Convergence and Error Bounds for Universal Prediction of Nonbinary Sequences

Marcus Hutter

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

Abstract. Solomonoff's uncomputable universal prediction scheme ξ allows to predict the next symbol x_k of a sequence $x_1...x_{k-1}$ for any Turing computable, but otherwise unknown, probabilistic environment μ. This scheme will be generalized to arbitrary environmental classes, which, among others, allows the construction of computable universal prediction schemes ξ. Convergence of ξ to μ in a conditional mean squared sense and with μ probability 1 is proven. It is shown that the average number of prediction errors made by the universal ξ scheme rapidly converges to those made by the best possible informed μ scheme. The schemes, theorems and proofs are given for general finite alphabet, which results in additional complications as compared to the binary case. Several extensions of the presented theory and results are outlined. They include general loss functions and bounds, games of chance, infinite alphabet, partial and delayed prediction, classification, and more active systems.

1 Introduction

The Bayesian framework is ideally suited for studying induction problems. The probability of observing x_k at time k, given past observations $x_1...x_{k-1}$, can be computed with Bayes' rule if the generating probability distribution μ, from which sequences $x_1x_2x_3...$ are drawn, is known. The problem, however, is that in many cases one does not even have a reasonable estimate of the true generating distribution. What is the true probability of weather sequences or stock charts? In order to overcome this problem we define a universal distribution ξ as a weighted sum of distributions $\mu_i \in M$, where M is any finite or countable set of distributions including μ. This is a generalization of Solomonoff induction, in which M is the set of all enumerable semi-measures [Sol64,Sol78]. We show that using the universal ξ as a prior is nearly as good as using the unknown generating distribution μ. In a sense, this solves the problem, that the generating distribution μ is not known, in a universal way. All results are obtained for general finite alphabet. Convergence of ξ to μ in a conditional mean squared sense and with μ probability 1 is proven. The number of errors E_{Θ_ξ} made by the universal prediction scheme Θ_ξ based on ξ minus the number of errors E_{Θ_μ} of

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

L. De Raedt and P. Flach (Eds.): ECML 2001, LNAI 2167, pp. 239–250, 2001.

the optimal informed prediction scheme Θ_μ based on μ is proven to be bounded by $O(\sqrt{E_{\Theta_\mu}})$.

Extensions to arbitrary loss functions, games of chance, infinite alphabet, partial and delayed prediction, classification, and more active systems are discussed (Section 5). The main new results are a generalization of the universal probability ξ [Sol64] to arbitrary probability classes and weights (Section 2), a generalization of the convergence [Sol78] $\xi \to \mu$ (Section 3) and the error bounds [Hut99] to arbitrary alphabet (Section 4). The non-binary setting causes substantial additional complications. Non-binary prediction cannot be (easily) reduced to the binary case. One may have in mind a binary coding of the symbols x_k in the sequence $x_1x_2...$. But this makes it necessary to predict a block of bits x_k, before receiving the true block of bits x_k, which differs from the bit-by-bit prediction considered in [Sol78,LV97,Hut99].

For an excellent introduction to Kolmogorov complexity and Solomonoff induction one should consult the book of Li and Vitányi [LV97] or the article [LV92] for a short course. Historical surveys of inductive reasoning and inference can be found in [AS83,Sol97].

2 Setup

2.1 Strings and Probability Distributions

We denote strings over a finite alphabet $\mathcal{A}$ by $x_1x_2...x_n$ with $x_k \in \mathcal{A}$. We further use the abbreviations $x_{n:m} := x_nx_{n+1}...x_{m-1}x_m$ and $x_{<n} := x_1...x_{n-1}$. We use Greek letters for probability distributions and underline their arguments to indicate that they are probability arguments. Let $\rho(\underline{x_1...x_n})$ be the probability that an (infinite) sequence starts with $x_1...x_n$:

$$\sum_{x_{1:n}\in\mathcal{A}^n} \rho(\underline{x}_{1:n}) = 1, \quad \sum_{x_n\in\mathcal{A}} \rho(\underline{x}_{1:n}) = \rho(\underline{x}_{<n}), \quad \rho(\epsilon) = 1. \tag{1}$$

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:

$$\rho(x_{<n}\underline{x}_n) = \rho(\underline{x}_{1:n})/\rho(\underline{x}_{<n}) \quad , \quad \rho(\underline{x_1...x_n}) = \rho(\underline{x}_1)\cdot\rho(x_1\underline{x}_2)\cdot...\cdot\rho(x_1...x_{n-1}\underline{x}_n). \tag{2}$$

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.

2.2 Universal Prior Probability Distribution

Most inductive inference problem can be brought into the following form: Given a string $x_{<k}$, take a guess at its continuation x_k. We will assume that the strings which have to be continued are drawn from a probability[1] distribution μ. The maximal prior information a prediction algorithm can possess is the exact knowledge of μ, but in many cases the generating distribution is not known. Instead, the prediction is based on a guess ρ of μ. We expect that a predictor based on ρ performs well, if ρ is close to μ or converges, in a sense, to μ. Let $M := \{\mu_1, \mu_2, ...\}$ be a finite or countable set of candidate probability distributions on strings. We define a weighted average on M

$$\xi(\underline{x}_{1:n}) \;:=\; \sum_{\mu_i \in M} w_{\mu_i} \cdot \mu_i(\underline{x}_{1:n}), \qquad \sum_{\mu_i \in M} w_{\mu_i} = 1, \quad w_{\mu_i} > 0. \tag{3}$$

It is easy to see that ξ is a probability distribution as the weights w_{μ_i} are positive and normalized to 1 and the $\mu_i \in M$ are probabilities. For finite M a possible choice for the w is to give all μ_i equal weight ($w_{\mu_i} = \frac{1}{|M|}$). We call ξ universal relative to M, as it multiplicatively dominates all distributions in M

$$\xi(\underline{x}_{1:n}) \;\geq\; w_{\mu_i} \cdot \mu_i(\underline{x}_{1:n}) \quad \text{for all} \quad \mu_i \in M. \tag{4}$$

In the following, we assume that M is known and contains[2] the true generating distribution, i.e. $\mu \in M$. We will see that this is not a serious constraint as we can always chose M to be sufficiently large. In the next section we show the important property of ξ converging to the generating distribution μ in a sense and, hence, might being a useful substitute for the true generating, but in general, unknown distribution μ.

2.3 Probability Classes

We get a rather wide class M if we include *all* computable probability distributions in M. In this case, the assumption $\mu \in M$ is very weak, as it only assumes that the strings are drawn from *any computable* distribution; and all valid physical theories (and, hence, all environments) *are* computable (in a probabilistic sense).

We will see that it is favorable to assign high weights w_{μ_i} to the μ_i. Simplicity should be favored over complexity, according to Occam's razor. In our context this means that a high weight should be assigned to simple μ_i. The prefix Kolmogorov complexity $K(\mu_i)$ is a universal complexity measure [Kol65,

[1] This includes deterministic environments, in which case the probability distribution μ is 1 for some sequence $x_{1:\infty}$ and 0 for all others. We call probability distributions of this kind *deterministic*.

[2] Actually all theorems remain valid for μ being a finite linear combination of $\mu_i \in L \subseteq M$ and $w_\mu := \min_{\mu_i \in L} w_{\mu_i}$ [Hut01].

ZL70,LV97]. It is defined as the length of the shortest self-delimiting program (on a universal Turing machine) computing $\mu_i(x_{1:n})$ given $x_{1:n}$. If we define

$$w_{\mu_i} := \frac{1}{\Omega} 2^{-K(\mu_i)} \quad , \quad \Omega := \sum_{\mu_i \in M} 2^{-K(\mu_i)}$$

then, distributions which can be calculated by short programs, have high weights. Besides ensuring correct normalization, Ω (sometimes called the number of wisdom) has interesting properties in itself [Cal98,Cha91]. If we enlarge M to include all enumerable semi-measures, we attain Solomonoff's universal probability, apart from normalization, which has to be treated differently in this case [Sol64,Sol78]. Recently, M has been further enlarged to include all cumulatively enumerable semi-measures [Sch00]. In all cases, ξ is not finitely computable, but can still be approximated to arbitrary but not pre-specifiable precision. If we consider *all* approximable (i.e. asymptotically computable) distributions, then the universal distribution ξ, although still well defined, is not even approximable [Sch00]. An interesting and quickly approximable distribution is the Speed prior S defined in [Sch00]. It is related to Levin complexity and Levin search [Lev73, Lev84], but it is unclear for now which distributions are dominated by S. If one considers only finite-state automata instead of general Turing machines, one can attain a quickly computable, universal finite-state prediction scheme related to that of Feder et al. [FMG92], which itself is related to the famous Lempel-Ziv data compression algorithm. If one has extra knowledge on the source generating the sequence, one might further reduce M and increase w. A detailed analysis of these and other specific classes M will be given elsewhere. Note that $\xi \in M$ in the enumerable and cumulatively enumerable case, but $\xi \notin M$ in the computable, approximable and finite-state case. If ξ is itself in M, it is called a universal element of M [LV97]. As we do not need this property here, M may be *any* finite or countable set of distributions. In the following we consider generic M and w.

3 Convergence

3.1 Upper Bound for the Relative Entropy

Let us define the relative entropy (also called Kullback Leibler divergence [Kul59]) between μ and ξ:

$$h_k(x_{<k}) := \sum_{x_k \in \mathcal{A}} \mu(x_{<k}\underline{x}_k) \ln \frac{\mu(x_{<k}\underline{x}_k)}{\xi(x_{<k}\underline{x}_k)}. \tag{5}$$

H_n is then defined as the sum-expectation, for which the following upper bound can be shown

$$H_n := \sum_{k=1}^{n} \sum_{x_{<k} \in \mathcal{A}^{k-1}} \mu(\underline{x}_{<k}) \cdot h_k(x_{<k}) = \sum_{k=1}^{n} \sum_{x_{1:k} \in \mathcal{A}^k} \mu(\underline{x}_{1:k}) \ln \frac{\mu(x_{<k}\underline{x}_k)}{\xi(x_{<k}\underline{x}_k)} = \tag{6}$$

$$= \sum_{x_{1:n}} \mu(\underline{x}_{1:n}) \ln \prod_{k=1}^{n} \frac{\mu(x_{<k}\underline{x}_k)}{\xi(x_{<k}\underline{x}_k)} = \sum_{x_{1:n}} \mu(\underline{x}_{1:n}) \ln \frac{\mu(\underline{x}_{1:n})}{\xi(\underline{x}_{1:n})} \leq \ln \frac{1}{w_\mu} =: d_\mu$$

In the first line we have inserted (5) 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 further 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. Using universality (4) of ξ, i.e. $\ln \mu(\underline{x}_{1:n})/\xi(\underline{x}_{1:n}) \leq \ln \frac{1}{w_\mu}$ for $\mu \in M$ yields the final inequality in (6). The proof given here is simplified version of those given in [Sol78] and [LV97].

3.2 Lower Bound for the Relative Entropy

We need the following inequality to lower bound H_n

$$\sum_{i=1}^{N} (y_i - z_i)^2 \leq \sum_{i=1}^{N} y_i \ln \frac{y_i}{z_i} \quad \text{for} \quad y_i \geq 0, \quad z_i \geq 0, \quad \sum_{i=1}^{N} y_i = 1 = \sum_{i=1}^{N} z_i. \quad (7)$$

The proof of the case $N = 2$

$$2(y-z)^2 \leq y \ln \frac{y}{z} + (1-y) \ln \frac{1-y}{1-z}, \quad 0 < y < 1, \quad 0 < z < 1 \quad (8)$$

will not be repeated here, as it is elementary and well known [LV97]. The proof of (7) is one point where the generalization from binary to arbitrary alphabet is not trivial.[3] We will reduce the general case $N > 2$ to the case $N = 2$. We do this by a partition $\{1, ..., N\} = G^+ \cup G^-$, $G^+ \cap G^- = \{\}$, and define $y^\pm := \sum_{i \in G^\pm} y_i$ and $z^\pm := \sum_{i \in G^\pm} z_i$. It is well known that the relative entropy is positive, i.e.

$$\sum_{i \in G^\pm} p_i \ln \frac{p_i}{q_i} \geq 0 \quad \text{for} \quad p_i \geq 0, \quad q_i \geq 0, \quad \sum_{i \in G^\pm} p_i = 1 = \sum_{i \in G^\pm} q_i. \quad (9)$$

Note that there are 4 probability distributions (p_i and q_i for $i \in G^+$ and $i \in G^-$). For $i \in G^\pm$, $p_i := y_i/y^\pm$ and $q_i := z_i/z^\pm$ satisfy the conditions on p and q. Inserting this into (9) and rearranging the terms we get $\sum_{i \in G^\pm} y_i \ln \frac{y_i}{z_i} \geq y^\pm \ln \frac{y^\pm}{z^\pm}$. If we sum this over $\pm$ and define $y \equiv y^+ = 1 - y^-$ and $z \equiv z^+ = 1 - z^-$ we get

$$\sum_{i=1}^{N} y_i \ln \frac{y_i}{z_i} \geq \sum_{\pm} y^\pm \ln \frac{y^\pm}{z^\pm} = y \ln \frac{y}{z} + (1-y) \ln \frac{1-y}{1-z}. \quad (10)$$

[3] We will not explicate every subtlety and only sketch the proofs. Subtleties regarding $y, z = 0/1$ have been checked but will be passed over. $0 \ln \frac{0}{z_i} := 0$ even for $z_i = 0$. Positive means ≥ 0.

For the special choice $G^{\pm} := \{i : y_i \gtrless z_i\}$, we can upper bound the quadratic term by

$$\sum_{i\in G^{\pm}} (y_i - z_i)^2 \;\leq\; \Big(\sum_{i\in G^{\pm}} |y_i - z_i|\Big)^2 \;=\; \Big(\sum_{i\in G^{\pm}} y_i - z_i\Big)^2 \;=\; (y^{\pm} - z^{\pm})^2.$$

The first equality is true, since all $y_i - z_i$ are positive/negative for $i \in G^{\pm}$ due to the special choice of $G^{\pm}$. Summation over $\pm$ gives

$$\sum_{i=1}^{N} (y_i - z_i)^2 \;\leq\; \sum_{\pm} (y^{\pm} - z^{\pm})^2 \;=\; 2(y-z)^2 \tag{11}$$

Chaining the inequalities (11), (8) and (10) proves (7). If we identify

$$\mathcal{A} = \{1, ..., N\}, \quad N = |\mathcal{A}|, \quad i = x_k, \quad y_i = \mu(x_{<k}\underline{x}_k), \quad z_i = \xi(x_{<k}\underline{x}_k) \tag{12}$$

multiply both sides of (7) with $\mu(\underline{x}_{<k})$ and take the sum over $x_{<k}$ and k we get

$$\sum_{k=1}^{n} \sum_{x_{1:k}} \mu(\underline{x}_{<k}) \Big(\mu(x_{<k}\underline{x}_k) - \xi(x_{<k}\underline{x}_k)\Big)^2 \;\leq\; \sum_{k=1}^{n} \sum_{x_{1:k}} \mu(\underline{x}_{1:k}) \ln \frac{\mu(x_{<k}\underline{x}_k)}{\xi(x_{<k}\underline{x}_k)}. \tag{13}$$

3.3 Convergence of ξ to μ

The upper (6) and lower (13) bounds on H_n allow us to prove the convergence of ξ to μ in a conditional mean squared sense and with μ probability 1.

Theorem 1 (Convergence). *Let there be sequences $x_1 x_2 ...$ over a finite alphabet $\mathcal{A}$ drawn with probability $\mu(\underline{x}_{1:n})$ for the first n symbols. The universal conditional probability $\xi(x_{<k}\underline{x}_k)$ of the next symbol x_k given $x_{<k}$ is related to the generating conditional probability $\mu(x_{<k}\underline{x}_k)$ in the following way:*

i) $\displaystyle\sum_{k=1}^{n} \sum_{x_{1:k}} \mu(\underline{x}_{<k}) \Big(\mu(x_{<k}\underline{x}_k) - \xi(x_{<k}\underline{x}_k)\Big)^2 \;\leq\; H_n \;\leq\; d_\mu \;=\; \ln \frac{1}{w_\mu} \;<\; \infty$

ii) $\xi(x_{<k}\underline{x}_k) \to \mu(x_{<k}\underline{x}_k)$ *for* $k \to \infty$ *with* μ *probability 1*

where H_n is the relative entropy (6), and w_μ is the weight (3) of μ in ξ.

(*i*) follows from (6) and (13). For $n \to \infty$ the l.h.s. of (*i*) is an infinite k-sum over positive arguments, which is bounded by the finite constant d_μ on the r.h.s. Hence the arguments must converge to zero for $k \to \infty$. Since the arguments are μ expectations of the squared difference of ξ and μ, this means that $\xi(x_{<k}\underline{x}_k)$ converges to $\mu(x_{<k}\underline{x}_k)$ with μ probability 1 or, more stringent, in a mean square sense. This proves (*ii*). The reason for the astonishing property of a single (universal) function ξ to converge to *any* $\mu_i \in M$ lies in the fact that the sets of μ-random sequences differ for different μ. Since the conditional probabilities are the basis of all prediction algorithms considered in this work, we expect a good prediction performance if we use ξ as a guess of μ. Performance measures are defined in the following sections.

4 Error Bounds

We now consider the following measure for the quality of a prediction: making a wrong prediction counts as one error, making a correct prediction counts as no error.

4.1 Total Expected Numbers of Errors

Let Θ_μ be the optimal prediction scheme when the strings are drawn from the probability distribution μ, i.e. the probability of x_k given $x_{<k}$ is $\mu(x_{<k}\underline{x}_k)$, and μ is known. Θ_μ predicts (by definition) $x_k^{\Theta_\mu}$ when observing $x_{<k}$. The prediction is erroneous if the true kth symbol is not $x_k^{\Theta_\mu}$. The probability of this event is $1-\mu(x_{<k}\underline{x}_k^{\Theta_\mu})$. It is minimized if $x_k^{\Theta_\mu}$ maximizes $\mu(x_{<k}\underline{x}_k^{\Theta_\mu})$. More generally, let Θ_ρ be a prediction scheme predicting $x_k^{\Theta_\rho} := \text{maxarg}_{x_k}\rho(x_{<k}\underline{x}_k)$ for some distribution ρ. Every deterministic predictor can be interpreted as maximizing some distribution. The μ probability of making a wrong prediction for the k^{th} symbol and the total μ-expected number of errors in the first n predictions of predictor Θ_ρ are

$$e_{k\Theta_\rho}(x_{<k}) \;:=\; 1-\mu(x_{<k}\underline{x}_k^{\Theta_\rho}) \quad , \quad E_{n\Theta_\rho} \;:=\; \sum_{k=1}^{n}\sum_{x_1\ldots x_{k-1}} \mu(\underline{x}_{<k})\cdot e_{k\Theta_\rho}(x_{<k}). \tag{14}$$

If μ is known, Θ_μ is obviously the best prediction scheme in the sense of making the least number of expected errors

$$E_{n\Theta_\mu} \;\leq\; E_{n\Theta_\rho} \quad \text{for any} \quad \Theta_\rho, \tag{15}$$

since $e_{k\Theta_\mu}(x_{<k}) = 1-\mu(x_{<k}\underline{x}_k^{\Theta_\mu}) = \min_{x_k}(1-\mu(x_{<k}\underline{x}_k)) \leq 1-\mu(x_{<k}\underline{x}_k^{\Theta_\rho}) = e_{k\Theta_\rho}(x_{<k})$ for any ρ.

4.2 Error Bound

Of special interest is the universal predictor Θ_ξ. As ξ converges to μ the prediction of Θ_ξ might converge to the prediction of the optimal Θ_μ. Hence, Θ_ξ may not make many more errors than Θ_μ and, hence, any other predictor Θ_ρ. Note that $x_k^{\Theta_\rho}$ is a discontinuous function of ρ and $x_k^{\Theta_\xi} \to x_k^{\Theta_\mu}$ can not be proved from $\xi \to \mu$. Indeed, this problem occurs in related prediction schemes, where the predictor has to be regularized so that it is continuous [FMG92]. Fortunately this is not necessary here. We prove the following error bound.

Theorem 2 (Error bound). *Let there be sequences $x_1x_2...$ over a finite alphabet $\mathcal{A}$ drawn with probability $\mu(\underline{x}_{1:n})$ for the first n symbols. The Θ_ρ-system predicts by definition $x_n^{\Theta_\rho} \in \mathcal{A}$ from $x_{<n}$, where $x_n^{\Theta_\rho}$ maximizes $\rho(x_{<n}\underline{x}_n)$. Θ_ξ is the universal prediction scheme based on the universal prior ξ. Θ_μ is the optimal*

informed prediction scheme. The total μ-expected number of prediction errors $E_{n\Theta_\xi}$ and $E_{n\Theta_\mu}$ of Θ_ξ and Θ_μ as defined in (14) are bounded in the following way

$$0 \;\leq\; E_{n\Theta_\xi} - E_{n\Theta_\mu} \;\leq\; H_n + \sqrt{4E_{n\Theta_\mu}H_n + H_n^2} \;\leq\; 2H_n + 2\sqrt{E_{n\Theta_\mu}H_n}$$

where $H_n \leq \ln \frac{1}{w_\mu}$ is the relative entropy (6), and w_μ is the weight (3) of μ in ξ.

First, we observe that the number of errors $E_{\infty\Theta_\xi}$ of the universal Θ_ξ predictor is finite if the number of errors $E_{\infty\Theta_\mu}$ of the informed Θ_μ predictor is finite. This is especially the case for deterministic μ, as $E_{n\Theta_\mu} \equiv 0$ in this case[4], i.e. Θ_ξ makes only a finite number of errors on deterministic environments. More precisely, $E_{\infty\Theta_\xi} \leq 2H_\infty \leq 2\ln \frac{1}{w_\mu}$. A combinatoric argument shows that there are M and $\mu \in M$ with $E_{\infty\Theta_\xi} \geq \log_2 |M|$. This shows that the upper bound $E_{\infty\Theta_\xi} \leq 2\ln |M|$ for uniform w must be rather tight. For more complicated probabilistic environments, where even the ideal informed system makes an infinite number of errors, the theorem ensures that the error excess $E_{n\Theta_\xi} - E_{n\Theta_\mu}$ is only of order $\sqrt{E_{n\Theta_\mu}}$. The excess is quantified in terms of the information content H_n of μ (relative to ξ), or the weight w_μ of μ in ξ. This ensures that the error densities E_n/n of both systems converge to each other. Actually, the theorem ensures more, namely that the quotient converges to 1, and also gives the speed of convergence $E_{n\Theta_\xi}/E_{n\Theta_\mu} = 1 + O(E_{n\Theta_\mu}^{-1/2}) \longrightarrow 1$ for $E_{n\Theta_\mu} \to \infty$.

4.3 Proof of Theorem 2

The first inequality in Theorem 2 has already been proved (15). The last inequality is a simple triangle inequality. For the second inequality, let us start more modestly and try to find constants A and B that satisfy the linear inequality

$$E_{n\Theta_\xi} \;\leq\; (A+1)E_{n\Theta_\mu} + (B+1)H_n. \tag{16}$$

If we could show

$$e_{k\Theta_\xi}(x_{<k}) \;\leq\; (A+1)e_{k\Theta_\mu}(x_{<k}) + (B+1)h_k(x_{<k}) \tag{17}$$

for all $k \leq n$ and all $x_{<k}$, (16) would follow immediately by summation and the definition of E_n and H_n. With the abbreviations (12) and the abbreviations $m = x_k^{\Theta_\mu}$ and $s = x_k^{\Theta_\xi}$ the various error functions can then be expressed by $e_{k\Theta_\xi} = 1-y_s$, $e_{k\Theta_\mu} = 1-y_m$ and $h_k = \sum_i y_i \ln \frac{y_i}{z_i}$. Inserting this into (17) we get

$$1-y_s \;\leq\; (A+1)(1-y_m) + (B+1)\sum_{i=1}^{N} y_i \ln \frac{y_i}{z_i}. \tag{18}$$

[4] Remember that we named a probability distribution *deterministic* if it is 1 for exactly one sequence and 0 for all others.

By definition of $x_k^{\Theta_\mu}$ and $x_k^{\Theta_\xi}$ we have $y_m \geq y_i$ and $z_s \geq z_i$ for all i. We prove a sequence of inequalities which show that

$$(B+1)\sum_{i=1}^{N} y_i \ln\frac{y_i}{z_i} + (A+1)(1-y_m) - (1-y_s) \;\geq\; ... \tag{19}$$

is positive for suitable $A \geq 0$ and $B \geq 0$, which proves (18). For $m = s$ (19) is obviously positive since the relative entropy is positive ($h_k \geq 0$). So we will assume $m \neq s$ in the following. We replace the relative entropy by the sum over squares (7) and further keep only contributions from $i = m$ and $i = s$.

$$... \;\geq\; (B+1)[(y_m - z_m)^2 + (y_s - z_s)^2] + (A+1)(1-y_m) - (1-y_s) \;\geq\; ...$$

By definition of y, z, m and s we have the constraints $y_m + y_s \leq 1$, $z_m + z_s \leq 1$, $y_m \geq y_s \geq 0$ and $z_s \geq z_m \geq 0$. From the latter two it is easy to see that the square terms (as a function of z_m and z_s) are minimized by $z_m = z_s = \frac{1}{2}(y_m + y_s)$. Furthermore, we define $x := y_m - y_s$ and eliminate y_s.

$$... \;\geq\; (B+1)\tfrac{1}{2}x^2 + A(1-y_m) - x \;\geq\; ... \tag{20}$$

The constraint on $y_m + y_s \leq 1$ translates into $y_m \leq \frac{x+1}{2}$, hence (20) is minimized by $y_m = \frac{x+1}{2}$.

$$... \;\geq\; \tfrac{1}{2}[(B+1)x^2 - (A+2)x + A] \;\geq\; ... \tag{21}$$

(21) is quadratic in x and minimized by $x^* = \frac{A+2}{2(B+1)}$. Inserting x^* gives

$$... \;\geq\; \frac{4AB - A^2 - 4}{8(B+1)} \;\geq\; 0 \quad \text{for} \quad B \geq \tfrac{1}{4}A + \tfrac{1}{A}, \quad A > 0, \quad (\Rightarrow B \geq 1). \tag{22}$$

Inequality (16) therefore holds for any $A > 0$, provided we insert $B = \frac{1}{4}A + \frac{1}{A}$. Thus we might minimize the r.h.s. of (16) w.r.t. A leading to the upper bound

$$E_{n\Theta_\xi} \;\leq\; E_{n\Theta_\mu} + H_n + \sqrt{4E_{n\mu}H_n + H_n^2} \qquad \text{for} \qquad A^2 = \frac{H_n}{E_{n\Theta_\mu} + \frac{1}{4}H_n}$$

which completes the proof of Theorem 2 □.

5 Generalizations

In the following we discuss several directions in which the findings of this work may be extended.

5.1 General Loss Function

A prediction is very often the basis for some decision. The decision results in an action, which itself leads to some reward or loss. To stay in the framework of (passive) prediction we have to assume that the action itself does not influence

the environment. Let $l^k_{x_k y_k}(x_{<k}) \in [l_{min}, l_{min}+l_\Delta]$ be the received loss when taking action $y_k \in \mathcal{Y}$ and $x_k \in \mathcal{A}$ is the kth symbol of the sequence. For instance, if we make a sequence of weather forecasts $\mathcal{A}=\{$sunny, rainy$\}$ and base our decision, whether to take an umbrella or wear sunglasses $\mathcal{Y}=\{$umbrella, sunglasses$\}$ on it, the action of taking the umbrella or wearing sunglasses does not influence the future weather (ignoring the butterfly effect). The error assignment of section 4 falls into this class. The action was just a prediction ($\mathcal{Y}=\mathcal{A}$) and a unit loss was assigned to an erroneous prediction ($l_{x_k y_k}=1$ for $x_k \neq y_k$) and no loss to a correct prediction ($l_{x_k x_k}=0$). In general, a Λ_ρ action/prediction scheme $y_k^{\Lambda_\rho} := \text{minarg}_{y_k} \sum_{x_k} \rho(x_{<k}\underline{x}_k) l_{x_k y_k}$ can be defined that minimizes the ρ-expected loss. Λ_ξ is the universal scheme based on the universal prior ξ. Λ_μ is the optimal informed scheme. In [Hut01] it is proven that the total μ-expected losses $L_{n\Lambda_\xi}$ and $L_{n\Lambda_\mu}$ of Λ_ξ and Λ_μ are bounded in the following way: $0 \leq L_{n\Lambda_\xi} - L_{n\Lambda_\mu} \leq l_\Delta H_n + \sqrt{4(L_{n\Lambda_\mu} - n l_{min}) l_\Delta H_n + l_\Delta^2 H_n^2}$. The loss bound has a similar form as the error bound of Theorem 2, but the proof is much more evolved.

5.2 Games of Chance

The general loss bound stated in the previous subsection can be used to estimate the time needed to reach the winning threshold in a game of chance (defined as a sequence of bets, observations and rewards). In step k we bet, depending on the history $x_{<k}$, a certain amount of money s_k, take some action y_k, observe outcome x_k, and receive reward r_k. Our profit, which we want to maximize, is $p_k = r_k - s_k \in [p_{max} - p_\Delta, p_{max}]$. The loss, which we want to minimize, can be identified with the negative profit, $l_{x_k y_k} = -p_k$. The Λ_ρ-system acts as to maximize the ρ-expected profit. Let $\bar{p}_{n\Lambda_\rho}$ be the average expected profit of the first n rounds. One can show that the average profit of the Λ_ξ system converges to the best possible average profit $\bar{p}_{n\Lambda_\mu}$ achieved by the Λ_μ scheme ($\bar{p}_{n\Lambda_\xi} - \bar{p}_{n\Lambda_\mu} = O(n^{-1/2}) \to 0$ for $n \to \infty$). If there is a profitable scheme at all, then asymptotically the universal Λ_ξ scheme will also become profitable with the same average profit. In [Hut01] it is further shown that $(\frac{2p_\Delta}{\bar{p}_{n\Lambda_\mu}})^2 \cdot d_\mu$ is an upper bound for the number of bets n needed to reach the winning zone. The bound is proportional to the relative entropy of μ and ξ.

5.3 Infinite Alphabet

In many cases the basic prediction unit is not a letter, but a number (for inducing number sequences), or a word (for completing sentences), or a real number or vector (for physical measurements). The prediction may either be generalized to a block by block prediction of symbols or, more suitably, the finite alphabet $\mathcal{A}$ could be generalized to countable (numbers, words) or continuous (real or vector) alphabet. The theorems should generalize to countably infinite alphabets by appropriately taking the limit $|\mathcal{A}| \to \infty$ and to continuous alphabets by a denseness or separability argument.

5.4 Partial Prediction, Delayed Prediction, Classification

The Λ_ρ schemes may also be used for partial prediction where, for instance, only every m^{th} symbol is predicted. This can be arranged by setting the loss l^k to zero when no prediction is made, e.g. if k is not a multiple of m. Classification could be interpreted as partial sequence prediction, where $x_{(k-1)m+1:km-1}$ is classified as x_{km}. There are better ways for classification by treating $x_{(k-1)m+1:km-1}$ as pure conditions in ξ, as has been done in [Hut00] in a more general context. Another possibility is to generalize the prediction schemes and theorems to delayed sequence prediction, where the true symbol x_k is given only in cycle $k+d$. A delayed feedback is common in many practical problems.

5.5 More Active Systems

Prediction means guessing the future, but not influencing it. A tiny step in the direction to more active systems, described in subsection 5.1, was to allow the Λ system to act and to receive a loss $l_{x_k y_k}$ depending on the action y_k and the outcome x_k. The probability μ is still independent of the action, and the loss function l^k has to be known in advance. This ensures that the greedy strategy is optimal. The loss function may be generalized to depend not only on the history $x_{<k}$, but also on the historic actions $y_{<k}$ with μ still independent of the action. It would be interesting to know whether the scheme Λ and/or the loss bounds generalize to this case. The full model of an acting agent influencing the environment has been developed in [Hut00], but loss bounds have yet to be proven.

5.6 Miscellaneous

Another direction is to investigate the learning aspect of universal prediction. Many prediction schemes explicitly learn and exploit a model of the environment. Learning and exploitation are melted together in the framework of universal Bayesian prediction. A separation of these two aspects in the spirit of hypothesis learning with MDL [VL00] could lead to new insights. Finally, the system should be tested on specific induction problems for specific M with computable ξ.

6 Summary

Solomonoff's universal probability measure has been generalized to arbitrary probability classes and weights. A wise choice of M widens the applicability by reducing the computational burden for ξ. Convergence of ξ to μ and error bounds have been proven for arbitrary finite alphabet. They show that the universal prediction scheme Λ_ξ is an excellent substitute for the best possible (but generally unknown) informed scheme Λ_μ. Extensions and applications, including general loss functions and bounds, games of chance, infinite alphabet, partial and delayed prediction, classification, and more active systems, have been discussed.

References

[AS83] D. Angluin and C. H. Smith. Inductive inference: Theory and methods. *ACM Computing Surveys*, 15(3):237–269, 1983.

[Cal98] C. S. Calude et al. Recursively enumerable reals and Chaitin Ω numbers. In *15th Annual Symposium on Theoretical Aspects of Computer Science*, volume 1373 of *lncs*, pages 596–606, Paris France, 1998. Springer.

[Cha91] G. J. Chaitin. Algorithmic information and evolution. *in O.T. Solbrig and G. Nicolis, Perspectives on Biological Complexity, IUBS Press*, pages 51–60, 1991.

[FMG92] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[Hut99] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Science, in press*, 1999. ftp://ftp.idsia.ch/pub/techrep/IDSIA-11-00.ps.gz.

[Hut00] M. Hutter. A theory of universal artificial intelligence based on algorithmic complexity. Technical report, 62 pages, 2000. http://arxiv.org/abs/cs.AI/0004001.

[Hut01] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. Technical Report IDSIA-09-01, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Manno(Lugano), Switzerland, 2001.

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

[Kul59] S. Kullback. *Information Theory and Statistics.* Wiley, 1959.

[Lev73] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9:265–266, 1973.

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

[LV92] M. Li and P. M. B. Vitányi. Inductive reasoning and Kolmogorov complexity. *Journal of Computer and System Sciences*, 44:343–384, 1992.

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

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

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

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

[Sol97] R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.

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

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