty \geq 2^s] \leq 2^{-s}\mathbf{E}[z_\infty] \leq 2^{-s}.$$

Substituting $KM \leq Km \rightsquigarrow Km_{U'}$ and $-\log\mu(x) = \ell(c(x))$ this shows that $\ell(c(\omega_{1:t})) \leq Km_{U'}(\omega_{1:t}) + s$ for almost all $t \in (s+1)I\!N$ with probability $\geq 1 - 2^{-s}$. Altogether this shows $l_t^{\Lambda_m}/l_t^{\Lambda_\mu} \overset{t\to\infty}{\not\to} 1$ w.h.p. □

7 Further Results

Predictive Properties of $\boldsymbol{k = 2^{-K}}$. We briefly discuss the predictive properties of the prefix Kolmogorov complexity K. We will be very brief, since K completely fails for predictive purposes, although K is close to KM within an additive logarithmic term.

Theorem 7 (Properties of $\boldsymbol{k = 2^{-K}}$) *For $b = k = 2^{-K}$, where K is the prefix Kolmogorov complexity, the following properties of Definition 2 are satisfied/violated: (o) $KM(x) \leq K(x) \leq KM(x) + 2\log K(x)$. (i),(ii),(iii) are violated. (iv) is satisfied only for $k. = k$ For $k. = k_|$ (iv) is only satisfied to logarithmic order. In any case (vi) and (vii) can be violated for deterministic as well as probabilistic $\mu \in \mathcal{M}_{comp}^{msr}$. (v) is satisfied.*

Proof sketch. (o) Similar to proof of Theorem 3.4 in [ZL70]. $\neg(i)$ for deterministic $\mu \in \mathcal{M}_{comp}^{msr}$ with $\mu(0_{1:n}) = 1$, we have $k(0_{1:n}) \to 0 \overset{\times}{\not\geq} 1 = \mu(0_{1:n})$, since $K(\omega_{1:n}) \overset{n\to\infty}{\longrightarrow} \infty\ \forall\omega$.

$\neg(ii)$, since $K(0_{1:n})\overset{+}{=}K(n)\geq\log n$ for most n, but $\overset{+}{\leq}2\mathrm{loglog}\,n$ for n being a power of 2. $\neg(ii)$ implies $\neg(iii)$. (iv) within log follows from [LV97, Thm.3.9.1]. $\neg(iv)$, since it does not even hold within an additive constant (see [LV97, p231]). (v) immediate from definition. $\neg(vii)$ Define a universal prefix Turing machine U via some other universal prefix Turing machine U' by $U(00p)=U'(p)0$, $U(1p)=U'(p)1$, $U(01)=\epsilon$. For this U we have $K(x0)=K(x1)+1\,\forall x$ $(K=K_U)$, which implies that Λ_k for the error loss always predicts 1. $\neg(vi)$ follows from $\neg(vii)$. $\square$

Also, $K(x|\ell(x))$ is a poor predictor, since $K(x0|\ell(x0))\overset{\pm}{=}K(x1|\ell(x1))$, and the additive constant can be chosen to ones need by an appropriate choice of U. Note that the larger a semimeasure, the more distributions it dominates, the better its predictive properties. This simple rule does not hold for non-semimeasures. Although M predicts better than m predicts better than k in accordance with (8), $2^{-K(x|\ell(x))}\overset{\times}{\geq}M(x)$ is a bad predictor disaccording with (8).

Simple MDL. There are other ways than m of using shortest programs for predictions. We have chosen the (in our opinion) most natural and promising way. A somewhat simpler version of MDL is to take the shortest (nonhalting) program p which outputs x, continue running p, and use the continuation y of x for prediction:

$$\widetilde{m}_|(x_t|x_{<t}):=1 \text{ if shortest program for } x_{<t}* \text{ computes } x_{<t}x_t*, \qquad \widetilde{m}_|(\bar{x}_t|x_{<t}):=0.$$

Theorem 8 (Properties of $\widetilde{m}$) *For the simple MDL predictor $\widetilde{m}_|(x_t|x_{<t})$ and $\widetilde{m}(x_{1:n}):=\prod_{t=1}^n\widetilde{m}_|(x_t|x_{<t})$, the following holds: $\widetilde{m}$ is a deterministic, (ii) monotone, (iii) measure, satisfying (iv) the chain rule (by definition), is $\neg(i)$ not universal w.r.t. $\mathcal{M}_{comp}^{msr}\cap\mathcal{M}_{det}$, and is $\neg(v)$ not enumerable, and is $\neg(vi)$ not convergent and $\neg(vii)$ not self-optimizing w.r.t. some $\mu\in\mathcal{M}_{comp}^{msr}$.*

Note that $\widetilde{m}_|$ contains more information than $\widetilde{m}$. $\widetilde{m}_|$ cannot be reconstructed from $\widetilde{m}$, since $\widetilde{m}_|(x_t'|x_{<t})$ is defined even if $\widetilde{m}(x_{<t})=0$. $\neg(vi)$ and $\neg(vii)$ follow from non-denseness $\{\widetilde{m}_|\}=\{0,1\}$. For $\neg(i)$ take $\omega=1^\infty$ in case $\widetilde{m}(1)=0$, and 0^∞ otherwise. We did not check the convergence properties for deterministic environments.

Another possibility is to define $m=f(Km)$ with f some monotone decreasing function other than $f(Km)=2^{-Km}$, since $m=2^{-Km}$ is not a semimeasure anyway. We do not expect exciting results.

8 Outlook and Open Problems

Speed of off-sequence convergence of m for computable environments. A more detailed analysis of the speed of convergence of $m(\bar{x}_t|x_{<t})$ to zero in deterministic environments would be interesting: How close are the off-sequence upper bound (vi_4) $\overset{\times}{=}Km^3$ and the lower bound $\neg(vi_5)$ 2^s-2. Can the lower bound be

improved to $2^s\cdot Km$? Maybe for the witnesses of $m\overset{\times}{\not=}M$? The upper bound can be improved to $\overset{\times}{=}Km^2\cdot\log Km$. Can the bound be improved to $\overset{\times}{=}Km$? Probably the most interesting open question is whether there exist universal Turing machines for which the multiplicative constant is of reasonable size. We expect that these hypothetical TMs, if they exist, are very natural in the sense that they also possess other convenient properties.

Non-self-optimization for general $\boldsymbol{U}$ and $\boldsymbol{\ell}$. Another open problem is whether for every non-degenerate loss-function, self-optimization of Λ_m can be violated. We have shown that this is the case for particular choices of the universal Turing machine U. If Λ_m were self-optimizing for some U and general loss, this would be an unusual situation in Algorithmic Information Theory, where properties typically hold for all or no U. So we expect Λ_m not to be self-optimizing for general loss and U (particular μ of course). A first step may be to try to prove that for all U there exists a computable sequence $x_{1:\infty}$ such that $K(x_{<t}\bar{x}_t)<K(x_{<t}x_t)$ for (infinitely) many t (which shows $\neg(vii)$ for K and error loss), and then try to generalize to probabilistic μ, Km, and general loss functions.

Other complexity measures. This work analyzed the predictive properties of the monotone complexity Km. This choice was motivated by the fact that m is the MDL approximation of the sum M, and Km is *very* close to KM. We expect all other (reasonable) alternative complexity measure to perform worse than Km. But we should be careful with precipitative conclusions, since closeness of unconditional predictive functions not necessarily implies good prediction performance, so distantness may not necessarily imply poor performance. Besides the discussed prefix Kolmogorov complexity K [Lev74, Gác74, Cha75], monotone complexity Km [Lev73a], and Solomonoff's universal prior $M=2^{-KM}$ [Sol64, Sol78, ZL70], one may investigate the predictive properties of the plain Kolmogorov complexity C [Kol65], process complexity [Sch73], Chaitin's complexity Kc [Cha75], extension semimeasure Mc [Cov74], uniform complexity [Lov69b, Lov69a], cumulative K^E and general K^G complexity and corresponding measures [Sch02a], predictive complexity KP [VW98], speed prior S [Sch02b], Levin complexity [Lev73b, Lev84], and several others. Most of them are described in [LV97]. Many properties and relations are known for the unconditional versions, but little relevant for prediction of the conditional versions is known.

Two-part MDL. We have approximated $M(x):=\sum_{p:U(p)=x*}2^{-\ell(p)}$ by its dominant contribution $m(x)=2^{-Km(x)}$, which we have interpreted as deterministic or one-part universal MDL. There is another representation of M due to Levin [ZL70] as a mixture over semimeasures: $M(x)=\sum_{\nu\in\mathcal{M}_{enum}^{semi}}2^{-K(\nu)}\nu(x)$ with dominant contribution $m_2(x)=2^{-Km_2(x)}$ and universal two-part MDL $Km_2(x):=\min_{\nu\in\mathcal{M}_{enum}^{semi}}\{-\log\nu(x)+K(\nu)\}$. MDL "lives" from the validity of this approximation. $K(\nu)$ is the complexity of the probabilistic model ν, and $-\log\nu(x)$ is the (Shannon-Fano) description length of data x in model ν. MDL usually refers to two-part MDL, and not to one-part MDL. A natural question is to ask about the predictive properties of m_2, similarly to m. m_2 is even closer to M than m is ($m_2 \overset{\times}{=} M$), but is also not a semimeasure. Drawing the analogy to m further, one may ask whether (slow) posterior convergence $m_2 \to \mu$ w.p.1 for computable probabilistic environments μ holds. In [PH04a, PH04b] we show, more generally, slow posterior convergence of two-part MDL w.p.1 in probabilistic environments μ. See also [BC91], for convergence results for two-part MDL in i.i.d. environments.

More abstract proofs showing that violation of some of the criteria $(i)-(iv)$ necessarily lead to violation of (vi) or (vii) may deal with a number of complexity measures simultaneously. For instance, we have seen that any non-dense posterior set $\{\tilde{k}(x_t|x_{<t})\}$ implies non-convergence and non-self-optimization in probabilistic environments; the particular structure of m did not matter. Maybe a probabilistic version of Theorem 4 on the convergence of universal non-semimeasures is possible under some (mild?) extra assumptions on b.

Extra conditions. Non-convergence or non-self-optimization of m do not necessarily mean that m fails in practice. Often one knows more than that the environment is (probabilistically) computable, or the environment possess certain additional properties, even if unknown. So one should find sufficient and/or necessary extra conditions on μ under which m converges / Λ_m self-optimizes rapidly. The results of this work have shown that for m-based prediction one *has* to make extra assumptions (as compared to M). It would be interesting to characterize the class of environments for which universal MDL alias m is a good predictive approximation to M. Deterministic computable environments were such a class, but a rather small one, and convergence can be slow.

References

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[Cha75] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.

[Cov74] T. M. Cover. Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. Technical Report 12, Statistics Department, Stanford University, Stanford, CA, 1974.

[Dem68] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, Series B 30:205–247, 1968.

[Doo53] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

[Gác74] P. Gács. On the symmetry of algorithmic information. *Soviet Mathematics Doklady*, 15:1477–1480, 1974.

[Gác83] P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theoretical Computer Science*, 22:71–93, 1983.

[Hut01a] M. Hutter. Convergence and error bounds for universal prediction of nonbinary sequences. In *Proc. 12th European Conf. on Machine Learning (ECML-2001)*, volume 2167 of *LNAI*, pages 239–250, Freiburg, 2001. Springer, Berlin.

[Hut01b] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

[Hut03a] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

[Hut03b] M. Hutter. Sequence prediction based on monotone complexity. In *Proc. 16th Annual Conf. on Learning Theory (COLT-2003)*, volume 2777 of *LNAI*, pages 506–521, Washington, DC, 2003. Springer, Berlin.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[Kol65] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1(1):1–7, 1965.

[KV86] P. R. Kumar and P. P. Varaiya. *Stochastic Systems: Estimation, Identification, and Adaptive Control.* Prentice Hall, Englewood Cliffs, NJ, 1986.

[Lev73a] L. A. Levin. On the notion of a random sequence. *Soviet Mathematics Doklady*, 14(5):1413–1416, 1973.

[Lev73b] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9:265–266, 1973.

[Lev74] L. A. Levin. Laws of information conservation (non-growth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10(3):206–210, 1974.

[Lev84] L. A. Levin. Randomness conservation inequalities: Information and independence in mathematical theories. *Information and Control*, 61:15–37, 1984.

[Lov69a] D. W. Loveland. On minimal-program complexity measures. In *Proc. 1st ACM Symposium on Theory of Computing*, pages 61–78. ACM Press, New York, 1969.

[Lov69b] D. W. Loveland. A variant of the Kolmogorov concept of complexity. *Information and Control*, 15(6):510–526, 1969.

[LV97] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer, Berlin, 2nd edition, 1997.

[PH04a] J. Poland and M. Hutter. Convergence of discrete MDL for sequential prediction. In *Proc. 17th Annual Conf. on Learning Theory (COLT-2004)*, volume 3120 of *LNAI*, pages 300–314, Banff, 2004. Springer, Berlin.

[PH04b] J. Poland and M. Hutter. On the convergence speed of MDL predictions for Bernoulli sequences. In *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT-2004)*, volume 3244 of *LNAI*, pages 294–308, Padova, 2004. Springer, Berlin.

[Sch73] C. P. Schnorr. Process complexity and effective random tests. *Journal of Computer and System Sciences*, 7(4):376–388, 1973.

[Sch00] J. Schmidhuber. Algorithmic theories of everything. Report IDSIA-20-00, quant-ph/0011122, IDSIA, Manno (Lugano), Switzerland, 2000.

[Sch02a] J. Schmidhuber. Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. *International Journal of Foundations of Computer Science*, 13(4):587–612, 2002.

[Sch02b] J. Schmidhuber. The speed prior: A new simplicity measure yielding near-optimal computable predictions. In *Proc. 15th Conf. on Computational Learning Theory (COLT-2002)*, volume 2375 of *LNAI*, pages 216–228, Sydney, 2002. Springer, Berlin.

[Sha76] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transaction on Information Theory*, IT-24:422–432, 1978.

[VL00] P. M. B. Vitányi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2):446–464, 2000.

[VW98] V. G. Vovk and C. Watkins. Universal portfolio selection. In *Proc. 11th Conf. on Computational Learning Theory (COLT-98)*, pages 12–23. ACM Press, New York, 1998.

[ZL70] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.