Optimality of Universal Bayesian Sequence Prediction

Marcus Hutter

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

14 September 2001

Keywords

Bayesian sequence prediction; Mixture distributions, Solomonoff induction; Kolmogorov complexity; learning; universal probability; tight loss and error bounds; Pareto-optimality.

Abstract

Various optimality properties of universal sequence prediction based on Bayes-mixtures in general, and Solomonoff's prediction scheme in particular will be studied. The probability of observing x_t at time t, given past observations $x_1...x_{t-1}$ can be computed with Bayes' rule if the true generating distribution μ of the sequences $x_1x_2x_3...$ is known. If μ is unknown, but known to belong to a class $\mathcal{M}$ one can base ones prediction on the Bayes-mix ξ defined as a w_ν weighted sum of distributions $\nu\in\mathcal{M}$. The number of additional prediction errors E_ξ made by the optimal universal prediction scheme based on ξ minus the number of errors E_μ of the optimal informed prediction scheme based on μ is bounded by $O(\sqrt{E_\mu})$. We show that the bound is tight and that no other predictor can lead to smaller bounds. Furthermore, for various performance measures we show Pareto-optimality of ξ in the sense that there is no other predictor which performs better or equal in all environments $\nu\in\mathcal{M}$ and strictly better in at least one. So, optimal predictors can (w.r.t. to most performance measures in expectation) be based on the mixture ξ. Finally we give an Occam's razor argument that Solomonoff's choice $w_\nu\sim 2^{-K(\nu)}$ for the weights is optimal, where $K(\nu)$ is the length of the shortest program describing ν.

1 Introduction

The universal prior: Most inductive inference problem can be brought into the following form: Given a string $x_1x_2...x_{t-1}$, take a guess at its continuation x_t. 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 true 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 $\mathcal{M} := \{\mu_1,\mu_2,...\}$ be a finite or countable set of candidate probability distributions on strings. We define a weighted average on $\mathcal{M}$.

$$\xi(x_1...x_n) \;:=\; \sum_{\nu\in\mathcal{M}} w_\nu\cdot\nu(x_1...x_n), \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu > 0. \tag{1}$$

It is easy to see that ξ is a probability distribution as the weights w_ν are positive and normalized to 1 and the $\nu\in\mathcal{M}$ are probabilities.[2] We call ξ universal relative to $\mathcal{M}$, as it multiplicatively dominates all distributions in $\mathcal{M}$

$$\xi(x_1...x_n) \;\geq\; w_\nu\cdot\nu(x_1...x_n) \quad \text{for all} \quad \nu\in\mathcal{M}. \tag{2}$$

In the following, we assume that $\mathcal{M}$ is known and contains the true distribution, i.e. $\mu\in\mathcal{M}$. In this work we show that predictors based on the Bayes-mix ξ in general, and Solomonoff's prediction scheme in particular are optimal w.r.t. various criteria. Furthermore we show that the error and loss bounds derived here and in [Hut01a, Hut01b, Hut02] are tight.

Contents: Apart from introducing notation we briefly discuss in Section 2 convergence and representation of the universal posterior $\xi(x_t|x_1...x_{t-1})$, Solomonoff's choice of $\mathcal{M}$ and w_μ, and the case $\mu\notin\mathcal{M}$. Previous bounds on the total expected number of errors E^{Θ_ξ} of the universal predictor Θ_ξ in terms of total expected number of errors E^{Θ_μ} of the optimal informed scheme Θ_μ are slightly improved in Section 3. In Section 4 we show that these improved bounds are optimal in the sense that there are $\mathcal{M}$ and $\mu\in\mathcal{M}$ and weights w_ν such that the derived error bounds are tight. Although also no other predictor can lead to better bounds there could still be a predictor with never larger error than E^{Θ_ξ} and smaller error for some μ. In Section 5 we show that there is no such predictor, proving ξ optimal in a Pareto-sense. Optimal predictors (w.r.t. to most performance measures in expectation) can be based on a mixture distributions ξ. This still leaves open how to choose the weights. In Section 6 we give an Occam's razor argument that Solomonoff's choice $w_\nu\sim 2^{-K(\nu)}$ is optimal, where $K(\nu)$ is the length of the shortest program describing ν.

[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] The weight w_ν may be interpreted as the initial belief in ν and $\xi(x_1...x_n)$ as the degree of belief in $x_1...x_n$. If the existence of true randomness is rejected on philosophical grounds one may consider $\mathcal{M}$ containing only deterministic environments. ξ still represents belief probabilities.

What is new? Theorem 3 on lower error bounds, Theorems 5, 6 and 7 on Pareto-optimality, and Theorem 8 on the optimal choice of weights are new. The error bound in Theorem 2 slightly improves [Hut01a]. The convergence Theorem 1 is a known result [Sol78, LV97, Hut01a].

Introductory references: There are good reviews/papers/books of Solomonoff sequence prediction [LV97], inductive inference [AS83, Sol97] in general, MDL and reasoning under uncertainty [Grü98], worst case (WM) approaches [CB97], Bayesian prediction approaches [HB92], and competitive online statistics [Vov99], which contain many further references.

2 Setup and Convergence

Notation: We denote strings over a finite alphabet $\mathcal{X}$ by $x_1x_2...x_n$ with $x_t \in \mathcal{X}$. We abbreviate $x_{n:m} := x_n x_{n+1}...x_{m-1}x_m$ and $x_{<n} := x_1...x_{n-1}$. We use Greek letters for probability distributions/measures, especially ρ for arbitrary ones, $\mu \in \mathcal{M}$ for the true (generating) one, $\nu \in \mathcal{M}$ for arbitrary ones in $\mathcal{M}$, and ξ for the mixture (1). Let $\rho(x_{1:t})$ be the probability that an (infinite) sequence starts with $x_1...x_t$. The conditional ρ probability $\rho(x_t|x_{<t}) = \rho(x_{1:t})/\rho(x_{<t})$ that a given string $x_1...x_{t-1}$ is continued by x_t is obtained by using Bayes' rule. The prediction schemes will be based on these posteriors. We abbreviate expectations w.r.t. x_t, $x_{1:n}$, and $x_{<t}$ by

$$\mathbf{E}_t[..] := \sum_{x_t \in \mathcal{X}} \mu(x_t|x_{<t})[..], \quad \mathbf{E}_{1:n}[..] := \sum_{x_{1:n} \in \mathcal{X}^n} \mu(x_{1:n})[..], \quad \mathbf{E}_{<t}[..] := \sum_{x_{<t} \in \mathcal{X}^{t-1}} \mu(x_{<t})[..]$$

Expectations $\mathbf{E}$ are *always* taken w.r.t. the true distribution μ. $\mathbf{E}_{1:n} = \mathbf{E}_{<n}\mathbf{E}_n$ by Bayes' rule. We abbreviate "with μ probability 1" by w.μ.p.1. Finally, we use the relative entropy and the Euclidian distance to measure the instantaneous and total distances between μ and ξ:

$$d_t(x_{<t}) \;:=\; \mathbf{E}_t \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})}, \quad D_n \;:=\; \sum_{t=1}^{n} \mathbf{E}_{<t} d_t(x_{<t}) \;=\; \mathbf{E}_{1:n} \ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \tag{3}$$

$$s_t(x_{<t}) \;:=\; \sum_{x_t} \Big(\mu(x_t|x_{<t}) - \xi(x_t|x_{<t})\Big)^2, \qquad S_n \;:=\; \sum_{t=1}^{n} \mathbf{E}_{<t} s_t(x_{<t}) \tag{4}$$

Theorem 1 (Convergence) *Let there be sequences $x_1x_2...$ over a finite alphabet $\mathcal{X}$ drawn with probability $\mu(x_{1:n})$ for the first n symbols. The universal conditional probability $\xi(x_t|x_{<t})$ of the next symbol x_t given $x_{<t}$ is related to the true conditional probability $\mu(x_t|x_{<t})$ in the following way:*

$$\sum_{t=1}^{n} \mathbf{E}_{<t} \sum_{x_t} \Big(\mu(x_t|x_{<t}) - \xi(x_t|x_{<t})\Big)^2 \;\equiv\; S_n \;\leq\; D_n \;\leq\; \ln w_\mu^{-1} \;<\; \infty$$

where d_t and D_n are the relative entropies (3), and w_μ is the weight (1) of μ in ξ.

A proof for binary alphabet can be found in [Sol78, LV97] and for general finite alphabet in [Hut01a]. The finiteness of S_∞ implies $\xi(x'_t|x_{<t})-\mu(x'_t|x_{<t})\to 0$ for $t\to\infty$ w.μ.p.1 for any x'_t. This convergence motivates the belief that predictions based on (the known) ξ are asymptotically as good as predictions based on (the unknown) μ with rapid convergence.

Universal posterior probability distribution: All prediction schemes in this work are based on the conditional probabilities $\rho(x_t|x_{<t})$. It is possible to express also the conditional probability $\xi(x_t|x_{<t})$ as a weighted average over the conditional $\nu(x_t|x_{<t})$ similarly to (1), but now with time dependent weights:

$$\xi(x_t|x_{<t}) = \sum_{\nu\in\mathcal{M}} w_\nu(x_{<t})\nu(x_t|x_{<t}), \quad w_\nu(x_{<t}) := w_\nu\frac{\nu(x_{<t})}{\xi(x_{<t})}. \tag{5}$$

This representation can be proven by dividing (1) by $\xi(x_{<t})$ and applying Bayes' rule. The expressions (5) can be used to give an intuitive, but non-rigorous, argument why $\xi(x_t|x_{<t})$ converges to $\nu(x_t|x_{<t})$: The weight w_ν of ν in ξ increases/decreases if ν assigns a high/low probability to the new symbol x_t, given $x_{<t}$. For a μ-random sequence $x_{1:t}$, $\mu(x_{1:t}) \gg \nu(x_{1:t})$ if ν (significantly) differs from μ. We expect the total weight for all ν consistent with μ to converges to 1, and all other weights converge to 0 for $t\to\infty$. Therefore we expect $\xi(x_t|x_{<t})$ to converge to $\mu(x_t|x_{<t})$ for μ-random strings $x_{1:n}$. In this form one sees how μ is "learned".

The case where $\mu\notin\mathcal{M}$: In the following we discuss two cases, where $\mu\notin\mathcal{M}$, but most parts of this work still apply. Actually all theorems remain valid for μ being a finite linear combination $\mu(x_{1:n})=\sum_{i\in\mathcal{L}} v_\nu\nu(x_{1:n})$ of μ's in $\mathcal{L}\subseteq\mathcal{M}$. Dominance $\xi(x_{1:n})\geq w_\mu\cdot\mu(x_{1:n})$ is still ensured with $w_\mu := \min_{i\in\mathcal{L}}\frac{w_\nu}{v_\nu} \geq \min_{i\in\mathcal{L}} w_\nu$. More general, if μ is an infinite linear combination, dominance is still ensured if w_ν itself dominate v_ν in the sense that $w_\nu\geq\alpha v_\nu$ for some $\alpha>0$ (then $w_\mu\geq\alpha$).

Another possibly interesting situation is when the true generating distribution $\mu\notin\mathcal{M}$, but a "nearby" distribution $\hat\mu$ with weight $w_{\hat\mu}$ is in $\mathcal{M}$. If we measure the distance of $\hat\mu$ to μ with the Kullback Leibler divergence $D_n(\mu||\hat\mu) := \sum_{x_{1:n}}\mu(x_{1:n})\ln\frac{\mu(x_{1:n})}{\hat\mu(x_{1:n})}$ and assume that it is bounded by a constant c, then

$$D_n \;=\; \mathbf{E}_{1:n}\ln\frac{\mu(x_{1:n})}{\xi(x_{1:n})} \;=\; \mathbf{E}_{1:n}\ln\frac{\hat\mu(x_{1:n})}{\xi(x_{1:n})} + \mathbf{E}_{1:n}\ln\frac{\mu(x_{1:n})}{\hat\mu(x_{1:n})} \;\leq\; \ln w_{\hat\mu}^{-1} + c.$$

So $D_n\leq\ln w_\mu^{-1}$ remains valid if we define $w_\mu := w_{\hat\mu}\cdot e^{-c}$. See [Grü98] for a more detailed discussion of this case in a related context.

Probability classes $\mathcal{M}$: For finite $\mathcal{M}$ a possible choice for the w is to give all ν equal weight ($w_\nu=\frac{1}{|\mathcal{M}|}$). In the following, we assume that $\mathcal{M}$ is known and contains the true distribution, i.e. $\mu\in M$. This is not a serious constraint if we include *all* computable probability distributions in $\mathcal{M}$ with a high weight assigned to simple ν. Solomonoff's

universal semi-measure is obtained if we take $\mathcal{M}$ to be the (multi)set enumerated by a Turing machine which enumerates all enumerable semi-measures and take weights $w_\nu \sim 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program for ν [Sol64, Sol78, LV97]. A discussion of various general purpose choices for $\mathcal{M}$ is given in [Hut01a]. From a constructive point of view the set containing all (finitely or asymptotically or semi-) computable ν is sufficiently rich. In Section 6 we give an Occam's razor argument that the choice $w_\nu \sim 2^{-K(\nu)}$ is not only good, but optimal.

If ξ is itself in $\mathcal{M}$, it is called a universal element of $\mathcal{M}$ [LV97]. As we do not need this property here, $\mathcal{M}$ may be *any* finite or countable set of distributions. In the following we consider generic $\mathcal{M}$ and w. Continuous classes $\mathcal{M}$ are considered in a companion paper [Hut02].

3 Error & Loss Bounds

We start with a very simple performance measure: making a wrong prediction counts as one error, making a correct prediction counts as no error. So the strategy Θ_μ which predicts the x_t given $x_{<t}$ which maximizes the posterior probability $\mu(x_t|x_{<t})$ minimizes the expected number of errors. More generally, let Θ_ρ be a prediction scheme predicting $x_t^{\Theta_\rho} := \text{argmax}_{x_t}\rho(x_t|x_{<t})$ for some distribution ρ.

The μ probability of making a wrong prediction for the t^{th} symbol and the total μ-expected number of errors in the first n predictions of predictor Θ_ρ are

$$e_t^{\Theta_\rho}(x_{<t}) \;:=\; 1-\mu(x_t^{\Theta_\rho}|x_{<t}) \quad , \quad E_n^{\Theta_\rho} \;:=\; \sum_{t=1}^{n} \mathbf{E}_{<t} e_t^{\Theta_\rho}(x_{<t}). \tag{6}$$

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

$$E_n^{\Theta_\mu} \;\leq\; E_n^{\Theta} \quad \text{for any} \quad \Theta, \tag{7}$$

where Θ is *any* prediction scheme. Of special interest is the universal predictor Θ_ξ for which the following error bound can be shown.

Theorem 2 (Error bound) *Let there be sequences $x_1x_2...$ over a finite alphabet $\mathcal{X}$ drawn with probability $\mu(x_{1:n})$ for the first n symbols. The Θ_ρ-system predicts by definition $x_t^{\Theta_\rho} \in \mathcal{X}$ from $x_{<t}$, where $x_t^{\Theta_\rho}$ maximizes $\rho(x_t|x_{<t})$. 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 (6) are bounded in the following way*

$$0 \;\leq\; E_n^{\Theta_\xi} - E_n^{\Theta_\mu} \;\leq\; \sqrt{2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})S_n} \;\leq\; S_n + \sqrt{4E_n^{\Theta_\mu}S_n + S_n^2} \;\leq\; 2S_n + 2\sqrt{E_n^{\Theta_\mu}S_n}$$

where $S_n \leq D_n \leq \ln w_\mu^{-1}$. S_n is the squared distance (4), D_n is the relative entropy (3), and w_μ is the weight (1) of μ in ξ.

The first bound actually contains $E_n^{\Theta_\xi}$ on the r.h.s., so it is not particularly useful, but this is the major bound we will prove, the others follow easily. Furthermore it has a somewhat nicer, more symmetric structure than the second bound. In Section 4 we show that the second bound is optimal. The last bound, which we discuss in the following, has the same asymptotics as the second bound. 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. For more complicated probabilistic environments, where even the ideal informed system makes an infinite number of errors, the Theorem ensures that the regret $E_n^{\Theta_\xi}-E_n^{\Theta_\mu}$ is only of order $\sqrt{E_n^{\Theta_\mu}}$, since $S_n \leq \ln w_\mu^{-1} = const.$ The Theorem shows that the ratio converges to 1, and also gives the speed of convergence $E_n^{\Theta_\xi}/E_n^{\Theta_\mu} = 1+O((E_n^{\Theta_\mu})^{-1/2}) \to 1$ for $E_n^{\Theta_\mu} \to \infty$. See [Hut01a] for a more detailed discussion.

The bounds given here are only a slight improvement over the bounds obtained in [Hut01a]. The proof is somewhat nicer, but the significance of this improvement stems from the fact that the new (second) bound is tight (cannot be improved in general) for finite n (i.e. even non-asymptotically).

Proof: The first inequality in Theorem 2 has already been proven (7). For the second inequality, let us start more modestly and try to find constants $A>0$ and $B>0$ that satisfy the linear inequality

$$E_n^{\Theta_\xi} - E_n^{\Theta_\mu} \;\leq\; A(E_n^{\Theta_\xi} + E_n^{\Theta_\mu}) + BS_n. \tag{8}$$

If we could show

$$e_t^{\Theta_\xi}(x_{<t}) - e_t^{\Theta_\xi}(x_{<t}) \;\leq\; A[e_t^{\Theta_\xi}(x_{<t}) + e_t^{\Theta_\mu}(x_{<t})] + Bs_t(x_{<t}) \tag{9}$$

for all $t \leq n$ and all $x_{<t}$, (8) would follow immediately by summation and the definition of E_n and S_n. With the abbreviations

$$\mathcal{X} = \{1,...,N\}, \quad N = |\mathcal{X}|, \quad i = x_t, \quad m = x_t^{\Theta_\mu}, \quad s = x_t^{\Theta_\xi},$$
$$y_i = \mu(x_t|x_{<t}), \quad z_i = \xi(x_t|x_{<t})$$

the various error functions can then be expressed by $e_t^{\Theta_\xi} = 1-y_s$, $e_t^{\Theta_\mu} = 1-y_m$ and $s_t = \sum_i (y_i - z_i)^2$. Inserting this into (9) we get

$$y_m - y_s \;\leq\; A[2-(y_m+y_s)] + B\sum_{i=1}^{N}(y_i - z_i)^2. \tag{10}$$

By definition of $x_t^{\Theta_\mu}$ and $x_t^{\Theta_\xi}$ 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

$$B\sum_{i=1}^{N}(y_i - z_i)^2 + A[2-(y_m+y_s)] - (y_m - y_s) \;\geq\; ... \tag{11}$$

is positive for suitable $A \geq 0$ and $B \geq 0$, which proves (10). For $m = s$ (11) is obviously positive. So we will assume $m \neq s$ in the following. From the square we keep only contributions from $i = m$ and $i = s$.

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

By definition of y, z, $\mathcal{M}$ 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 replace (y_m+y_s) by 1.

$$... \;\geq\; \tfrac{1}{2}Bx^2+A-x \;\geq\; ... \tag{12}$$

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

$$... \;\geq\; A-\frac{1}{2B} \;\geq\; 0 \quad \text{for} \quad 2AB\geq 1. \tag{13}$$

Inequality (8) therefore holds for any $A>0$, provided we insert $B=\frac{1}{2A}$. Thus we might minimize the r.h.s. of (8) w.r.t. A leading to the upper bound

$$E_n^{\Theta_\xi}-E_n^{\Theta_\mu} \;\leq\; \sqrt{2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})S_n} \qquad \text{for} \qquad A^2=\frac{S_n}{2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})} \tag{14}$$

which is the first bound in Theorem 2. For the second bound we have to prove

$$\sqrt{2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})S_n}-S_n \;\leq\; \sqrt{4E_n^{\Theta_\mu}S_n+S_n^2} \tag{15}$$

If we square both sides of this expressions and simplify we just get (14). Hence, (14) implies (15). The last inequality in Theorem 2 is a simple triangle inequality. This completes the proof of Theorem 2 $\Box$.

Note that also the second bound implies the first one:

$$E_n^{\Theta_\xi}-E_n^{\Theta_\mu} \;\leq\; \sqrt{2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})S_n} \quad\Leftrightarrow\quad (E_n^{\Theta_\xi}-E_n^{\Theta_\mu})^2 \;\leq\; 2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})S_n \quad\Leftrightarrow$$

$$\Leftrightarrow\quad (E_n^{\Theta_\xi}-E_n^{\Theta_\mu}-S_n)^2 \;\leq\; 4E_n^{\Theta_\mu}S_n+S_n^2 \quad\Leftrightarrow\quad E_n^{\Theta_\xi}-E_n^{\Theta_\mu}-S_n \;\leq\; \sqrt{4E_n^{\Theta_\mu}S_n+S_n^2}$$

where we only have used $E_n^{\Theta_\xi}\geq E_n^{\Theta_\mu}$. Nevertheless the bounds are not equal.

General loss function: The setup and result can be generalized to arbitrary loss functions. Let $\ell_{x_ty_t}\in\!\!R$ be the received loss when predicting y_t, but x_t is the actual outcome. The error function of the previous subsection is a special case which assigns unit loss to an erroneous prediction ($\ell_{x_ty_t}=1$ for $x_t\neq y_t$) and no loss to a correct prediction ($\ell_{x_tx_t}=0$). The true probability of the next symbol being x_t, given $x_{<t}$, is $\mu(x_t|x_{<t})$. The expected loss when predicting y_t is $\mathbf{E}_t\ell_{x_ty_t}$. The goal is to minimize the expected loss. More generally we define the Λ_ρ prediction scheme

$$y_t^{\Lambda_\rho} \;:=\; \arg\min_{y_t\in\mathcal{Y}}\sum_{x_t}\rho(x_t|x_{<t})\ell_{x_ty_t} \tag{16}$$

which minimizes the ρ-expected loss.[3] As the true distribution is μ, the actual μ-expected loss when Λ_ρ predicts the t^{th} symbol and the total μ-expected loss in the first n predictions are

$$l_t^{\Lambda_\rho}(x_{<t}) \;:=\; \mathbf{E}_t \ell_{x_t y_t^{\Lambda_\rho}} \quad , \quad L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^{n} \mathbf{E}_{<t} l_t^{\Lambda_\rho}(x_{<t}). \tag{17}$$

Let Λ be *any* (causal) prediction scheme (deterministic or probabilistic does not matter) with no constraint at all, predicting *any* $y_t^\Lambda \in \mathcal{Y}$ with losses l_t^Λ and L_n^Λ similarly defined as (17). If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss

$$L_n^{\Lambda_\mu} \;\leq\; L_n^\Lambda \quad \text{for any} \quad \Lambda \tag{18}$$

The following loss bound for the universal Λ_ξ predictor is proven in the companion paper [Hut02].

$$0 \;\leq\; L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} \;\leq\; D_n + \sqrt{4L_n^{\Lambda_\mu} D_n + D_n^2} \;\leq\; 2D_n + 2\sqrt{L_n^{\Lambda_\mu} D_n} \tag{19}$$

The loss bounds have the same form as the error bounds when substituting $S_n \leq\; D_n$ in Theorem 2. One can show that (19) with D_n replaced by S_n gives an invalid bound, so the general bound is slightly weaker. Example loss functions including the absolute, square, logarithmic, and Hellinger loss are discussed in [Hut02]. Instantaneous error/loss bounds are also be proven:

$$e_n^\Theta(x_{<t}) - e_n^{\Theta_\mu}(x_{<t}) \;\leq\; \sqrt{2s_t(x_{<t})}, \quad l_{n\Lambda}(x_{<t}) - l_{n\Lambda_\mu}(x_{<t}) \;\leq\; \sqrt{2d_t(x_{<t})}.$$

4 Lower Error Bound

We want to show that there exists a class $\mathcal{M}$ of distributions such that *any* predictor Θ not knowing from which distribution $\mu \in \mathcal{M}$ the observed sequence is sampled from must make some minimal additional number of errors as compared to the best informed predictor Θ_μ. For deterministic environments a lower bound can easily be obtained by a combinatoric argument. Consider a class $\mathcal{M}$ containing 2^n binary sequences such that each prefix of length n occurs exactly once. Assume any deterministic predictor Θ (not knowing the sequence in advance), then for every prediction x_t^Θ of Θ at times $t \leq n$ there exists a sequence with opposite symbol $x_t = 1 - x_t^\Theta$. Hence, $E_\infty^\Theta \geq E_n^\Theta = \log_2|\mathcal{M}|$ is a lower worst case bound for every predictor Θ, (this includes Θ_ξ, of course). The upper bound $E_\infty^{\Theta_\xi} \leq 2\ln|\mathcal{M}|$ from Theorem 2 (obtained by inserting $E_\infty^{\Theta_\mu} = 0$ and $D_n \leq \ln|\mathcal{M}|$) is not sharp but can easily improved to $E_\infty^{\Theta_\xi} \leq log_2|\mathcal{M}|$ for deterministic environments, matching the lower bound. In the general probabilistic case we can show by a similar argument that the upper bound of Theorem 2 is sharp.

[3] $\text{argmin}_y(\cdot)$ is defined as the y which minimizes the argument. A tie is broken arbitrarily. In general, the prediction space $\mathcal{Y}$ is allowed to differ from $\mathcal{X}$. If $\mathcal{Y}$ is finite, then $y_t^{\Lambda_\rho}$ always exists. For infinite action space $\mathcal{Y}$ we assume that a minimizing $y_t^{\Lambda_\rho} \in \mathcal{Y}$ exists, although even this assumption may be removed.

Theorem 3 (Lower Error Bound) *Let* Θ *be* any deterministic *predictor not knowing from which distribution* $\mu\in\mathcal{M}$ *the observed sequence* $x_1x_2...$ *is sampled from.* Θ *knows (depends on)* $\mathcal{M}$ *and has at time* t *access to the previous outcomes* $x_{<t}$*. Then there is for every* n *an* $\mathcal{M}$ *and* $\mu\in\mathcal{M}$ *and weights* w_ν *such that*

$$e_n^\Theta - e_n^{\Theta_\mu} \;=\; \sqrt{2s_t(x_{<t})} \quad \textit{and} \quad E_n^\Theta - E_n^{\Theta_\mu} \;=\; S_n + \sqrt{4E_n^{\Theta_\mu}S_n + S_n^2}$$

where E_n^Θ *and* $E_n^{\Theta_\mu}$ *are the total expected number of errors of* Θ *and* Θ_μ*, and* s_t *and* S_n *are defined in (4). The equalities especially hold for the universal predictor* Θ_ξ*.*

Proof: The proof parallels and generalizes the deterministic case. Consider a class $\mathcal{M}$ of 2^n distributions (over binary alphabet) indexed by $a\equiv a_1...a_n\in\{0,1\}^n$. For each t we want a distribution with posterior probability $\frac{1}{2}(1+\varepsilon)$ for $x_t=1$ and one with posterior probability $\frac{1}{2}(1-\varepsilon)$ for $x_t=1$ independent of the past $x_{<t}$ with $\varepsilon>0$ to be determined later. That is

$$\mu_a(x_1...x_n)=\mu_{a_1}(x_1)\cdot...\cdot\mu_{a_n}(x_n),\quad \text{where}\quad \mu_{a_t}(x_t)=\begin{cases}\frac{1}{2}(1+\varepsilon) & \text{for} \quad a_t=x_t\\ \frac{1}{2}(1-\varepsilon) & \text{for} \quad a_t\neq x_t\end{cases}$$

We are not interested in predictions beyond time n but for completeness we may define μ_a to assign probability 1 to $x_t=1$ for all $t>n$. If $\mu=\mu_a$, the informed scheme Θ_μ always predicts the bit which has highest μ-probability, i.e. $y_t^{\Theta_\mu}=a_t$

$$\Longrightarrow\quad e_t^{\Theta_\mu}=1-\mu_{a_t}(y_t^{\Theta_\mu})=\tfrac{1}{2}(1-\varepsilon)\quad\Longrightarrow\quad E_n^{\Theta_\mu}=\tfrac{n}{2}(1-\varepsilon).$$

Since $E_n^{\Theta_\mu}$ is the same for all a we seek to maximize E_n^Θ for a given predictor Θ in the following. Assume Θ predicts y_t^Θ (possibly depending on the history $x_{<t}$). Since we want lower bounds we seek for a worst case μ. A success $y_t^\Theta=x_t$ has lowest possible probability $\frac{1}{2}(1-\varepsilon)$ if $a_t=1-y_t^\Theta$.

$$\Longrightarrow\quad e_t^{\Theta}=1-\mu_{a_t}(y_t^{\Theta})=\tfrac{1}{2}(1+\varepsilon)\quad\Longrightarrow\quad E_n^{\Theta}=\tfrac{n}{2}(1+\varepsilon).$$

So we have $e_t^\Theta-e_t^{\Theta_\mu}=\varepsilon$ and $E_n^\Theta-E_n^{\Theta_\mu}=n\varepsilon$ for the regrets. We need to eliminate n and ε in favor of s_t, S_n, and $E_n^{\Theta_\mu}$. If we assume uniform weights $w_{\mu_a}=2^{-n}$ for all μ_a we get

$$\xi(x_{1:n}) \;=\; \sum_a w_{\mu_a}\mu_a(x_{1:n}) \;=\; 2^{-n}\prod_{t=1}^n\sum_{a_t\in\{0,1\}}\mu_{a_t}(x_t) \;=\; 2^{-n}\prod_{t=1}^n 1 \;=\; 2^{-n},$$

i.e. ξ is an unbiased Bernoulli sequence ($\xi(x_t|x_{<t})=\frac{1}{2}$).

$$\Longrightarrow\quad s_t(x_{<t}) \;=\; \sum_{x_t}(\tfrac{1}{2}-\mu_{a_t}(x_t))^2 \;=\; \tfrac{1}{2}\varepsilon^2 \quad\text{and}\quad S_n=\tfrac{n}{2}\varepsilon^2.$$

So we have $\varepsilon=\sqrt{2s_t}$ which proves the instantaneous regret formula $e_t^\Theta-e_t^{\Theta_\mu}=\sqrt{2s_t}$. Inserting $\varepsilon=\sqrt{\frac{2}{n}S_n}$ into $E_n^{\Theta_\mu}$ and solving w.r.t. $\sqrt{2n}$ we get $\sqrt{2n}=\sqrt{S_n}+\sqrt{4E_n^{\Theta_\mu}+S_n}$. So we finally get

$$E_n^\Theta - E_n^{\Theta_\mu} \;=\; n\varepsilon \;=\; \sqrt{S_n}\sqrt{2n} \;=\; S_n+\sqrt{4E_n^{\Theta_\mu}S_n+S_n^2}$$

which proves the total regret formula of Theorem 3. □

Since $d_t/s_t = 1+O(\varepsilon^2)$ we have $D_n/S_n \to 1$ for $\varepsilon \to 0$. Hence the error bound of Theorem 2 with S_n replaced by D_n is asymptotically tight for $E_n^{\Theta_\mu}/D_n \to \infty$ (which implies $\varepsilon \to 0$). This shows that without restrictions on the loss function which exclude the error loss, the loss bounds (19) are asymptotically tight. An n independent set $\mathcal{M}$ leading to a good (but not tight) lower bound is $\mathcal{M} = \{\mu_1,\mu_2\}$ with $\mu_{1/2}(1|x_{<t}) = \frac{1}{2} \pm \varepsilon_t$ with $\varepsilon_t = \min\{\frac{1}{2}, \sqrt{\ln w_{\mu_1}^{-1}}/\sqrt{t}\ln t\}$. For $w_{\mu_1} \ll w_{\mu_2}$ and $n \to \infty$ one can show that $E_n^{\Theta_\xi} - E_n^{\Theta_\mu} \sim \frac{1}{\ln n}\sqrt{E_n^{\Theta_\mu} D_n}$.

5 Pareto Optimality of ξ

In this subsection we want to establish a different kind of optimality property of ξ. Let $\mathcal{F}(\mu,\rho)$ be any of the performance measures of ρ relative to μ considered in the previous sections (e.g. s_t, or D_n, or L_n, ...). It is easy to find ρ more tailored towards μ such that $\mathcal{F}(\mu,\rho) < \mathcal{F}(\mu,\xi)$. This improvement may be achieved by increasing w_μ, but probably at the expense of increasing $\mathcal{F}$ for other ν, i.e. $\mathcal{F}(\nu,\rho) > \mathcal{F}(\nu,\xi)$ for some $\nu \in \mathcal{M}$. Since we do not know μ in advance we may ask whether there exists a ρ with better or equal performance for *all* $\nu \in \mathcal{M}$ and strictly better performance for one $\nu \in \mathcal{M}$. This would clearly render ξ suboptimal w.r.t. to $\mathcal{F}$. We show that there is no such ρ for all performance measures studied in this work.

Definition 4 (Pareto Optimality) *Let $\mathcal{F}(\mu,\rho)$ be any performance measure of ρ relative to μ. The universal prior ξ is called Pareto-optimal w.r.t. $\mathcal{F}$ if there is no ρ with $\mathcal{F}(\nu,\rho) \leq \mathcal{F}(\nu,\xi)$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one ν.*

Theorem 5 (Pareto Optimality) *The universal prior ξ is Pareto-optimal w.r.t. the instantaneous and total squared distances s_t and S_n (4), entropy distances d_t and D_n (3), errors e_t and E_n (6), and losses l_t and L_n (17).*

Proof: We first proof Theorem 5 for the instantaneous expected loss l_t. We need the more general ρ expected instantaneous losses

$$l_{t\rho}^{\Lambda}(x_{<t}) \;:=\; \sum_{x_t} \rho(x_t|x_{<t}) \ell_{x_t y_t^{\Lambda}} \tag{20}$$

for a predictor Λ. We want to arrive at a contradiction by assuming that ξ is not Pareto-optimal, i.e. by assuming the existence of a predictor[4] Λ with $l_{t\nu}^{\Lambda} \leq l_{t\nu}^{\Lambda_\xi}$ for all $\nu \in \mathcal{M}$ and

[4]According to definition 4 we should look for a ρ, but for each deterministic predictor Λ there exists a ρ with $\Lambda = \Lambda_\rho$.

strict inequality for some ν. Implicit to this assumption is the assumption that $l^\Lambda_{t\nu}$ and $l^{\Lambda_\xi}_{t\nu}$ exist. $l^\Lambda_{t\nu}$ exists iff $\nu(x_t|x_{<t})$ exists iff $\nu(x_{<t})>0$ iff $w_\nu(x_{<t})>0$.

$$l^\Lambda_{t\xi} \;=\; \sum_\nu w_\nu(x_{<t}) l^\Lambda_{t\nu} \;<\; \sum_\nu w_\nu(x_{<t}) l^{\Lambda_\xi}_{t\nu} \;=\; l^{\Lambda_\xi}_{t\xi} \;\leq\; l^\Lambda_{t\xi}$$

The two equalities follow from inserting (5) into (20). The strict inequality follows from the assumption and $w_\nu(x_{<t})>0$. The last inequality follows from the fact that Λ_ξ minimizes by definition (16) the ξ-expected loss (similarly to (18)). The contradiction $l^\Lambda_{t\xi}<l^\Lambda_{t\xi}$ proves Pareto-optimality of ξ w.r.t. l_t.

In the same way we can prove Pareto-optimality of ξ w.r.t. the total loss L_n by defining the ρ expected total losses

$$L^\Lambda_{n\rho} \;:=\; \sum_{t=1}^n \sum_{x_{<t}} \rho(x_{<t}) l^\Lambda_{t\rho}(x_{<t}) \;=\; \sum_{t=1}^n \sum_{x_{1:t}} \rho(x_{1:t}) \ell_{x_t y^\Lambda_t} \tag{21}$$

for a predictor Λ, and by assuming $L^\Lambda_{n\nu} \leq L^{\Lambda_\xi}_{n\nu}$ for all ν and strict inequality for some ν, from which we get the contradiction $L^\Lambda_{n\xi} = \sum_\nu w_\nu L^\Lambda_{n\nu} < \sum_\nu w_\nu L^{\Lambda_\xi}_{n\nu} = L^{\Lambda_\xi}_{n\xi} \leq L^\Lambda_{n\xi}$ with the help of (1). The instantaneous and total expected errors e_t and E_n can be considered as special loss functions.

Pareto-optimality of ξ w.r.t. s_t (and hence S_n) can be understood from geometrical insight. A formal proof for s_t goes as follows: With the abbreviations $i=x_t$, $y_{\nu i}=\nu(x_t|x_{<t})$, $z_i=\xi(x_t|x_{<t})$, $r_i=\rho(x_t|x_{<t})$, and temporarily $w_\nu=w_\nu(x_{<t})\geq 0$ we ask for a vector $\mathbf{r}$ with $\sum_i(y_{\nu i}-r_i)^2 \leq \sum_i(y_{\nu i}-z_i)^2 \;\forall\nu$. This implies

$$\begin{aligned} 0 \;&\geq\; \sum_\nu w_\nu \Big[\sum_i (y_{\nu i}-r_i)^2 - \sum_i (y_{\nu i}-z_i)^2 \Big] \;= \\ &=\; \sum_\nu w_\nu \Big[\sum_i -2y_{\nu i}r_i + r_i^2 + 2y_{\nu i}z_i - z_i^2 \Big] \;= \\ &=\; \sum_i -2z_i r_i + r_i^2 + 2z_i z_i - z_i^2 \;=\; \sum_i (r_i - z_i)^2, \end{aligned}$$

where we have used $\sum_\nu w_\nu = 1$ and $\sum_\nu w_\nu y_{\nu i} = z_i$ (5). $\sum_i (r_i-z_i)^2 \leq 0$ implies $\mathbf{r}=\mathbf{z}$ proving unique Pareto-optimality of ξ w.r.t. s_t. Similarly for d_t the assumption $\sum_i y_{\nu i} \ln \frac{y_{\nu i}}{r_i} \leq \sum_i y_{\nu i} \ln \frac{y_{\nu i}}{z_i} \;\forall\nu$ implies

$$0 \;\geq\; \sum_\nu w_\nu \Big[\sum_i y_{\nu i} \ln \frac{y_{\nu i}}{r_i} - y_{\nu i} \ln \frac{y_{\nu i}}{z_i} \Big] \;=\; \sum_\nu w_\nu \sum_i y_{\nu i} \ln \frac{z_i}{r_i} \;=\; \sum_i z_i \ln \frac{z_i}{r_i} \;\geq\; 0$$

which implies $\mathbf{r}=\mathbf{z}$ proving unique Pareto-optimality of ξ w.r.t. d_t. The proofs for S_n and D_n are similar. □

We have proven that ξ is *uniquely* Pareto-optimal w.r.t. s_t, S_n, d_t and D_n. In the case of e_t, E_n, l_t and L_n there are other $\rho\neq\xi$ with $\mathcal{F}(\nu,\rho)=\mathcal{F}(\nu,\xi)\forall\nu$, but the actions/predictions they invoke are unique ($y^{\Lambda_\rho}_t = y^{\Lambda_\xi}_t$) (if ties in argmax_{y_t} are broken in a consistent way), and this is all what counts.

For all measures which are relevant from a decision theoretic point of view, i.e. for all loss functions l_t and L_n, ξ has the welcomed property of being Pareto-optimal, but ξ is *not* Pareto-optimal w.r.t. to all thinkable performance measures.

Theorem 6 ((Non)Pareto-optimality) *ξ is Pareto-optimal w.r.t.*

- *the α-norm $||\cdot||_\alpha$ for $\alpha \geq 1$,*
- *positive linear combinations of α_i-norms with all $\alpha_i \geq 1$,*
- *a power of $\mathcal{F}$ if Pareto-optimal w.r.t. $\mathcal{F}$, i.e. esp. w.r.t. $||\cdot||_\alpha^\alpha$.*

ξ is (in general) not *Pareto-optimal w.r.t.*

- *the α-norm $||\cdot||_\alpha$ for $\alpha < 1$,*
- *positive linear combinations of $||\cdot||_{\alpha_i}^{\alpha_i}$ with all $\alpha_i \geq 1$.*
- *positive linear combinations of $\mathcal{F}_i$ even if Pareto-optimal w.r.t. all $\mathcal{F}_i$.*

Intuition on this problem can be gained by considering probability vectors $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v} \in \Delta \subset \mathbb{R}^3$, where Δ is the 2d probability triangle, and $\mathbf{z} = w\mathbf{x} + (1-w)\mathbf{y}$ is a mixture of $\mathbf{x}$ and $\mathbf{y}$. Consider the sets $M_\mathbf{x} := \{\mathbf{r} : \mathcal{F}(\mathbf{x},\mathbf{r}) \leq \mathcal{F}(\mathbf{x},\mathbf{z})\}$ and analogously $M_\mathbf{y}$. $M_\mathbf{x} \cap M_\mathbf{y}$ is not empty; it contains $\mathbf{z}$. If $M_\mathbf{x} \cap M_\mathbf{y}$ has an interior, then $\mathbf{z}$ is not Pareto-optimal. Visualize the 1d boundaries of the 2d areas $M_\mathbf{x}$ and $M_\mathbf{y}$ qualitatively for the various performance measures $\mathcal{F}$. This gives some intuition of how to prove Pareto-optimality and to construct counter-examples. Proofs will be given elsewhere.

Pareto-optimality should be regarded as a necessary condition for a prediction scheme aiming to be optimal. From a practical point of view a significant decrease of $\mathcal{F}$ for many ν may be desirable even if this causes a small increase of $\mathcal{F}$ for a few other ν. The impossibility of such a "balanced" improvement is a more demanding condition on ξ than pure Pareto-optimality. The next theorem shows that Λ_ξ is also balanced-Pareto-optimal. We only consider the performance measure L_n and suppress the index n for convenience.

Theorem 7 (Balanced Pareto Optimality w.r.t. L)

$$\Delta_\nu := L_\nu^{\tilde{\Lambda}} - L_\nu^{\Lambda_\xi}, \quad \Delta := \sum_{\nu \in \mathcal{M}} w_\nu \Delta_\nu \quad \Rightarrow \quad \Delta \geq 0.$$

This implies the following: Assume $\tilde{\Lambda}$ has larger loss than Λ_ξ on environments $\mathcal{L}$ by a total weighted amount of $\Delta_\mathcal{L} := \sum_{\lambda \in \mathcal{L}} w_\lambda \Delta_\lambda$. Then $\tilde{\Lambda}$ can have smaller loss on $\eta \in \mathcal{H} := \mathcal{M} \setminus \mathcal{L}$, but the improvement is bounded by $\Delta_\mathcal{H} := |\sum_{\eta \in \mathcal{H}} w_\eta \Delta_\eta| \leq \Delta_\mathcal{L}$. Especially $|\Delta_\eta| \leq w_\eta^{-1} \max_{\lambda \in \mathcal{L}} \Delta_\lambda$.

This means that a weighted loss decrease $\Delta_\mathcal{H}$ by using $\tilde{\Lambda}$ instead of Λ_ξ is compensated by an at least as large weighted increase $\Delta_\mathcal{L}$ on other environments. If the increase is small, the decrease can also only be small. In the special case of only a single environment with increased loss Δ_λ, the decrease is bound by $\Delta_\eta \leq \frac{w_\lambda}{w_\eta} |\Delta_\lambda|$, i.e. an increase by an amount Δ_λ can only cause a decrease by at most the same amount times a factor $\frac{w_\lambda}{w_\eta}$. A increase can only cause a smaller decrease in simpler environments, but a scaled decrease in more complex environments. Finally note that pure Pareto-optimality (5) follows from balanced Pareto-optimality in the special case of no increase $\Delta_\mathcal{L} \equiv 0$.

Proof: $\Delta \geq 0$ follows from $\Delta = \sum_\nu w_\nu [L_\nu^{\tilde{\Lambda}} - L_\nu^{\Lambda_\xi}] = L_\xi^{\tilde{\Lambda}} - L_\xi^{\Lambda_\xi} \geq 0$, where we have used linearity of L_ρ in ρ and $L_\xi^{\Lambda_\xi} \leq L_\xi^\Lambda$. The remainder of Theorem 7 is obvious from $0 \leq \Delta = \Delta_\mathcal{L} - \Delta_\mathcal{H}$ and by bounding the weighted average Δ_η by its maximum.□

6 On the Optimal Choice of Weights

In the following we indicate the dependency of ξ on w explicitly by writing ξ_w. We have shown that the Λ_{ξ_w} prediction schemes are (balanced) Pareto optimal, i.e. that *no* prediction scheme Λ (whether based on a Bayes mix or not) can be uniformly better. Least assumptions on the environment are made for $\mathcal{M}$ which are as large as possible. In Section 2 we have discussed the set $\mathcal{M}$ of all enumerable semimeasures which we regarded as sufficiently large (see [Sch02] for even larger sets, but which are still in the computational realm). Agreeing on this $\mathcal{M}$ still leaves open the question of how to choose the weights (prior believes) w_ν, since every ξ_w with $w_\nu > 0 \forall \nu$ is Pareto-optimal and leads asymptotically to optimal predictions.

We have derived bounds for the mean squared sum $S_n \leq \ln w_\mu^{-1}$ and for the loss excess $L_{n\nu}^{\Lambda_{\xi_w}} - L_{n\nu}^{\Lambda_\nu} \leq 2\ln w_\nu^{-1} + 2\sqrt{\ln w_\nu^{-1} L_{n\nu}^{\Lambda_\nu}}$. All bounds monotonically decrease with increasing w_ν. So it is desirable to assign high weights to all $\nu \in \mathcal{M}$. Due to the (semi)probability constraint $\sum_\nu w_\nu \leq 1$ one has to find a compromise.[5] In the following we will argue that in the class of enumerable weight functions with short program there is an optimal compromise, namely $w_\nu = 2^{-K(\nu)}$ which gives Solomonoff's prior.

Consider the class of enumerable weight function $\mathcal{V} := \{v_{(.)} : \mathcal{M} \to \mathbb{R}^+ \text{ with } \sum_\nu v_\nu \leq 1 \text{ and } K(v) = O(1)\}$. Let $w_\nu := 2^{-K(\nu)}$ and $v_{(.)} \in \mathcal{V}$. Corollary 4.3.1 of [LV97, p255] says that $K(x) \leq -\log_2 P(x) + K(P) + O(1)$ for all x if P is an enumerable discrete semimeasure. Identifying P with v and x with (the program index describing) ν we get

$$\ln w_\nu^{-1} \ \leq \ln v_\nu^{-1} + O(1).$$

This means that the bounds for ξ_w depending on $\ln w_\nu^{-1}$ are at most $O(1)$ larger than the bounds for ξ_v depending on $\ln v_\nu^{-1}$. So we lose at most an additive constant of order 1 in the bounds when using ξ_w instead of ξ_v. In using Solomonoff's prior ξ_w we are on the safe side, getting (within $O(1)$) best bounds for *all* environments.

Theorem 8 (Optimality of universal weights) *Within the set $\mathcal{V}$ of enumerable weight functions with short program, the universal weights $w_\nu = 2^{-K(\nu)}$ lead to the smallest performance bounds within an additive (to $\ln w_\mu^{-1}$) constant in all enumerable environments.*

Since this justifies the use of Solomonoff's prior and Solomonoff's prior assigns high probability to an environment if and only if it has low (Kolmogorov) complexity, one may

[5] All results in this paper haven stated and proven for probability measures μ, ξ and w_ν, i.e. $\sum_{x_{1:t}} \xi(x_{1:t}) = \sum_{x_{1:t}} \mu(x_{1:t}) = \sum_\nu w_\nu = 1$. On the other hand, the class $\mathcal{M}$ considered here is the class of all enumerable semimeasures and $\sum_\nu w_\nu < 1$. In general, each of the following 4 items could be semi ($<$) or not ($=$): (ξ, μ, $\mathcal{M}$, w_ν), where $\mathcal{M}$ is semi if some elements are semi. Six out of the 2^4 combinations make sense. Convergence (1), the error bound (Theorem 2), the loss bound (19), as well as most other statements hold for ($<$,$=$,$<$,$<$), but not for ($<$,$<$,$<$,$<$). Nevertheless, $\xi \to \mu$ holds also for ($<$,$<$,$<$,$<$) with maximal μ semi-probability, i.e. fails with μ semi-probability 0.

interpret the result as a justification of Occam's razor[6]. But note that this is more of a bootstrap argument, since we implicitly used Occam's razor to justify the restriction to enumerable semimeasures. We also considered only weight functions v with low complexity $K(v)=O(1)$. What did not enter as an assumption but came out as a result is that the specific universal weights $w_\nu=2^{-K(\nu)}$ are optimal.

We want to conclude this chapter with a remark on Occam's razor versus No Free Lunches. We do not regard the (balanced) Pareto-optimality result as a "No Free Lunch" (NFL) theorem [WM97]. Since most environments are completely random, a small concession on the loss in each of these completely uninteresting environments provides enough margin $\Delta_{\mathcal{H}}$ to yield distinguished performance on the few non-random (interesting) environments. Indeed, we would interpret the NFL theorems for optimization and search in [WM97] as balanced Pareto-optimality results. Interestingly, whereas for prediction only Bayes-mixes are Pareto-optimal, for search and optimization every algorithm is Pareto-optimal. There is a strong religious war between believers in Occam's razor and believers in no free lunches which we cannot go into here [Sto01].

7 Conclusions

Various optimality properties of universal prediction based on Bayes-mixtures in general, and Solomonoff's prediction scheme in particular have been studied in this work. We have shown that there are $\mathcal{M}$ and $\mu\in\mathcal{M}$ and weights w_ν such that the error and loss bounds derived in here and in [Hut01a, Hut01b] cannot be improved in general, i.e. without making extra assumptions on ℓ, $\mathcal{M}$, or w_ν, and this is true for any μ independent predictor. We have also shown Pareto-optimality of ξ in the sense that there is no other predictor which performs better or equal in all environments $\nu\in\mathcal{M}$ and strictly better in at least one. Optimal predictors can (in most cases) be based on a mixture distributions ξ. Finally we gave an Occam's razor argument that Solomonoff's choice $w_\nu=2^{-K(\nu)}$ for the weights, where $K(\nu)$ is the length of the shortest program describing ν, is optimal. Of course, optimality always depends on the setup, the assumptions, and the chosen criteria. For instance, the universal predictor was not always Pareto-optimal, but at least for many popular, and for all decision theoretic performance measures. Bayes predictors are also not necessarily optimal under worst case criteria [CBL01]. See [Hut01b] for references and further discussions on the duality between the Bayes and worst case (WM) approaches and results, classification tasks, games of chances, infinite alphabet, active systems influencing the environment ...

Acknowledgements

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

[6]The *only if* direction can be shown by a more easy and direct argument [Sch02].

References

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

[CB97] N. Cesa-Bianchi et al. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[CBL01] N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43(3):247–264, 2001.

[Grü98] P. Grünwald. *The Minimum Discription Length Principle and Reasoning under Uncertainty*. PhD thesis, Universiteit van Amsterdam, 1998.

[HB92] D. Haussler and A. Barron. How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values. *Proc. 3rd NEC Symposium on Computation and Cognition, SIAM*, pages 74–100, 1992.

[Hut01a] M. Hutter. Convergence and error bounds of universal prediction for general alphabet. *Proceedings of the 12th Eurpean Conference on Machine Learning (ECML-2001)*, pages 239–250, 2001.

[Hut01b] M. Hutter. General loss bounds for universal sequence prediction. *Proceedings of the 18^{th} International Conference on Machine Learning (ICML-2001)*, pages 210–217, 2001.

[Hut02] M. Hutter. Finite loss bounds for universal Bayesian sequence prediction. *submitted*, 2002. http://www.idsia.ch/∼marcus/ai/spupper.ps.

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

[Sch02] J. Schmidhuber. Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. *International Journal of Foundations of Computer Science*, 2002. In press.

[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.

[Sto01] D. Stork. Foundations of Occam's razor and parsimony in learning. *NIPS 2001 Workshop*, 2001. http://www.rii.ricoh.com/∼stork/OccamWorkshop.html.

[Vov99] V. G. Vovk. Competitive on-line statistics. Technical report, CLRC and DoCS, University of London, 1999.

[WM97] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.