Available online at www.sciencedirect.com
ScienceDirect
Information and Computation 205 (2007) 242–261
Information and Computation
www.elsevier.com/locate/ic

Algorithmic complexity bounds on future prediction errors☆

Alexey Chernov [a,d], Marcus Hutter [a,c,*], Jürgen Schmidhuber [a,b]

[a] *IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland*
[b] *TU Munich, Boltzmannstr. 3, 85748 Garching, München, Germany*
[c] *RSISE/ANU/NICTA, Canberra, ACT 0200, Australia*
[d] *LIF, CMI, 39 rue Joliot Curie, 13453 Marseille cedex 13, France*

Received 18 July 2005; revised 12 September 2006
Available online 19 December 2006

Abstract

We bound the future loss when predicting any (computably) stochastic sequence online. Solomonoff finitely bounded the total deviation of his universal predictor M from the true distribution μ by the algorithmic complexity of μ. Here we assume that we are at a time $t>1$ and have already observed $x=x_1\cdots x_t$. We bound the future prediction performance on $x_{t+1}x_{t+2}\cdots$ by a new variant of algorithmic complexity of μ given x, plus the complexity of the randomness deficiency of x. The new complexity is monotone in its condition in the sense that this complexity can only decrease if the condition is prolonged. We also briefly discuss potential generalizations to Bayesian model classes and to classification problems.

Keywords: Kolmogorov complexity; Posterior bounds; Online sequential prediction; Solomonoff prior; Monotone conditional complexity; Total error; Future loss; Randomness deficiency

☆ This work was supported by SNF Grants 200020-107590/1, 2100-67712, and 200020-107616. A shorter version appeared in the proceedings of the ALT'05 conference [1].
* Corresponding author.
E-mail addresses: alexey@idsia.ch (A. Chernov), marcus@idsia.ch (M. Hutter), juergen@idsia.ch (J. Schmidhuber).
URLs: http://www.idsia.ch/~alexey (A. Chernov), http://www.idsia.ch/~marcus (M. Hutter), http://www.idsia.ch/~juergen (J. Schmidhuber).

0890-5401/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ic.2006.10.004

1. Introduction

We consider the problem of online=sequential predictions. We assume that the sequences $x=x_1x_2x_3\cdots$ are drawn from some "true" but unknown probability distribution μ. Bayesians proceed by considering a class $\mathcal{M}$ of models=hypotheses=distributions, sufficiently large such that $\mu\in\mathcal{M}$, and a prior over $\mathcal{M}$. Solomonoff considered the truly large class that contains all computable probability distributions [2]. He showed that his universal distribution M converges rapidly to μ [3], i.e., predicts well in any environment as long as it is computable or can be modeled by a computable probability distribution (all physical theories are of this sort). $M(x)$ is roughly $2^{-K(x)}$, where $K(x)$ is the length of the shortest description of x, called the Kolmogorov complexity of x. Since K and M are incomputable, they have to be approximated in practice. See e.g., [4–7] and references therein. The universality of M also precludes useful statements about the prediction quality at particular time instances n [5, p. 62], as opposed to simple classes like i.i.d. sequences (data) of size n, where accuracy is typically $O(n^{-1/2})$. Luckily, bounds on the expected *total*=cumulative loss (e.g., number of prediction errors) for M can be derived [3,8–10], which is often sufficient in an online setting. The bounds are in terms of the (Kolmogorov) complexity of μ. For instance, for deterministic μ, the number of errors is (in a sense tightly) bounded by $K(\mu)$ which measures in this case the information (in bits) in the observed infinite sequence x.

What's new. In this paper we assume we are at a time $t>1$ and have already observed $x=x_1\cdots x_t$. Hence we are interested in the future prediction performance on $x_{t+1}x_{t+2}\cdots$, since typically we do not care about past errors. If the total loss is finite, the future loss must necessarily be small for large t. In a sense the paper intends to quantify this apparent triviality. If the complexity of μ bounds the total loss, a natural guess is that something like the conditional complexity of μ given x bounds the future loss. (If x contains a lot of (or even all) information about μ, we should make fewer (no) errors anymore.) Indeed, we prove two bounds of this kind but with additional terms describing structural properties of x. These additional terms appear since the total loss is bounded only in expectation, and hence the future loss is small only for "most" $x_1\cdots x_t$. In the first bound (Theorem 1), the additional term is the complexity of the length of x (a kind of worst-case estimation). The second bound (Theorem 7) is finer: the additional term is the complexity of the randomness deficiency of x. The advantage is that the deficiency is small for "typical" x and bounded on average (in contrast to the length). But in this case the conventional conditional complexity turned out to be unsuitable. So we introduce a new natural modification of conditional Kolmogorov complexity, which is monotone as a function of condition. Informally speaking, we require programs (=descriptions) to be consistent in the sense that if a program generates some μ given x, then it must generate the same μ given any prolongation of x. The new posterior bounds also significantly improve upon the previous total bounds.

Contents. The paper is organized as follows. Some basic notation and definitions are given in Sections 2 and 3. In Section 4 we prove and discuss the length-based bound Theorem 1. In Section 5 we show why a new definition of complexity is necessary and formulate the deficiency-based bound Theorem 7. We discuss the definition and basic properties of the new complexity in Section 6, and prove Theorem 7 in Section 7. We briefly discuss potential generalizations to general model classes $\mathcal{M}$ and classification in the concluding Section 8.

2. Notation and definitions

We essentially follow the notation of [5,6].

2.1. Strings and natural numbers

We write $\mathcal{X}^*$ for the set of finite strings over a finite alphabet $\mathcal{X}$, and $\mathcal{X}^\infty$ for the set of infinite sequences. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. We use letters i,k,l,n,t for natural numbers, u,v,x,y,z for finite strings, ϵ for the empty string, and $\alpha=\alpha_{1:\infty}$ etc. for infinite sequences. For a string x of length $\ell(x)=n$ we write $x_1x_2\cdots x_n$ with $x_t\in\mathcal{X}$ and further abbreviate $x_{k:n}:=x_kx_{k+1}\cdots x_{n-1}x_n$ and $x_{<n}:=x_1\cdots x_{n-1}$. For $x_t\in\mathcal{X}$, denote by $\bar{x}_t$ an (arbitrary) element from $\mathcal{X}$ such that $\bar{x}_t\neq x_t$. For binary alphabet $\mathcal{X}=\{0,1\}$, the $\bar{x}_t$ is uniquely defined. We occasionally identify strings with natural numbers.

2.2. Prefix sets

A string x is called a (proper) prefix of y if there is a $z(\neq\epsilon)$ such that $xz=y$; y is called a prolongation of x. We write $x*=y$ in this case, where $*$ is a wildcard for a string, and similarly for the case where y is an infinite sequence. A set of strings is called prefix free if no element is a proper prefix of another. Any prefix-free set $\mathcal{P}$ has the important property of satisfying Kraft's inequality $\sum_{x\in\mathcal{P}}|\mathcal{X}|^{-\ell(x)}\leqslant 1$.

2.3. Asymptotic notation

We write $f(x)\stackrel{\times}{\leqslant}g(x)$ for $f(x)=O(g(x))$ and $f(x)\stackrel{+}{\leqslant}g(x)$ for $f(x)=g(x)+O(1)$. Equalities $\stackrel{\times}{=}$, $\stackrel{+}{=}$ are defined similarly: they hold if the corresponding inequalities hold in both directions.

2.4. (Semi)measures

We call $\rho:\mathcal{X}^*\to[0,1]$ a semimeasure *iff* $\sum_{x_n\in\mathcal{X}}\rho(x_{1:n})\leqslant\rho(x_{<n})$ and $\rho(\epsilon)\leqslant 1$, and a measure *iff* both non-strict inequalities are equalities. $\rho(x)$ is interpreted as the ρ-probability of sampling a sequence which starts with x. The conditional probability (posterior)

$$\rho(y|x):=\frac{\rho(xy)}{\rho(x)} \tag{1}$$

is the ρ-probability that a string x is followed by (continued with) y. If $\rho(x)=0$, $\rho(y|x)$ is defined arbitrarily and every such function is called a version of conditional probability. We call ρ deterministic if $\exists\alpha:\rho(\alpha_{1:n})=1\ \forall n$. In this case we identify ρ with α.

2.5. Random events and expectations

We assume that sequence $\omega=\omega_{1:\infty}$ is sampled from the "true" measure μ, i.e., $\mathbf{P}[\omega_{1:n}=x_{1:n}]=\mu(x_{1:n})$. We denote expectations w.r.t. μ by $\mathbf{E}$, i.e., for a function $f:\mathcal{X}^n\to\mathbb{R}$, $\mathbf{E}[f]=\mathbf{E}[f(\omega_{1:n})]=\sum_{x_{1:n}}\mu(x_{1:n})f(x_{1:n})$. We abbreviate $\mu_t:=\mu(\cdot|\omega_{<t})$.

2.6. Enumerable sets and functions

A set of strings (or naturals, or other constructive objects) is called *enumerable* if it is the range of some computable function. A function $f:\mathcal{X}^* \to \mathbb{R}$ is called *(co-)enumerable* if the set of pairs $\{\langle x,\frac{k}{n}\rangle \mid f(x) \overset{(<)}{>} \frac{k}{n}\}$ is enumerable. A measure μ is called *computable* if it is enumerable and co-enumerable and the set $\{x \mid \mu(x)=0\}$ is decidable (i. e. enumerable and co-enumerable).

To simplify the statements of the theorems below, we assume that for every computable measure μ, there is one fixed computable version of conditional probability $\mu(y|x)$, for example, $\mu(y|x)$ is the uniform measure on y's for $\mu(x)=0$.

2.7. Prefix Kolmogorov complexity

The conditional prefix complexity $K(y|x) := \min\{\ell(p)|U(p,x)=y\}$ is the length of the shortest binary (self-delimiting) program $p \in \{0,1\}^*$ on a universal prefix Turing machine U with output $y \in \mathcal{X}^*$ and input $x \in \mathcal{X}^*$ [6]. $K(x) := K(x|\epsilon)$. For non-string objects o we define $K(o) := K(\langle o\rangle)$, where $\langle o\rangle \in \mathcal{X}^*$ is some standard code for o. In particular, if $(f_i)_{i=1}^{\infty}$ is an enumeration of all (co-)enumerable functions, we define $K(f_i) := K(i)$. We need the following properties: The co-enumerability of K, the upper bounds $K(x|\ell(x)) \overset{+}{\leqslant} \ell(x)\log_2|\mathcal{X}|$ and $K(n) \overset{+}{\leqslant} 2\log_2 n$, Kraft's inequality $\sum_x 2^{-K(x)} \leqslant 1$, the lower bound $K(x) \geqslant l(x)$ for "most" x and $K(n) \overset{n\to\infty}{\longrightarrow} \infty$, extra information bounds $K(x|y) \overset{+}{\leqslant} K(x) \overset{+}{\leqslant} K(x,y)$, subadditivity $K(xy) \overset{+}{\leqslant} K(x,y) \overset{+}{\leqslant} K(y)+K(x|y)$, information non-increase $K(f(x)) \overset{+}{\leqslant} K(x)+K(f)$ for computable $f:\mathcal{X}^* \to \mathcal{X}^*$, and coding relative to a probability distribution (MDL): if $P:\mathcal{X}^* \to [0,1]$ is enumerable and $\sum_x P(x) \leqslant 1$, then $K(x) \overset{+}{\leqslant} -\log_2 P(x)+K(P)$.

2.8. Monotone and Solomonoff complexity

The monotone complexity $Km(x) := \min\{\ell(p)|U(p)=x*\}$ is the length of the shortest binary (possibly non-halting) program $p \in \{0,1\}^*$ on a universal monotone Turing machine U which outputs a string starting with x. Solomonoff's prior $M(x) := \sum_{p:U(p)=x*} 2^{-\ell(p)} =: 2^{-KM(x)}$ is the probability that U outputs a string starting with x if provided with fair coin flips on the input tape. Most complexities coincide within an additive term $O(\log\ell(x))$, e.g., $K(x|\ell(x)) \overset{+}{\leqslant} KM(x) \leqslant Km(x) \leqslant K(x)$, hence similar relations as for K hold.

3. Setup

3.1. Convergent predictors

We assume that μ is a "true"[1] sequence generating measure, also called an 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

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

estimated probability[2] of x_t, given $x_{<t}$. Closeness of $\rho(x_t|x_{<t})$ to $\mu(x_t|x_{<t})$ is desirable as a goal in itself or when performing a Bayes decision y_t that has the minimal ρ-expected loss $l_t^\rho(x_{<t}) := \min_{y_t}\sum_{x_t}\mathrm{Loss}(x_t,y_t)\rho(x_t|x_{<t})$. 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

$$\rho(x_t'|x_{<t})-\mu(x_t'|x_{<t}) \stackrel{\text{(fast)}}{\longrightarrow} 0 \quad \text{as } t\to\infty$$

seems reasonable.

3.2. Convergence in mean sum

We can quantify the deviation of ρ_t from μ_t, e.g., by the squared difference

$$s_t(\omega_{<t}) := \sum_{x_t\in\mathcal{X}}(\rho(x_t|\omega_{<t})-\mu(x_t|\omega_{<t}))^2 \equiv \sum_{x_t}(\rho_t-\mu_t)^2$$

Alternatively one may also use the squared absolute distance $s_t := \frac{1}{2}\big(\sum_{x_t}|\rho_t-\mu_t|\big)^2$, the Hellinger distance $s_t := \sum_{x_t}\big(\sqrt{\rho_t}-\sqrt{\mu_t}\big)^2$, the KL-divergence $s_t := \sum_{x_t}\mu_t\ln\frac{\mu_t}{\rho_t}$, or the squared Bayes regret $s_t := \frac{1}{2}\big(l_t^\rho-l_t^\mu\big)^2$ for $l_t\in[0,1]$. For all these distances one can show [5,9,11] that their cumulative expectation from l to n is bounded as follows:

$$0 \leqslant \mathbf{E}\left[\sum_{t=l}^{n} s_t \middle| \omega_{<l}\right] \leqslant \mathbf{E}\left[\ln\frac{\mu(\omega_{l:n}|\omega_{<l})}{\rho(\omega_{l:n}|\omega_{<l})}\middle|\omega_{<l}\right] =: D_{l:n}(\omega_{<l}). \tag{2}$$

$D_{l:n}$ is increasing in n, hence $D_{l:\infty}\in[0,\infty]$ exists [5,12]. A sequence of random variables like s_t is said to converge to zero with probability 1 if the set $\{\omega|s_t(\omega)\stackrel{t\to\infty}{\longrightarrow}0\}$ has measure 1. s_t is said to converge to zero in mean sum if $\sum_{t=1}^{\infty}\mathbf{E}\big[|s_t|\big]\leqslant c<\infty$, which implies convergence with probability 1 (rapid if c is of reasonable size). Therefore a small finite bound on $D_{1:\infty}$ would imply rapid convergence of the s_t defined above to zero, hence $\rho_t\to\mu_t$ and $l_t^\rho\to l_t^\mu$ fast. So the crucial quantities to consider and bound (in expectation) are $\ln\frac{\mu(x)}{\rho(x)}$ if $l=1$ and $\ln\frac{\mu(y|x)}{\rho(y|x)}$ for $l>1$. For illustration we will sometimes loosely interpret $D_{1:\infty}$ and other quantities as the number of prediction errors, as for the error-loss they are closely related to it [8,12].

3.3. Bayes mixtures

A Bayesian considers a class of distributions $\mathcal{M} := \{\nu_1,\nu_2,...\}$, large enough to contain μ, and uses the Bayes mixture

$$\xi(x) := \sum_{\nu\in\mathcal{M}} w_\nu\cdot\nu(x), \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu>0 \tag{3}$$

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

for prediction, where w_ν can be interpreted as the prior of (or initial belief in) ν. The dominance

$$\xi(x) \geqslant w_\mu \cdot \mu(x) \quad \forall x \in \mathcal{X}^* \tag{4}$$

is its most important property. Using $\rho = \xi$ for prediction, this implies $D_{1:\infty} \leqslant \ln w_\mu^{-1} < \infty$, hence $\xi_t \to \mu_t$. If $\mathcal{M}$ is chosen sufficiently large, then $\mu \in \mathcal{M}$ is not a serious constraint.

3.4. Solomonoff prior

So we consider the largest (from a computational point of view) relevant class, the class $\mathcal{M}_U$ of all enumerable semimeasures (which includes all computable probability distributions) and choose $w_\nu = 2^{-K(\nu)}$ which is biased towards simple environments (Occam's razor). This gives us Solomonoff–Levin's prior M [2,13] (this definition coincides within an irrelevant multiplicative constant with the one in Section 2). In the following we assume $\mathcal{M} = \mathcal{M}_U$, $\rho = \xi = M$, $w_\nu = 2^{-K(\nu)}$, and $\mu \in \mathcal{M}_U$ being a computable (proper) measure, hence $M(x) \geqslant 2^{-K(\mu)}\mu(x) \forall x$ by (4).

3.5. Prediction of deterministic environments

Consider a computable sequence $\alpha = \alpha_{1:\infty}$ "sampled from $\mu \in \mathcal{M}$" with $\mu(\alpha) = 1$, i.e., μ is deterministic, then from (4) we get

$$\sum_{t=1}^{\infty} |1 - M(\alpha_t|\alpha_{<t})| \leqslant -\sum_{t=1}^{\infty} \ln M(\alpha_t|\alpha_{<t}) = -\ln M(\alpha_{1:\infty}) \leqslant K(\mu)\ln 2 < \infty, \tag{5}$$

which implies that $M(\alpha_t|\alpha_{<t})$ converges rapidly to 1 and hence $M(\bar{\alpha}_t|\alpha_{<t}) \to 0$, i.e., asymptotically M correctly predicts the next symbol. The number of prediction errors is of the order of the complexity $K(\mu) \stackrel{+}{=} Km(\alpha)$ of the sequence.

For binary alphabet this is the best we can expect, since at each time-step only a single bit can be learned about the environment, and only after we "know" the environment we can predict correctly. For non-binary alphabet, $K(\mu)$ still measures the information in μ in bits, but feedback per step can now be $\log_2|\mathcal{X}|$ bits, so we may expect a better bound $K(\mu)/\log_2|\mathcal{X}|$. But in the worst case all $\alpha_t \in \{0,1\} \subseteq \mathcal{X}$. So without structural assumptions on μ the bound cannot be improved even if $\mathcal{X}$ is huge. We will see how our posterior bounds can help in this situation.

3.6. Individual randomness (deficiency)

Let us now consider a general (not necessarily deterministic) computable measure $\mu \in \mathcal{M}$. The Shannon–Fano code of x w.r.t. μ has code-length $\lceil -\log_2 \mu(x) \rceil$, which is "optimal" for "typical/random" x sampled from μ. Further, $-\log_2 M(x) \approx K(x)$ is the length of an "optimal" code for x. Hence $-\log_2 \mu(x) \approx -\log_2 M(x)$ for "μ-typical/random" x. This motivates the definition of *μ-randomness deficiency*

$$d_\mu(x) := \log_2 \frac{M(x)}{\mu(x)}$$

which is small for "typical/random" x. Formally, a sequence α is called (Martin–Löf) random iff $d_\mu(\alpha) := \sup_n d_\mu(\alpha_{1:n}) < \infty$, i.e., iff its Shannon–Fano code is "optimal" (note that $d_\mu(\alpha) \geqslant -K(\mu) > -\infty$ for all sequences), i.e., iff

$$\sup_n \Big| \sum_{t=1}^{n} \log \frac{\mu(\alpha_t|\alpha_{<t})}{M(\alpha_t|\alpha_{<t})} \Big| \equiv \sup_n \Big| \log \frac{\mu(\alpha_{1:n})}{M(\alpha_{1:n})} \Big| < \infty.$$

Unfortunately this does not imply $M_t \to \mu_t$ on the μ-random α, since M_t may oscillate around μ_t, which indeed can happen [14]. But if we take the expectation, Solomonoff [3,5,12] showed

$$0 \leqslant \sum_{t=1}^{\infty} \mathbf{E} \sum_{x_t} (M_t - \mu_t)^2 \leqslant D_{1:\infty} = \lim_{n\to\infty} \mathbf{E}[-d_\mu(\omega_{1:n})] \ln 2 \leqslant K(\mu) \ln 2 < \infty, \tag{6}$$

hence, $M_t \to \mu_t$ with μ-probability 1. So in any case, $d_\mu(x)$ is an important quantity, since the smaller $-d_\mu(x)$ (at least in expectation) is, the better M predicts.

4. Posterior bounds

4.1. Posterior bounds

Both bounds (5) and (6) bound the total (cumulative) discrepancy (error) between M_t and μ_t. Since the discrepancy sum $D_{1:\infty}$ is finite, we know that after sufficiently long time $t = l$, we will make few further errors, i.e., the future error sum $D_{l:\infty}$ is small. The main goal of this paper is to quantify this asymptotic statement. So we need bounds on $\log_2 \frac{\mu(y|x)}{M(y|x)}$, where x are the past and y the future observations. Since $\log_2 \frac{\mu(y)}{M(y)} \leqslant K(\mu)$ and $\mu(y|x)/M(y|x)$ are conditional versions of true/universal distributions, it seems natural that the unconditional bound $K(\mu)$ also simply conditionalizes to $\log_2 \frac{\mu(y|x)}{M(y|x)} \leqslant K(\mu|x)$. The more information the past observation x contains about μ, the easier it is to code μ, i.e., the smaller $K(\mu|x)$ is, and hence the less future predictions errors $D_{l:\infty}$ we should make. Once x contains all information about μ, i.e., $K(\mu|x) \stackrel{+}{=} 0$, we should make no errors anymore. More formally, optimally coding x, then $\mu|x$, and finally $y|\mu,x$ by Shannon–Fano gives a code for xy, hence $K(xy) \lesssim K(x) + K(\mu|x) - \log_2 \mu(y|x)$. Since $K(z) \approx -\log_2 M(z)$, this implies $\log_2 \frac{\mu(y|x)}{M(y|x)} \lesssim K(\mu|x)$, but with a logarithmic fudge that tends to infinity as $\ell(y) \to \infty$, which is unacceptable. The y-independent bound we need was first stated in [5, Prob. 2.6 (iii)]:

Theorem 1. *For any computable measure μ and any $x, y \in \mathcal{X}^*$ it holds*

$$\log_2 \frac{\mu(y|x)}{M(y|x)} \stackrel{+}{\leqslant} K(\mu|x) + K(\ell(x)).$$

Proof. For each l we define the following function of $z \in \mathcal{X}^*$. For $\ell(z) \geqslant l$,

$$\psi^l(z) := \sum_{\nu\in\mathcal{M}} 2^{-K(\nu|z_{1:l})} M(z_{1:l}) \nu(z_{l+1:\ell(z)}).$$

For $\ell(z) < l$, we extend ψ^l by defining $\psi^l(z) := \sum_{u:\ell(u)=l-\ell(z)} \psi^l(zu)$. It is easy to see that ψ^l is an enumerable semimeasure. By the definition of M, we have

$$M(z) \geqslant 2^{-K(\psi^l)} \psi^l(z)$$

for any l and z. Now let $l = \ell(x)$ and $z = xy$. Let us define an enumerable measure $\mu^x(y) := \mu(y|x)$. Then

$$M(xy) \geqslant 2^{-K(\psi^l)} \psi^l(xy) \geqslant 2^{-K(\psi^l)} 2^{-K(\mu^x|x)} M(x) \mu^x(y).$$

Taking the logarithm, after trivial transformations, we get

$$\log_2 \frac{\mu(y|x)}{M(y|x)} \leqslant K(\mu^x|x) + K(\psi^l).$$

To complete the proof, let us note that $K(\psi^l) \stackrel{+}{\leqslant} K(l)$ and $K(\mu^x|x) \stackrel{+}{\leqslant} K(\mu|x)$. □

Corollary 2. *The future deviation of M_t from μ_t is bounded by*

$$\sum_{t=l+1}^{\infty} \mathbf{E}[s_t|\omega_{1:l}] \leqslant D_{l+1:\infty}(\omega_{1:l}) \stackrel{+}{\leqslant} (K(\mu|\omega_{1:l}) + K(l)) \ln 2 \qquad (i)$$

For s_t being squared (absolute) distance, Hellinger distance, or squared Bayes regret, the total deviation of M_t from μ_t is bounded by

$$\sum_{t=1}^{\infty} \mathbf{E}[s_t] \stackrel{+}{\leqslant} \min_l \{\mathbf{E}[K(\mu|\omega_{1:l}) + K(l)] \ln 2 + 2l\} \qquad (ii)$$

Proof. (*i*) The first inequality is (2) and the second follows by taking the conditional expectation $\mathbf{E}[\cdot|\omega_{1:l}]$ in Theorem 1. (*ii*) follows from (*i*) by taking the unconditional expectation and from $\sum_{t=1}^{l} \mathbf{E}[s_t] \leqslant 2l$, since $s_t \leqslant 2$ for these distances [5]. □

4.2. Examples and more motivation

The bounds Theorem 1 and Corollary 2(*i*) prove and quantify the intuition that the more we know about the environment, the better our predictions. We show the usefulness of the new bounds for some deterministic environments $\mu \widehat{=} \alpha$.

Assume all observations are identical, i.e., $\alpha = x_1 x_1 x_1 \cdots$. Further assume that $\mathcal{X}$ is huge and $K(x_1) = \log_2 |\mathcal{X}|$, i.e., x_1 is a typical/random/complex element of $\mathcal{X}$. For instance if x_1 is a 256^3 color 512×512 pixel image, then $|\mathcal{X}| = 256^{3 \times 512 \times 512}$. Hencethe standard bound (6) on the number of errors $D_{1:\infty}/\ln 2 \leqslant K(\mu) \stackrel{+}{=} K(x_1) = 3 \cdot 2^{21}$ is huge. Of course, interesting pictures are not purely random, but their complexity is often only a factor 10..100 less, so still large. On the other hand, any reasonable prediction scheme observing a few (rather than several thousands) identical images, should predict that the next image will be the same. This is what our posterior bound gives, $D_{2:\infty}(x_1) \stackrel{+}{\leqslant} (K(\mu|x_1) + K(1)) \ln 2 \stackrel{+}{=} 0$, hence indeed M makes only $\sum_{t=1}^{\infty} \mathbf{E}[s_t] = \mathrm{O}(1)$ errors by Corollary 2(*ii*), significantly improving upon Solomonoff's bound $K(x_1) \ln 2$.

More generally, assume $\alpha = x\omega$, where the initial part $x = x_{1:l}$ contains all information about the remainder, i.e., $K(\mu|x) \stackrel{+}{=} K(\omega|x) \stackrel{+}{=} 0$. For instance, x may be a binary program for π or e, and ω its $|\mathcal{X}|$-ary expansion. Sure, given the algorithm for some number sequence, it should be perfectly predictable. Indeed, Theorem 1 implies $D_{l+1:\infty} \stackrel{+}{\leqslant} K(l)$, which can be exponentially smaller than Solomonoff's bound $K(\mu)$ ($\stackrel{+}{=} l$ if $K(x) \stackrel{+}{=} \ell(x)$). On the other hand, $K(l) \geqslant \log_2 l$ for most l, i.e., is larger than the O(1) that one might hope for.

4.3. Logarithmic versus constant accuracy

There is one blemish in the bound. There is an additive correction of logarithmic size in the length of x. Many theorems in algorithmic information theory hold to within an additive constant, sometimes this is easily reached, sometimes with difficulty, sometimes one needs a suitable complexity variant, and sometimes the logarithmic accuracy cannot be improved [6]. The latter is the case with Theorem 1.

Lemma 3. *For $\mathcal{X} = \{0,1\}$, for any positive computable measure μ, there exists a computable sequence $\alpha \in \{0,1\}^\infty$ such that for any $l \in \mathbb{N}$*

$$D_{l:\infty}(\alpha_{<l}) \geqslant D_{l:l}(\alpha_{<l}) \equiv \sum_{b\in\{0,1\}} \mu(b|\alpha_{<l}) \ln \frac{\mu(b|\alpha_{<l})}{M(b|\alpha_{<l})} \stackrel{+}{\geqslant} \tfrac{1}{3} K(l).$$

Proof. Let us construct such computable sequence $\alpha \in \{0,1\}^\infty$ by induction. Assume that $\alpha_{<l}$ is constructed. Since μ is a measure, either $\mu(0|\alpha_{<l}) > c$ or $\mu(1|\alpha_{<l}) > c$ for $c := [3\ln 2]^{-1} < \frac{1}{2}$. Since μ is computable, we can find (effectively) $b \in \{0,1\}$ such that $\mu(b|\alpha_{<l}) > c$. Put $\alpha_l = \bar{b}$.

Let us estimate $M(\bar{\alpha}_l|\alpha_{<l})$. Since α is computable, $M(\alpha_{<l}) \stackrel{\times}{\geqslant} 1$. We claim that $M(\alpha_{<l}\bar{\alpha}_l) \stackrel{\times}{\leqslant} 2^{-K(l)}$. Actually, consider the set $\{\alpha_{<l}\bar{\alpha}_l \mid l > 0\}$. This set is prefix free and decidable. Therefore $P(l) = M(\alpha_{<l}\bar{\alpha}_l)$ is an enumerable function with $\sum_l P(l) \leqslant 1$, and the claim follows from the coding theorem. Thus, we have $M(\bar{\alpha}_l|\alpha_{<l}) \stackrel{\times}{\leqslant} 2^{-K(l)}$ for any l. Since $\mu(\bar{\alpha}_l|\alpha_{<l}) > c$, we get

$$\sum_{b\in\{0,1\}} \mu(b|\alpha_{<l}) \ln \frac{\mu(b|\alpha_{<l})}{M(b|\alpha_{<l})} \stackrel{+}{\geqslant} \mu(\bar{\alpha}_l|\alpha_{<l}) \ln \frac{c}{2^{-K(l)}} + \min_{p\in[0,1-c]} p \ln \frac{p}{M(\alpha_l|\alpha_{<l})} \stackrel{+}{\geqslant} cK(l)\ln 2 \qquad \square$$

A constant fudge is generally preferable to a logarithmic one for quantitative and aesthetical reasons. It also often leads to particular insight and/or interesting new complexity variants (which will be the case here). Though most complexity variants coincide within logarithmic accuracy (see [15,16] for exceptions), they can have very different other properties. For instance, Solomonoff complexity $KM(x) = -\log_2 M(x)$ is an excellent predictor, but monotone complexity Km can be exponentially worse and prefix complexity K fails completely [17,18].

4.4. Exponential bounds

Bayes is often approximated by MAP or MDL. In our context this means approximating KM by Km with exponentially worse bounds (in deterministic environments) [17]. (Intuitively, since an error with Bayes eliminates half of the environments, while MAP/MDL may eliminate only one.)

Also for more complex "reinforcement" learning problems, bounds can be $2^{K(\mu)}$ rather than $K(\mu)$ due to sparser feedback. For instance, for a sequence $x_1x_1x_1\cdots$ if we do not observe x_1 but only receive a reward if our prediction was correct, then the only way a universal predictor can find x_1 is by trying out all $|\mathcal{X}|$ possibilities and making (in the worst case) $|\mathcal{X}|-1\stackrel{\times}{=}2^{K(\mu)}$ errors. Posterization allows us to boost such gross bounds to useful bounds $2^{K(\mu|x_1)}=\mathrm{O}(1)$. But in general, additive logarithmic corrections as in Theorem 1 also exponentiate and lead to bounds polynomial in l which may be quite sizeable. Here the advantage of a constant correction becomes even more apparent [5, Problems 2.6, 3.13, 6.3 and Section 5.3.3].

5. More bounds and new complexity measure

Lemma 3 shows that the bound in Theorem 1 is attained for some binary strings. But for other binary strings the bound may be very rough. (Similarly, $K(x)$ is greater than $\ell(x)$ infinitely often, but $K(x)\ll\ell(x)$ for many "interesting" x.) Let us try to find a new bound, which does not depend on $\ell(x)$.

First observe that, in contrast to the unconditional case (6), $K(\mu)$ is not an upper bound (again by Lemma 3). Informally speaking, the reason is that M can predict the future very badly if the past is not "typical" for the environment (such past x have low μ-probability, therefore in the unconditional case their contribution to the expected loss is small). So, it is natural to bound the loss in terms of randomness deficiency $d_\mu(x)$, which is a quantitative measure of "typicalness".

Theorem 4. *For any computable measure μ and any $x,y\in\{0,1\}^*$ it holds*

$$\log_2\frac{\mu(y|x)}{M(y|x)}\equiv d_\mu(x)-d_\mu(xy)\stackrel{+}{\leqslant}K(\mu)+K(\lceil d_\mu(x)\rceil).$$

Theorem 4 is a variant of the "deficiency conservation theorem" from [19]. We do not know who was the first to discover this statement and whether it was published (the special case where μ is the uniform measure was proved by An. Muchnik as an auxiliary lemma for one of his unpublished results; then A. Shen placed a generalized statement to the (unfinished) book [19]).

Now, our goal is to replace $K(\mu)$ in the last bound by a conditional complexity of μ. Unfortunately, the conventional conditional prefix complexity is not suitable:

Lemma 5. *Let $\mathcal{X}=\{0,1\}$. There is a constant C_0 such that for any $l\in\mathbb{N}$, there are a computable measure μ and $x\in\{0,1\}^l$ such that*

$$K(\mu|x)\leqslant C_0,\quad |d_\mu(x)|\leqslant C_0,\quad \textit{and}$$

$$D_{l+1:l+1}(x)\equiv\sum_{b\in\{0,1\}}\mu(b|x)\ln\frac{\mu(b|x)}{M(b|x)}\stackrel{+}{\geqslant}K(l)\ln 2.$$

Proof. For $l\in\mathbb{N}$, define a deterministic measure μ^l such that μ^l is equal to 1 on the prefixes of 0^l1^∞ and is equal to 0 otherwise.

Let $x=0^l$. Then $\mu^l(x)=1$, $\mu^l(x0)=0$, and $\mu^l(x1)=1$. Also $1\geqslant M(x)\geqslant M(x0)\geqslant M(0^\infty)\stackrel{\times}{=}1$ and (as in the proof of Lemma 3) $M(x1)\stackrel{\times}{\leqslant}2^{-K(l)}$. Trivially, $d_{\mu^l}(x)=\log_2 M(x)\stackrel{\times}{=}1$, and $K(\mu^l|x)\stackrel{+}{=}K(\mu^l|l)\stackrel{+}{=}0$. Thus, $K(\mu^l|x)$ and $d_{\mu^l}(x)$ are bounded by a constant C_0 independent of l. On the other hand

$$\sum_{b\in\{0,1\}}\mu^l(b|x)\ln\frac{\mu^l(b|x)}{M(b|x)}=\ln\frac{1}{M(1|x)}\stackrel{+}{\geqslant}K(l)\ln 2.$$

(One can obtain the same result also for non-deterministic μ, for example, taking μ^l mixed with the uniform measure.) □

Informally speaking, in Lemma 5 we exploit the fact that $K(y|x)$ can use the information about the length of the condition x. Hence $K(y|x)$ can be small for a certain x and is large for some (actually almost all) prolongations of x. But in our case of sequence prediction, the length of x grows taking all intermediate values and cannot contain any relevant information. Thus we need a new kind of conditional complexity.

Consider a Turing machine T with two input tapes. Inputs are provided without delimiters, so the size of the input is determined by the machine itself. Let us call such a machine *twice prefix*. We write that $T(x,y)=z$ if machine T, given a sequence beginning with x on the first tape and a sequence beginning with y on the second tape, halts after reading exactly x and y and prints z to the output tape. (Obviously, if $T(x,y)=z$, then the computation does not depend on the contents of the input tapes after x and y.) We define

$$C_T(y|x):=\min\{\ell(p)\,|\,\exists k\leqslant\ell(x):T(p,x_{1:k})=y\}.$$

Clearly, $C_T(y|x)$ is an enumerable from above function of T, x, and y. Using a standard argument [6], one can show that there exists an optimal twice prefix machine U in the sense that for any twice prefix machine T

$$C_U(y|x)\stackrel{+}{\leqslant}C_T(y|x).$$

Definition 6. *Complexity monotone in conditions* is defined for some fixed optimal twice prefix machine U as

$$K_*(y|x*):=C_U(y|x)=\min\{\ell(p)\,|\,\exists k\leqslant\ell(x):U(p,x_{1:k})=y\}.$$

Here $*$ in $x*$ is a syntactical part of the complexity notation $K_*(\cdot|\cdot*)$, though one may think of $K_*(y|x*)$ as of the minimal length of a program that produces y given any $z=x*$.

Theorem 7. *For any computable measure μ and any $x,y\in\mathcal{X}^*$ it holds*

$$\log_2\frac{\mu(y|x)}{M(y|x)}\stackrel{+}{\leqslant}K_*(\mu|x*)+K(\lceil d_\mu(x)\rceil).$$

Note. One can get slightly stronger variants of Theorems 1 and 7 by replacing the complexity of a standard code of μ by more sophisticated values. First, in any effective encoding there are

many codes for every μ, and in all the upper bounds (including Solomonoff's one) one can take the minimum of the complexities of all the codes for μ. Moreover, in Theorem 1 it is sufficient to take the complexity of $\mu^x=\mu(\cdot|x)$ (and it is sufficient that μ^x is enumerable, while μ can be incomputable). For Theorem 7 one can prove a similar strengthening: The complexity of μ is replaced by the complexity of any computable function that is equal to μ on all prefixes and prolongations of x.

To demonstrate the usefulness of the new bound, let us again consider some deterministic environment $\mu\mathrel{\widehat{=}}\alpha$. For $\mathcal{X}=\{0,1\}$ and $\alpha=x^\infty$ with $x=0^n1$, Theorem 1 gives the bound $K(\mu|n)+K(n)\stackrel{+}{=}K(n)$. Consider the new bound $K_*(\mu|x*)+K(\lceil d_\mu(x)\rceil)$. Since μ is deterministic, we have $d_\mu(x)=\log_2 M(x)\stackrel{+}{=}-K(n)$, and $K(\lceil d_\mu(x)\rceil)\stackrel{+}{=}K(K(n))$. To estimate $K_*(\mu|x*)$, let us consider a machine T that reads only its second tape and outputs the number of 0s before the first 1. Clearly, $C_T(n|x)=0$, hence $K_*(\mu|x*)\stackrel{+}{=}0$. Finally, $K_*(\mu|x*)+K(\lceil d_\mu(x)\rceil)\stackrel{+}{\leqslant}K(K(n))$, which is much smaller than $K(n)$.

6. Properties of the new complexity

The above definition of K_* is based on computations of some Turing machine. Such definitions are quite visual, but are often not convenient for formal proofs. We will give an alternative definition in terms of enumerable sets (see [20] for definitions of unconditional complexities in this style), which summarizes the properties we actually need for the proof of Theorem 7.

An enumerable set E of triples of strings is called *K_*-correct* if it satisfies the following requirements:

1. if $\langle p,x,y_1\rangle\in E$ and $\langle p,x,y_2\rangle\in E$, then $y_1=y_2$;
2. if $\langle p,x,y\rangle\in E$, then $\langle p',x',y\rangle\in E$ for all p' being prolongations of p and all x' being prolongations of x;
3. if $\langle p,x',y\rangle\in E$ and $\langle p',x,y\rangle\in E$, and p is a prefix of p' and x is a prefix of x', then $\langle p,x,y\rangle\in E$.

A complexity of y under a condition x w.r.t. a set E is

$$C_E(y|x)=\min\{\ell(p)\mid\langle p,x,y\rangle\in E\}.$$

A K_*-correct set E is called *optimal* if

$$C_E(y|x)\stackrel{+}{\leqslant}C_{E'}(y|x)$$

for any K_*-correct set E'. One can easily construct an enumeration of all K_*-correct sets, and an optimal set exists by the standard argument.

It is easy to see that a twice prefix Turing machine T can be transformed to a set E such that $C_T(y|x)=C_E(y|x)$. The set E is constructed as follows: T is run on all possible inputs, and if $T(p,x)=y$, then pairs $\langle p',x',y\rangle$ are added to E for all p' being prolongations of p and all x' being prolongations of x. Evidently, E is enumerable, and the second requirement of K_*-correctness is satisfied. To verify the other requirements, let us consider arbitrary $\langle p'_1,x'_1,y_1\rangle\in E$ and $\langle p'_2,x'_2,y_2\rangle\in E$ such that p'_1 and p'_2, x'_1 and x'_2 are comparable (one is a prefix of the other). By construction of E, there are p_i being prefixes of p'_i and x_i being prefixes of x'_i such that $T(p_i,x_i)=y_i$ for $i=1,2$. Clearly, p_1 and p_2, x_1 and

x_2 are comparable too. Since replacing the unused part of the inputs does not affect the running of the machine T and comparable words have a common prolongation, we get $p_1 = p_2$, $x_1 = x_2$, and $y_1 = y_2$. Thus E is a K_*-correct set.

The transformation in the other direction is impossible in some cases: the set $E = \{\langle 0^{h(n)}p, 0^n 1 q, 0\rangle \mid n \in \mathbb{N}, p,q \in \{0,1\}^*\}$, where $h(n)$ is 0 if the n-th Turing machine halts and 1 otherwise, is K_*-correct, but does not have a corresponding machine T. (Assume that such a machine T exists. If the n-th machine halts, then $\langle \epsilon, 0^n 1, 0\rangle \in E$ and thus T does not read the input tape at all. If the n-th machine does not halt, then $\langle 0, 0^n 1, 0\rangle \in E$ and $\langle 1, 0^n 1, 0\rangle \notin E$ and thus T has to read first symbol on the input tape. Therefore, one can use T to solve the halting problem.) However, we conjecture (but cannot prove) that for every set E there exists a machine T such that $C_T(x|y) \stackrel{+}{=} C_E(x|y)$.

Probably, the requirements on E can be even weaker, namely, the third requirement might be superfluous. Let us notice that the first requirement of K_*-correctness allows us to consider the set E as a partial computable function: $E(p,x) = y$ iff $\langle p,x,y\rangle \in E$. The second requirement says that E becomes a continuous function if we take the topology of prolongations (any neighborhood of $\langle p,x\rangle$ contains the cone $\{\langle p*,x*\rangle\}$) on the arguments and the discrete topology ($\{y\}$ is a neighborhood of y) on values. It is known (see [20] for references) that different complexities (plain, prefix, decision) can be naturally defined in a similar "topological" fashion. We conjecture the same is true in our case: an optimal enumerable set satisfying the requirements (1) and (2) (obviously, it exists) specifies the same complexity (up to an additive constant) as an optimal twice prefix machine.

It follows immediately from the definition(s) that $K_*(y|x*)$ is monotone as a function of x: $K_*(y|xz*) \leqslant K_*(y|x*)$ for all x, y, z.

The following lemma provides bounds for $K_*(x|y*)$ in terms of prefix complexity K. The lemma holds for all the definitions of $K_*(x|y*)$ above.

Lemma 8. *For any $x,y \in \mathcal{X}^*$ it holds*

$$K(x|y) \stackrel{+}{\leqslant} K_*(x|y*) \stackrel{+}{\leqslant} \min_{l \leqslant \ell(y)} \{K(x|y_{1:l}) + K(l)\} \stackrel{+}{\leqslant} K(x).$$

In general, none of the bounds are equal to $K_*(x|y*)$ even within a $o(K(x))$ term, but they are attained for certain y: For every x there is a y such that

$$K(x|y) \stackrel{+}{=} 0 \quad \textit{and} \quad K_*(x|y*) \stackrel{+}{=} \min_{l \leqslant \ell(y)} \{K(x|y_{1:l}) + K(l)\} \stackrel{+}{=} K(x),$$

and for every x there is a y such that

$$K(x|y) \stackrel{+}{=} K_*(x|y*) \stackrel{+}{=} 0 \quad \textit{and} \quad \min_{l \leqslant \ell(y)} \{K(x|y_{1:l}) + K(l)\} \stackrel{+}{=} K(x).$$

Proof. The first inequality is trivial (any twice-prefix machine is also a prefix machine in the first argument), as well as the last one (consider $l = 0$). Let us describe a twice prefix machine that provides $K_*(x|y*) \stackrel{+}{\leqslant} \min_{l \leqslant \ell(y)} \{K(x|y_{1:l}) + K(l)\}$. The first tape contains a prefix code p_l of l followed by a prefix code p for x under condition $y_{1:l}$, and the second tape contains y. The machine reads the p_l on the first tape and reconstructs the number l, then reads l bits from the second tape, and then reads p using these bits as the condition. Thus, $K_*(x|y*) \stackrel{+}{\leqslant} \ell(p_l) + \ell(p) \stackrel{+}{\leqslant} K(l) + K(x|y_{1:l})$.

Let us show that the bounds are attained.

Let us observe that $K(x) \stackrel{+}{\leqslant} K_*(x|0^n*)$ for all x and n. Actually, let $P(x) = \max\{2^{-\ell(p)} \mid \exists n \langle p, 0^n, x\rangle \in E\}$ (which implies $-\log_2 P(x) \leqslant K_*(x|0^n*)$ for all n). Obviously, $P(x)$ is enumerable. Further, $\sum_x P(x) \leqslant 1$ since $\sum_x P(x)$ is a sum of $2^{-\ell(p)}$ over a prefix-free set of p. (Assume the converse, p is a prefix of q, and $\langle p, 0^n, x\rangle \in E$, $\langle q, 0^m, y\rangle \in E$ for some n, m, and different x, y. By the second requirement of K_*-correctness, $\langle q, 0^{\max\{m,n\}}, x\rangle \in E$, $\langle q, 0^{\max\{m,n\}}, y\rangle \in E$. By the first requirement, $x = y$, contradiction.) Thus, by the coding theorem, $K(x) \stackrel{+}{\leqslant} -\log_2 P(x) \stackrel{+}{\leqslant} K_*(x|0^n*)$.

To get the first example, for arbitrary x, let us take $y = 0^n$ such that n is the number of x in some ordering of all binary strings. Then $K(x|y) \stackrel{+}{=} K(x|n) \stackrel{+}{=} 0$, $K_*(x|y*) \stackrel{+}{=} K(x)$, and we have $\min_l\{K(x|y_{1:l}) + K(l)\} \stackrel{+}{=} K(x)$ since $K_*(x|y*) \stackrel{+}{\leqslant} \min_l\{K(x|y_{1:l}) + K(l)\} \stackrel{+}{\leqslant} K(x|n) + K(n) \stackrel{+}{=} K(x)$.

To get the second example, for an arbitrary x let us take n such that $K(l) \geqslant K(x)$ for all $l \geqslant n$. Then put $y = 0^n 1 \tilde{x}$, where $\tilde{x}$ is any prefix code of x (e. g., $\tilde{x} = 0^{\ell(x)} 1 x$). Obviously, $K(x|y) \stackrel{+}{=} 0$ and $K_*(x|y*) \stackrel{+}{=} 0$. Consider $K(x|y_{1:l}) + K(l)$. If $l \leqslant n$, then it is equal to $K(x|0^l) + K(l) \stackrel{+}{\geqslant} K(\langle x, l\rangle) \stackrel{+}{\geqslant} K(x)$. If $l > n$, then $K(l) \geqslant K(x)$ by definition of n. □

Corollary 9. *The future deviation of M_t from μ_t is bounded by*

$$\sum_{t=l+1}^{\infty} \mathbf{E}[s_t|\omega_{1:l}] \stackrel{+}{\leqslant} [K_*(\mu|\omega_{1:l}*) + K(\lceil d_\mu(\omega_{1:l})\rceil)] \ln 2$$
$$\stackrel{+}{\leqslant} [\min_{i \leqslant l}\{K(\mu|\omega_{1:i}) + K(i)\} + K(\lceil d_\mu(\omega_{1:l})\rceil)] \ln 2\,.$$

Let us note that if ω is μ-random, then $K(\lceil d_\mu(\omega_{1:l})\rceil) \stackrel{+}{\leqslant} K(\lceil d_\mu(\omega_{1:\infty})\rceil) + K(K(\mu))$, and therefore we get the bound, which does not increase with l, in contrast to the bound (i) in Corollary 2.

Finally, let us point out one more approach to defining the complexity K_*. The survey [20] provides "encoding-free" definitions of the main complexities. In a similar fashion, K_* could be defined as a minimal (up to an additive constant) function with the following properties:

1. The function $K_*(y|x*)$ is non-negative and co-enumerable;
2. $K_*(y|xz*) \leqslant K_*(y|x*)$ for all x, y, z;
3. $\sum_y 2^{-K_*(y|x*)} \leqslant 1$ for all x.

Probably, condition 2 expressing strict monotonicity is superfluous, and both conditions 2 and 3 can be replaced by

2′. For any set $A = \{\langle x, y\rangle\}$ such that all the first elements x of the pairs from A have a common prolongation and the second elements y are different for all pairs from A, it holds $\sum_{\langle x,y\rangle \in A} 2^{-K_*(y|x*)} \leqslant 1$.

It is easy to check that these properties are satisfied for all the previously defined "versions" of K_*. We conjecture that all the definitions are equivalent, though we cannot prove this.

7. Proof of Theorem 7

If $\mu(x)=0$, then $d_\mu(x)=\infty$ and the bound trivially holds. Below assume that $\mu(x)\neq 0$ and thus $d_\mu(x)$ is finite.

The plan is to get a statement of the form $2^d\mu(xy)\stackrel{\times}{\leqslant} M(xy)$, where $d\approx d_\mu(x)=\log_2\frac{M(x)}{\mu(x)}$. To this end, we define a new semimeasure ν: we take the set $S=\{z|d_\mu(z)>d\}$ and put ν to be $2^d\mu$ on prolongations of $z\in S$; this is possible since S has μ-measure 2^{-d}. Then we have $\nu(z)\leqslant C\cdot M(z)$ by universality of M. However, the constant C depends on μ and also on d. To make the dependence explicit, we repeat the above construction for all numbers d and all semimeasures μ^T, obtaining semimeasures $\nu_{d,T}$, and take $\nu=\sum 2^{-K(d)}\cdot 2^{-K(T)}\nu_{d,T}$. This construction would give us the term $K(\mu)$ in the right-hand side of Theorem 7. To get $K_*(\mu|x*)$, we need a more complicated strategy: instead of a sum of semimeasures $\nu_{d,T}$, for every fixed d we sum "pieces" of $\nu_{d,T}$ at each point z, with coefficients depending on z as well as on d and T.

Now proceed with the formal proof. Let $\{\mu^T\}_{T\in\mathbb{N}}$ be any (effective) enumeration of all enumerable semimeasures. For any integer d and any T, put

$$S_{d,T}:=\left\{z\left|\sum_{v\in\mathcal{X}^{\ell(z)}\setminus\{z\}}\mu^T(v)+2^{-d}M(z)>1\right.\right\}.$$

The set $S_{d,T}$ is enumerable given d and T.

Let E be the optimal K_*-correct set (satisfying all three requirements). $E(p,z)$ is the corresponding partial computable function. For any $z\in\mathcal{X}^*$ and T, put

$$\lambda_{d,T}(z):=\max\{2^{-\ell(p)}\,|\,\exists k\leqslant\ell(z):z_{1:k}\in S_{d,T}\text{ and }E(p,z_{1:k})=T\}$$

(if there is no such p, then $\lambda_{d,T}(z)=0$). Put

$$\tilde\nu_d(z):=\sum_T\lambda_{d,T}(z)\cdot 2^d\mu^T(z).$$

Obviously, this value is enumerable. It is not a semimeasure, but it has the following property:

Claim 10. *For any prefix-free set A,*

$$\sum_{z\in A}\tilde\nu_d(z)\leqslant 1.$$

This implies that there exists an enumerable semimeasure ν_d such that $\nu_d(z)\geqslant\tilde\nu_d(z)$ for all z. Actually, to enumerate ν_d, one enumerates $\tilde\nu_d(z)$ for all z, and at each step the current approximation of $\nu_d(z)$ is the maximum of the current approximations of $\tilde\nu_d(z)$ and $\sum_{u\in\mathcal{X}}\nu_d(zu)$. Trivially, this provides $\nu_d(z)\geqslant\sum_{u\in\mathcal{X}}\nu_d(zu)$. To show that $\nu_d(\epsilon)\leqslant 1$, let us note that at any step of enumeration the current approximation of $\nu_d(\epsilon)$ is the sum of current approximations of $\tilde\nu_d(z)$ over some prefix-free set, and thus is bounded by 1. Put

$$\nu(z):=\sum_d 2^{-K(d)}\nu_d(z).$$

Clearly, ν is an enumerable semimeasure, thus $\nu(z) \stackrel{\times}{\leqslant} M(z)$. Let μ be an arbitrary computable measure, and $x,y \in \mathcal{X}^*$. Let $p \in \{0,1\}^*$ be a string such that $K_*(\mu|x*) = \ell(p)$, $E(p,x) = T$, and $\mu = \mu^T$. Put $d = \lceil d_\mu(x) \rceil - 1$, i.e., $d_\mu(x) - 1 \leqslant d < d_\mu(x)$. Hence $\mu(x) < 2^{-d} M(x)$. Since $\mu = \mu^T$ is a measure, we have $\sum_{v \in \mathcal{X}^{\ell(x)}} \mu^T(v) = 1$, and therefore $x \in S_{d,T}$. By definition, $\lambda_{d,T}(xy) \geqslant 2^{-\ell(p)}$, thus $\tilde{\nu}_d(xy) \geqslant 2^{-\ell(p)} 2^d \mu(xy)$, and

$$2^{-K(d)} 2^{-\ell(p)} 2^d \mu(xy) \leqslant \nu(xy) \stackrel{\times}{\leqslant} M(xy).$$

Replacing 2^d in the left-hand side by a smaller value $2^{d_\mu(x)-1}$, after trivial transformations we get

$$\log_2 \frac{\mu(y|x)}{M(y|x)} \stackrel{+}{\leqslant} K_*(\mu|x*) + K(d),$$

which completes the proof of Theorem 7.

Proof of Claim 10. First observe that for all $z \in S_{d,T}$

$$M(z) > 2^d \mu^T(z),$$

since

$$\sum_{v \in \mathcal{X}^{\ell(z)} \setminus \{z\}} \mu^T(v) + 2^{-d} M(z) > 1 \quad \text{and} \quad \sum_{v \in \mathcal{X}^{\ell(z)}} \mu^T(v) \leqslant 1$$

by definition of $S_{d,T}$ and by the semimeasure property, respectively. To prove the claim we will group items with the same μ^T, replace sums of μ^T-measures of several z by the μ^T-measure of their common prefix from $S_{d,T}$, change μ^T to M using the inequality above, and finally show (using "prefix-free" properties of K_*) that the coefficients of $M(z)$ in the sum are small. By definition,

$$\sum_{z \in A} \tilde{\nu}_d(z) = \sum_{z \in A} \sum_T \lambda_{d,T}(z) \cdot 2^d \mu^T(z) = \sum_T \sum_{z \in A} \lambda_{d,T}(z) \cdot 2^d \mu^T(z).$$

Let us estimate the inner sum. Let $\pi_{d,T}(z)$ be the string p that gives the maximum in the definition of $\lambda_{d,T}(z)$ (if there are several such p we always take, say, the lexicographically first), that is $\lambda_{d,T}(z) = 2^{-\ell(p)}$ and there exists z' being a prefix of z such that $z' \in S_{d,T}$ and $E(p,z') = T$. Let $\zeta_{d,T}(z)$ be the shortest of such z'. It is easy to see that $\zeta_{d,T}(\zeta_{d,T}(z)) = \zeta_{d,T}(z)$ and $\lambda_{d,T}(\zeta_{d,T}(z)) = \lambda_{d,T}(z)$

$$\begin{aligned}
\sum_{z \in A} \lambda_{d,T}(z) \cdot 2^d \mu^T(z) &= \sum_v \sum_{z \in A : \zeta_{d,T}(z) = v} \lambda_{d,T}(z) \cdot 2^d \mu^T(z) \;=\; \sum_v \sum_{z \in A : \zeta_{d,T}(z) = v} \lambda_{d,T}(v) \cdot 2^d \mu^T(z) \\
&\leqslant \sum_{v : \exists z \in A : \zeta_{d,T}(z) = v} \lambda_{d,T}(v) \cdot 2^d \mu^T(v) \;\leqslant\; \sum_{v : \zeta_{d,T}(v) = v} \lambda_{d,T}(v) \cdot 2^d \mu^T(v) \\
&< \sum_{v : \zeta_{d,T}(v) = v} \lambda_{d,T}(v) M(v)\,.
\end{aligned}$$

In the first inequality we used that $\zeta_{d,T}(z)$ is a prefix of z, that the set A is prefix free, and summed the $\mu^T(z)$ to $\mu^T(v)$. Now we can forget about A. If $\zeta_{d,T}(z) = v$ for some z, then $\zeta_{d,T}(v) = \zeta_{d,T}(\zeta_{d,T}(z)) = v$,

and we get the second inequality. The last inequality holds since $\zeta_{d,T}(v)$ belongs to $S_{d,T}$. Thus, we need to bound the sum

$$\sum_T \sum_{v: v=\zeta_{d,T}(v)} \lambda_{d,T}(v) M(v) = \sum_v \left(\sum_{T: v=\zeta_{d,T}(v)} \lambda_{d,T}(v) \right) M(v).$$

We say that a function $f: \mathcal{X}^* \to [0,1]$ is *unit-summable along any sequence* if for any $z \in \mathcal{X}^*$

$$\sum_{i=1}^{\ell(z)} f(z_{1:i}) \leqslant 1.$$

Claim 11. *The function* $f(v) = \sum_{T: v=\zeta_{d,T}(v)} \lambda_{d,T}(v)$ *is unit-summable along any sequence.*

Lemma 12. *Let* ν *be a semimeasure. If a function* f *is unit-summable along any sequence, then*

$$\sum_{z \in \mathcal{X}^*} f(z)\nu(z) \leqslant 1.$$

This concludes the proof of Claim 10. □

Proof of Lemma 12. Since $f(z)$ and $\nu(z)$ are non-negative, it is sufficient to prove $\sum_{\ell(z) \leqslant n} f(z)\nu(z) \leqslant 1$ for all n. Also we can assume that ν is a measure (the sum does not decrease, if ν is increased to a measure).

$$\sum_{\ell(z) \leqslant n} f(z)\nu(z) = \sum_{\ell(z) \leqslant n} f(z) \sum_{\substack{\ell(v) = n, \\ z \text{ prefix of } v}} \nu(v) = \sum_{\ell(v)=n} \sum_{\substack{\ell(z) \leqslant n, \\ z \text{ prefix of } v}} f(z)\nu(v)$$

$$= \sum_{\ell(v)=n} \sum_{i=1}^{n} f(v_{1:i})\nu(v) \leqslant \sum_{\ell(v)=n} \nu(v) \leqslant 1. \qquad \square$$

Proof of Claim 11. Take any $z \in \mathcal{X}^*$. Let us show that

$$\sum_{\substack{v \text{ prefix of } z, \\ T: v=\zeta_{d,T}(v)}} \lambda_{d,T}(v) \leqslant 1.$$

Recall that if $\lambda_{d,T}(v) \neq 0$, then $\lambda_{d,T}(v) = 2^{-\ell(\pi_{d,T}(v))}$. We will show that the set $B(z) = \{\pi_{d,T}(v) \mid v = \zeta_{d,T}(v),\ v \text{ is a prefix of } z\}$ is prefix free, and if $\pi_{d,T_1}(v_1) = \pi_{d,T_2}(v_2) \in B(z)$, then $v_1 = v_2$ and $T_1 = T_2$. Consequently,

$$\sum_{\substack{v \text{ prefix of } z, \\ T: v=\zeta_{d,T}(v)}} \lambda_{d,T}(v) = \sum_{p \in B(z)} 2^{-\ell(p)} \leqslant 1.$$

Assume the converse, that there exist different v_i, T_i, $i=1,2$, such that $p_1=\pi_{d,T_1}(v_1)$ is a prefix (proper or not) of $p_2=\pi_{d,T_2}(v_2)$, v_1 and v_2 are prefixes of z, and $v_i=\zeta_{d,T_i}(v_i)$.

By definition of ζ, we have $v_i \in S_{d,T_i}$ and $T_i = E(p_i,v_i)$. Hence, by the second requirement of K_*-correctness, $T_1=E(p_1,v_1)=E(p_2,z)=E(p_2,v_2)=T_2$. Let $T=T_1=T_2$.

Let us show that $v_1=v_2$ too. Since they both are prefixes of z, one of them is a prefix of the other. Suppose v_1 is a prefix of v_2: By the second requirement of K_*-correctness, $E(p_2,v_1)=E(p_1,v_1)=T$. By definition, $\zeta_{d,T}(v_2)$ is the shortest prefix of v_2 belonging to $S_{d,T}$ and such that $E(p_2,\cdot)=T$, therefore $\zeta_{d,T}(v_2)$ is a prefix of v_1, and thus $v_1=v_2$. Suppose v_2 is a prefix of v_1. Since $E(p_1,v_1)=T$ and $E(p_2,v_2)=T$, we have $E(p_1,v_2)=T$ by the third requirement of K_*-correctness. As before, we get $\zeta_{d,T}(v_1)$ is a prefix of v_2, and $v_1=v_2$. □

8. Discussion

8.1. Conclusion

We evaluated the quality of predicting a stochastic sequence at an intermediate time, when some beginning of the sequence has been already observed, estimating the future loss of the universal Solomonoff predictor M. We proved general upper bounds for the discrepancy between conditional values of the predictor M and the true environment μ, and demonstrated a kind of tightness for these bounds. One of the bounds is based on a new variant of conditional algorithmic complexity K_*, which has interesting properties in its own. In contrast to standard prefix complexity K, K_* is a monotone function of conditions: $K_*(y|xz*) \leqslant K_*(y|x*)$.

8.2. General Bayesian posterior bounds

A natural question is whether posterior bounds for general Bayes mixtures based on general $\mathcal{M} \ni \mu$ could also be derived. The mixture representation (3) can be written as a posterior representation

$$\xi(y|x) = \sum_{\nu\in\mathcal{M}} w_\nu(x)\nu(y|x) \geqslant w_\mu(x)\mu(y|x), \quad \text{where} \quad w_\nu(x) := w_\nu \frac{\nu(x)}{\xi(x)}$$

is the posterior belief in ν after observing x (and w_ν is the prior). This immediately implies the bound $D_{l:\infty} \leqslant \ln w_\mu(\omega_{<l})^{-1}$. Strangely enough, for $\mathcal{M}=\mathcal{M}_U$, $\log_2 w_\nu^{-1} := K(\nu)$ does *not* imply $\log_2 w_\mu(x)^{-1} = K(\mu|x)$, not even within logarithmic accuracy, so it was essential to consider $D_{l:\infty}$. It would be interesting to derive bounds on $D_{l:\infty}$ or $\ln w_\mu(x)^{-1}$ for general $\mathcal{M}$ similar to the ones derived here for $\mathcal{M}=\mathcal{M}_U$.

8.3. Online classification

All considered distributions $\rho(x)$ (in particular ξ, M, and μ) may be replaced everywhere by distributions $\rho(x|z)$ additionally conditioned on some z. The z-conditions nowhere cause problems as they can essentially be thought of as fixed (or as oracles or spectators). An (i.i.d.) classification

problem is a typical example: At time t one arranges an experiment z_t (or observes data z_t), then tries to make a prediction, and finally observes the true outcome x_t with probability $\mu(x_t|z_t)$. In this case $\mathcal{M} = \{\nu(x_{1:n}|z_{1:n}) = \nu(x_1|z_1)\cdots\nu(x_n|z_n)\}$. (Note that ξ is not i.i.d). Solomonoff's bound $K(\mu)\ln 2$ in (6) holds unchanged. Compared to the sequence prediction case we have extra information z, so we may wonder whether some improved bound $K(\mu|z)$ or so, holds. For a fixed z this can be achieved by also replacing $2^{-K(\mu)}$ in (3) by $2^{-K(\mu|z)}$. But if at time t only $z_{1:t}$ is known like in the classification example, this leads to difficulties (ξ is no longer a (semi)measure, which sometimes can be corrected [21]). Alternatively we could keep definition (3) but apply it to the (chronologically correctly ordered) sequence $z_1x_1z_2x_2\cdots$, condition by (1) to $z_{1:t}$, and try to derive improved bounds.

8.4. More open problems

Since $D_{1:\infty}$ is finite, one may expect that the tails $D_{l:\infty}$ tend to 0 as $l\to\infty$. However, as Lemma 3 implies, this holds only with probability 1: for some special α we have even $D_{l:\infty}(\alpha_{<l}) \stackrel{+}{\geqslant} \frac{1}{3}K(l) \stackrel{l\to\infty}{\longrightarrow} \infty$. It would be very interesting to find a wide class of α such that $D_{l:\infty}(\alpha_{<l})\to 0$. The natural conjecture is that one should take μ-random α. Another (probably, closely related) task is to study the asymptotic behavior of $K_*(\mu|\alpha_{<l}*)$. It is natural to expect that $K_*(\mu|\alpha_{<l}*)$ is bounded by an absolute constant (independent of μ) for "most" α and for sufficiently large l. Finally, (dis)proving our conjectured equality of the various definitions of K_* we gave, would be interesting and useful.

Acknowledgments

The authors are grateful to Andrej Muchnik, Alexander Shen, and Nikolai Vereshchagin for discussing the history of the deficiency conservation theorem (Theorem 4), and to anonymous referees for useful comments.

References

[1] A. Chernov, M. Hutter, Monotone conditional complexity bounds on future prediction errors, in: Proc. 16th Int. Conf. Algorithmic Learning Theory (ALT'05), LNAI, vol. 3734, Springer, Berlin, Singapore, 2005, pp. 414–428, http://arxiv.org/abs/cs.LG/0507041.

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

[3] R.J. Solomonoff, Complexity-based induction systems: comparisons and convergence theorems, IEEE Trans. Inf. Theory IT-24 (1978) 422–432.

[4] J. Schmidhuber, The Speed Prior: a new simplicity measure yielding near-optimal computable predictions, in: Proc. 15th Annual Conf. Computational Learning Theory (COLT 2002), Lecture Notes in Artificial Intelligence, Springer, Sydney, Australia, 2002, pp. 216–228.

[5] M. Hutter, Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability, Springer, Berlin, 2005, 300 p.,http://www.idsia.ch/~marcus/ai/uaibook.htm.

[6] M. Li, P.M.B. Vitányi, An Introduction to Kolmogorov Complexity and Its Applications, second ed., Springer, Berlin, 1997.

[7] R. Cilibrasi, P.M.B. Vitányi, Clustering by compression, IEEE Trans. Information Theory 51 (4) (2005) 1523–1545.

[8] M. Hutter, New error bounds for Solomonoff prediction, J. Comput. Syst. Sci. 62 (4) (2001) 653–667,http://arxiv.org/abs/cs.AI/9912008.

[9] M. Hutter, Convergence and loss bounds for Bayesian sequence prediction, IEEE Trans. Inf. Theory 49 (8) (2003) 2061–2067.http://arxiv.org/abs/cs.LG/0301014.
[10] M. Hutter, Optimality of universal Bayesian prediction for general loss and alphabet, J. Mach. Learn. Res. 4 (2003) 971–1000.http://arxiv.org/abs/cs.LG/0311014.
[11] M. Hutter, General loss bounds for universal sequence prediction, in: Proc. 18th Intl. Conf. Machine Learning (ICML-2001), 2001, pp. 210–217,http://arxiv.org/abs/cs.AI/0101019.
[12] M. Hutter, Convergence and error bounds for universal prediction of nonbinary sequences, in: Proc. 12th European Conf. Machine Learning (ECML-2001), 2001,pp. 239–250. http://arxiv.org/abs/cs.LG/0106036.
[13] A.K. Zvonkin, 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, Russ. Math. Surveys 25 (6) (1970) 83–124.
[14] M. Hutter, A.A. Muchnik, Universal convergence of semimeasures on individual random sequences, in: Proc. 15th Int. Conf. Algorithmic Learning Theory (ALT'04), LNAI, vol. 3244, Springer, Berlin, Padova, 2004, pp. 234–248,http://arxiv.org/abs/cs.LG/0407057.
[15] J. Schmidhuber, Algorithmic theories of everything, Report IDSIA-20-00, quant-ph/0011122, IDSIA, Manno (Lugano), Switzerland, 2000.
[16] J. Schmidhuber, Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit, Int. J. Foundations Comput. Sci. 13 (4) (2002) 587–612.
[17] M. Hutter, Sequence prediction based on monotone complexity, in: Proc. 16th Annual Conf. Learning Theory (COLT'03), LNAI, vol. 2777, Springer, Berlin, 2003, pp. 506–521. http://arxiv.org/abs/cs.AI/0306036.
[18] M. Hutter, Sequential predictions based on algorithmic complexity, J. Comput. Syst. Sci. 72 (2006) 95–117,http://arxiv.org/abs/cs.IT/0508043.
[19] N.K. Vereshchagin, A. Shen, V.A. Uspensky, Lecture Notes on Kolmogorov Complexity, Unpublished, http://lpcs.math.msu.su/∼ver/kolm-book (2005).
[20] V.A. Uspensky, A. Shen, Relations between varieties of Kolmogorov complexities, Math. Systems Theory 29 (1996) 271–292.
[21] J. Poland, M. Hutter, Convergence of discrete MDL for sequential prediction, in: Proc. 17th Annual Conf. Learning Theory (COLT'04), LNAI, vol. 3120, Springer, Berlin, Banff, 2004, pp. 300–314, http://arxiv.org/abs/cs.LG/0404057.