Journal of Machine Learning Research 4 (2003) 971-1000 Submitted 2/03; Published 11/03

Optimality of Universal Bayesian Sequence Prediction for General Loss and Alphabet

Marcus Hutter MARCUS@IDSIA.CH
IDSIA, Galleria 2
6928 Manno-Lugano, Switzerland

Editor: Peter Bartlett

Abstract

Various optimality properties of universal sequence predictors 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 the chain 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 countable or continuous class $\mathcal{M}$ one can base ones prediction on the Bayes-mixture ξ defined as a w_ν-weighted sum or integral of distributions $\nu\in\mathcal{M}$. The cumulative expected loss of the Bayes-optimal universal prediction scheme based on ξ is shown to be close to the loss of the Bayes-optimal, but infeasible prediction scheme based on μ. We show that the bounds are tight and that no other predictor can lead to significantly smaller bounds. Furthermore, for various performance measures, we show Pareto-optimality of ξ and give an Occam's razor argument that the choice $w_\nu\sim 2^{-K(\nu)}$ for the weights is optimal, where $K(\nu)$ is the length of the shortest program describing ν. The results are applied to games of chance, defined as a sequence of bets, observations, and rewards. The prediction schemes (and bounds) are compared to the popular predictors based on expert advice. Extensions to infinite alphabets, partial, delayed and probabilistic prediction, classification, and more active systems are briefly discussed.

Keywords: Bayesian sequence prediction; mixture distributions; Solomonoff induction; Kolmogorov complexity; learning; universal probability; tight loss and error bounds; Pareto-optimality; games of chance; classification.

1. Introduction

Many problems are of the induction type in which statements about the future have to be made, based on past observations. What is the probability of rain tomorrow, given the weather observations of the last few days? Is the Dow Jones likely to rise tomorrow, given the chart of the last years and possibly additional newspaper information? Can we reasonably doubt that the sun will rise tomorrow? Indeed, one definition of science is to predict the future, where, as an intermediate step, one tries to understand the past by developing theories and finally to use the prediction as the basis for some decision. Most induction problems can be studied in the Bayesian framework. The probability of observing x_t at time t, given the observations $x_1...x_{t-1}$ can be computed with the chain rule, if we know the true probability distribution, which generates the observed sequence $x_1x_2x_3...$. The problem is that in many cases we do not even have a reasonable guess of the true distribution μ. What is the true probability of weather sequences, stock charts, or sunrises?

In order to overcome the problem of the unknown true distribution, one can define a mixture distribution ξ as a weighted sum or integral over distributions $\nu\in\mathcal{M}$, where $\mathcal{M}$ is any discrete or continuous (hypothesis) set including μ. $\mathcal{M}$ is assumed to be known and to contain the true distribution, i.e. $\mu\in\mathcal{M}$. Since the probability ξ can be shown to converge rapidly to the true probability μ in a conditional sense, making decisions based on ξ is often nearly as good as the infeasible optimal decision based on the unknown μ (Merhav and Feder, 1998). Solomonoff (1964) had the idea to define a universal mixture as a weighted average over deterministic programs. Lower weights were assigned to longer programs. He unified Epicurus' principle of multiple explanations and Occam's razor [simplicity] principle into one formal theory (See Li and Vitányi 1997 for this interpretation of Solomonoff 1964). Inspired by Solomonoff's idea, Levin (1970) defined the closely related universal prior ξ_U as a weighted average over *all* semi-computable probability distributions. If the environment possesses some effective structure at all, Solomonoff-Levin's posterior "finds" this structure (Solomonoff, 1978), and allows for a good prediction. In a sense, this solves the induction problem in a universal way, i.e. without making problem specific assumptions.

Section 2 explains notation and defines the *universal or mixture distribution* ξ as the w_ν-weighted sum of probability distributions ν of a set $\mathcal{M}$, which includes the true distribution μ. No structural assumptions are made on the ν. ξ multiplicatively dominates all $\nu\in\mathcal{M}$, and the relative entropy between μ and ξ is bounded by $\ln w_\mu^{-1}$. Convergence of ξ to μ in a mean squared sense is shown in Theorem 1. The representation of the universal posterior distribution and the case $\mu\notin\mathcal{M}$ are briefly discussed. Various standard sets $\mathcal{M}$ of probability measures are discussed, including computable, enumerable, cumulatively enumerable, approximable, finite-state, and Markov (semi)measures.

Section 3 is essentially a generalization of the deterministic error bounds found by Hutter (2001b) from the binary alphabet to a general finite alphabet $\mathcal{X}$. Theorem 2 bounds $E^{\Theta_\xi}-E^{\Theta_\mu}$ by $O(\sqrt{E^{\Theta_\mu}})$, where E^{Θ_ξ} is the expected number of errors made by the optimal universal predictor Θ_ξ, and E^{Θ_μ} is the expected number of errors made by the optimal informed prediction scheme Θ_μ. The non-binary setting cannot be reduced to the binary case! One might think of a binary coding of the symbols $x_t\in\mathcal{X}$ in the sequence x_1x_2.... But this makes it necessary to predict a block of bits x_t, before one receives the true block of bits x_t, which differs from the bit by bit prediction scheme considered by Solomonoff (1978) and Hutter (2001b). The framework generalizes to the case where an action $y_t\in\mathcal{Y}$ results in a loss $\ell_{x_ty_t}$ if x_t is the next symbol of the sequence. Optimal universal Λ_ξ and optimal informed Λ_μ prediction schemes are defined for this case, and loss bounds similar to the error bounds of the last section are stated. No assumptions on ℓ have to be made, besides boundedness.

Section 4 applies the loss bounds to games of chance, defined as a sequence of bets, observations, and rewards. The average profit $\bar{p}_n^{\Lambda_\xi}$ achieved by the Λ_ξ scheme rapidly converges to the best possible average profit $\bar{p}_n^{\Lambda_\mu}$ achieved by the Λ_μ scheme ($\bar{p}_n^{\Lambda_\xi}-\bar{p}_n^{\Lambda_\mu}=O(n^{-1/2})$). If there is a profitable scheme at all ($\bar{p}_n^{\Lambda_\mu}>\varepsilon>0$), asymptotically the universal Λ_ξ scheme will also become profitable. Theorem 3 bounds the time needed to reach the winning zone. It is proportional to the relative entropy of μ and ξ with a factor depending on the profit range and on $\bar{p}_n^{\Lambda_\mu}$. An attempt is made to give an information theoretic interpretation of the result.

Section 5 discusses the quality of the universal predictor and the bounds. We show that there are $\mathcal{M}$ and $\mu\in\mathcal{M}$ and weights w_ν such that the derived error bounds are tight. This shows that the error bounds cannot be improved in general. We also show Pareto-optimality of ξ in the sense that there is no other predictor which performs at least as well in all environments $\nu\in\mathcal{M}$ and strictly better in at least one. Optimal predictors can always be based on mixture distributions ξ. This still leaves open how to choose the weights. We give an Occam's razor argument that the choice $w_\nu=2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program describing ν is optimal.

Section 6 generalizes the setup to continuous probability classes $\mathcal{M}=\{\mu_\theta\}$ consisting of continuously parameterized distributions μ_θ with parameter $\theta\in I\!R^d$. Under certain smoothness and regularity conditions a bound for the relative entropy between μ and ξ, which is central for all presented results, can still be derived. The bound depends on the Fisher information of μ and grows only logarithmically with n, the intuitive reason being the necessity to describe θ to an accuracy $O(n^{-1/2})$. Furthermore, two ways of using the prediction schemes for partial sequence prediction, where not every symbol needs to be predicted, are described. Performing and predicting a sequence of independent experiments and online learning of classification tasks are special cases. We also compare the universal prediction scheme studied here to the popular predictors based on expert advice (PEA) (Littlestone and Warmuth, 1989, Vovk, 1992, Littlestone and Warmuth, 1994, Cesa-Bianchi et al., 1997, Haussler et al., 1998, Kivinen and Warmuth, 1999). Although the algorithms, the settings, and the proofs are quite different, the PEA bounds and our error bound have the same structure. Finally, we outline possible extensions of the presented theory and results, including infinite alphabets, delayed and probabilistic prediction, active systems influencing the environment, learning aspects, and a unification with PEA.

Section 7 summarizes the results.

There are good introductions and surveys of Solomonoff sequence prediction (Li and Vitányi, 1992, 1997), inductive inference in general (Angluin and Smith, 1983, Solomonoff, 1997, Merhav and Feder, 1998), reasoning under uncertainty (Grünwald, 1998), and competitive online statistics (Vovk, 1999), with interesting relations to this work. See Section 6.3 for some more details.

2. Setup and Convergence

In this section we show that the mixture ξ converges rapidly to the true distribution μ. After defining basic notation in Section 2.1, we introduce in Section 2.2 the *universal or mixture distribution* ξ as the w_ν-weighted sum of probability distributions ν of a set $\mathcal{M}$, which includes the true distribution μ. No structural assumptions are made on the ν. ξ multiplicatively dominates all $\nu\in\mathcal{M}$. A posterior representation of ξ with incremental weight update is presented in Section 2.3. In Section 2.4 we show that the relative entropy between μ and ξ is bounded by $\ln w_\mu^{-1}$ and that ξ converges to μ in a mean squared sense. The case $\mu\notin\mathcal{M}$ is briefly discussed in Section 2.5. The section concludes with Section 2.6, which discusses various standard sets $\mathcal{M}$ of probability measures, including computable, enumerable, cumulatively enumerable, approximable, finite-state, and Markov (semi)measures.

2.1 Random Sequences

We denote strings over a finite alphabet $\mathcal{X}$ by $x_1x_2...x_n$ with $x_t\in\mathcal{X}$ and $t,n,N\in I\!N$ and $N=|\mathcal{X}|$. We further use the abbreviations ϵ for the empty string, $x_{t:n}:=x_tx_{t+1}...x_{n-1}x_n$ for $t\leq n$ and ϵ for $t>n$, and $x_{<t}:=x_1...x_{t-1}$. We use Greek letters for probability distributions (or measures). Let $\rho(x_1...x_n)$ be the probability that an (infinite) sequence starts with $x_1...x_n$:

$$\sum_{x_{1:n}\in\mathcal{X}^n}\rho(x_{1:n})=1,\quad \sum_{x_t\in\mathcal{X}}\rho(x_{1:t})=\rho(x_{<t}),\quad \rho(\epsilon)=1.$$

We also need conditional probabilities derived from the chain rule:

$$\begin{aligned}\rho(x_t|x_{<t}) &= \rho(x_{1:t})/\rho(x_{<t}),\\ \rho(x_1...x_n) &= \rho(x_1)\!\cdot\!\rho(x_2|x_1)\!\cdot\!...\!\cdot\!\rho(x_n|x_1...x_{n-1}).\end{aligned}$$

The first equation states that the probability that a string $x_1...x_{t-1}$ is followed by x_t is equal to the probability that a string starts with $x_1...x_t$ divided by the probability that a string starts with $x_1...x_{t-1}$. For convenience we define $\rho(x_t|x_{<t})=0$ if $\rho(x_{<t})=0$. The second equation is the first, applied n times. Whereas ρ might be any probability distribution, μ denotes the true (unknown) generating distribution of the sequences. We denote probabilities by $\mathbf{P}$, expectations by $\mathbf{E}$ and further abbreviate

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

Probabilities $\mathbf{P}$ and expectations $\mathbf{E}$ are *always* w.r.t. the true distribution μ. $\mathbf{E}_{1:n}=\mathbf{E}_{<n}\mathbf{E}_n$ by the chain rule and $\mathbf{E}[...]=\mathbf{E}_{<t}[...]$ if the argument is independent of $x_{t:\infty}$, and so on. We abbreviate "with μ-probability 1" by w.μ.p.1. We say that z_t converges to z_* *in mean sum* (i.m.s.) if $c:=\sum_{t=1}^\infty\mathbf{E}[(z_t-z_*)^2]<\infty$. One can show that convergence in mean sum implies convergence with probability 1.[1] Convergence i.m.s. is very strong: it provides a "rate" of convergence in the sense that the expected number of times t in which z_t deviates more than ε from z_* is finite and bounded by c/ε^2 and the probability that the number of ε-deviations exceeds $\frac{c}{\varepsilon^2\delta}$ is smaller than δ.

2.2 Universal Prior Probability Distribution

Every inductive inference problem can be brought into the following form: Given a string $x_{<t}$, take a guess at its continuation x_t. We will assume that the strings which have to be continued are drawn from a probability[2] distribution μ. The maximal prior information a prediction algorithm can possess is the exact knowledge of μ, but in many cases (like the

[1] Convergence in the mean, i.e. $\mathbf{E}[(z_t-z_*)^2]\stackrel{t\to\infty}{\longrightarrow}0$, only implies convergence in probability, which is weaker than convergence with probability 1.

[2] 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*.

probability of sun tomorrow) the true 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 $\mathcal{M}:=\{\nu_1,\nu_2,...\}$ be a countable set of candidate probability distributions on strings. Results are generalized to continuous sets $\mathcal{M}$ in Section 6.1. We define a weighted average on $\mathcal{M}$

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

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.[3] For a finite $\mathcal{M}$ a possible choice for the w is to give all ν equal weight ($w_\nu=\frac{1}{|\mathcal{M}|}$). We call ξ universal relative to $\mathcal{M}$, as it multiplicatively dominates all distributions in $\mathcal{M}$

$$\xi(x_{1:n}) \;\geq\; w_\nu\!\cdot\!\nu(x_{1:n}) \quad \text{for all} \quad \nu\in\mathcal{M}. \tag{2}$$

In the following, we assume that $\mathcal{M}$ is known and contains the true distribution, i.e. $\mu\!\in\!\mathcal{M}$. If $\mathcal{M}$ is chosen sufficiently large, then $\mu\!\in\!\mathcal{M}$ is not a serious constraint.

2.3 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 $\nu(x_t|x_{<t})$, but now with time dependent weights:

$$\xi(x_t|x_{<t}) = \sum_{\nu\in\mathcal{M}} w_\nu(x_{<t})\nu(x_t|x_{<t}), \quad w_\nu(x_{1:t}) := w_\nu(x_{<t})\frac{\nu(x_t|x_{<t})}{\xi(x_t|x_{<t})}, \quad w_\nu(\epsilon) := w_\nu. \tag{3}$$

The denominator just ensures correct normalization $\sum_\nu w_\nu(x_{1:t})\!=\!1$. By induction and the chain rule we see that $w_\nu(x_{<t})\!=\!w_\nu\nu(x_{<t})/\xi(x_{<t})$. Inserting this into $\sum_\nu w_\nu(x_{<t})\nu(x_t|x_{<t})$ using (1) gives $\xi(x_t|x_{<t})$, which proves the equivalence of (1) and (3). The expressions (3) can be used to give an intuitive, but non-rigorous, argument why $\xi(x_t|x_{<t})$ converges to $\mu(x_t|x_{<t})$: 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 converge to 1, and all other weights to 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 $x_{1:\infty}$.

Expressions (3) seem to be more suitable than (1) for studying convergence and loss bounds of the universal predictor ξ, but it will turn out that (2) is all we need, with the sole exception in the proof of Theorem 6. Probably (3) is useful when one tries to understand the learning aspect in ξ.

[3] The weight w_ν may be interpreted as the initial degree of 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.

2.4 Convergence of ξ to μ

We use the relative entropy and the squared Euclidian/absolute distance to measure the instantaneous and total distances between μ and ξ:

$$d_t(x_{<t}) \;:=\; \mathbf{E}_t \ln\frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})}, \qquad D_n \;:=\; \sum_{t=1}^n \mathbf{E}_{<t} d_t(x_{<t}) \;=\; \mathbf{E}_{1:n}\ln\frac{\mu(x_{1:n})}{\xi(x_{1:n})} \tag{4}$$

$$s_t(x_{<t}) \;:=\; \sum_{x_t}\Big(\mu(x_t|x_{<t})-\xi(x_t|x_{<t})\Big)^2, \qquad S_n \;:=\; \sum_{t=1}^n \mathbf{E}_{<t} s_t(x_{<t}) \tag{5}$$

$$a_t(x_{<t}) \;:=\; \sum_{x_t}\Big|\mu(x_t|x_{<t})-\xi(x_t|x_{<t})\Big|, \qquad V_n \;:=\; \frac{1}{2}\sum_{t=1}^n \mathbf{E}_{<t} a_t^2(x_{<t}) \tag{6}$$

One can show that $s_t \le \frac{1}{2}a_t^2 \le d_t$ (Hutter, 2001a, Sec.3.2) (Cover and Thomas, 1991, Lem.12.6.1), hence $S_n\le V_n\le D_n$ (for binary alphabet, $s_t=\frac{1}{2}a_t^2$, hence $S_n=V_n$). So bounds in terms of S_n are tightest, while the (implied) looser bounds in terms of V_n as a referee pointed out have an advantage in case of continuous alphabets (not considered here) to be reparametrization-invariant. The weakening to D_n is used, since D_n can easily be bounded in terms of the weight w_μ.

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:*

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

where d_t and D_n are the relative entropies (4), and w_μ is the weight (1) of μ in ξ.

A proof for binary alphabet can be found in works by Solomonoff (1978) or Li and Vitányi (1997) and for a general finite alphabet in work by Hutter (2001a). The finiteness of S_∞ implies $\xi(x'_t|x_{<t})-\mu(x'_t|x_{<t})\to 0$ for $t\to\infty$ i.m.s., and hence w.μ.p.1 for any x'_t. There are other convergence results, most notably $\xi(x_t|x_{<t})/\mu(x_t|x_{<t})\to 1$ for $t\to\infty$ w.μ.p.1 (Li and Vitányi, 1997, Hutter, 2003a). These convergence results motivate the belief that predictions based on (the known) ξ are asymptotically as good as predictions based on (the unknown) μ with rapid convergence.

2.5 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 $\mu(x_{1:n})=\sum_{\nu\in\mathcal{L}}v_\nu\nu(x_{1:n})$ of ν's in $\mathcal{L}\subseteq\mathcal{M}$. Dominance $\xi(x_{1:n})\ge w_\mu\cdot\mu(x_{1:n})$ is still ensured with $w_\mu:=\min_{\nu\in\mathcal{L}}\frac{w_\nu}{v_\nu}\ge\min_{\nu\in\mathcal{L}}w_\nu$. More generally, if μ is an infinite linear combination, dominance is still ensured if w_ν itself dominates v_ν in the sense that $w_\nu\ge\alpha v_\nu$ for some $\alpha>0$ (then $w_\mu\ge\alpha$).

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 $w_\mu:=w_{\hat\mu}\cdot e^{-c}$.

2.6 Probability Classes $\mathcal{M}$

In the following we describe some well-known and some less known probability classes $\mathcal{M}$. This relates our setting to other works in this area, embeds it into the historical context, illustrates the type of classes we have in mind, and discusses computational issues.

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

We will see that it is favorable to assign high weights w_ν to the ν. 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 ν. The prefix Kolmogorov complexity $K(\nu)$ is a universal complexity measure (Kolmogorov, 1965, Zvonkin and Levin, 1970, Li and Vitányi, 1997). It is defined as the length of the shortest self-delimiting program (on a universal Turing machine) computing $\nu(x_{1:n})$ given $x_{1:n}$. If we define

$$w_\nu \;:=\; 2^{-K(\nu)}$$

then distributions which can be calculated by short programs, have high weights. The relative entropy is bounded by the Kolmogorov complexity of μ in this case ($D_n\leq K(\mu)\cdot\ln 2$). Levin's universal semi-measure ξ_U is obtained if we take $\mathcal{M}=\mathcal{M}_U$ to be the (multi)set enumerated by a Turing machine which enumerates all enumerable semi-measures (Zvonkin and Levin, 1970, Li and Vitányi, 1997). Recently, $\mathcal{M}$ has been further enlarged to include all cumulatively enumerable semi-measures (Schmidhuber, 2002a). In the enumerable and cumulatively enumerable 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 (Hutter, 2003b). An interesting and quickly approximable distribution is the Speed prior S defined by Schmidhuber (2002b). It is related to Levin complexity and Levin search (Levin, 1973, 1984), but it is unclear for now, which distributions are dominated by S. If one considers only finite-state automata instead of general Turing machines, ξ is related to the quickly computable, universal finite-state prediction scheme of Feder et al. (1992), 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 $\mathcal{M}$ and increase w. A detailed analysis of these and other specific classes $\mathcal{M}$ will be given elsewhere. Note that $\xi\in\mathcal{M}$ in the enumerable and cumulatively enumerable case, but $\xi\notin\mathcal{M}$ in the computable, approximable and finite-state case. If ξ is itself in $\mathcal{M}$, it is called a universal element of $\mathcal{M}$ (Li and Vitányi, 1997). As we do not need this property here, $\mathcal{M}$ may be *any* countable set of distributions. In the following sections we consider generic $\mathcal{M}$ and w.

We have discussed various discrete classes $\mathcal{M}$, which are sufficient from a constructive or computational point of view. On the other hand, it is convenient to also allow for continuous classes $\mathcal{M}$. For instance, the class of *all* Bernoulli processes with parameter $\theta\in[0,1]$ and uniform prior $w_\theta\equiv 1$ is much easier to deal with than computable θ only, with prior $w_\theta=2^{-K(\theta)}$. Other important continuous classes are the class of i.i.d. and Markov processes. Continuous classes $\mathcal{M}$ are considered in more detail in Section 6.1.

3. Error Bounds

In this section we prove error bounds for predictors based on the mixture ξ. Section 3.1 introduces the concept of Bayes-optimal predictors Θ_ρ, minimizing ρ-expected error. In Section 3.2 we bound $E^{\Theta_\xi}-E^{\Theta_\mu}$ by $O(\sqrt{E^{\Theta_\mu}})$, where E^{Θ_ξ} is the expected number of errors made by the optimal universal predictor Θ_ξ, and E^{Θ_μ} is the expected number of errors made by the optimal informed prediction scheme Θ_μ. The proof is deferred to Section 3.3. In Section 3.4 we generalize the framework to the case where an action $y_t\in\mathcal{Y}$ results in a loss $\ell_{x_t y_t}$ if x_t is the next symbol of the sequence. Optimal universal Λ_ξ and optimal informed Λ_μ prediction schemes are defined for this case, and loss bounds similar to the error bounds are presented. No assumptions on ℓ have to be made, besides boundedness.

3.1 Bayes-Optimal Predictors

We start with a very simple measure: making a wrong prediction counts as one error, making a correct prediction counts as no error. Hutter (2001b) has proven error bounds for the binary alphabet $\mathcal{X}=\{0,1\}$. The following generalization to an arbitrary alphabet involves only minor additional complications, but serves as an introduction to the more complicated model with arbitrary loss function. Let Θ_μ be the optimal prediction scheme when the strings are drawn from the probability distribution μ, i.e. the probability of x_t given $x_{<t}$ is $\mu(x_t|x_{<t})$, and μ is known. Θ_μ predicts (by definition) $x_t^{\Theta_\mu}$ when observing $x_{<t}$. The prediction is erroneous if the true t^{th} symbol is not $x_t^{\Theta_\mu}$. The probability of this event is $1-\mu(x_t^{\Theta_\mu}|x_{<t})$. It is minimized if $x_t^{\Theta_\mu}$ maximizes $\mu(x_t^{\Theta_\mu}|x_{<t})$. 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. Every deterministic predictor can be interpreted as maximizing some distribution.

3.2 Total Expected Numbers of Errors

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

$$e_t^{\Theta_\rho}(x_{<t}) \;:=\; 1-\mu(x_t^{\Theta_\rho}|x_{<t}) \quad,\quad E_n^{\Theta_\rho} \;:=\; \sum_{t=1}^n \mathbf{E}_{<t}e_t^{\Theta_\rho}(x_{<t}). \tag{7}$$

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

$$E_n^{\Theta_\mu} \;\leq\; E_n^{\Theta_\rho} \quad \text{for any} \quad \Theta_\rho, \tag{8}$$

since

$$e_t^{\Theta_\mu}(x_{<t}) \;=\; 1-\mu(x_t^{\Theta_\mu}|x_{<t}) \;=\; \min_{x_t}\{1-\mu(x_t|x_{<t})\} \;\leq\; 1-\mu(x_t^{\Theta_\rho}|x_{<t}) \;=\; e_t^{\Theta_\rho}(x_{<t})$$

for any ρ. 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 $x_t^{\Theta_\rho}$ is a discontinuous function of ρ and $x_t^{\Theta_\xi}\to x_t^{\Theta_\mu}$ cannot be proven from $\xi\to\mu$. Indeed, this problem occurs in related prediction schemes, where the predictor has to be regularized so that it is continuous (Feder et al., 1992). 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 $\mathcal{X}$ 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 (7) are bounded in the following way*

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

where $Q_n=\sum_{t=1}^n\mathbf{E}_{<t}q_t$ (with $q_t(x_{<t}):=1-\delta_{x_t^{\Theta_\xi}x_t^{\Theta_\mu}}$) is the expected number of non-optimal predictions made by Θ_ξ and $S_n\leq V_n\leq D_n\leq \ln w_\mu^{-1}$, where S_n is the squared Euclidian distance (5), V_n half of the squared absolute distance (6), D_n the relative entropy (4), and w_μ the weight (1) of μ in ξ.

The first two bounds have a nice structure, but the r.h.s. actually depends on Θ_ξ, so they are not particularly useful, but these are the major bounds we will prove, the others follow easily. In Section 5 we show that the third bound is optimal. The last bound, which we discuss in the following, has the same asymptotics as the third bound. Note that the bounds hold for any (semi)measure ξ; only $D_n\leq\ln_\mu w^{-1}$ depends on ξ dominating μ with domination constant w_μ.

First, we observe that Theorem 2 implies 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. In particular, this is 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. This can also be proven by elementary means. Assume $x_1x_2...$ is the sequence generated by μ and Θ_ξ makes a wrong prediction $x_t^{\Theta_\xi}\neq x_t$. Since $\xi(x_t^{\Theta_\xi}|x_{<t})\geq\xi(x_t|x_{<t})$, this implies $\xi(x_t|x_{<t})\leq\frac{1}{2}$. Hence

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

$e_t^{\Theta_\xi} = 1 \leq -\ln\xi(x_t|x_{<t})/\ln 2 = d_t/\ln 2$. If Θ_ξ makes a correct prediction $e_t^{\Theta_\xi} = 0 \leq d_t/\ln 2$ is obvious. Using (4) this proves $E_\infty^{\Theta_\xi} \leq D_\infty/\ln 2 \leq \log_2 w_\mu^{-1}$. A combinatoric argument given in Section 5 shows that there are $\mathcal{M}$ and $\mu\in\mathcal{M}$ with $E_\infty^{\Theta_\xi} \geq \log_2|\mathcal{M}|$. This shows that the upper bound $E_\infty^{\Theta_\xi} \leq \log_2|\mathcal{M}|$ for uniform w is sharp. From Theorem 2 we get the slightly weaker bound $E_\infty^{\Theta_\xi} \leq 2S_\infty \leq 2D_\infty \leq 2\ln w_\mu^{-1}$. For more complicated probabilistic environments, where even the ideal informed system makes an infinite number of errors, the theorem ensures that the error regret $E_n^{\Theta_\xi} - E_n^{\Theta_\mu}$ is only of order $\sqrt{E_n^{\Theta_\mu}}$. The regret is quantified in terms of the information content D_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 $E_n^{\Theta_\mu} \to \infty$. If we increase the first occurrence of $E_n^{\Theta_\mu}$ in the theorem to E_n^{Θ} and the second to $E_n^{\Theta_\xi}$ we get the bound $E_n^{\Theta} \geq E_n^{\Theta_\xi} - 2\sqrt{E_n^{\Theta_\xi} S_n}$, which shows that *no* (causal) predictor Θ whatsoever makes significantly less errors than Θ_ξ. In Section 5 we show that the third bound for $E_n^{\Theta_\xi} - E_n^{\Theta_\mu}$ given in Theorem 2 can in general not be improved, i.e. for every predictor Θ (particularly Θ_ξ) there exist $\mathcal{M}$ and $\mu\in\mathcal{M}$ such that the upper bound is essentially achieved. See Hutter (2001b) for some further discussion and bounds for binary alphabet.

3.3 Proof of Theorem 2

The first inequality in Theorem 2 has already been proven (8). For the second inequality, let us start more modestly and try to find constants $A>0$ and $B>0$ that satisfy the linear inequality

$$E_n^{\Theta_\xi} - E_n^{\Theta_\mu} \;\leq\; AQ_n + BS_n. \tag{9}$$

If we could show

$$e_t^{\Theta_\xi}(x_{<t}) - e_t^{\Theta_\mu}(x_{<t}) \;\leq\; Aq_t(x_{<t}) + Bs_t(x_{<t}) \tag{10}$$

for all $t \leq n$ and all $x_{<t}$, (9) would follow immediately by summation and the definition of E_n, Q_n and S_n. With the abbreviations

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

the various error functions can then be expressed by $e_t^{\Theta_\xi} = 1 - y_s$, $e_t^{\Theta_\mu} = 1 - y_m$, $q_t = 1 - \delta_{ms}$ and $s_t = \sum_i (y_i - z_i)^2$. Inserting this into (10) we get

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

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

$$B\sum_{i=1}^N (y_i - z_i)^2 + A[1-\delta_{ms}] - (y_m - y_s) \;\geq\; ... \tag{12}$$

is positive for suitable $A\geq 0$ and $B\geq 0$, which proves (11). For $m=s$ (12) is obviously positive. So we will assume $m\neq s$ in the following. From the square we keep only contributions from $i=m$ and $i=s$.

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

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 $z_m=z_s=\frac{1}{2}(y_m+y_s)$. Together with the abbreviation $x:=y_m-y_s$ we get

$$... \;\geq\; \tfrac{1}{2}Bx^2+A-x \;\geq\; ... \tag{13}$$

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

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

Inequality (9) therefore holds for any $A>0$, provided we insert $B=\frac{1}{2A}$. Thus we might minimize the r.h.s. of (9) w.r.t. A leading to the upper bound

$$E_n^{\Theta_\xi}-E_n^{\Theta_\mu} \;\leq\; \sqrt{2Q_nS_n} \qquad\text{for}\qquad A^2=\frac{S_n}{2Q_n}$$

which is the first bound in Theorem 2. For the second bound we have to show $Q_n\leq E_n^{\Theta_\xi}+E_n^{\Theta_\mu}$, which follows by summation from $q_t\leq e_t^{\Theta_\xi}+e_t^{\Theta_\mu}$, which is equivalent to $1-\delta_{ms}\leq 1-y_s+1-y_m$, which holds for $m=s$ as well as $m\neq s$. For the third bound we have to prove

$$\sqrt{2(E_n^{\Theta_\xi}+E_n^{\Theta_\mu})S_n}-S_n \;\leq\; \sqrt{4E_n^{\Theta_\mu}S_n+S_n^2}. \tag{14}$$

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

Note that also the third bound implies the second 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.

3.4 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. If the action itself can influence the environment we enter the domain of acting agents which has been analyzed in the context of universal probability by Hutter (2001c). To stay in the framework of (passive) prediction we have to assume that the action itself does not influence the environment. Let $\ell_{x_ty_t}\in I\!R$ be the received loss when taking action $y_t\in\mathcal{Y}$ and $x_t\in\mathcal{X}$ is the t^{th} symbol of the sequence. We make the assumption that ℓ is bounded. Without loss of generality we normalize ℓ by linear scaling such that $0\leq\ell_{x_ty_t}\leq 1$. For instance, if we make a sequence of weather forecasts $\mathcal{X}=\{$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 losses might be

Loss	sunny	rainy
umbrella	0.1	0.3
sunglasses	0.0	1.0

Note the loss assignment even when making the right decision to take an umbrella when it rains because sun is still preferable to rain.

In many cases the prediction of x_t can be identified or is already the action y_t. The forecast *sunny* can be identified with the action *wear sunglasses*, and *rainy* with *take umbrella*. $\mathcal{X}\equiv\mathcal{Y}$ in these cases. The error assignment of the previous subsections falls into this class together with a special loss function. It 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 ($\ell_{x_tx_t}=0$).

For convenience we name an action a prediction in the following, even if $\mathcal{X}\neq\mathcal{Y}$. 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

$$y_t^{\Lambda_\rho} \;:=\; \arg\min_{y_t\in\mathcal{Y}}\sum_{x_t}\rho(x_t|x_{<t})\ell_{x_ty_t} \tag{15}$$

which minimizes the ρ-expected loss.[5] 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

$$l_t^{\Lambda_\rho}(x_{<t}) \;:=\; \mathbf{E}_t\ell_{x_ty_t^{\Lambda_\rho}} \quad,\quad L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^n \mathbf{E}_{<t}l_t^{\Lambda_\rho}(x_{<t}). \tag{16}$$

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 (16). If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss

$$L_n^{\Lambda_\mu} \;\leq\; L_n^\Lambda \quad\text{for any}\quad \Lambda. \tag{17}$$

The following loss bound for the universal Λ_ξ predictor is proven by Hutter (2003a).

$$0 \;\leq\; L_n^{\Lambda_\xi}-L_n^{\Lambda_\mu} \;\leq\; D_n+\sqrt{4L_n^{\Lambda_\mu}D_n+D_n^2} \;\leq\; 2D_n+2\sqrt{L_n^{\Lambda_\mu}D_n}. \tag{18}$$

The loss bounds have the same form as the error bounds when substituting $S_n\leq D_n$ in Theorem 2. For a comparison to Merhav's and Feder's (1998) loss bound, see Hutter

[5] $\text{argmin}_y(\cdot)$ 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 an 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.

(2003a). Replacing D_n by S_n or V_n in (18) gives an invalid bound, so the general bound is slightly weaker. For instance, for $\mathcal{X}=\{0,1\}$, $\ell_{00}=\ell_{11}=0$, $\ell_{10}=1$, $\ell_{01}=c<\frac{1}{4}$, $\mu(1)=0$, $\nu(1)=2c$, and $w_\mu=w_\nu=\frac{1}{2}$ we get $\xi(1)=c$, $s_1=2c^2$, $y_1^{\Lambda_\mu}=0$, $l_1^{\Lambda_\mu}=\ell_{00}=0$, $y_1^{\Lambda_\xi}=1$, $l_1^{\Lambda_\xi}=\ell_{01}=c$, hence $L_1^{\Lambda_\xi}-L_1^{\Lambda_\mu}=c\not\leq 4c^2=2S_1+2\sqrt{L_1^{\Lambda_\mu}S_1}$. Example loss functions including the absolute, square, logarithmic, and Hellinger loss are discussed in work by Hutter (2003a). Instantaneous error/loss bounds can also be proven:

$$e_t^{\Theta_\xi}(x_{<t})-e_t^{\Theta_\mu}(x_{<t}) \;\leq\; \sqrt{2s_t(x_{<t})}, \quad l_t^{\Lambda_\xi}(x_{<t})-l_t^{\Lambda_\mu}(x_{<t}) \;\leq\; \sqrt{2d_t(x_{<t})}.$$

4. Application to Games of Chance

This section applies the loss bounds to games of chance, defined as a sequence of bets, observations, and rewards. After a brief introduction in Section 4.1 we show in Section 4.2 that if there is a profitable scheme at all, asymptotically the universal Λ_ξ scheme will also become profitable. We bound the time needed to reach the winning zone. It is proportional to the relative entropy of μ and ξ with a factor depending on the profit range and the average profit. Section 4.3 presents a numerical example and Section 4.4 attempts to give an information theoretic interpretation of the result.

4.1 Introduction

Consider investing in the stock market. At time t an amount of money s_t is invested in portfolio y_t, where we have access to past knowledge $x_{<t}$ (e.g. charts). After our choice of investment we receive new information x_t, and the new portfolio value is r_t. The best we can expect is to have a probabilistic model μ of the behavior of the stock-market. The goal is to maximize the net μ-expected profit $p_t=r_t-s_t$. Nobody knows μ, but the assumption of all traders is that there *is* a computable, profitable μ they try to find or approximate. From Theorem 1 we know that Levin's universal prior $\xi_U(x_t|x_{<t})$ converges to any computable $\mu(x_t|x_{<t})$ with probability 1. If there is a computable, asymptotically profitable trading scheme at all, the Λ_ξ scheme should also be profitable in the long run. To get a practically useful, computable scheme we have to restrict $\mathcal{M}$ to a finite set of computable distributions, e.g. with bounded Levin complexity Kt (Li and Vitányi, 1997). Although convergence of ξ to μ is pleasing, what we are really interested in is whether Λ_ξ is asymptotically profitable and how long it takes to become profitable. This will be explored in the following.

4.2 Games of Chance

We use the loss bound (18) to estimate the time needed to reach the winning threshold when using Λ_ξ in a game of chance. We assume a game (or a sequence of possibly correlated games) which allows a sequence of bets and observations. In step t we bet, depending on the history $x_{<t}$, a certain amount of money s_t, take some action y_t, observe outcome x_t, and receive reward r_t. Our profit, which we want to maximize, is $p_t=r_t-s_t\in[p_{min},p_{max}]$, where $[p_{min},p_{max}]$ is the [minimal,maximal] profit per round and $p_\Delta:=p_{max}-p_{min}$ the profit range. The loss, which we want to minimize, can be defined as the negative scaled profit, $\ell_{x_t y_t}=(p_{max}-p_t)/p_\Delta\in[0,1]$. The probability of outcome x_t, possibly depending on the history $x_{<t}$, is $\mu(x_t|x_{<t})$. The total μ-expected profit when using scheme Λ_ρ is $P_n^{\Lambda_\rho}=np_{max}-p_\Delta L_n^{\Lambda_\rho}$. If

we knew μ, the optimal strategy to maximize our expected profit is just Λ_μ. We assume $P_n^{\Lambda_\mu}>0$ (otherwise there is no winning strategy at all, since $P_n^{\Lambda_\mu}\geq P_n^\Lambda\,\forall\Lambda$). Often we are not in the favorable position of knowing μ, but we know (or assume) that $\mu\in\mathcal{M}$ for some $\mathcal{M}$, for instance that μ is a computable probability distribution. From bound (18) we see that the average profit per round $\bar{p}_n^{\Lambda_\xi}:=\frac{1}{n}P_n^{\Lambda_\xi}$ of the universal Λ_ξ scheme converges to the average profit per round $\bar{p}_n^{\Lambda_\mu}:=\frac{1}{n}P_n^{\Lambda_\mu}$ of the optimal informed scheme, i.e. asymptotically we can make the same money even without knowing μ, by just using the universal Λ_ξ scheme. Bound (18) allows us to lower bound the universal profit $P_n^{\Lambda_\xi}$

$$P_n^{\Lambda_\xi} \;\geq\; P_n^{\Lambda_\mu}-p_\Delta D_n-\sqrt{4(np_{max}-P_n^{\Lambda_\mu})p_\Delta D_n+p_\Delta^2D_n^2}. \tag{19}$$

The time needed for Λ_ξ to perform well can also be estimated. An interesting quantity is the expected number of rounds needed to reach the winning zone. Using $P_n^{\Lambda_\mu}>0$ one can show that the r.h.s. of (19) is positive if, and only if

$$n \;>\; \frac{2p_\Delta(2p_{max}-\bar{p}_n^{\Lambda_\mu})}{(\bar{p}_n^{\Lambda_\mu})^2}\cdot D_n. \tag{20}$$

Theorem 3 (Time to Win) *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. In step t we make a bet, depending on the history $x_{<t}$, take some action y_t, and observe outcome x_t. Our net profit is $p_t\in[p_{max}-p_\Delta,p_{max}]$. The Λ_ρ-system (15) acts as to maximize the ρ-expected profit. $P_n^{\Lambda_\rho}$ is the total and $\bar{p}_n^{\Lambda_\rho}=\frac{1}{n}P_n^{\Lambda_\rho}$ is the average expected profit of the first n rounds. For the universal Λ_ξ and for the optimal informed Λ_μ prediction scheme the following holds:*

$$\begin{array}{ll} i) & \bar{p}_n^{\Lambda_\xi} \;=\; \bar{p}_n^{\Lambda_\mu}-O(n^{-1/2})\longrightarrow \bar{p}_n^{\Lambda_\mu} \quad \textit{for} \quad n\to\infty \\ ii) & n\;>\left(\frac{2p_\Delta}{\bar{p}_n^{\Lambda_\mu}}\right)^2\cdot b_\mu \quad\wedge\quad \bar{p}_n^{\Lambda_\mu}>0 \quad\Longrightarrow\quad \bar{p}_n^{\Lambda_\xi}>0 \end{array}$$

where $b_\mu=\ln w_\mu^{-1}$ with w_μ being the weight (1) of μ in ξ in the discrete case (and b_μ as in Theorem 8 in the continuous case).

By dividing (19) by n and using $D_n\leq b_\mu$ (4) we see that the leading order of $\bar{p}_n^{\Lambda_\xi}-\bar{p}_n^{\Lambda_\mu}$ is bounded by $\sqrt{4p_\Delta p_{max}b_\mu/n}$, which proves ($i$). The condition in ($ii$) is actually a weakening of (20). $P_n^{\Lambda_\xi}$ is trivially positive for $p_{min}>0$, since in this wonderful case *all* profits are positive. For negative p_{min} the condition of (ii) implies (20), since $p_\Delta>p_{max}$, and (20) implies positive (19), i.e. $P_n^{\Lambda_\xi}>0$, which proves (ii).

If a winning strategy Λ with $\bar{p}_n^\Lambda>\varepsilon>0$ exists, then Λ_ξ is asymptotically also a winning strategy with the same average profit.

4.3 Example

Let us consider a game with two dice, one with two black and four white faces, the other with four black and two white faces. The dealer who repeatedly throws the dice uses one or the other die according to some deterministic rule, which correlates the throws (e.g. the first die could be used in round t iff the t^{th} digit of π is 7). We can bet on black or white; the stake s is 3$ in every round; our return r is 5$ for every correct prediction.

The profit is $p_t = r\delta_{x_t y_t} - s$. The coloring of the dice and the selection strategy of the dealer unambiguously determine μ. $\mu(x_t|x_{<t})$ is $\frac{1}{3}$ or $\frac{2}{3}$ depending on which die has been chosen. One should bet on the more probable outcome. If we knew μ the expected profit per round would be $\bar{p}_n^{\Lambda_\mu} = p_n^{\Lambda_\mu} = \frac{2}{3}r - s = \frac{1}{3}\$ > 0$. If we don't know μ we should use Levin's universal prior with $D_n \leq b_\mu = K(\mu)\cdot\ln 2$, where $K(\mu)$ is the length of the shortest program coding μ (see Section 2.6). Then we know that betting on the outcome with higher ξ probability leads asymptotically to the same profit (Theorem 3(i)) and Λ_ξ reaches the winning threshold no later than $n_{thresh} = 900\ln 2\cdot K(\mu)$ (Theorem 3(ii)) or sharper $n_{thresh} = 330\ln 2\cdot K(\mu)$ from (20), where $p_{max} = r - s = 2\$$ and $p_\Delta = r = 5\$$ have been used.

If the die selection strategy reflected in μ is not too complicated, the Λ_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 (see next subsection). Furthermore, this is only a bound for the turnaround value of n_{thresh}. The true expected turnaround n might be smaller. However, for every game for which there exists a computable winning strategy with $\bar{p}_n^\Lambda > \varepsilon > 0$, Λ_ξ is guaranteed to get into the winning zone for some $n \sim K(\mu)$.

4.4 Information-Theoretic Interpretation

We try to give an intuitive explanation of Theorem 3(ii). We know that $\xi(x_t|x_{<t})$ converges to $\mu(x_t|x_{<t})$ for $t \to \infty$. In a sense Λ_ξ learns μ from past data $x_{<t}$. The information content in μ relative to ξ is $D_\infty/\ln 2 \leq b_\mu/\ln 2$. One might think of a Shannon-Fano prefix code of $\nu \in \mathcal{M}$ of length $\lceil b_\nu/\ln 2\rceil$, which exists since the Kraft inequality $\sum_\nu 2^{-\lceil b_\nu/\ln 2\rceil} \leq \sum_\nu w_\nu \leq 1$ is satisfied. $b_\mu/\ln 2$ bits have to be learned before Λ_ξ can be as good as Λ_μ. In the worst case, the only information conveyed by x_t is in form of the received profit p_t. Remember that we always know the profit p_t before the next cycle starts.

Assume that the distribution of the profits in the interval $[p_{min}, p_{max}]$ is mainly due to noise, and there is only a small informative signal of amplitude $\bar{p}_n^{\Lambda_\mu}$. To reliably determine the sign of a signal of amplitude $\bar{p}_n^{\Lambda_\mu}$, disturbed by noise of amplitude p_Δ, we have to resubmit a bit $O((p_\Delta/\bar{p}_n^{\Lambda_\mu})^2)$ times (this reduces the standard deviation below the signal amplitude $\bar{p}_n^{\Lambda_\mu}$). To learn μ, $b_\mu/\ln 2$ bits have to be transmitted, which requires $n \geq O((p_\Delta/\bar{p}_n^{\Lambda_\mu})^2)\cdot b_\mu/\ln 2$ cycles. This expression coincides with the condition in (ii). Identifying the signal amplitude with $\bar{p}_n^{\Lambda_\mu}$ is the weakest part of this consideration, as we have no argument why this should be true. It may be interesting to make the analogy more rigorous, which may also lead to a simpler proof of (ii) not based on bounds (18) with their rather complex proofs.

5. Optimality Properties

In this section we discuss the quality of the universal predictor and the bounds. In Section 5.1 we show that there are $\mathcal{M}$ and $\mu \in \mathcal{M}$ and weights w_ν such that the derived error bounds are tight. This shows that the error bounds cannot be improved in general. In Section 5.2 we show Pareto-optimality of ξ in the sense that there is no other predictor which performs at least as well in all environments $\nu\in\mathcal{M}$ and strictly better in at least one. Optimal predictors can always be based on mixture distributions ξ. This still leaves open how to choose the weights. In Section 5.3 we give an Occam's razor argument that the choice $w_\nu=2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program describing ν is optimal.

5.1 Lower Error Bound

We want to show that there exists a class $\mathcal{M}$ of distributions such that *any* predictor Θ ignorant of the distribution $\mu\in\mathcal{M}$ from which the observed sequence is sampled 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, $E_\infty^\Theta\geq E_n^\Theta=n=\log_2|\mathcal{M}|$ is a lower worst case bound for every predictor Θ, (this includes Θ_ξ, of course). This shows that the upper bound $E_\infty^{\Theta_\xi}\leq\log_2|\mathcal{M}|$ for uniform w obtained in the discussion after Theorem 2 is sharp. In the general probabilistic case we can show by a similar argument that the upper bound of Theorem 2 is sharp for Θ_ξ and "static" predictors, and sharp within a factor of 2 for general predictors. We do not know whether the factor two gap can be closed.

Theorem 4 (Lower Error Bound) *For every n there is an $\mathcal{M}$ and $\mu\in\mathcal{M}$ and weights w_ν such that*

$$(i) \qquad e_t^{\Theta_\xi}-e_t^{\Theta_\mu} \;=\; \sqrt{2s_t} \quad \textit{and} \quad E_n^{\Theta_\xi}-E_n^{\Theta_\mu} \;=\; S_n+\sqrt{4E_n^{\Theta_\mu}S_n+S_n^2}$$

where $E_n^{\Theta_\xi}$ and $E_n^{\Theta_\mu}$ are the total expected number of errors of Θ_ξ and Θ_μ, and s_t and S_n are defined in (5). More generally, the equalities hold for any "static" *deterministic predictor θ for which y_t^Θ is independent of $x_{<t}$. For every n and* arbitrary *deterministic predictor Θ, there exists an $\mathcal{M}$ and $\mu\in\mathcal{M}$ such that*

$$(ii) \qquad e_t^{\Theta}-e_t^{\Theta_\mu} \;\geq\; \tfrac{1}{2}\sqrt{2s_t(x_{<t})} \quad \textit{and} \quad E_n^{\Theta}-E_n^{\Theta_\mu} \;\geq\; \tfrac{1}{2}[S_n+\sqrt{4E_n^{\Theta_\mu}S_n+S_n^2}]$$

Proof. (i) 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 $0<\varepsilon\leq\frac{1}{2}$. That is

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

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 (independent of history $x_{<t}$). Since we want lower bounds we seek 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 $e_t^\Theta - e_t^{\Theta_\mu}=\sqrt{2s_t}$ for static Θ. 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 in (i) for static Θ. We can choose[6] $y_t^{\Theta_\xi}\equiv 0$ to be a static predictor. Together this shows (i).

(ii) For non-static predictors, $a_t=1-y_t^\Theta$ in the proof of (i) depends on $x_{<t}$, which is not allowed. For general, but fixed a_t we have $e_t^\Theta(x_{<t})=1-\mu_{a_t}(y_t^\Theta)$. This quantity may assume any value between $\frac{1}{2}(1-\varepsilon)$ and $\frac{1}{2}(1+\varepsilon)$, when averaged over $x_{<t}$, and is, hence of little direct help. But if we additionally average the result also over all environments μ_a, we get

$$< E_n^\Theta >_a \;=< \sum_{t=1}^n \mathbf{E}[e_t^\Theta(x_{<t})] >_a \;=\; \sum_{t=1}^n \mathbf{E}[< e_t^\Theta(x_{<t}) >_a] \;=\; \sum_{t=1}^n \mathbf{E}[\tfrac{1}{2}] = \tfrac{1}{2}n$$

whatever Θ is chosen: a sort of No-Free-Lunch theorem (Wolpert and Macready, 1997), stating that on *uniform* average all predictors perform equally well/bad. The expectation of E_n^Θ w.r.t. a can only be $\frac{1}{2}n$ if $E_n^\Theta\geq\frac{1}{2}n$ for some a. Fixing such an a and choosing $\mu=\mu_a$ we get $E_n^\Theta - E_n^{\Theta_\mu}\geq\frac{1}{2}n\varepsilon=\frac{1}{2}[S_n+\sqrt{4E_n^{\Theta_\mu}S_n+S_n^2}]$, and similarly $e_n^\Theta - e_n^{\Theta_\mu}\geq\frac{1}{2}\varepsilon=\frac{1}{2}\sqrt{2s_t(x_{<t})}$. □

[6] This choice may be made unique by slightly non-uniform $w_{\mu_a}=\prod_{t=1}^n[\frac{1}{2}+(\frac{1}{2}-a_t)\delta]$ with $\delta\ll 1$.

Since for binary alphabet $s_t = \frac{1}{2}a_t^2$, Theorem 4 also holds with s_t replaced by $\frac{1}{2}a_t^2$ and S_n replaced by V_n. 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 bound (18) can also not be improved. Note that the bounds are tight even when $\mathcal{M}$ is restricted to Markov or i.i.d. environments, since the presented counterexample is i.i.d.

A set $\mathcal{M}$ independent of n 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 $E_n^{\Theta_\xi} - E_n^{\Theta_{\mu_1}} \sim \frac{1}{\ln n}\sqrt{E_n^{\Theta_\mu}\ln w_{\mu_1}^{-1}}$.

Unfortunately there are many important special cases for which the loss bound (18) is not tight. For continuous $\mathcal{Y}$ and logarithmic or quadratic loss function, for instance, one can show that the regret $L_\infty^{\Lambda_\xi} - L_\infty^{\Lambda_\mu} \leq \ln w_\mu^{-1} < \infty$ is finite (Hutter, 2003a). For arbitrary loss function, but μ bounded away from certain critical values, the regret is also finite. For instance, consider the special error-loss, binary alphabet, and $|\mu(x_t|x_{<t}) - \frac{1}{2}| > \varepsilon$ for all t and x. Θ_μ predicts 0 if $\mu(0|x_{<t}) > \frac{1}{2}$. If also $\xi(0|x_{<t}) > \frac{1}{2}$, then Θ_ξ makes the same prediction as Θ_μ, for $\xi(0|x_{<t}) < \frac{1}{2}$ the predictions differ. In the latter case $|\xi(0|x_{<t}) - \mu(0|x_{<t})| > \varepsilon$. Conversely for $\mu(0|x_{<t}) < \frac{1}{2}$. So in any case $e_t^{\Theta_\xi} - e_t^{\Theta_\mu} \leq \frac{1}{\varepsilon^2}[\xi(x_t|x_{<t}) - \mu(x_t|x_{<t})]^2$. Using (7) and Theorem 1 we see that $E_\infty^{\Theta_\xi} - E_\infty^{\Theta_\mu} \leq \frac{1}{\varepsilon^2}\ln w_\mu^{-1} < \infty$ is finite too. Nevertheless, Theorem 4 is important as it tells us that bound (18) can only be strengthened by making further assumptions on ℓ or $\mathcal{M}$.

5.2 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 a 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 most performance measures studied in this work.

Definition 5 (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 6 (Pareto Optimality) *The universal prior ξ is Pareto-optimal w.r.t. the instantaneous and total squared distances s_t and S_n (5), entropy distances d_t and D_n (4), errors e_t and E_n (7), and losses l_t and L_n (16).*

Proof. We first prove Theorem 6 for the instantaneous expected loss l_t. We need the more general ρ-expected instantaneous losses

$$l_{t\rho}^{\Lambda}(x_{<t}) \;:=\; \sum_{x_t}\rho(x_t|x_{<t})\ell_{x_t y_t^\Lambda} \tag{21}$$

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[7] Λ with $l_{t\nu}^{\Lambda}\leq l_{t\nu}^{\Lambda_\xi}$ for all $\nu\in\mathcal{M}$ and strict inequality for some ν. Implicit to this assumption is the assumption that $l_{t\nu}^{\Lambda}$ and $l_{t\nu}^{\Lambda_\xi}$ exist. $l_{t\nu}^{\Lambda}$ exists iff $\nu(x_t|x_{<t})$ exists iff $\nu(x_{<t})>0$ iff $w_\nu(x_{<t})>0$.

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

The two equalities follow from inserting (3) into (21). 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 (15) the ξ-expected loss (similarly to (17)). The contradiction $l_{t\xi}^{\Lambda}<l_{t\xi}^{\Lambda}$ 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

$$L_{n\rho}^{\Lambda} \;:=\; \sum_{t=1}^{n}\sum_{x_{<t}}\rho(x_{<t})l_{t\rho}^{\Lambda}(x_{<t}) \;=\; \sum_{t=1}^{n}\sum_{x_{1:t}}\rho(x_{1:t})\ell_{x_t y_t^\Lambda}$$

for a predictor Λ, and by assuming $L_{n\nu}^{\Lambda}\leq L_{n\nu}^{\Lambda_\xi}$ for all ν and strict inequality for some ν, from which we get the contradiction $L_{n\xi}^{\Lambda}=\sum_\nu w_\nu L_{n\nu}^{\Lambda}<\sum_\nu w_\nu L_{n\nu}^{\Lambda_\xi}=L_{n\xi}^{\Lambda_\xi}\leq L_{n\xi}^{\Lambda}$ 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 $w_\nu=w_\nu(x_{<t})\geq 0$ 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

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

where we have used $\sum_\nu w_\nu=1$ and $\sum_\nu w_\nu y_{\nu i}=z_i$ (3). $0\geq\sum_i(r_i-z_i)^2\geq 0$ 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

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

[7] According to Definition 5 we should look for a ρ, but for each deterministic predictor Λ there exists a ρ with $\Lambda=\Lambda_\rho$.

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 ($y_t^{\Lambda_\rho}=y_t^{\Lambda_\xi}$) (if ties in argmax_{y_t} are broken in a consistent way), and this is all that counts.

Note that ξ is *not* Pareto-optimal w.r.t. to *all* performance measures. Counterexamples can be given for $\mathcal{F}(\nu,\xi)=\sum_{x_t}|\nu(x_t|x_{<t})-\xi(x_t|x_{<t})|^\alpha$ for $\alpha\neq 2$, e.g. a_t and V_n. Nevertheless, for all performance 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 welcome property of being Pareto-optimal.

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. One can show that such a "balanced" improvement is (not) possible in the following sense: For instance, by using $\tilde\Lambda$ instead of Λ_ξ, the w_ν-expected loss may increase or decrease, i.e. $L_{n\nu}^{\tilde\Lambda}\lessgtr L_{n\nu}^{\Lambda_\xi}$, but on average, the loss can not decrease, since $\sum_\nu w_\nu[L_{n\nu}^{\tilde\Lambda}-L_{n\nu}^{\Lambda_\xi}]=L_{n\xi}^{\tilde\Lambda}-L_{n\xi}^{\Lambda_\xi}\geq 0$, where we have used linearity of $L_{n\rho}$ in ρ and $L_{n\xi}^{\Lambda_\xi}\leq L_{n\xi}^{\Lambda}$. In particular, a loss increase by an amount Δ_λ in only a single environment λ, can cause a decrease by at most the same amount times a factor $\frac{w_\lambda}{w_\eta}$ in some other environment η, i.e. a loss increase can only cause a smaller decrease in simpler environments, but a scaled decrease in more complex environments. We do not regard this as a "No Free Lunch" (NFL) theorem (Wolpert and Macready, 1997). Since most environments are completely random, a small concession on the loss in each of these completely uninteresting environments provides enough margin to yield distinguished performance on the few non-random (interesting) environments. Indeed, we would interpret the NFL theorems for optimization and search by Wolpert and Macready (1997) as balanced Pareto-optimality results. Interestingly, whereas for prediction only Bayes-mixes are Pareto-optimal, for search and optimization every algorithm is Pareto-optimal.

The term *Pareto-optimal* has been taken from the economics literature, but there is the closely related notion of unimprovable strategies (Borovkov and Moullagaliev, 1998) or admissible estimators (Ferguson, 1967) in statistics for parameter estimation, for which results similar to Theorem 6 exist. Furthermore, it would be interesting to show under which conditions, the class of *all* Bayes-mixtures (i.e. with all possible values for the weights) is complete in the sense that *every* Pareto-optimal strategy can be based on a Bayes-mixture. Pareto-optimality is sort of a minimal demand on a prediction scheme aiming to be optimal. A scheme which is not even Pareto-optimal cannot be regarded as optimal in any reasonable sense. Pareto-optimality of ξ w.r.t. most performance measures emphasizes the distinctiveness of Bayes-mixture strategies.

5.3 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 2.6 we have discussed the set $\mathcal{M}$ of all enumerable semimeasures which we regarded as sufficiently large from a computational point of view (see Schmidhuber 2002a, Hutter 2003b 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 beliefs) 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 $S_{n\nu}^{\xi_w} \leq \ln w_\nu^{-1}$ and for the loss regret $L_{n\nu}^{\Lambda_{\xi_w}} - L_{n\nu}^{\Lambda_\nu} \leq 2\ln w_\nu^{-1} + 2\sqrt{\ln w_\nu^{-1} L_{n\nu}^{\Lambda_\nu}}$. 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.[8] In the following we will argue that in the class of enumerable weight functions with short program there is an optimal compromise, namely $w_\nu = 2^{-K(\nu)}$.

Consider the class of enumerable weight functions with short programs, namely $\mathcal{V} := \{v_{(.)} : \mathcal{M} \to I\!R^+ \text{ with } \sum_\nu v_\nu \leq 1 \text{ and } K(v) = O(1)\}$. Let $w_\nu := 2^{-K(\nu)}$ and $v_{(.)} \in \mathcal{V}$. Corollary 4.3.1 of Li and Vitányi (1997, 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 one in the bounds when using ξ_w instead of ξ_v. In using ξ_w we are on the safe side, getting (within $O(1)$) best bounds for *all* environments.

Theorem 7 (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 loss bounds within an additive (to $\ln w_\mu^{-1}$) constant in all enumerable environments.*

Since the above justifies the use of ξ_w, and ξ_w assigns high probability to an environment if and only if it has low (Kolmogorov) complexity, one may interpret the result as a justification of Occam's razor.[9] 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.

On the other hand, this choice for w_ν is not unique (even not within a constant factor). For instance, for $0 < v_\nu = O(1)$ for $\nu = \xi_w$ and v_ν arbitrary (e.g. 0) for all other ν, the obvious dominance $\xi_\nu \geq v_\nu \nu$ can be improved to $\xi_\nu \geq c \cdot w_\nu \nu$, where $0 < c = O(1)$ is a universal constant. Indeed, formally every choice of weights $v_\nu > 0\, \forall\nu$ leads within a multiplicative constant to the same universal distribution, but this constant is not necessarily of "acceptable" size. Details will be presented elsewhere.

[8] All results in this paper have been 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 (Theorem 1), the error bound (Theorem 2), the loss bound (18), 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.

[9] The *only if* direction has been shown by a more easy and direct argument by Schmidhuber (2002a).

6. Miscellaneous

This section discusses miscellaneous topics. Section 6.1 generalizes the setup to continuous probability classes $\mathcal{M}=\{\mu_\theta\}$ consisting of continuously parameterized distributions μ_θ with parameter $\theta\in I\!R^d$. Under certain smoothness and regularity conditions a bound for the relative entropy between μ and ξ, which is central for all presented results, can still be derived. The bound depends on the Fisher information of μ and grows only logarithmically with n, the intuitive reason being the necessity to describe θ to an accuracy $O(n^{-1/2})$. Section 6.2 describes two ways of using the prediction schemes for partial sequence prediction, where not every symbol needs to be predicted. Performing and predicting a sequence of independent experiments and online learning of classification tasks are special cases. In Section 6.3 we compare the universal prediction scheme studied here to the popular predictors based on expert advice (PEA) (Littlestone and Warmuth, 1989, Vovk, 1992, Littlestone and Warmuth, 1994, Cesa-Bianchi et al., 1997, Haussler et al., 1998, Kivinen and Warmuth, 1999). Although the algorithms, the settings, and the proofs are quite different, the PEA bounds and our error bound have the same structure. Finally, in Section 6.4 we outline possible extensions of the presented theory and results, including infinite alphabets, delayed and probabilistic prediction, active systems influencing the environment, learning aspects, and a unification with PEA.

6.1 Continuous Probability Classes $\mathcal{M}$

We have considered thus far countable probability classes $\mathcal{M}$, which makes sense from a computational point of view as emphasized in Section 2.6. On the other hand in statistical parameter estimation one often has a continuous hypothesis class (e.g. a Bernoulli(θ) process with unknown $\theta\in[0,1]$). Let

$$\mathcal{M} \;:=\; \{\mu_\theta : \theta\in\Theta\subseteq I\!R^d\}$$

be a family of probability distributions parameterized by a d-dimensional continuous parameter θ. Let $\mu\equiv\mu_{\theta_0}\in\mathcal{M}$ be the true generating distribution and θ_0 be in the interior of the compact set Θ. We may restrict $\mathcal{M}$ to a countable dense subset, like $\{\mu_\theta\}$ with computable (or rational) θ. If θ_0 is itself a computable real (or rational) vector then Theorem 1 and bound (18) apply. From a practical point of view the assumption of a computable θ_0 is not so serious. It is more from a traditional analysis point of view that one would like quantities and results depending smoothly on θ and not in a weird fashion depending on the computational complexity of θ. For instance, the weight $w(\theta)$ is often a continuous probability density

$$\xi(x_{1:n}) \;:=\; \int_\Theta d\theta\, w(\theta)\cdot\mu_\theta(x_{1:n}), \qquad \int_\Theta d\theta\, w(\theta) \;=\; 1, \qquad w(\theta) \;\geq\; 0. \tag{22}$$

The most important property of ξ used in this work was $\xi(x_{1:n})\geq w_\nu\cdot\nu(x_{1:n})$ which has been obtained from (1) by dropping the sum over ν. The analogous construction here is to restrict the integral over Θ to a small vicinity N_δ of θ. For sufficiently smooth μ_θ and $w(\theta)$ we expect $\xi(x_{1:n})\gtrsim|N_{\delta_n}|\cdot w(\theta)\cdot\mu_\theta(x_{1:n})$, where $|N_{\delta_n}|$ is the volume of N_{δ_n}. This in turn leads to $D_n\lesssim\ln w_\mu^{-1}+\ln|N_{\delta_n}|^{-1}$, where $w_\mu:=w(\theta_0)$. N_{δ_n} should be the largest possible region in which lnμ_θ is approximately flat on average. The averaged instantaneous, mean, and total curvature matrices of lnμ are

$$\begin{aligned} \jmath_t(x_{<t}) &:= \mathbf{E}_t \nabla_\theta \ln\mu_\theta(x_t|x_{<t})\nabla_\theta^T \ln\mu_\theta(x_t|x_{<t})_{|\theta=\theta_0}, \qquad \bar\jmath_n \;:=\; \tfrac{1}{n}J_n \\ J_n &:= \sum_{t=1}^n \mathbf{E}_{<t}\jmath_t(x_{<t}) \;=\; \mathbf{E}_{1:n}\nabla_\theta \ln\mu_\theta(x_{1:n})\nabla_\theta^T \ln\mu_\theta(x_{1:n})_{|\theta=\theta_0} \end{aligned}$$

They are the Fisher information of μ and may be viewed as measures of the parametric complexity of μ_θ at $\theta=\theta_0$. The last equality can be shown by using the fact that the μ-expected value of $\nabla\ln\mu\cdot\nabla^T\ln\mu$ coincides with $-\nabla\nabla^T\ln\mu$ (since $\mathcal{X}$ is finite) and a similar equality as in (4) for D_n.

Theorem 8 (Continuous Entropy Bound) *Let μ_θ be twice continuously differentiable at $\theta_0\in\Theta\subseteq I\!R^d$ and $w(\theta)$ be continuous and positive at θ_0. Furthermore we assume that the inverse of the mean Fisher information matrix $(\bar\jmath_n)^{-1}$ exists, is bounded for $n\to\infty$, and is uniformly (in n) continuous at θ_0. Then the relative entropy D_n between $\mu\equiv\mu_{\theta_0}$ and ξ (defined in (22)) can be bounded by*

$$D_n \;:=\; \mathbf{E}_{1:n}\ln\tfrac{\mu(x_{1:n})}{\xi(x_{1:n})} \;\leq\; \ln w_\mu^{-1} + \tfrac{d}{2}\ln\tfrac{n}{2\pi} + \tfrac{1}{2}\ln\det\bar\jmath_n + o(1) \;=:\; b_\mu$$

where $w_\mu\equiv w(\theta_0)$ is the weight density (22) of μ in ξ and $o(1)$ tends to zero for $n\to\infty$.

Proof sketch. For independent and identically distributed distributions $\mu_\theta(x_{1:n})=\mu_\theta(x_1)\cdot...\cdot\mu_\theta(x_n)\,\forall\theta$ this bound has been proven by Clarke and Barron (1990, Theorem 2.3). In this case $J^{[CB90]}(\theta_0)\equiv\bar\jmath_n\equiv\jmath_n$ independent of n. For stationary (k^{th}-order) Markov processes $\bar\jmath_n$ is also constant. The proof generalizes to arbitrary μ_θ by replacing $J^{[CB90]}(\theta_0)$ with $\bar\jmath_n$ everywhere in their proof. For the proof to go through, the vicinity $N_{\delta_n}:=\{\theta:||\theta-\theta_0||_{\bar\jmath_n}\leq\delta_n\}$ of θ_0 must contract to a point set $\{\theta_0\}$ for $n\to\infty$ and $\delta_n\to 0$. $\bar\jmath_n$ is always positive semi-definite as can be seen from the definition. The boundedness condition of $\bar\jmath_n^{-1}$ implies a strictly positive lower bound independent of n on the eigenvalues of $\bar\jmath_n$ for all sufficiently large n, which ensures $N_{\delta_n}\to\{\theta_0\}$. The uniform continuity of $\bar\jmath_n$ ensures that the remainder $o(1)$ from the Taylor expansion of D_n is independent of n. Note that twice continuous differentiability of D_n at θ_0 (Clarke and Barron, 1990, Condition 2) follows for finite $\mathcal{X}$ from twice continuous differentiability of μ_θ. Under some additional technical conditions one can even prove an equality $D_n=\ln w_\mu^{-1}+\frac{d}{2}\ln\frac{n}{2\pi e}+\frac{1}{2}\ln\det\bar\jmath_n+o(1)$ for the i.i.d. case (Clarke and Barron, 1990, (1.4)), which is probably also valid for general μ. □

The $\ln w_\mu^{-1}$ part in the bound is the same as for countable $\mathcal{M}$. The $\frac{d}{2}\ln\frac{n}{2\pi}$ can be understood as follows: Consider $\theta\in[0,1)$ and restrict the continuous $\mathcal{M}$ to θ which are finite binary fractions. Assign a weight $w(\theta)\approx 2^{-l}$ to a θ with binary representation of length l. $D_n\lesssim l\cdot\ln 2$ in this case. But what if θ is not a finite binary fraction? A continuous parameter can typically be estimated with accuracy $O(n^{-1/2})$ after n observations. The data do not allow to distinguish a $\tilde\theta$ from the true θ if $|\tilde\theta-\theta|<O(n^{-1/2})$. There is such a $\tilde\theta$ with binary representation of length $l=\log_2 O(\sqrt{n})$. Hence we expect $D_n\lesssim\frac{1}{2}\ln n+O(1)$ or $\frac{d}{2}\ln n+O(1)$ for a d-dimensional parameter space. In general, the $O(1)$ term depends on the parametric complexity of μ_θ and is explicated by the third $\frac{1}{2}\ln\det\bar\jmath_n$ term in Theorem 8. See

Clarke and Barron (1990, p454) for an alternative explanation. Note that a uniform weight $w(\theta)=\frac{1}{|\Theta|}$ does not lead to a uniform bound unlike the discrete case. A uniform bound is obtained for Bernando's (or in the scalar case Jeffreys') reference prior $w(\theta)\sim\sqrt{\det\bar{\jmath}_\infty(\theta)}$ if $\jmath_\infty$ exists (Rissanen, 1996).

For a finite alphabet $\mathcal{X}$ we consider throughout the paper, $j_t^{-1}<\infty$ independent of t and $x_{<t}$ in case of i.i.d. sequences. More generally, the conditions of Theorem 8 are satisfied for the practically very important class of stationary (k-th order) finite-state Markov processes ($k=0$ is i.i.d.).

Theorem 8 shows that Theorems 1 and 2 are also applicable to the case of continuously parameterized probability classes. Theorem 8 is also valid for a mixture of the discrete and continuous cases $\xi=\sum_a\int d\theta\, w^a(\theta)\,\mu_\theta^a$ with $\sum_a\int d\theta\, w^a(\theta)=1$.

6.2 Further Applications

Partial sequence prediction. There are (at least) two ways to treat partial sequence prediction. With this we mean that not every symbol of the sequence needs to be predicted, say given sequences of the form $z_1x_1...z_nx_n$ we want to predict the $x's$ only. The first way is to keep the Λ_ρ prediction schemes of the last sections mainly as they are, and use a time dependent loss function, which assigns zero loss $\ell_{zy}^t\equiv 0$ at the z positions. Any dummy prediction y is then consistent with (15). The losses for predicting x are generally non-zero. This solution is satisfactory as long as the $z's$ are drawn from a probability distribution. The second (preferable) way does not rely on a probability distribution over the z. We replace all distributions $\rho(x_{1:n})$ ($\rho=\mu$, ν, ξ) everywhere by distributions $\rho(x_{1:n}|z_{1:n})$ conditioned on $z_{1:n}$. The $z_{1:n}$ conditions cause nowhere problems as they can essentially be thought of as fixed (or as oracles or spectators). So the bounds in Theorems 1...8 also hold in this case for all individual z's.

Independent experiments and classification. A typical experimental situation is a sequence of independent (i.i.d) experiments, predictions and observations. At time t one arranges an experiment z_t (or observes data z_t), then tries to make a prediction, and finally observes the true outcome x_t. Often one has a parameterized class of models (hypothesis space) $\mu_\theta(x_t|z_t)$ and wants to infer the true θ in order to make improved predictions. This is a special case of partial sequence prediction, where the hypothesis space $\mathcal{M}=\{\mu_\theta(x_{1:n}|z_{1:n})=\mu_\theta(x_1|z_1)\cdot...\cdot\mu_\theta(x_n|z_n)\}$ consists of i.i.d. distributions, but note that ξ is not i.i.d. This is the same setting as for on-line learning of classification tasks, where a $z\in\mathcal{Z}$ should be classified as an $x\in\mathcal{X}$.

6.3 Prediction with Expert Advice

There are two schools of universal sequence prediction: We considered expected performance bounds for Bayesian prediction based on mixtures of environments, as is common in information theory and statistics (Merhav and Feder, 1998). The other approach are predictors based on expert advice (PEA) with worst case loss bounds in the spirit of Littlestone, Warmuth, Vovk and others. We briefly describe PEA and compare both approaches. For a more comprehensive comparison see Merhav and Feder (1998). In the following we focus on topics not covered by Merhav and Feder (1998). PEA was invented by Littlestone and Warmuth (1989, 1994) and Vovk (1992) and further developed by Cesa-Bianchi et al. (1997), Haussler et al. (1998), Kivinen and Warmuth (1999) and by many others. Many variations known by many names (prediction/learning with expert advice, weighted majority/average, aggregating strategy, hedge algorithm, ...) have meanwhile been invented. Early works in this direction are Dawid (1984), Rissanen (1989). See Vovk (1999) for a review and further references. We describe the setting and basic idea of PEA for binary alphabet. Consider a finite binary sequence $x_1x_2...x_n\in\{0,1\}^n$ and a finite set $\mathcal{E}$ of experts $e\in\mathcal{E}$ making predictions x_t^e in the unit interval $[0,1]$ based on past observations $x_1x_2...x_{t-1}$. The loss of expert e in step t is defined as $|x_t-x_t^e|$. In the case of binary predictions $x_t^e\in\{0,1\}$, $|x_t-x_t^e|$ coincides with our error measure (7). The PEA algorithm $p_{\beta n}$ combines the predictions of all experts. It forms its own prediction[10] $x_t^p\in[0,1]$ according to some weighted average of the expert's predictions x_t^e. There are certain update rules for the weights depending on some parameter β. Various bounds for the total loss $L_p(\mathbf{x}):=\sum_{t=1}^n|x_t-x_t^p|$ of PEA in terms of the total loss $L_\varepsilon(\mathbf{x}):=\sum_{t=1}^n|x_t-x_t^\varepsilon|$ of the best expert $\varepsilon\in\mathcal{E}$ have been proven. It is possible to fine tune β and to eliminate the necessity of knowing n in advance. The first bound of this kind has been obtained by Cesa-Bianchi et al. (1997):

$$L_p(\mathbf{x}) \;\leq\; L_\varepsilon(\mathbf{x})+2.8\ln|\mathcal{E}|+4\sqrt{L_\varepsilon(\mathbf{x})\ln|\mathcal{E}|}. \tag{23}$$

The constants 2.8 and 4 have been improved by Auer and Gentile (2000), Yaroshinsky and El-Yaniv (2001). The last bound in Theorem 2 with $S_n\leq D_n\leq\ln|\mathcal{M}|$ for uniform weights and with $E_n^{\Theta_\mu}$ increased to E_n^Θ reads

$$E_n^{\Theta_\xi} \;\leq\; E_n^\Theta+2\ln|\mathcal{M}|+2\sqrt{E_n^\Theta\ln|\mathcal{M}|}.$$

It has a quite similar structure as (23), although the algorithms, the settings, the proofs, and the interpretation are quite different. Whereas PEA performs well in any environment, but only relative to a given set of experts $\mathcal{E}$, our Θ_ξ predictor competes with the best possible Θ_μ predictor (and hence with any other Θ predictor), but only in expectation and for a given set of environments $\mathcal{M}$. PEA depends on the set of experts, Θ_ξ depends on the set of environments $\mathcal{M}$. The basic $p_{\beta n}$ algorithm has been extended in different directions: incorporation of different initial weights ($|\mathcal{E}|\rightsquigarrow w_\nu^{-1}$) (Littlestone and Warmuth, 1989, Vovk, 1992), more general loss functions (Haussler et al., 1998), continuous valued outcomes (Haussler et al., 1998), and multi-dimensional predictions (Kivinen and Warmuth, 1999) (but not yet for the absolute loss). The work of Yamanishi (1998) lies somewhat in between PEA and this work; "PEA" techniques are used to prove expected loss bounds (but only for sequences of independent symbols/experiments and limited classes of loss functions). Finally, note that the predictions of PEA are continuous. This is appropriate for weather forecasters which announce the probability of rain, but the *decision* to wear sunglasses or to take an umbrella is binary, and the suffered loss depends on this binary decision, and not on the probability estimate. It is possible to convert the continuous prediction of PEA into a probabilistic binary prediction by predicting 1 with probability $x_t^p\in[0,1]$. $|x_t-x_t^p|$ is then the probability of making an error. Note that the expectation

[10] The original PEA version (Littlestone and Warmuth, 1989) had discrete prediction $x_t^p\in\{0,1\}$ with (necessarily) twice as many errors as the best expert and is only of historical interest any more.

is taken over the probabilistic prediction, whereas for the deterministic Θ_ξ algorithm the expectation is taken over the environmental distribution μ. The multi-dimensional case (Kivinen and Warmuth, 1999) could then be interpreted as a (probabilistic) prediction of symbols over an alphabet $\mathcal{X}=\{0,1\}^d$, but error bounds for the absolute loss have yet to be proven. In work by Freund and Schapire (1997) the regret is bounded by $\ln|\mathcal{E}|+\sqrt{2\tilde{L}\ln|\mathcal{E}|}$ for arbitrary unit loss function and alphabet, where $\tilde{L}$ is an upper bound on L_ε, which has to be known in advance. It would be interesting to generalize PEA and bound (23) to arbitrary alphabet and weights and to general loss functions with probabilistic interpretation.

6.4 Outlook

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

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 $\mathcal{X}$ could be generalized to countable (numbers, words) or continuous (real or vector) alphabets. The presented theorems are independent of the size of $\mathcal{X}$ and hence should generalize to countably infinite alphabets by appropriately taking the limit $|\mathcal{X}|\to\infty$ and to continuous alphabets by a denseness or separability argument. Since the proofs are also independent of the size of $\mathcal{X}$ we may directly replace all finite sums over $\mathcal{X}$ by infinite sums or integrals and carefully check the validity of each operation. We expect all theorems to remain valid in full generality, except for minor technical existence and convergence constraints.

An infinite prediction space $\mathcal{Y}$ was no problem at all as long as we assumed the existence of $y_t^{\Lambda_\rho}\in\mathcal{Y}$ (15). In case $y_t^{\Lambda_\rho}\in\mathcal{Y}$ does not exist one may define $y_t^{\Lambda_\rho}\in\mathcal{Y}$ in a way to achieve a loss at most $\varepsilon_t=o(t^{-1})$ larger than the infimum loss. We expect a small finite correction of the order of $\varepsilon=\sum_{t=1}^\infty\varepsilon_t<\infty$ in the loss bounds somehow.

Delayed & probabilistic prediction. The Λ_ρ schemes and theorems may be generalized to delayed sequence prediction, where the true symbol x_t is given only in cycle $t+d$. A delayed feedback is common in many practical problems. We expect bounds with D_n replaced by $d\cdot D_n$. Further, the error bounds for the probabilistic suboptimal ξ scheme defined and analyzed by Hutter (2001b) can also be generalized to arbitrary alphabet.

More active systems. Prediction means guessing the future, but not influencing it. A small step in the direction of more active systems was to allow the Λ system to act and to receive a loss $\ell_{x_t y_t}$ depending on the action y_t and the outcome x_t. The probability μ is still independent of the action, and the loss function ℓ^t has to be known in advance. This ensures that the greedy strategy (15) is optimal. The loss function may be generalized to depend not only on the history $x_{<t}$, but also on the historic actions $y_{<t}$ 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 by Hutter (2001c). Pareto-optimality and asymptotic bounds are proven by Hutter (2002), but a lot remains to be done in the active case.

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 (Vitányi and Li, 2000) could lead to new insights. Also, the separation of noise from useful data, usually an important issue (Gács et al., 2001), did not play a role here. The attempt at an information theoretic interpretation of Theorem 3 may be made more rigorous in this or another way. In the end, this may lead to a simpler proof of Theorem 3 and maybe even for the loss bounds. A unified picture of the loss bounds obtained here and the loss bounds for predictors based on expert advice (PEA) could also be fruitful. Yamanishi (1998) used PEA methods to prove expected loss bounds for Bayesian prediction, so maybe the proof technique presented here could be used *vice versa* to prove more general loss bounds for PEA. Maximum-likelihood or MDL predictors may also be studied. For instance, $2^{-K(x)}$ (or some of its variants) is a close approximation of ξ_U, so one may think that predictions based on (variants of) K may be as good as predictions based on ξ_U, but it is easy to see that K completely fails for predictive purposes. Also, more promising variants like the monotone complexity Km and universal two-part MDL, both extremely close to ξ_U, fail in certain situations (Hutter, 2003c). Finally, the system should be applied to specific induction problems for specific $\mathcal{M}$ with computable ξ.

7. Summary

We compared universal predictions based on Bayes-mixtures ξ to the infeasible informed predictor based on the unknown true generating distribution μ. Our main focus was on a decision-theoretic setting, where each prediction $y_t \in \mathcal{X}$ (or more generally action $y_t \in \mathcal{Y}$) results in a loss $\ell_{x_t y_t}$ if x_t is the true next symbol of the sequence. We have shown that the Λ_ξ predictor suffers only slightly more loss than the Λ_μ predictor. We have shown that the derived error and loss bounds cannot be improved in general, i.e. without making extra assumptions on ℓ, μ, $\mathcal{M}$, or w_ν. Within a factor of 2 this is also 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 at least as well in all environments $\nu \in \mathcal{M}$ and strictly better in at least one. Optimal predictors can (in most cases) be based on mixture distributions ξ. Finally we gave an Occam's razor argument that the universal prior with weights $w_\nu = 2^{-K(\nu)}$ is optimal, where $K(\nu)$ is the Kolmogorov complexity of ν. 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 (Cesa-Bianchi and Lugosi, 2001). We also derived a bound for the relative entropy between ξ and μ in the case of a continuously parameterized family of environments, which allowed us to generalize the loss bounds to continuous $\mathcal{M}$. Furthermore, we discussed the duality between the Bayes-mixture and expert-mixture (PEA) approaches and results, classification tasks, games of chances, infinite alphabet, active systems influencing the environment, and others.

References

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

P. Auer and C. Gentile. Adaptive and self-confident on-line learning algorithms. In *Proceedings of the 13th Conference on Computational Learning Theory*, pages 107–117. Morgan Kaufmann, San Francisco, 2000.

A. A. Borovkov and A. Moullagaliev. *Mathematical Statistics.* Gordon & Breach, 1998.

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

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

B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–471, 1990.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

A. P. Dawid. Statistical theory. The prequential approach. *J.R. Statist. Soc. A*, 147:278–292, 1984.

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

T. S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York, 3rd edition, 1967.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

P. Gács, J. Tromp, and P. M. B. Vitányi. Algorithmic statistics. *IEEE Transactions on Information Theory*, 47(6):2443–2463, 2001.

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

D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

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, 2001a.

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

M. Hutter. Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. *Proceedings of the 12^{th} Eurpean Conference on Machine Learning (ECML-2001)*, pages 226–238, 2001c.

M. Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, Lecture Notes in Artificial Intelligence, pages 364–379, Sydney, Australia, 2002. Springer.

M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003a.

M. Hutter. On the existence and convergence of computable universal priors. In R. Gavaldá, K. P. Jantke, and E. Takimoto, editors, *Proceedings of the 14th International Conference on Algorithmic Learning Theory (ALT-2003)*, volume 2842 of *LNAI*, pages 298–312, Berlin, 2003b. Springer.

M. Hutter. Sequence prediction based on monotone complexity. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT-2003)*, Lecture Notes in Artificial Intelligence, pages 506–521, Berlin, 2003c. Springer.

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In P. Fischer and H. U. Simon, editors, *Proceedings of the 4th European Conference on Computational Learning Theory (Eurocolt-99)*, volume 1572 of *LNAI*, pages 153–167, Berlin, 1999. Springer.

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

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

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

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

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

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261, Research Triangle Park, North Carolina, 1989. IEEE.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6): 2124–2147, 1998.

J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publ. Co., 1989.

J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans on Information Theory*, 42(1):40–47, January 1996.

J. Schmidhuber. Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. *International Journal of Foundations of Computer Science*, 13 (4):587–612, 2002a.

J. Schmidhuber. The Speed Prior: a new simplicity measure yielding near-optimal computable predictions. In J. Kivinen and R. H. Sloan, editors, *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, Lecture Notes in Artificial Intelligence, pages 216–228, Sydney, Australia, July 2002b. Springer.

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

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

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

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.

V. G. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992.

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

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

K. Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*, 44:1424–1439, 1998.

R. Yaroshinsky and R. El-Yaniv. Smooth online learning of expert advice. Technical report, Technion, Haifa, Israel, 2001.

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