SEQUENCE PREDICTION BASED ON MONOTONE COMPLEXITY*

Marcus Hutter

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

6 June 2003

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 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 behavior is unclear. In probabilistic environments, neither the posterior nor the losses converge, in general.

Keywords

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

*This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

1 Introduction

Complexity based sequence prediction. 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 $\tilde{k}(x)$. For simplicity we only consider binary alphabet $\mathcal{X}=\{0,1\}$ in this work.

The most well-studied complexity regarding its predictive properties is $KM(x)= -\log M(x)$, where $M(x)$ is Solomonoff's universal prior [Sol64, Sol78]. Solomonoff has shown that the posterior $M(x_t|x_1...x_{t-1})$ rapidly converges to the true data generating distribution. In [Hut01b, Hut02] 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 $M(x_{1:t})$, 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 [Lev73, 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's 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 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 4* 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-optimizingness. Some later needed normalization issues are also discussed. *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 behavior is unclear. In probabilistic environments, m neither converges, nor is it self-optimizing, in general. The proofs are presented in *Section 7*. *Section 8* contains an outlook and a list of open questions.

2 Notation and Setup

Strings and natural numbers. We write $\mathcal{X}^*$ for the set of finite strings over binary alphabet $\mathcal{X}=\{0,1\}$, 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, $l(x)$ 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 $\sum_{x\in\mathcal{P}}2^{-l(x)}\leq 1$.

Asymptotic notation. We abbreviate $\lim_{t\to\infty}[f(t)-g(t)]=0$ by $f(t)\stackrel{t\to\infty}{\longrightarrow}g(t)$ and say f converges to g, without implying that $\lim_{t\to\infty}g(t)$ itself exists. We write $f(x)\stackrel{\times}{\leq}g(x)$ for $f(x)=O(g(x))$ and $f(x)\stackrel{+}{\leq}g(x)$ 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 $a_t\stackrel{t\to\infty}{\longrightarrow}0$. 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 $\#\mathcal{S}$.

(Semi)measures. We call $\rho:\mathcal{X}^*\to[0,1]$ a (semi)measure *iff* $\sum_{x_n\in\mathcal{X}}\rho(x_{1:n})\stackrel{(\leq)}{=}\rho(x_{<n})$ and $\rho(\epsilon)\stackrel{(\leq)}{=}1$. $\rho(x)$ is interpreted as the ρ-probability of sampling a sequence which starts with x. The conditional probability (posterior)

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

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 "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

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

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 $\{\omega: z_t(\omega)\stackrel{t\to\infty}{\longrightarrow}0\}$ 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 $x_{1:n}$

[1] Environments of interest have reasonable complexity K, but 2^K is not of reasonable size.
[2] Also called *objective* or *aleatory* probability or *chance*.
[3] Also called *subjective* or *belief* or *epistemic* probability.

by means of a loss function, for instance $\ell_{x_t y_t} := (x_t - y_t)^2$, which does not involve μ. More generally, let $\ell_{x_t y_t} \in [0,1] \subset I\!R$ 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

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

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

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

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

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

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

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

We define a prefix Turing machine T as a Turing machine with binary unidirectional input and output tapes, 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 Turing-machines $\{T_1, T_2, ...\}$ can be effectively enumerated. There exists a universal prefix 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]

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

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 [Sol78] is

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

The sum is over all (possibly non-halting) 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, Hut02] (the numbering will become clearer later):

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) $\sum_{t=1}^\infty \mathbf{E}[\sum_{x'_t}(M(x'_t|x_{<t})-\mu(x'_t|x_{<t}))^2] \overset{+}{\leq} \ln 2\cdot K(\mu)$, *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) $\sum_{t=1}^\infty \mathbf{E}[(l_t^{\Lambda_M}-l_t^{\Lambda_\mu})^2] \overset{+}{\leq} 2\ln 2\cdot K(\mu)$, *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 which are closely related to M also lead to good predictors. The prefix Kolmogorov complexity K is closely related to KM ($K(x)=KM(x)+O(\log l(x))$). $K(x)$ is defined as the length of the shortest halting program on U with output x:

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

In Section 8 we briefly discuss that K completely fails for predictive purposes. More promising is to approximate $M(x)=\sum_{p:U(p)=x*}2^{-l(p)}$ by the dominant contribution in the sum, which is given by

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

Km is called *monotone complexity* and has been shown to be *very* close to KM [Lev73, Gác83] (see also Theorem 5(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:

$$0\ \leq\ K(x|l(x))\ \overset{+}{\leq}\ KM(x)\ \leq\ Km(x)\ \leq\ K(x)\ \overset{+}{\leq}\ l(x)+2\log l(x). \tag{8}$$

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 the 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-optimizingness (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:*

[5] Usually written without index $|$.

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

i) Universality: $b\stackrel{\times}{\geq}\mathcal{M}$, *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: $b(x_{1:t})=b.(x_t|x_{<t})b(x_{<t})$

v) Enumerability: *b is lower semi-computable*

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

vii) Self-optimizingness: $l_t^{\Lambda_{b.}}\stackrel{t\to\infty}{\longrightarrow}l_t^{\Lambda_\mu}$ *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-optimizingness (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-optimizingness i.m.s./w.p.1.*

Proof sketch. (a) follows trivially from dropping the sum in (ii), (b) and (c) are Solomonoff's major results [Sol78, LV97, Hut01a], (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 $\ell\in[0,1]$ [Hut02, Th.4(ii)]. $\square$

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

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

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

Whatever b we start with, b_{norm} is a measure, i.e. (iii) is satisfied with equality. Convergence and self-optimizingness 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 $M_{norm}\geq M\overset{\times}{\geq}\mathcal{M}_{enum}^{semi}$ is universal and converges (vi) with 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.

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

We can now state which predictive properties of m hold, and which not. In order not to overload the reader, we first summarize the qualitative predictive properties of m in Corollary 4, and subsequently present detailed quantitative results in Theorem 5, followed by an item-by-item explanation and discussion. The proofs are deferred to the next section.

Corollary 4 (Properties of $m\!=\!2^{-Km}$) *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).

Theorem 5 (Detailed properties of $m\!=\!2^{-Km}$) *For $b = m = 2^{-Km}$, where $Km(x):=\min_p\{l(p):U(p)=x*\}$ is the monotone Kolmogorov complexity, the following properties of Definition 2 are satisfied / violated:*

- (o) $\forall\mu\in\mathcal{M}_{comp}^{msr}\,\forall\mu$*-random* $\omega\,\exists c_\omega: Km(\omega_{1:n})\leq KM(\omega_{1:n})+c_\omega\,\forall n,$ *[Lev73]*
  $KM(x)\leq Km(x)\leq KM(x)+2\log Km(x)+O(1)\,\forall x.$ *[ZL70, Th.3.4]*
- $\neg(o)$ $\forall c: Km(x)-KM(x)\geq c$ *for infinitely many* x. *[Gác83]*
- (i) $Km(x)\overset{+}{\leq}-\log\mu(x)+K(\mu)$ *if* $\mu\in\mathcal{M}_{comp}^{msr}$, *[LV97, Th.4.5.4]*
  $m\overset{\times}{\geq}\mathcal{M}_{comp}^{msr}$, *but* $m\overset{\times}{\not\geq}\mathcal{M}_{enum}^{semi}$ *(unlike* $M\overset{\times}{\geq}\mathcal{M}_{enum}^{semi}$*).*
- (ii) $Km(xy)\geq Km(x)\in I\!\!N_0, \quad 0<m(xy)\leq m(x)\in 2^{-I\!\!N_0}\leq 1.$

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

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

$\neg(iv)$ *if* $m_|(x|y) := 2^{-\min_p\{l(p):U(p,y)=x*\}}$, *then* $\exists x,y : m(yx) \neq m_|(x|y) \cdot m(y)$,
$Km(yx) = Km_|(x|y) + Km(y) \pm O(\log l(xy))$.

(v) *m is enumerable, i.e. lower semi-computable.*

(vi) $\sum_{t=1}^n |1 - m(x_t|x_{<t})| \leq \frac{1}{2} Km(x_{1:n})$, $\quad m(x_t|x_{<t}) \stackrel{fast}{\longrightarrow} 1$ *for comp.* $x_{1:\infty}$,
Indeed, $m(x_t|x_{<t}) \neq 1$ *at most* $Km(x_{1:\infty})$ *times,*
$\sum_{t=1}^n m(\bar{x}_t|x_{<t}) \leq 2^{Km(x_{1:n})}$, $\quad m(\bar{x}_t|x_{<t}) \stackrel{slow?}{\longrightarrow} 0$ *for computable* $x_{1:\infty}$.

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

(vii) $l_t^{\Lambda_m}(x_{<t}) \stackrel{slow?}{\longrightarrow} l_t^{\Lambda_\omega} := \text{argmin}_{y_t} \ell_{x_t y_t}$ *if* $\omega \equiv x_{1:\infty}$ *is computable.*
$\Lambda_m = \Lambda_{m_{norm}}$, *i.e.* $y_t^{\Lambda_m} = y_t^{\Lambda_{m_{norm}}}$ *and* $l_t^{\Lambda_m} = l_t^{\Lambda_{m_{norm}}}$.

$\neg(vii)$ $\forall |\mathcal{Y}| > 2 \, \exists \ell,\mu : l_t^{\Lambda_m}/l_t^{\Lambda_\mu} = c > 1 \, \forall t \quad$ ($c = \frac{6}{5} - \varepsilon$ *possible*).
$\forall$ *non-degenerate* ℓ $\exists U,\mu : l_t^{\Lambda_m}/l_t^{\Lambda_\mu} \stackrel{t\to\infty}{\not\longrightarrow} 1$ *with high probability.*

Explanation and discussion. (o) The first line 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 (second line).

$\neg(o)$ A later result by Gács, indeed, implies that $Km - KM$ is unbounded (for infinite alphabet it can even increase logarithmically).

(i) The first line can be interpreted as a "continuous" coding theorem for Km and recursive μ. It implies (by exponentiation) that m dominates all computable measures (second line). 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.

(ii) The monotonicity property is obvious from the definition of Km and is the origin of calling Km monotone complexity.

$\neg(iii)$ shows and quantifies how the crucial semimeasure property is violated for m in an essential way, where *almost all* n means "all but finitely many," and *most* n means "all but an asymptotically vanishing fraction.".

(iv) 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.

$\neg(iv)$ 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.

(v) m shares the obvious enumerability property with M.

(vi) (first line) 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/complex $x_{1:\infty}$. 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 perfect on-sequence prediction could trivially be achieved by always predicting 1 ($b.\equiv 1$). Since we do not know the true outcome x_t in advance, we need to predict $m(x'_t|x_{<t})$ 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 $M(x_t|x_{<t}) \stackrel{fast}{\longrightarrow} 1$ implies $M(\bar{x}_t|x_{<t}) \stackrel{fast}{\longrightarrow} 0$, since $\sum_{x'_t} M(x'_t|x_{<t}) \leq 1$. Consider an environment $x_{1:\infty}$ describable in 500 bits, then bound (vi) line 2 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.

$\neg(vi)$ 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 $x_{1:\infty}$.

(vii) Since (vi) implies (vii) by continuity, we have convergence of the instantaneous losses for computable environments $x_{1:\infty}$, but since we do not know the speed of convergence off-sequence, we do not know how fast the losses converge to optimum.

$\neg(vii)$ Non-convergence $\neg(vi)$ does not necessarily imply that Λ_m is not self-optimizing, since different predictive functions can lead to the same predictor Λ. But it turns out that Λ_m is not self-optimizing even in Bernoulli environments μ for particular losses ℓ (first line). A similar result holds for *any non-degenerate loss function* (especially for the error-loss, cf. (9)), for specific choices of the universal Turing-machine U (second line). 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).

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

7 Proof of Theorem 5

(o) The first two properties are due to Levin and are proven in [Lev73] and [ZL70, Th.3.4], respectively. The third property is an easy corollary from Gács result [Gác83], which says that if g is some monotone co-enumerable function for which $Km(x)-KM(x)\leq g(l(x))$ holds for all x, then $g(n)$ must be $\stackrel{+}{\geq} K(n)$. Assume $Km(x)-KM(x)\geq\log l(x)$ only for finitely many x only. Then there exists a c such that $Km(x)-KM(x)\leq\log l(x)+c$ for *all* x. Gács' theorem now implies $\log n+c\stackrel{+}{\geq} K(n)\,\forall n$, which is wrong due to Kraft's inequality $\sum_n 2^{-K(n)}\leq 1$.

(i) The first line is proven in [LV97, Th.4.5.4]. Exponentiating this result gives $m(x)\geq c_\mu\mu(x)\,\forall x,\mu\in\mathcal{M}_{comp}^{msr}$, i.e. $m\stackrel{\times}{\geq}\mathcal{M}_{comp}^{msr}$. Exponentiation of $\neg(o)$ gives $m(x)\leq M(x)/l(x)$, which implies $m(x)\stackrel{\times}{\not\geq}M(x)\in\mathcal{M}_{enum}^{semi}$, i.e. $m\stackrel{\times}{\not\geq}\mathcal{M}_{enum}^{semi}$.

(ii) is obvious from the definition of Km and m.

$\neg$**(iii)** Simple violation of the semimeasure property can be inferred indirectly from (i),(iv),$\neg(vi)$ and Theorem 3b. To prove $\neg(iii)$ 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. $\neg(iii)$ now follows from $\#\{t\leq n:\sum_{x_t}m(x_{1:t})\leq m(x_{<t})\}\leq\#\{t\leq n:Km(x_{1:t})\neq Km(x_{<t})\}\leq\sum_{t=1}^n Km(x_{1:t})-Km(x_{<t})=Km(x_{1:n})$, where we exploited (ii) in the last inequality.

(iv) immediate from (ii).

$\neg$**(iv)** (first line) follows from the fact that equality does not even hold within an additive constant, i.e. $Km(yx)\stackrel{+}{\neq}Km(x|y)+Km(y)$. The proof of the latter is similar to the one for K (see [LV97]). $\neg(iv)$ (second line) follows within log from $Km=K+O(\log)$ and $K(yx)=K(x|y)+K(y)+O(\log)$ [LV97].

(v) immediate from definition.

(vi) $\#\{t\leq n:m(x_t|x_{<t})\neq 1\}\leq\sum_{t=1}^n 2|1-m(x_t|x_{<t})|\leq-\sum_{t=1}^n\log m(x_t|x_{<t})=-\log m(x_{1:n})=Km(x_{1:n})<\infty$. In the first inequality we used $m:=m(x_t|x_{<t})\in 2^{-I\!N_0}$, hence $1\leq 2|1-m|$ for $m\neq 1$. In the second inequality we used $1-m\leq-\frac{1}{2}\log m$, valid for $m\in[0,\frac{1}{2}]\cup\{1\}$. In the first equality we used (the log of) the chain rule n times. For computable $x_{1:\infty}$ we have $\sum_{t=1}^\infty|1-m(x_t|x_{<t})|\leq\frac{1}{2}Km(x_{1:\infty})<\infty$, 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). The last line is shown as follows: Fix a sequence $x_{1:\infty}$ and define $\mathcal{Q}:=\{x_{<t}\bar{x}_t:t\in I\!N,\bar{x}_t\neq x_t\}$. $\mathcal{Q}$ is a prefix-free set of finite strings. For any such $\mathcal{Q}$ and any semimeasure μ, one can show that $\sum_{x\in\mathcal{Q}}\mu(x)\leq 1$.[8] Since M is a

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

semimeasure lower bounded by m we get

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

With this, and using monotonicity of m we get

$$\sum_{t=1}^{n} m(\bar{x}_t|x_{<t}) = \sum_{t=1}^{n} \frac{m(x_{<t}\bar{x}_t)}{m(x_{<t})} \leq \sum_{t=1}^{n} \frac{m(x_{<t}\bar{x}_t)}{m(x_{1:n})} \leq \frac{1}{m(x_{1:n})} = 2^{Km(x_{1:n})}$$

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

$\neg$**(vi)** follows from the non-denseness of the range of $m_{(norm)}$: We choose $\mu(1|x_{<t})=\frac{3}{8}$, hence $\mu(0|x_{<t})=\frac{5}{8}$. Since $m(x_t|x_{<t})\in 2^{-I\!N_0}=\{1,\frac{1}{2},\frac{1}{4},\frac{1}{8},...\}$, we have $|m(x_t|x_{<t})-\mu(x_t|x_{<t})|\geq\frac{1}{8}\;\forall t,\forall x_{1:\infty}$. Similarly for

$$m_{norm}(x_t|x_{<t}) \;=\; \frac{m(x_t|x_{<t})}{m(0|x_{<t})+m(1|x_{<t})} \;\in\; \{\tfrac{2^{-n}}{2^{-n}+2^{-m}}:n,m\in I\!N_0\} \;=$$

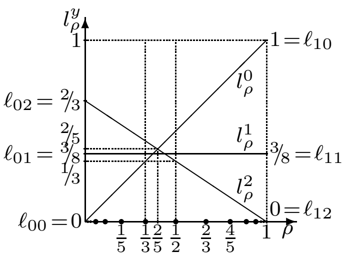
$$=\; \{\tfrac{1}{1+2^z}:z\in Z\!\!Z\} \;=\; \tfrac{1}{1+2^{Z\!\!Z}} \;=\; \{...,\tfrac{1}{9},\tfrac{1}{5},\tfrac{1}{3},\tfrac{1}{2},\tfrac{2}{3},\tfrac{4}{5},\tfrac{8}{9},...\}$$

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

(vii) The first line follows from (vi) and Theorem 3d. 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)** Non-convergence of m does not necessarily imply non-convergence of the losses. For instance, for $\mathcal{Y}=\{0,1\}$, 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 $y\in\mathcal{Y}=\{0,1,2\}$. To prove the first line of $\neg$(vii) 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 $l_\rho^y:=\sum_{x_t=0}^{1}\rho(x_t|x_{<t})\ell_{x_t y}$; $l_\rho^0=\rho$, $l_\rho^1=\frac{3}{8}$, $l_\rho^2=\frac{2}{3}(1-\rho)$ with $\rho:=\rho(1|x_{<t})$ (see Figure). 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. With this extension, the analysis of the $|\mathcal{Y}|=3$ 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 $|\mathcal{Y}|\geq 3$.

We now construct a monotone universal Turing machine U satisfying $\neg(vii)$ (second line). In case where ambiguities in the choice of y in $\arg\min_y\ell_{xy}$ matter we consider the set of solutions $\{\arg\min_y\ell_{xy}\}:=\{\tilde y:\ell_{x\tilde y}=\min_y\ell_{xy}\}\neq\{\}$. We define a one-to-one (onto A) decoding function $d:\{0,1\}^s\to A$ with $A=\{0^{s+1}\}\cup 1\{0,1\}^s\setminus 1\{0^s\}\subset\mathcal{X}^{s+1}$ as $d(0_{1:s})=0_{1:s+1}$ and $d(x_{1:s})=1x_{1:s}$ for $x_{1:s}\neq 0_{1:s}$ with a large $s\in I\!N$ to be determined later. We extend d to $d:(\{0,1\}^s)^*\to A^*$ 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 $c:A\to\{0,1\}^s$ and its extension $c:A^*\to(\{0,1\}^s)^*$ by $c=d^{-1}$. Roughly, U is defined as $U(1p_{1:sn}0_{1:s})=d(p_{1:sn})0_{1:s+1}$. 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 $0_{1:s}$ 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 $x\in A^*$ we have

$$\begin{array}{rcll} Km(x0)= & Km(x0_{1:s+1}) & = l(c(x))+s+1 & \text{and} \\ Km(x1)= & Km(x1z0_{1:s+1}) & = l(c(x))+2s+1, & \end{array} \tag{10}$$

where z is any string of length s. Hence, $m_{norm}(0|x)=[1+2^{-s}]^{-1}\stackrel{s\to\infty}{\longrightarrow}1$ and $m_{norm}(1|x)=[1+2^s]^{-1}\stackrel{s\to\infty}{\longrightarrow}0$. For $t-1\in(s+1)I\!N$ we get $l_m^{y_t}:=\sum_{x_t}m_{norm}(x_t|x_{<t})\ell_{x_ty_t}\stackrel{s\to\infty}{\longrightarrow}\ell_{0y_t}$. This implies

$$y_t^{\Lambda_m}\in\{\arg\min_{y_t}l_m^{y_t}\}\subseteq\{\arg\min_y\ell_{0y}\}\quad\text{for sufficiently large finite } s. \tag{11}$$

We now define $\mu(z)=|A|^{-1}=2^{-s}$ for $z\in A$ and $\mu(z)=0$ for $z\in\mathcal{X}^{s+1}\setminus A$, 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 $I\!N\ni t\leq n\in(s+1)I\!N$. For $x\in A^*$ we have $\mu(0|x)=\mu(0)=\mu(0_{1:s+1})=2^{-s}\stackrel{s\to\infty}{\longrightarrow}0$ and $\mu(1|x)=\mu(1)=\sum_{y\in\mathcal{X}^s}\mu(1y)=\sum_{z\in A\setminus\{0^{s+1}\}}\mu(z)=(2^s-1)\cdot 2^{-s}=1-2^{-s}\stackrel{s\to\infty}{\longrightarrow}1$. For $t-1\in(s+1)I\!N$ we get $l_\mu^{y_t}:=\sum_{x_t}\mu(x_t|x_{<t})\ell_{x_ty_t}\stackrel{s\to\infty}{\longrightarrow}\ell_{1y_t}$. This implies

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

By definition, ℓ is non-degenerate iff $\{\arg\min_y\ell_{0y}\}\cap\{\arg\min_y\ell_{1y}\}=\{\}$. This, together with (11) and (12) implies $y_t^{\Lambda_m}\neq y_t^{\Lambda_\mu}$, 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, $l(c(x)) \leq Km_T(x)+s$ 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 $\min_p\{l(0p):U(0p)=x*\}=3s+1+Km_{U'}(x) \geq 3s+1+l(c(x))-s > l(c(x))+s+1 \geq \min_p\{l(1p):U(1p)=x*\}$, where the first $\geq$ holds with high probability $(1-2^{-s})$. This shows that the expressions (10) for Km are with high probability not affected by the extension of U. Altogether this shows $l_t^{\Lambda_m}/l_t^{\Lambda_\mu} \stackrel{t\to\infty}{\not\longrightarrow} 1$ with high probability. □

8 Outlook and Open Problems

Speed of off-sequence convergence of m for computable environments. The probably most interesting open question is how fast $m(\bar{x}_t|x_{<t})$ converges to zero in the deterministic case.

Non-self-optimizingness for general U and ℓ. Another open problem is whether for every non-degenerate loss-function, self-optimizingness 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 $K_U(x_{<t}\bar{x}_t) < K_U(x_{<t}x_t)$ 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. What is easy to see is that $K(x)$ (and $K(x|l(x))$) are completely unsuitable for prediction, since $K(x0) \stackrel{+}{=} K(x1)$ (and $K(x0|l(x0)) \stackrel{+}{=} K(x1|l(x1))$), which implies that the predictive functions do not even converge for deterministic computable environments. Note that the larger a semimeasures, 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), $2^{-K(x|l(x))} \stackrel{\times}{\geq} M(x)$ is a bad predictor disaccording with (8). Besides the discussed prefix Kolmogorov complexity K, monotone complexity Km, and Solomonoff's universal prior $M=2^{-KM}$, one may investigate the predictive properties of the historically first plain Kolmogorov complexity C, Schnorr's process complexity, Chaitin's complexity Kc, Cover's extension semimeasure Mc, Loveland's uniform complexity, Schmidhuber's cumulative K^E and general K^G complexity and corresponding measures, Vovk's predictive complexity KP, Schmidhuber's speed prior S, Levin complexity Kt, and several others [LV97, VW98, Sch00]. 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 $M(x):=\sum_{p:U(p)=x*}2^{-l(p)}$ 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 semi-measures: $M(x)=\sum_{\nu\in\mathcal{M}_{enum}^{semi}}2^{-K(\nu)}\nu(x)$ 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 ($m_2\stackrel{\times}{=}M$), but is also not a semi-measure. Drawing the analogy to m further, we conjecture slow posterior convergence $m_2\to\mu$ w.p.1 for computable probabilistic environments μ. In [BC91], MDL has been shown to converge for computable 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-optimizingness; the particular structure of m did not matter.

Extra conditions. Non-convergence or non-self-optimizingness 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 is possibly slow.

References

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

[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 of universal prediction for general alphabet. *Proceedings of the 12th Eurpean Conference on Machine Learning (ECML-2001)*, pages 239–250, 2001.

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

[Hut02] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. Technical Report IDSIA-09-01, IDSIA, Manno(Lugano), CH, 2002. http://arxiv.org/abs/cs.LG/0301014.

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

[Lev73] L. A. Levin. On the notion of a random sequence. *Soviet Math. Dokl.*, 14(5):1413–1416, 1973.

[LV97] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.

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

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

[Sol78] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory*, IT-24:422–432, 1978.

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

[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.