Technical Report IDSIA-11-00, 14. November 2000

New Error Bounds for Solomonoff Prediction

Marcus Hutter

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

Keywords

Induction; Solomonoff, Bayesian, deterministic prediction; algorithmic probability, Kolmogorov complexity.

Abstract

Solomonoff sequence prediction is a scheme to predict digits of binary strings without knowing the underlying probability distribution. We call a prediction scheme informed when it knows the true probability distribution of the sequence. Several new relations between universal Solomonoff sequence prediction and informed prediction and general probabilistic prediction schemes will be proved. Among others, they show that the number of errors in Solomonoff prediction is finite for computable distributions, if finite in the informed case. Deterministic variants will also be studied. The most interesting result is that the deterministic variant of Solomonoff prediction is optimal compared to any other probabilistic or deterministic prediction scheme apart from additive square root corrections only. This makes it well suited even for difficult prediction problems, where it does not suffice when the number of errors is minimal to within some factor greater than one. Solomonoff's original bound and the ones presented here complement each other in a useful way.

1 Introduction

Induction is the process of predicting the future from the past or, more precisely, it is the process of finding rules in (past) data and using these rules to guess future data. The induction principle has been subject to long philosophical controversies. Highlights are Epicurus' principle of multiple explanations, Occams' razor (simplicity) principle and Bayes' rule for conditional probabilities [2]. In 1964, Solomonoff [8] elegantly unified all these aspects into one formal theory of inductive inference. The theory allows the prediction of digits of binary sequences without knowing their true probability distribution in contrast to what we call an informed scheme, where the true distribution is known. A first error estimate was also given by Solomonoff 14 years later in [9]. It states that the total means squared distance of the prediction probabilities of Solomonoff and informed prediction is bounded by the Kolmogorov complexity of the true distribution. As a corollary, this theorem ensures that Solomonoff prediction converges to informed prediction for computable sequences in the limit. This is the key result justifying the use of Solomonoff prediction for long sequences of low complexity.

Another natural question is to ask for relations between the total number of expected errors E_ξ in Solomonoff prediction and the total number of prediction errors E_μ in the informed scheme. Unfortunately [9] does not bound E_ξ in terms of E_μ in a satisfactory way. For example it does not exclude the possibility of an infinite E_ξ even if E_μ is finite. Here we want to prove upper bounds to E_ξ in terms of E_μ ensuring as a corollary that the above case cannot happen. On the other hand, our theorem does not say much about the convergence of Solomonoff to informed prediction. So Solomonoff's and our bounds complement each other in a nice way.

In the preliminary Section 2 we give some notations for strings and conditional probability distributions on strings. Furthermore, we introduce Kolmogorov complexity and the universal probability, where we take care to make the latter a true probability measure.

In Section 3 we define the general probabilistic prediction scheme (ρ) and Solomonoff (ξ) and informed (μ) prediction as special cases. We will give several error relations between these prediction schemes. A bound for the error difference $|E_\xi - E_\mu|$ between Solomonoff and informed prediction is the central result. All other relations are then simple, but interesting consequences or known results such as the Euclidean bound.

In Section 4 we study deterministic variants of Solomonoff (Θ_ξ) and informed (Θ_μ) prediction. We will give similar error relations as in the probabilistic case between these prediction schemes. The most interesting consequence is that the Θ_ξ system is optimal compared to any other probabilistic or deterministic prediction scheme apart from additive square root corrections only.

In the Appendices A, B and C we prove the inequalities (18), (20) and (26), which are the central parts for the proofs of the Theorems 1 and 2.

For an excellent introduction to Kolmogorov complexity and Solomonoff induction one should consult the book of Li and Vitányi [7] or the article [6] for a short course. Historical surveys of inductive reasoning/inference can be found in [1, 10].

2 Preliminaries

Throughout the paper we will consider binary sequences/strings and conditional probability measures on strings.

We will denote strings over the binary alphabet $\{0,1\}$ by $s = x_1 x_2 ... x_n$ with $x_k \in \{0,1\}$ and their lengths with $l(s) = n$. ϵ is the empty string, $x_{n:m} := x_n x_{n+1} ... x_{m-1} x_m$ for $n \leq m$ and ϵ for $n > m$. Furthermore, $x_{<n} := x_1 ... x_{n-1}$.

We use Greek letters for probability measures and underline their arguments to indicate that they are probability arguments. Let $\rho_n(\underline{x_1 ... x_n})$ be the probability that an (infinite) sequence starts with $x_1 ... x_n$. We drop the index on ρ if it is clear from its arguments:

$$\sum_{x_n \in \{0,1\}} \rho(\underline{x_{1:n}}) = \sum_{x_n} \rho_n(\underline{x_{1:n}}) = \rho_{n-1}(\underline{x_{<n}}) = \rho(\underline{x_{<n}}), \quad \rho(\epsilon) = \rho_0(\epsilon) = 1. \tag{1}$$

We also need conditional probabilities derived from Bayes' rule. We prefer a notation which preserves the order of the words in contrast to the standard notation $\rho(\cdot|\cdot)$ which flips it. We extend the definition of ρ to the conditional case with the following convention for its arguments: An underlined argument $\underline{x_k}$ is a probability variable and other non-underlined arguments x_k represent conditions. With this convention, Bayes' rule has the following look:

$$\begin{gathered} \rho(x_{<n}\underline{x_n}) = \rho(\underline{x_{1:n}}) / \rho(\underline{x_{<n}}) \quad \text{and} \\ \rho(\underline{x_1 ... x_n}) = \rho(\underline{x_1}) \cdot \rho(x_1 \underline{x_2}) \cdot ... \cdot \rho(x_1 ... x_{n-1} \underline{x_n}). \end{gathered} \tag{2}$$

The first equation states that the probability that a string $x_1 ... x_{n-1}$ is followed by x_n is equal to the probability that a string starts with $x_1 ... x_n$ divided by the probability that a string starts with $x_1 ... x_{n-1}$. The second equation is the first, applied n times.

Let us choose some universal monotone Turing machine U with unidirectional input and output tapes and a bidirectional work tape. We can then define the prefix Kolmogorov complexity [3, 5] as the length of the shortest program p, for which U outputs string s:

$$K(s) \ := \ \min_p \{ l(p) : U(p) = s \}. \tag{3}$$

The universal semi-measure $M(s)$ is defined as the probability that the output of the universal Turing machine U starts with s when provided with fair coin flips on the input tape. It is easy to see that this is equivalent to the formal definition

$$M(s) \ := \sum_{p \,:\, \exists \omega : U(p) = s\omega} 2^{-l(p)}, \tag{4}$$

where the sum is over minimal programs p for which U outputs a string starting with s. U might be non-terminating. M has the important universality property [12] that it majorizes every computable probability measure ρ up to a multiplicative factor depending only on ρ but not on s:

$$\rho(\underline{s}) \leq 2^{K(\rho) + O(1)} M(s). \tag{5}$$

The Kolmogorov complexity of a function like ρ is defined as the length of the shortest self-delimiting coding of a Turing machine computing this function. Unfortunately M itself is *not* a probability measure on the binary strings. We have $M(s0)+M(s1)<M(s)$ because there are programs p which output just s, followed neither by 0 nor by 1; they just stop after printing s or continue forever without any further output. This drawback can easily be corrected[1][9]. Let us define the universal probability measure ξ by defining first the conditional probabilities

$$\xi(s\underline{x}) \;:=\; \frac{M(sx)}{M(s0)+M(s1)} \quad , \quad x\in\{0,1\} \quad , \quad \xi(\epsilon):=1 \tag{6}$$

and then by using (2) to get $\xi(\underline{x_1...x_n})$. It is easily verified by induction that ξ is indeed a probability measures and universal

$$\rho(\underline{s}) \leq 2^{K(\rho)+O(1)}\xi(\underline{s}). \tag{7}$$

The latter follows from $\xi(\underline{s}) \geq M(s)$ and (5). The universality property (7) is all we need to know about ξ in the following.

3 Probabilistic Sequence Prediction

Every inductive inference problem can be brought into the following form: Given a string x, give a guess for its continuation y. We will assume that the strings which have to be continued are drawn according to a probability distribution[2]. In this section we consider probabilistic predictors of the next bit of a string. So let $\mu(\underline{x_1...x_n})$ be the true probability measure of string $x_{1:n}$, $x_k\in\{0,1\}$ and $\rho(x_{<n}\underline{x}_n)$ be the probability that the system predicts x_n as the successor of $x_1...x_{n-1}$. We are not interested here in the probability of the next bit itself. We want our system to output either 0 or 1. Probabilistic strategies are useful in game theory where they are called mixed strategies. We keep μ fixed and compare different ρ. Interesting quantities are the probability of making an error when predicting x_n, given $x_{<n}$. If $x_n=0$, the probability of our system to predict 1 (making an error) is $\rho(x_{<n}\underline{1})=1-\rho(x_{<n}\underline{0})$. That x_n is 0 happens with probability $\mu(x_{<n}\underline{0})$. Analogously for $0\leftrightarrow 1$. So the probability of making a wrong prediction in the n^{th} step ($x_{<n}$ fixed) is

$$e_{n\rho}(x_{<n}) \;:=\; \sum_{x_n\in\{0,1\}} \mu(x_{<n}\underline{x}_n)[1-\rho(x_{<n}\underline{x}_n)]. \tag{8}$$

The total μ-expected number of errors in the first n predictions is

$$E_{n\rho} \;:=\; \sum_{k=1}^{n}\sum_{x_1...x_{k-1}} \mu(\underline{x}_{<k})\cdot e_{k\rho}(x_{<k}). \tag{9}$$

[1]Another popular way is to keep M and sacrifice some of the axioms of probability theory. The reason for doing this is that M, although not computable [7, 9], is at least enumerable. On the other hand, we are interested in conditional probabilities, derived from M, which are no longer enumerable anyway, so there is no reason for us to stick to M. ξ is still computable in the limit or approximable.

[2]This probability measure μ might be 1 for some sequence $x_{1:\infty}$ and 0 for all others. In this case, $K(\mu_n)$ is equal to $K(x_{1:n})$ (up to terms of order 1).

If μ is known, a natural choice for ρ is $\rho=\mu$. This is what we call an informed prediction scheme. If the probability of x_n is high (low), the system predicts x_n with high (low) probability. If μ is unknown, one could try the universal distribution ξ for ρ as defined in (4) and (6). This is known as Solomonoff prediction [8].

What we are most interested in is an upper bound for the μ-expected number of errors $E_{n\xi}$ of the ξ-predictor. One might also be interested in the probability difference of predictions at step n of the μ- and ξ-predictor or the total absolute difference to some power α (α-norm in n-space).

$$d_k^\alpha(x_{<k}) \;:=\; \sum_{x_k}\mu(x_{<k}\underline{x}_k)\cdot\Big|\xi(x_{<k}\underline{x}_k)-\mu(x_{<k}\underline{x}_k)\Big|^\alpha \;=\; \Big|\xi(x_{<k}\underline{0})-\mu(x_{<k}\underline{0})\Big|^\alpha$$

$$\Delta_n^{(\alpha)} \;:=\; \sum_{k=1}^n\sum_{x_{<k}}\mu(\underline{x}_{<k})\cdot d_k^\alpha(x_{<k}),\quad \alpha=1,2 \tag{10}$$

For $\alpha=2$ there is the well known-result [9]

$$\Delta_n^{(2)}<\tfrac{1}{2}\ln 2\cdot K(\mu)<\infty \quad \text{for computable } \mu. \tag{11}$$

One reason to directly study relations between $E_{n\xi}$ and $E_{n\mu}$ is that from (11) alone it does not follow that $E_{\infty\xi}$ is finite, if $E_{\infty\mu}$ is finite. Assume that we could choose μ such that $e_{n\mu}\sim 1/n^2$ and $e_{n\xi}\sim 1/n$. Then $E_{\infty\mu}$ would be finite, but $E_{\infty\xi}$ would be infinite, without violating (11). There are other theorems, the most prominent being $\xi(x_{<n}\underline{x}_n)/\mu(x_{<n}\underline{x}_n)\stackrel{n\to\infty}{\longrightarrow}1$ with μ probability 1 (see [7] page 332). However, neither of them settles the above question. In the following we will show that a finite $E_{\infty\mu}$ causes a finite $E_{\infty\xi}$.

Let us define the Kullback Leibler distance [4] or relative entropy between μ and ξ:

$$h_n(x_{<n}) := \sum_{x_n}\mu(x_{<n}\underline{x}_n)\ln\frac{\mu(x_{<n}\underline{x}_n)}{\xi(x_{<n}\underline{x}_n)}. \tag{12}$$

H_n is then defined as the sum-expectation for which the following can be shown [9]

$$H_n \;:=\; \sum_{k=1}^n\sum_{x_{<k}}\mu(\underline{x}_{<k})\cdot h_k(x_{<k}) \;=\; \sum_{k=1}^n\sum_{x_{1:k}}\mu(\underline{x}_{1:k})\ln\frac{\mu(x_{<k}\underline{x}_k)}{\xi(x_{<k}\underline{x}_k)} \;= \tag{13}$$

$$=\;\sum_{x_{1:n}}\mu(\underline{x}_{1:n})\ln\prod_{k=1}^n\frac{\mu(x_{<k}\underline{x}_k)}{\xi(x_{<k}\underline{x}_k)} \;=\; \sum_{x_{1:n}}\mu(\underline{x}_{1:n})\ln\frac{\mu(\underline{x}_{1:n})}{\xi(\underline{x}_{1:n})} \;<\; \ln 2\cdot K(\mu_n)+O(1)$$

In the first line we have inserted (12) and used Bayes rule $\mu(\underline{x}_{<k})\cdot\mu(x_{<k}\underline{x}_k)=\mu(\underline{x}_{1:k})$. Due to (1) we can replace $\sum_{x_{1:k}}\mu(\underline{x}_{1:k})$ by $\sum_{x_{1:n}}\mu(\underline{x}_{1:n})$ as the argument of the logarithm is independent of $x_{k+1:n}$. The k sum can now be exchanged with the $x_{1:n}$ sum and transforms to a product inside the logarithm. In the last equality we have used the second form of Bayes rule (2) for μ and ξ. If we use universality (7) of ξ, i.e. $\ln\mu(\underline{x}_{1:n})/\xi(\underline{x}_{1:n})<\ln 2\cdot K(\mu_n)+O(1)$, the final inequality in (13) is yielded, which is the basis of all error estimates.

We now come to our first theorem:

THEOREM 1. *Let there be binary sequences $x_1x_2...$ drawn with probability $\mu_n(\underline{x}_{1:n})$ for the first n bits. A ρ-system predicts by definition x_n from $x_{<n}$ with probability $\rho(x_{<n}\underline{x}_n)$. $e_{n\rho}(x_{<n})$ is the error probability in the n^{th} prediction (8) and $E_{n\rho}$ is the μ-expected total number of errors in the first n predictions (9). The following error relations hold between universal Solomonoff ($\rho=\xi$), informed ($\rho=\mu$) and general (ρ) predictions:*

$$
\begin{array}{rllll}
i) & |E_{n\xi}-E_{n\mu}| \;\leq\; \Delta_n^{(1)} & < & H_n+\sqrt{2E_{n\mu}H_n} & \\
ii) & \Delta_n^{(2)} \;\leq\; \frac{1}{2}H_n & & & \\
iii) & E_{n\xi} \;>\; \Delta_n^{(2)}+\frac{1}{2}E_{n\mu} & & & \\
iv) & E_{n\xi} \;>\; E_{n\mu}+H_n-\sqrt{2E_{n\mu}H_n} & > & H_n & \text{for} \quad E_{n\mu}>2H_n \\
v) & E_{n\mu} \;\leq\; 2E_{n\rho} \quad , \quad e_{n\mu}\;\leq\; 2e_{n\rho} & & \text{for any } \rho & \\
vi) & E_{n\xi} \;<\; 2E_{n\rho}+H_n+\sqrt{4E_{n\rho}H_n} & & \text{for any } \rho, &
\end{array}
$$

where $H_n<\ln 2\cdot K(\mu)+O(1)$ is the relative entropy (13) and $K(\mu)$ is the Kolmogorov complexity of μ (3).

COROLLARY 1. *For computable μ, i.e. for $K(\mu)<\infty$, the following statements immediately follow from Theorem 1:*

$$
\begin{array}{rl}
vii) & \text{if } E_{\infty\mu} \text{ is finite, then } E_{\infty\xi} \text{ is finite} \\
viii) & E_{n\xi}/E_{n\mu} \;=\; 1+O(E_{n\mu}^{-1/2}) \stackrel{E_{n\mu}\to\infty}{\longrightarrow} 1 \\
ix) & E_{n\xi}-E_{n\mu} = \quad O(\sqrt{E_{n\mu}}) \\
x) & E_{n\xi}/E_{n\rho} \;\leq\; 2+O(E_{n\rho}^{-1/2}).
\end{array}
$$

Relation (i) is the central new result. It is best illustrated for computable μ by the corollary. Statements (vii), $(viii)$ and (ix) follow directly from (i) and the finiteness of H_∞. Statement (x) follows from (vi).

First of all, (vii) ensures finiteness of the number of errors of Solomonoff prediction, if the informed prediction makes only a finite number of errors. This is especially the case for deterministic μ, as $E_{n\mu}=0$ in this case[3]. Solomonoff prediction makes only a finite number of errors on computable sequences. For more complicated probabilistic environments, where even the ideal informed system makes an infinite number of errors, (ix) ensures that the error excess of Solomonoff prediction is only of order $\sqrt{E_{n\mu}}$. This ensures that the error densities E_n/n of both systems converge to each other, but (ix) actually says more than this. It ensures that the quotient converges to 1 and also gives the speed of convergence $(viii)$.

Relation (ii) is the well-known Euclidean bound [9]. It is the only upper bound in Theorem 1 which remains finite for $E_{n\mu/\rho}\to\infty$. It ensures convergence of the individual prediction probabilities $\xi(x_{<n}\underline{x}_n)\to\mu(x_{<n}\underline{x}_n)$. Relation (iii) shows that the ξ system makes at least half of the errors of the μ system. Relation (iv) improves the lower bounds of (i) and (iii). Together with the upper bound in (i) it says that the excess of ξ errors as compared to μ

[3] We call a probability measure deterministic if it is 1 for exactly one sequence and 0 for all others.

errors is given by H_n apart from $O(\sqrt{E_{n\mu}H_n})$ corrections. The excess is neither smaller nor larger. This result is plausible, since knowing μ means additional information, which saves making some of the errors. The information content of μ (relative to ξ) is quantified in terms of the relative entropy H_n.

Relation (v) states that no prediction scheme can have less than half of the errors of the μ system, whatever we take for ρ. This ensures the optimality of μ apart from a factor of 2. Combining this with (i) ensures optimality of Solomonoff prediction, apart from a factor of 2 and additive (inverse) square root corrections (vi), (x). Note that even when comparing ξ with ρ, the computability of μ is what counts, whereas ρ might be any, even an uncomputable, probabilistic predictor. The optimality within a factor of 2 might be sufficient for some applications, especially for finite $E_{\infty\mu}$ or if $E_{n\mu}/n\to 0$, but is inacceptable for others. More about this in the next section, where we consider deterministic prediction, where no factor 2 occurs.

Proof of Theorem 1. The first inequality in (i) follows directly from the definition of E_n and Δ_n and the triangle inequality. For the second inequality, let us start more modestly and try to find constants A and B which satisfy the linear inequality

$$\Delta_n^{(1)} \;<\; A\!\cdot\! E_{n\mu} + B\!\cdot\! H_n \tag{14}$$

If we could show

$$d_k(x_{<k}) \;<\; A\!\cdot\! e_{k\mu}(x_{<k}) + B\!\cdot\! h_k(x_{<k}) \tag{15}$$

for all $k\leq n$ and all $x_{<k}$, (14) would follow immediately by summation and the definition of Δ_n, E_n and H_n. With k, $x_{<k}$, μ, ξ fixed now, we abbreviate

$$\begin{array}{rcl} y := \mu(x_{<k}\underline{1}) & , & 1-y = \mu(x_{<k}\underline{0}) \\ z := \xi(x_{<k}\underline{1}) & , & 1-z = \xi(x_{<k}\underline{0}) \\ r := \rho(x_{<k}\underline{1}) & , & 1-r = \rho(x_{<k}\underline{0}). \end{array} \tag{16}$$

The various error functions can then be expressed by y, z and r

$$\begin{array}{rcl} e_{k\mu} & = & 2y(1-y) \\ e_{k\xi} & = & y(1-z)+(1-y)z \\ e_{k\rho} & = & y(1-r)+(1-y)r \\ d_k & = & |y-z| \\ h_k & = & y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z}. \end{array} \tag{17}$$

Inserting this into (15) we get

$$|y-z| \;<\; A\!\cdot\! 2y(1-y) + B\!\cdot\!\left[y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z}\right]. \tag{18}$$

In Appendix A we will show that this inequality is true for $B\geq\frac{1}{2A}+1$, $A>0$. Inequality (14) therefore holds for any $A>0$, provided we insert $B=\frac{1}{2A}+1$. Thus we might minimize the r.h.s of (14) w.r.t A. The minimum is at $A=\sqrt{H_n/2E_{n\mu}}$ leading to the upper bound

$$\Delta_n^{(1)} \;<\; H_n+\sqrt{2E_{n\mu}H_n}$$

which completes the proof of (i).

Bound (ii) is well known [9]. It is already linear and is proved by showing $d_n^2 \leq \frac{1}{2}h_n$. Inserting the abbreviations (17) we get

$$2(y-z)^2 \;\leq\; y\ln\frac{y}{z} + (1-y)\ln\frac{1-y}{1-z} \tag{19}$$

This lower bound for the Kullback Leibler distance is well known [4].

Relation (iii) does not involve H_n at all and is elementary. It is reduced to $e_{n\xi} > d_n^2 + \frac{1}{2}e_{n\mu}$, equivalent to $z(1-y)+y(1-z) > (y-z)^2 + y(1-y)$, equivalent to $z(1-z) > 0$, which is obviously true.

The second inequality of (iv) is trivial and the first is proved similarly to (i). Again we start with a linear inequality $-E_{n\xi} < (A-1)E_{n\mu} + (B-1)H_n$, which is further reduced to $-e_{k\xi} < (A-1)e_{k\mu} + (B-1)h_k$. Inserting the abbreviations (17) we get

$$-y(1-z) - z(1-y) \;<\; (A-1)2y(1-y) + (B-1)\left[y\ln\frac{y}{z} + (1-y)\ln\frac{1-y}{1-z}\right]. \tag{20}$$

In Appendix B this inequality is shown to hold for $2AB \geq 1$, when $B > 1$. If we insert $B = 1/2A$ and minimize w.r.t. A, the minimum is again at $A = \sqrt{H_n/2E_{n\mu}}$ leading to the upper bound $-E_{n\xi} \leq -E_{n\mu} - H_n + \sqrt{2E_{n\mu}H_n}$ restricted to $E_{n\mu} > 2H_n$, which completes the proof of (iv).

Statement (v) is satisfied because $2y(1-y) \leq 2[y(1-r)+(1-y)r]$. Statement (vi) is a direct consequence of (i) and (v). This completes the proof of Theorem 1. $\square$

4 Deterministic Sequence Prediction

In the last section several relations were derived between the number of errors of the universal ξ-system, the informed μ-system and arbitrary ρ-systems. All of them were probabilistic predictors in the sense that given $x_{<n}$ they output 0 or 1 with certain probabilities. In this section, we are interested in systems whose output on input $x_{<n}$ is deterministically 0 or 1. Again we can distinguish between the case where the true distribution μ is known or unknown. In the probabilistic scheme we studied the μ and the ξ system. Given any probabilistic predictor ρ it is easy to construct a deterministic predictor Θ_ρ from it in the following way: If the probability of predicting 0 is larger than $\frac{1}{2}$, the deterministic predictor always chooses 0. Analogously for $0 \leftrightarrow 1$. We define[4]

$$\Theta_\rho(x_{<n}\underline{x}_n) \;:=\; \Theta(\rho(x_{<n}\underline{x}_n) - \tfrac{1}{2}) \;:=\; \left\{\begin{array}{lll} 0 & \text{for} & \rho(x_{<n}\underline{x}_n) < \frac{1}{2} \\ 1 & \text{for} & \rho(x_{<n}\underline{x}_n) > \frac{1}{2}. \end{array}\right.$$

Note that every deterministic predictor can be written in the form Θ_ρ for some ρ and that although $\Theta_\rho(\underline{x_1...x_n})$, defined via Bayes' rule (2), takes only values in $\{0,1\}$, it may still

[4] All results will be independent of the choice for $\rho = \frac{1}{2}$, so one might choose 0 for definiteness.

be interpreted as a probability measure. Deterministic prediction is just a special case of probabilistic prediction. The two models Θ_μ and Θ_ξ will be studied now.

Analogously to the last section we draw binary strings randomly with distribution μ and define the probability that the Θ_ρ system makes an erroneous prediction in the n^{th} step and the total μ-expected number of errors in the first n predictions as

$$\begin{aligned} e_{n\Theta_\rho}(x_{<n}) &:= \sum_{x_n} \mu(x_{<n}\underline{x}_n)[1-\Theta_\rho(x_{<n}\underline{x}_n)] \\ E_{n\Theta_\rho} &:= \sum_{k=1}^{n}\sum_{x_{<k}} \mu(\underline{x}_{<k})\cdot e_{k\Theta_\rho}(x_{<k}). \end{aligned} \tag{21}$$

The definitions (12) and (13) of h_n and H_n remain unchanged (ξ is *not* replaced by Θ_ξ).

The following relations will be derived:

THEOREM 2. *Let there be binary sequences drawn with probability $\mu_n(\underline{x}_{1:n})$ for the first n bits. A ρ-system predicts by definition x_n from $x_{<n}$ with probability $\rho(x_{<n}\underline{x}_n)$. A deterministic system Θ_ρ always predicts 1 if $\rho(x_{<n}\underline{x}_n)>\frac{1}{2}$ and 0 otherwise. If $e_{n\rho}(x_{<n})$ is the error probability in the n^{th} prediction, $E_{n\rho}$ the total μ-expected number of errors in the first n predictions (9), the following relations hold:*

$$\begin{aligned} &i) \quad 0 \leq E_{n\Theta_\xi} - E_{n\Theta_\mu} = \textstyle\sum_{x_k}\mu(\underline{x}_{<k})|e_{n\Theta_\xi}-e_{n\Theta_\mu}| < H_n + \sqrt{4E_{n\Theta_\mu}H_n + H_n^2} \\ &ii) \quad E_{n\Theta_\mu} \leq E_{n\rho} \quad , \quad e_{n\Theta_\mu} \leq e_{n\rho} \qquad \textit{for any } \rho \\ &iii) \quad E_{n\Theta_\xi} < E_{n\rho} + H_n + \sqrt{4E_{n\rho}H_n + H_n^2} \qquad \textit{for any } \rho, \end{aligned}$$

where $H_n < \ln 2\cdot K(\mu)+O(1)$ is the relative entropy (13), which is finite for computable μ.

No other useful bounds have been found, especially no bounds for the analogue of Δ_n.

COROLLARY 2. *For computable μ, i.e. for $K(\mu)<\infty$, the following statements immediately follow from Theorem 2:*

$$\begin{aligned} &vii) \quad \textit{if } E_{\infty\Theta_\mu} \textit{ is finite, then } E_{\infty\Theta_\xi} \textit{ is finite} \\ &viii) \quad E_{n\Theta_\xi}/E_{n\Theta_\mu} = 1 + O(E_{n\Theta_\mu}^{-1/2}) \longrightarrow 1 \quad \textit{for} \quad E_{n\Theta_\mu}\to\infty \\ &ix) \quad E_{n\Theta_\xi} - E_{n\Theta_\mu} = O(\sqrt{E_{n\Theta_\mu}}) \\ &x) \quad E_{n\Theta_\xi}/E_{n\rho} \leq 1 + O(E_{n\rho}^{-1/2}). \end{aligned}$$

Most of what we said in the probabilistic case remains valid here, as the Theorems and Corollaries 1 and 2 parallel each other. For this reason we will only highlight the differences.

The last inequality of (i) is the central new result in the deterministic case. Again, it is illustrated in the corollary, which follows trivially from Theorem 2.

From (ii) we see that Θ_μ is the best prediction scheme possible, compared to any other probabilistic or deterministic prediction ρ. The error expectation $e_{n\Theta_\mu}$ is smaller in every single step and hence, the total number of errors are also. This itself is not surprising and nearly obvious, as the Θ_μ system always predicts the bit of highest probability. So, for known μ, the Θ_μ system should always be preferred to any other prediction scheme, even to the informed μ prediction system.

Combining (i) and (ii) leads to a bound (iii) on the number of prediction errors of the deterministic variant of Solomonoff prediction. For computable μ, no prediction scheme can have fewer errors than that of the Θ_ξ system, whatever we take for ρ, apart from some additive correction of order $\sqrt{E_{n\Theta_\mu}}$. No factor 2 occurs as in the probabilistic case. Together with the quick convergence $E_{n\rho}^{-1/2}$ stated in (x), the Θ_ξ model should be sufficiently good in many applications.

Example. Let us consider a critical example. We want to predict the outcome of a die colored black (=0) and white (=1). Two faces should be white and the other 4 should be black. The game becomes more interesting by having a second complementary die with two black and four white sides. The dealer who throws the dice uses one or the other die according to some deterministic rule. The stake s is \$3 in every round; our return r is \$5 for every correct prediction.

The coloring of the dice and the selection strategy of the dealer unambiguously determine μ. $\mu(x_{<n}\underline{0})$ is $\frac{2}{3}$ for die 1 or $\frac{1}{3}$ for die 2. If we use ρ for prediction, we will have made $E_{n\rho}$ incorrect and $n-E_{n\rho}$ correct predictions in the first n rounds. The expected profit will be

$$P_{n\rho} \;:=\; (n-E_{n\rho})r - ns \;=\; (2n-5E_{n\rho})\$. \tag{22}$$

The winning threshold $P_{n\rho}>0$ is reached if $E_{n\rho}/n<1-s/r=\frac{2}{5}$.

If we knew μ, we could use the best possible prediction scheme Θ_μ. The error (21) and profit (22) expectations per round in this case are

$$e_{\Theta_\mu} \;:=\; e_{n\Theta_\mu}(x_{<n}) \;=\; \frac{1}{3} \;=\; \frac{E_{n\Theta_\mu}}{n} \;<\; \frac{2}{5} \quad , \quad \frac{P_{n\Theta_\mu}}{n} \;=\; \frac{1}{3}\$ > 0 \tag{23}$$

so we can make money from this game. If we predict according to the probabilistic μ prediction scheme (8) we would lose money in the long run:

$$e_{n\mu}(x_{<n}) \;=\; 2\cdot\frac{1}{3}\cdot\frac{2}{3} \;=\; \frac{4}{9} \;=\; \frac{E_{n\mu}}{n} \;>\; \frac{2}{5} \quad , \quad \frac{P_{n\mu}}{n} \;=\; -\frac{2}{9}\$ < 0$$

In the more interesting case where we do not know μ we can use Solomonoff prediction ξ or its deterministic variant Θ_ξ. From $(viii)$ of Corollaries 1 and 2 we know that

$$P_{n\xi}/P_{n\mu} \;=\; 1+O(n^{-1/2}) \;=\; P_{n\Theta_\xi}/P_{n\Theta_\mu},$$

so asymptotically the ξ system provides the same profit as the μ system and the Θ_ξ system the same as the Θ_μ system. Using the ξ system is a losing strategy, while using the Θ_ξ

system is a winning strategy. Let us estimate the number of rounds we have to play before reaching the winning zone with the Θ_ξ system. $P_{n\Theta_\xi}>0$ if $E_{n\Theta_\xi}<(1-s/r)n$ if

$$E_{n\Theta_\mu}+H_n+\sqrt{4E_{n\Theta_\mu}H_n+H_n^2} \;<\; (1-s/r)\cdot n$$

by Theorem 2 (i). Solving w.r.t. H_n we get

$$H_n \;<\; \frac{(1-s/r-E_{n\Theta_\mu}/n)^2}{2\cdot(1-s/r+E_{n\Theta_\mu}/n)}\cdot n.$$

Using $H_n<\ln 2\cdot K(\mu)+O(1)$ and (23) we expect to be in the winning zone for

$$n \;>\; \frac{2\cdot(1-s/r+e_{\Theta_\mu})}{(1-s/r-e_{\Theta_\mu})^2}\cdot\ln 2\cdot K(\mu)+O(1) \;=\; 330\ln 2\cdot K(\mu)+O(1).$$

If the die selection strategy reflected in μ is not too complicated, the Θ_ξ prediction system reaches the winning zone after a few thousand rounds. The number of rounds is not really small because the expected profit per round is one order of magnitude smaller than the return. This leads to a constant of two orders of magnitude size in front of $K(\mu)$. Stated otherwise, it is due to the large stochastic noise, which makes it difficult to extract the signal, i.e. the structure of the rule μ. Furthermore, this is only a bound for the turnaround value of n. The true expected turnaround n might be smaller.

However, every game for which there exists a winning strategy ρ with $P_{n\rho}\sim n$, Θ_ξ is guaranteed to get into the winning zone for some $n\sim K(\mu)$, i.e. $P_{n\Theta_\xi}>0$ for sufficiently large n. This is *not* guaranteed for the ξ-system, due to the factor 2 in the bound (x) of Corollary 1.

Proof of Theorem 2. The method of proof is the same as in the previous section, so we will keep it short. With the abbreviations (16) we can write $e_{k\Theta_\xi}$ and $e_{k\Theta_\mu}$ in the forms

$$\begin{aligned} e_{k\Theta_\xi} &= y(1-\Theta(z-\tfrac12))+(1-y)\Theta(z-\tfrac12) \;=\; |y-\Theta(z-\tfrac12)| \\ e_{k\Theta_\mu} &= y(1-\Theta(y-\tfrac12))+(1-y)\Theta(y-\tfrac12) \;=\; \min\{y,1-y\}. \end{aligned} \tag{24}$$

With these abbreviations, (ii) is equivalent to $\min\{y,1-y\}\leq y(1-r)+(1-y)r$, which is true, because the minimum of two numbers is always smaller than their weighted average.

The first inequality and equality of (i) follow directly from (ii). To prove the last inequality, we start once again with a linear model

$$E_{n\Theta_\xi} \;<\; (A+1)E_{n\Theta_\mu}+(B+1)H_n. \tag{25}$$

Inserting the definition of E_n and H_n, using (24), and omitting the sums we have to find A and B, which satisfy

$$|y-\Theta(z-\tfrac12)| \;<\; (A+1)\min\{y,1-y\}+(B+1)\left[y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z}\right]. \tag{26}$$

In Appendix C we will show that the inequality is satisfied for $B\geq\frac14 A+\frac1A$ and $A>0$. Inserting $B=\frac14 A+\frac1A$ into (25) and minimizing the r.h.s. w.r.t. A, we get the upper bound

$$E_{n\Theta_\xi} \;<\; E_{n\Theta_\mu}+H_n+\sqrt{4E_{n\mu}H_n+H_n^2} \qquad\text{for}\qquad A^2=\frac{H_n}{E_{n\Theta_\mu}+\frac14 H_n}.$$

Statement (iii) is a direct consequence of (i) and (ii). This completes the proof of Theorem 2. □

5 Conclusions

We have proved several new error bounds for Solomonoff prediction in terms of informed prediction and in terms of general prediction schemes. Theorem 1 and Corollary 1 summarize the results in the probabilistic case and Theorem 2 and Corollary 2 for the deterministic case. We have shown that in the probabilistic case $E_{n\xi}$ is asymptotically bounded by twice the number of errors of any other prediction scheme. In the deterministic variant of Solomonoff prediction this factor 2 is absent. It is well suited, even for difficult prediction problems, as the error probability E_{Θ_ξ}/n converges rapidly to that of the minimal possible error probability E_{Θ_μ}/n.

Acknowledgments: I thank Ray Solomonoff and Jürgen Schmidhuber for proofreading this work and for numerous discussions.

A Proof of Inequality (18)

[5]With the definition

$$f(y,z;A,B) \;:=\; A\cdot 2y(1-y)+B\cdot\left[y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z}\right]-|y-z|$$

we have to show $f(y,z;A,B)>0$ for $0<y<1$, $0<z<1$ and suitable A and B. We do this by showing that $f>0$ at all extremal values, 'at' boundaries and at non-analytical points. $f\to+\infty$ for $z\to 0/1$, if we choose $B>0$. Moreover, at the non-analytic point $z=y$ we have $f(y,y;A,B)=2Ay(1-y)\geq 0$ for $A\geq 0$. The extremal condition $\partial f/\partial z=0$ for $z\neq y$ (keeping y fixed) leads to

$$y \;=\; y^* \;:=\; z\cdot[1-\frac{s}{B}(1-z)], \quad s \;:=\; \mathrm{sign}(z-y) \;=\; \pm 1.$$

Inserting y^* into the definition of f and omitting the positive term $B[\ldots]$, we get

$$f(y^*,z;A,B) \;>\; 2Ay^*(1-y^*)-|z-y^*| \;=\; \tfrac{1}{B^2}z(1-z)\cdot g(z;A,B)$$

$$g(z;A,B) \;:=\; 2A(B-s(1-z))(B+sz)-sB\cdot$$

We have reduced the problem to showing $g\geq 0$. Since $s=\pm 1$, we have $g(z;A,B)>2A(B-1+z)(B-z)-B$ for $B>1$. The latter is quadratic in z and symmetric in $z\leftrightarrow 1-z$ with a maximum at $\frac{1}{2}$. Thus it is sufficient to check the boundary values $g(0;A,B)=g(1;A,B)=2A(B-1)B-B$. They are non-negative for $2A(B-1)\geq 1$. Putting everything together, we have proved that $f>0$ for $B\geq\frac{1}{2A}+1$ and $A>0$. □

[5]The proofs are a bit sketchy. We will be a little sloppy about boundary values $y=0/1$, $z=\frac{1}{2}$, $\Theta(0)$, $\geq$ versus $>$, and *approaching* versus *at* the boundary. All subtleties have been checked and do not spoil the results. As $0<\xi<1$, therefore $0<z<1$ is strict.

B Proof of Inequality (20)

The proof of this inequality is similar to the previous one. With the definition

$$f(y,z;A,B) \;:=\; (A-1)2y(1-y)+(B-1)\left[y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z}\right]+y(1-z)+z(1-y)$$

we have to show $f(y,z;A,B)>0$ for $0<y<1$, $0<z<1$ and suitable A and B. Again, we do this by showing that $f>0$ at all extremal values and 'at' the boundary. $f\to+\infty$ for $z\to 0,1$, if we choose $B>1$. The extremal condition $\partial f/\partial z=0$ (keeping y fixed) leads to

$$y \;=\; y^* \;:=\; z\cdot\frac{z-B}{1-B-2z(1-z)}, \quad 0<y^*<1.$$

Inserting y^* into the definition of f and omitting the positive term $(B-1)[\ldots]$, we get

$$f(y^*,z;A,B) \;>\; 2Ay^*(1-y^*)-(2y^*-1)(z-y^*) \;=\; \tfrac{z(1-z)}{[1-B-2z(1-z)]^2}\cdot g(z;A,B)$$

$$g(z;A,B) \;:=\; 2A(z-B)(1-z-B)-(B-1)(2z-1)^2.$$

We have reduced the problem to showing $g\geq 0$. This is easy, since g is quadratic in z and symmetric in $z\leftrightarrow 1-z$. The extremal value $g(\frac{1}{2};A,B)=2A(B-\frac{1}{2})^2$ is positive for $A>0$. The boundary values $g(0;A,B)=g(1;A,B)=(2AB-1)(B-1)$ are ≥ 0 for $2AB\geq 1$. Putting everything together, we have proved that $f>0$ for $2AB\geq 1$ and $B>1$. □

C Proof of Inequality (26)

We want to show that

$$|y-\Theta(z-\tfrac{1}{2})| \;<\; (A+1)\min\{y,1-y\}+(B+1)\left[y\ln\tfrac{y}{z}+(1-y)\ln\tfrac{1-y}{1-z}\right]$$

The formula is symmetric w.r.t. $y\leftrightarrow 1-y$ and $z\leftrightarrow 1-z$ simultaneously, so we can restrict ourselves to $0<y<1$ and $0<z<\frac{1}{2}$. Furthermore, let $B>-1$. Using (19), it is enough to prove

$$f(y,z;A,B) \;:=\; (A+1)\min\{y,1-y\}+(B+1)2(y-z)^2-y \;>\; 0$$

f is quadratic in z; thus for $y<\frac{1}{2}$ it takes its minimum at $z=y$. Since $f(y,y;A,B)=Ay>0$ for $A>0$, we can concentrate on the case $y\geq\frac{1}{2}$. In this case, the minimum is reached at the boundary $z=\frac{1}{2}$.

$$f(y,\tfrac{1}{2};A,B) \;=\; (A+1)(1-y)+(B+1)2(y-\tfrac{1}{2})^2-y$$

This is now quadratic in y with minimum at

$$y^* \;=\; \frac{A+2B+4}{4(B+1)}, \quad f(y^*,\tfrac{1}{2};A,B) \;=\; \frac{4AB-A^2-4}{8(B+1)} \;\geq\; 0$$

for $B\geq\frac{1}{4}A+\frac{1}{A}$, $A>0$, $(\Rightarrow B\geq 1)$. □

References

[1] D. Angluin, C. H. Smith: *Inductive inference: Theory and methods*; Comput. Surveys, 15:3, (1983), 237–269 .

[2] T. Bayes: *An essay towards solving a problem in the doctrine of chances*; Philos. Trans. Roy. Soc., 53 (1763) 376–398.

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

[4] S. Kullback: *Information Theory and Statistics*; Wiley, New York (1959).

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

[6] M. Li and P.M.B. Vitányi: *Inductive reasoning and Kolmogorov complexity*; J. Comput. System Sci., 44:2 (1992), 343–384.

[7] M. Li and P.M.B. Vitányi: *An Introduction to Kolmogorov Complexity and its Applications*; Springer-Verlag, New York, 2nd Edition, 1997.

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

[9] R.J Solomonoff: *Complexity-based induction systems: comparisons and convergence theorems*; IEEE Trans. Inform. Theory, IT-24:4, (1978), 422–432.

[10] R.J. Solomonoff: *The discovery of algorithmic probability*; J. Comput. System Sci. 55 (1997), 73–88.

[11] D.G. Willis: *Computational complexity and probability constructions*; J. Assoc. Comput. Mach., 4 (1970), 241–259.

[12] 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 Math. Surveys, 25:6 (1970), 83–124.