Technical Report IDSIA-02-02 – 18 February 2002

Optimality of Universal Bayesian Sequence Prediction for General Loss and Alphabet

Marcus Hutter

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

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

Abstract

The Bayesian framework is ideally suited for induction problems. The probability of observing x_t at time t, given past observations $x_1...x_{t-1}$ can be computed with Bayes' rule if the true generating distribution μ of the sequences $x_1x_2x_3...$ is known. The problem, however, is that in many cases one does not even have a reasonable guess of the true distribution. In order to overcome this problem a universal (or mixture) distribution ξ is defined as a weighted sum or integral of distributions $\nu \in \mathcal{M}$, where $\mathcal{M}$ is any countable or continuous set of distributions including μ. This is a generalization of Solomonoff induction, in which $\mathcal{M}$ is the set of all enumerable semi-measures. It is shown for several performance measures that using the universal ξ as a prior is nearly as good as using the unknown true distribution μ. In a sense, this solves the problem of the unknown prior in a universal way. All results are obtained for general finite alphabet. Convergence of ξ to μ in a conditional mean squared sense and of $\xi/\mu \to 1$ with μ probability 1 is proven. The number of additional errors E_ξ made by the optimal universal prediction scheme based on ξ minus the number of errors E_μ of the optimal informed prediction scheme based on μ is proven to be bounded by $O(\sqrt{E_\mu})$. The prediction framework is generalized to arbitrary loss functions. A system is allowed to take an action y_t, given $x_1...x_{t-1}$ and receives loss $\ell_{x_t y_t}$ if x_t is the next symbol of the sequence. No assumptions on ℓ are necessary, besides boundedness. Optimal universal Λ_ξ and optimal informed Λ_μ prediction schemes are defined and the total loss of Λ_ξ is bounded in terms of the total loss of Λ_μ, similar to the error bounds. We show that the bounds are tight and that no other predictor can lead to smaller bounds. Furthermore, for various performance measures we show Pareto-optimality of ξ in the sense that there is no other predictor which performs better or equal in all environments $\nu \in \mathcal{M}$ and strictly better in at least one. So, optimal predictors can (w.r.t. to most performance measures in expectation) be based on the mixture ξ. Finally we give an Occam's razor argument that Solomonoff's choice $w_\nu \sim 2^{-K(\nu)}$ for the weights is optimal, where $K(\nu)$ is the length of the shortest program describing ν. Furthermore, games of chance, defined as a sequence of bets, observations, and rewards are studied. The average profit achieved by the Λ_ξ scheme rapidly converges to the best possible profit. The time needed to reach the winning zone is proportional to the relative entropy of μ and ξ. The prediction schemes presented here are compared to the weighted majority algorithm(s). Although the algorithms, the settings, and the proofs are quite different the bounds of both schemes have a very similar structure. Extensions to infinite alphabets, partial, delayed and probabilistic prediction, classification, and more active systems are briefly discussed.

Contents

1 Introduction

1.1 Induction

Many problems are of 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, as a consequence of prediction, one tries to manipulate the future. All induction problems may 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 Bayes' 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?

1.2 Universal Sequence Prediction

In order to overcome the problem of the unknown true distribution, one can define a mixture distribution ξ as w_ν 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 [MF98]. Solomonoff [Sol64] had the idea to define a universal prior ξ as a weighted average over *all* (semi)computable probability distributions. Lower weights were assigned to more complex distributions. He unified Epicurus' principle of multiple explanations, Occams' razor [simplicity] principle and Bayes' rule into an elegant formal theory. If the environment possesses some effective structure at all, Solomonoff's posterior "finds" this structure, 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.

1.3 Contents

The main **new contributions** of this work are to

- generalize the convergence [Sol78, LV97] of ξ to μ (Section 3),
- derive general error and loss bounds measuring the performance of ξ relative to μ (Section 4), improving upon previous results [Hut01a, MF98],
- apply the results to games of chance (Section 5),

- show that the error/loss bounds are tight and that Solomonoff's universal prior is optimal (Section 6),
- generalize the bound in [CB90] on the relative entropy between ξ and μ for continuous i.i.d. probability classes $\mathcal{M}$ to the non-i.i.d. case (Section 7),
- compare the universal prediction scheme and its loss bounds to the weighted majority scheme and its loss bounds [Ces97] (Section 9).

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 2. Furthermore, an elementary proof of $\xi/\mu\to 1$ (not based on semi-martingales) including the convergence rate is given. 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 and finite-state (semi)measures.

Section 3 is essentially a generalization of the deterministic error bounds found in [Hut01a] from binary alphabet to a general finite alphabet $\mathcal{X}$. Theorem 3 bounds the number of additional errors ($E^{\Theta_\xi}-E^{\Theta_\mu}$) made by optimal universal predictor Θ_ξ, as compared to optimal informed prediction scheme Θ_μ by $O(\sqrt{E^{\Theta_\mu}})$. 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 in [Sol78, Hut01a].

Section 4 generalizes the prediction framework 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 proved. No assumptions on ℓ have to be made, besides boundedness. For unit loss ($0\leq\ell_{x_ty_t}\leq 1$) the loss bounds in Theorem 4 are essentially the same as the error bounds of Theorem 3 with *error* replaced by *loss*, but the proofs are much more involved. The bounds are compared to the loss bound obtained in [MF98]. Theorem 7 generalizes the bounds to non-unit and non-static loss functions. Convergence of the instantaneous losses are also studied. Some popular loss functions, including the absolute, square, logarithmic, and Hellinger loss are discussed.

Section 5 applies Theorem 7 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 8 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 6 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 better or equal in all environments $\nu \in \mathcal{M}$ and strictly better in at least one. Optimal predictors can always be based on a mixture distributions ξ. This still leaves open how to choose the weights. We give an Occam's razor argument that Solomonoff's choice $w_\nu = 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program describing ν is optimal.

Section 7 generalizes the setup to continuous probability classes $\mathcal{M} = \{\mu_\theta\}$ consisting of continuously parameterized distributions μ_θ with parameter $\theta \in \mathbb{R}^d$. 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 8 discusses further applications. 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.

Section 9 compares the universal prediction scheme studied here to the weighted majority (WM) algorithm(s) [LW89, Vov92, LW94, Ces97, HKW98, KW99]. WM combines forecasts of experts $e \in \mathcal{E}$ to form its own prediction. The number of prediction errors of WM are compared to the best expert in $\mathcal{E}$. No assumption is made on the distribution of the strings – the bounds are worst case bounds. Although the algorithms, the settings, and the proofs are quite different, the WM bounds and the last bound of Theorem 4 have the same structure.

Section 10 outlines possible extensions of the presented theory and results. They include infinite alphabets, delayed and probabilistic prediction, active systems influencing the environment, learning aspects, and a unification with WM.

Section 11 summarizes the results.

Appendices A-D contain some technical proofs.

1.4 Introductory References

There are good introductions and surveys of Solomonoff sequence prediction [LV92, LV97], inductive inference in general [AS83, Sol97, MF98], reasoning under uncertainty [Grü98], and competitive online statistics [Vov99], with interesting relations to this work. See Section 9 for some more details. This paper is more or less self-contained. Exceptions are Subsections 2.7 and 6.4 on Solomonoff mixtures, Section 7 on continuous classes $\mathcal{M}$, and Section 9 on WM.

2 Setup and Convergence

2.1 Random sequences

We denote strings over a finite alphabet $\mathcal{X}$ by $x_1x_2...x_n$ with $x_t \in \mathcal{X}$. We further use the abbreviations $x_{n:m} := x_n x_{n+1}...x_{m-1}x_m$ and $x_{<n} := x_1...x_{n-1}$. We use Greek letters for probability distributions (or measures). Let $\rho(x_1...x_n)$ be the probability that an (infinite) sequence starts with $x_1...x_n$:

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

where ϵ is the empty string. We also need conditional probabilities derived from Bayes' rule:

$$\rho(x_t|x_{<t}) = \rho(x_{1:t})/\rho(x_{<t}), \tag{2}$$

$$\rho(x_1...x_n) = \rho(x_1)\cdot\rho(x_2|x_1)\cdot...\cdot\rho(x_n|x_1...x_{n-1}). \tag{3}$$

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 Bayes' 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 $\sum_{t=1}^{\infty}\mathbf{E}[(z_t-z_*)^2] < \infty$. One can show that convergence in mean sum implies convergence with probability 1.[1] Actually it allows a much stronger conclusion; it gives the "speed" of convergence in the sense that the expected number of times t in which z_t deviates more than ε from z_* is finitely bounded $\mathbf{E}[(z_t-z_*)^2]/\varepsilon^2$.

In a more statistical language we have a sample space $\Omega=\mathcal{X}^\infty$ with elements $\omega=\omega_1\omega_2\omega_3...\in\Omega$ being infinite sequences over the finite alphabet $\mathcal{X}$. The cylinder sets $\Gamma_{x_{1:n}} := \{\omega : \omega_{1:n} = x_{1:n}\}$ are events. We define the σ-algebra $\mathcal{F}$ as the set generated from the cylinder sets by countable union. A probability measure μ is uniquely defined by giving its values $\mu(\Gamma_{x_{1:n}})$ on the cylinder sets, which we abbreviate by $\mu(x_{1:n})$. See [LV97, Doo53] or any other statistics book for a more thorough treatment.

Some expressions (like conditional or inverse probabilities) are undefined when μ gets zero. In this case one should restrict the analysis to the set of strings with non-zero μ-probability.

[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.

If we define the critical set $Z := \{\omega \in \mathcal{X}^\infty : \exists t : \mu(\omega_{1:t}) = 0\} = \bigcup_{t=1}^{\infty} \bigcup_{x_{1:t}:\mu(x_{1:t})=0} \Gamma_{x_{1:t}}$. Since Z is a countable (for discrete alphabet) union of cylinder sets $\Gamma_{x_{1:t}}$ of measure zero, Z itself is measurable with μ-measure zero. So all theorems proven with μ-probability 1 on $\Omega \setminus Z$ still hold on Ω with μ-probability 1, since $\mu(Z)=0$. In critical situations, sums over x have to be restricted to exclude Z. Some measures on Ω, especially ξ, defined in the next paragraph, deteriorate to semimeasures on $\Omega \setminus Z$. In order to keep the presentation simple, we will usually simply ignore these subtleties and proceed as if μ (and ξ) were always non-zero. Only in critical cases we use $\sum'$ do indicate a sum restricted to $\Omega \setminus Z$ and exploit

$$\mathbf{E}_t[..] = \sum_{x_t \in \mathcal{X}}{}' \mu(x_t|x_{<t})[..] \quad \text{with } \mu \text{ probability 1 (w.}\mu\text{.p.1).} \tag{4}$$

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 (as for the sunrise example) the true distribution is not known. Instead, the prediction is based on a guess ρ of μ. We expect that a predictor based on ρ performs well, if ρ is close to μ or converges, in a sense, to μ. Let $\mathcal{M} := \{\mu_1, \mu_2, ...\}$ be a finite or countable set of candidate probability distributions on strings. Results are generalized to continuous sets $\mathcal{M}$ in Section 7. We define a weighted average on $\mathcal{M}$

$$\xi(x_{1:n}) \;:=\; \sum_{\nu \in \mathcal{M}} w_\nu \cdot \nu(x_{1:n}), \qquad \sum_{\nu \in \mathcal{M}} w_\nu = 1, \quad w_\nu > 0. \tag{5}$$

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 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 $\mathcal{M}$

$$\xi(x_{1:n}) \;\geq\; w_\nu \cdot \nu(x_{1:n}) \quad \text{for all} \quad \nu \in \mathcal{M}. \tag{6}$$

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. Generic classes, especially where $\mathcal{M}$ contains *all* computable probability distributions, are discussed in Subsection 2.7. Generalizations to the case where $\mathcal{M}$ does not contain μ are briefly discussed in Subsection 2.6. In the next Subsection we motivate and in Subsection 2.5 we show the important property of ξ converging to the true distribution $\mu \in \mathcal{M}$ in a sense and, hence, might being a useful substitute for the true, but in general, unknown distribution μ.

[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*.

[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.

2.3 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:

$$\xi(x_t|x_{<t}) = \sum_{\nu\in\mathcal{M}} w_\nu(x_{<t})\nu(x_t|x_{<t}), \quad w_\nu(x_{1:t}) := w_\nu(x_{<t})\frac{\nu(x_t|x_{<t})}{\xi(x_t|x_{<t})}, \quad w_\nu(\varepsilon) := w_\nu. \quad (7)$$

The denominator just ensures correct normalization $\sum_\nu w_\nu(x_{1:t}) = 1$. By induction and Bayes' 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 (5) gives $\xi(x_t|x_{<t})$, which proves the equivalence of (5) and (7). The expressions (7) can be used to give an intuitive, but non-rigorous, argument why $\xi(x_t|x_{<t})$ converges to $\nu(x_t|x_{<t})$: The weight w_ν of ν in ξ increases/decreases if ν assigns a high/low probability to the new symbol x_t, given $x_{<t}$. For a μ-random sequence $x_{1:t}$, $\mu(x_{1:t}) \gg \nu(x_{1:t})$ if ν (significantly) differs from μ. We expect the total weight for all ν consistent with μ to converges to 1, and all other weights converge to 0 for $t\to\infty$. Therefore we expect $\xi(x_t|x_{<t})$ to converge to $\mu(x_t|x_{<t})$ for μ-random strings $x_{1:n}$.

Expressions (7) seem to be more suitable than (5) for studying convergence and loss bounds of the universal predictor ξ, but it will turn out that (6) is all we need, with the sole exception in the proof of Theorem 11. Probably (7) is useful when one tries to understand the learning aspect in ξ.

2.4 Distance Measures between Probability Distributions

We need several distance measures between vectors $\mathbf{y} = (y_i)$ and $\mathbf{z} = (z_i)$ in general, and probability distributions for which $y_i \geq 0$, $z_i \geq 0$, and $\sum_i y_i = \sum_i z_i = 1$ in particular, $i = \{1,...,N\}$. The absolute distance a, the quadratic or Euclidian distance s, the Hellinger distance h, and the relative entropy or Kullback-Leibler divergence d are defined as follows:

$$a(\mathbf{y},\mathbf{z}) := \sum_i |y_i - z_i|, \qquad s(\mathbf{y},\mathbf{z}) := \sum_i (y_i - z_i)^2, \quad (8)$$
$$h(\mathbf{y},\mathbf{z}) := \sum_i (\sqrt{y_i} - \sqrt{z_i})^2, \qquad d(\mathbf{y},\mathbf{z}) := \sum_i y_i \ln\frac{y_i}{z_i}.$$

The relative entropy is not a true distance measure, but for probability distributions, for which it is defined, it is at least non-negative and zero if and only if $\mathbf{y} = \mathbf{z}$. All bounds we prove in this work heavily rely on the following inequalities:

Lemma 1 (Entropy Inequalities) *Let $\{y_i\}$ and $\{z_i\}$ be two probability distributions, i.e. $y_i \geq 0$, $z_i \geq 0$, and $\sum_i y_i = \sum_i z_i = 1$ and f be a convex and even ($f(x) = f(-x)$) function with $f(0) \leq 0$, then the following inequalities hold:*

$$\tfrac{1}{2}\sum_i f(y_i - z_i) \overset{(f)}{\leq} f\Big(\sqrt{\tfrac{1}{2}\sum_i y_i \ln\frac{y_i}{z_i}}\Big), \qquad \sum_i (y_i - z_i)^2 \overset{(s)}{\leq} \sum_i y_i \ln\frac{y_i}{z_i}$$

$$\sum_i |y_i - z_i| \overset{(a)}{\leq} \sqrt{2\sum_i y_i \ln\frac{y_i}{z_i}}, \qquad \sum_i (\sqrt{y_i} - \sqrt{z_i})^2 \overset{(h)}{\leq} \sum_i y_i \ln\frac{y_i}{z_i}$$

Proofs are given in Appendix A. Inequality (Lemma 1s) is a generalization of the binary $N=2$ case used in [Sol78, Hut01a, LV97]. If we insert

$$\mathcal{X}=\{1,...,N\}, \quad N=|\mathcal{X}|, \quad i=x_t, \quad y_i=\mu(x_t|x_{<t}), \quad z_i=\xi(x_t|x_{<t}) \tag{9}$$

into (8) we get various *instantaneous distances* (at time t) between μ and ξ. If we take the expectation over $x_{<t}$ and sum over $t=1..n$, ($\sum_{t=1}^{n}\mathbf{E}_{<t}[...]$) we get various *total distances* between μ and ξ:

$$a_t(x_{<t}) := \sum_{x_t}\Big|\mu(x_t|x_{<t})-\xi(x_t|x_{<t})\Big|, \qquad A_n := \sum_{t=1}^{n}\mathbf{E}_{<t}a_t(x_{<t}) \tag{10}$$

$$s_t(x_{<t}) := \sum_{x_t}\Big(\mu(x_t|x_{<t})-\xi(x_t|x_{<t})\Big)^2, \qquad S_n := \sum_{t=1}^{n}\mathbf{E}_{<t}s_t(x_{<t}) \tag{11}$$

$$h_t(x_{<t}) := \sum_{x_t}\Big(\sqrt{\mu(x_t|x_{<t})}-\sqrt{\xi(x_t|x_{<t})}\Big)^2, \quad H_n := \sum_{t=1}^{n}\mathbf{E}_{<t}h_t(x_{<t}) \tag{12}$$

$$d_t(x_{<t}) := \sum_{x_t}\mu(x_t|x_{<t})\ln\frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})}, \qquad D_n := \sum_{t=1}^{n}\mathbf{E}_{<t}d_t(x_{<t}) \tag{13}$$

For D_n the following can be shown [Sol78, LV97]

$$D_n = \sum_{t=1}^{n}\mathbf{E}_{<t}d_t(x_{<t}) = \sum_{t=1}^{n}\mathbf{E}_{1:t}\ln\frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} = \tag{14}$$

$$= \mathbf{E}_{1:n}\ln\prod_{t=1}^{n}\frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} = \mathbf{E}_{1:n}\ln\frac{\mu(x_{1:n})}{\xi(x_{1:n})} \leq \ln w_\mu^{-1} =: b_\mu$$

In the first line we have inserted (13) and used Bayes' rule $\mu(x_{<t})\cdot\mu(x_t|x_{<t})=\mu(x_{1:t})$. Due to (1), we can further replace $\mathbf{E}_{1:t}$ by $\mathbf{E}_{1:n}$ as the argument of the logarithm is independent of $x_{t+1:n}$. The t sum can now be exchanged with the $\mathbf{E}_{1:n}$ expectation and transforms to a product inside the logarithm. In the last equality we have used the second form of Bayes' rule (3) for μ and ξ. Using universality (6) of ξ, i.e. $\ln\mu(x_{1:n})/\xi(x_{1:n})\leq\ln w_\mu^{-1}$ for $\mu\in\mathcal{M}$ yields the final inequality in (14).

2.5 Convergence of ξ to μ

Theorem 2 (Convergence of ξ to μ) *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:*

$$i)\quad \sum_{t=1}^{n}\mathbf{E}_{<t}\sum_{x'_t}\Big(\mu(x'_t|x_{<t})-\xi(x'_t|x_{<t})\Big)^2 \;\equiv\; S_n \;\leq\; D_n \;\leq\; \ln w_\mu^{-1} \;<\; \infty$$

$$ii)\quad \sum_{x'_t}\Big(\mu(x'_t|x_{<t})-\xi(x'_t|x_{<t})\Big)^2 \;\equiv\; s_t(x_{<t}) \;\leq\; d_t(x_{<t})\to 0 \quad \textit{for } t\to\infty \textit{ w.}\mu\textit{.p.1}$$

$$iii)\quad \xi(x'_t|x_{<t})-\mu(x'_t|x_{<t})\to 0 \quad \textit{for } t\to\infty \textit{ w.}\mu\textit{.p.1 (and i.m.s) for any } x'_t$$

$$iv)\quad \sum_{t=1}^{n}\mathbf{E}\left[\left(\sqrt{\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}}-1\right)^2\right] \;\leq\; H_n \;\leq\; D_n \;\leq\; \ln w_\mu^{-1} \;<\; \infty$$

$$v)\quad \sqrt{\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}}\to 1 \quad \textit{i.m.s} \quad \textit{and} \quad \frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}\to 1 \quad \textit{w.}\mu\textit{.p.1} \quad \textit{for } t\to\infty$$

$$vi)\quad a_t(x_{<t}) \;\leq\; \sqrt{2d_t(x_{<t})}, \qquad A_n \;\leq\; \sqrt{2nD_n},$$

where d_t *and* D_n *are the relative entropies (12), and* w_μ *is the weight (5) of* μ *in* ξ*.*

Proof: Inequality (ii) follows from the definitions (11) and (13) and from the entropy inequality $(1s)$. From the definition and finiteness of D_∞ (14), and from $d_t(x_{<t})\geq 0$ one sees that $d_t(x_{<t})\to 0$ for $t\to\infty$ w.μ.p.1. The inequality (i) follows from (ii) by taking the $\mathbf{E}_{<t}$ expectation and the $\sum_{t=1}^{n}$ sum. (iii) is a direct consequence of $(ii)/(i)$. The reason for the astonishing property of a single (universal) function ξ to converge to *any* $\mu\in\mathcal{M}$ lies in the fact that the sets of μ-random sequences differ for different μ. (iv) and (v) are related (but incomparable) convergence results to (i) and (iii). To prove (iv) we use the abbreviations $y_t=\mu(x_t|x_{<t})$ and $z_t=\xi(x_t|x_{<t})$.

$$\mathbf{E}_t\Big[\Big(\sqrt{\frac{z_t}{y_t}}-1\Big)^2\Big]=\sum_{x_t}{}'\mu(x_t|x_{<t})\Big(\sqrt{\frac{z_t}{y_t}}-1\Big)^2=\sum_{x_t}{}'(\sqrt{z_t}-\sqrt{y_t})^2\leq h_t(x_{<t})\leq d_t(x_{<t}). \quad (15)$$

The first equality holds w.μ.p.1 (4), the last two inequalities follow from (12) and $(1h)$. (iv) now follows by taking the $\mathbf{E}_{<t}$ expectation and the $\sum_{t=1}^{n}$ sum. (v) follows from (iv) by the definition of convergence i.m.s., which implies convergence w.μ.p.1. In (vi), $a_t\leq\sqrt{2d_t}$ immediately follows from inequality $(1a)$ and Definitions (10) and (13). $A_n\leq\sqrt{2nD_n}$ follows from

$$\tfrac{1}{n}A_n\equiv\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{<t}[a_t] \;\leq\; \frac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{<t}[\sqrt{2d_t}] \;\leq\; \frac{1}{n}\sum_{t=1}^{n}\sqrt{\mathbf{E}_{<t}[2d_t]} \;\leq\; \sqrt{\frac{1}{n}\sum_{t=1}^{n}\mathbf{E}_{<t}[2d_t]} \;\equiv\; \sqrt{\tfrac{2}{n}D_n} \quad (16)$$

where we have used Jensen's inequality for exchanging the averages ($\frac{1}{n}\sum_{t=1}^{n}$ and $\mathbf{E}_{<t}$) with the convex functions ($\sqrt{\ }$). □

Since the conditional probabilities are the basis of all prediction algorithms considered in this work, we expect a good prediction performance if we use ξ as a guess of μ. Performance measures are defined in the following sections.

The elementary proof for (v) given here does not rely on the semi-martingale convergence Theorem [Doo53, pp. 324–325] as the proof given in [LV97]. Furthermore, (iv) gives the "speed" of convergence. Note the subtle difference between (iii) and (v). If $x_{1:\infty}$ is a μ-random sequence, and $x'_{1:\infty}$ is *any* (possibly constant and not necessarily μ-random) sequence then $\mu(x'_t|x_{<t})-\xi(x'_t|x_{<t})$ converges to zero, but no statement is possible for $\xi(x'_t|x_{<t})/\mu(x'_t|x_{<t})$, since $\liminf\mu(x'_t|x_{<t})$ could be zero. On the other hand, if we stay *on* the μ-random sequence ($x'_{1:\infty}=x_{1:\infty}$), (v) shows that $\xi(x_t|x_{<t})/\mu(x_t|x_{<t})\to 1$ (whether $\inf\mu(x_t|x_{<t})$ tends to zero or not does not matter). Indeed, it is easy to give an example where $\xi(x'_t|x_{<t})/\mu(x'_t|x_{<t})$ diverges. If we choose

$$\mathcal{M}=\{\mu_1,\mu_2\},\quad \mu\equiv\mu_1,\quad \mu_1(1|x_{<t})=\tfrac{1}{2}t^{-3}\quad\text{and}\quad \mu_2(1|x_{<t})=\tfrac{1}{2}t^{-2}$$

the contribution of μ_2 to ξ causes ξ to fall off like $\mu_2\sim t^{-2}$, much slower than $\mu\sim t^{-3}$ causing the quotient to diverge:

$$\begin{aligned}
\mu_1(0_{1:n}) &= \prod_{t=1}^{n}(1-\tfrac{1}{2}t^{-3}) \stackrel{n\to\infty}{\longrightarrow} c_1=0.450...>0 \quad\Rightarrow\quad 0_{1:\infty}\text{ is a }\mu\text{-random sequence,}\\
\mu_2(0_{1:n}) &= \prod_{t=1}^{n}(1-\tfrac{1}{2}t^{-2}) \stackrel{n\to\infty}{\longrightarrow} c_2=0.358...>0 \quad\Rightarrow\quad \xi(0_{1:n})\to w_1c_1+w_2c_2=:c_\xi>0.\\
\xi(0_{<t}1) &= w_1\mu_1(1|0_{<t})\mu_1(0_{<t})+w_2\mu_2(1|0_{<t})\mu_2(0_{<t})\to\tfrac{1}{2}w_2c_2t^{-2}
\end{aligned}$$

$$\Longrightarrow\quad \xi(1|0_{<t})=\frac{\xi(0_{<t}1)}{\xi(0_{<t})}\to\frac{w_2c_2}{2c_\xi}t^{-2}\quad\Longrightarrow\quad \frac{\xi(1|0_{<t})}{\mu(1|0_{<t})}\to\frac{w_2c_2}{c_\xi}t\to\infty\quad\text{diverges.}$$

Further interesting convergence results can be found in [Vov87].

2.6 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 $\xi(x_{1:n})\geq w_\mu\cdot\mu(x_{1:n})$ is still ensured with $w_\mu:=\min_{\nu\in\mathcal{L}}\frac{w_\nu}{v_\nu}\geq\min_{\nu\in\mathcal{L}}w_\nu$. More generally, if μ is an infinite linear combination, dominance is still ensured if w_ν itself dominate v_ν in the sense that $w_\nu\geq\alpha v_\nu$ for some $\alpha>0$ (then $w_\mu\geq\alpha$).

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

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

So $D_n\leq\ln w_\mu^{-1}$ remains valid if we define $w_\mu:=w_{\hat{\mu}}\cdot e^{-c}$.

2.7 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* 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 computable* distribution; and all valid physical theories (and, hence, all environments) *are* computable (in a probabilistic sense).

We will see that it is favorable to assign high weights w_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 [Kol65, ZL70, LV97]. 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

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

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 ($D_n \leq K(\mu)\cdot\ln 2$). Solomonoff's universal semi-measure[4] is obtained if we take $\mathcal{M}$ to be the (multi)set enumerated by a Turing machine which enumerates all enumerable semi-measures [Sol64, Sol78, LV97]. In this case, Ω (sometimes called the number of wisdom) has interesting properties in itself [Cal98, Cha75, Cha91]. Recently, $\mathcal{M}$ has been further enlarged to include all cumulatively enumerable semi-measures [Sch00, Sch02]. 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 [Sch00]. An interesting and quickly approximable distribution is the Speed prior S defined in [Sch00]. It is related to Levin complexity and Levin search [Lev73, Lev84], but it is unclear for now, which distributions are dominated by S. If one considers only finite-state automata instead of general Turing machines, ξ is related to the quickly computable, universal finite-state prediction scheme of Feder et al. [FMG92], which itself is related to the famous Lempel-Ziv data compression algorithm. If one has extra knowledge on the source generating the sequence, one might further reduce $\mathcal{M}$ 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 $\mathcal{M}$ [LV97]. As we do not need this property here, $\mathcal{M}$ may be *any* finite or countable set of distributions. In the following we consider generic $\mathcal{M}$ and w. Continuous classes $\mathcal{M}$ are considered in Section 7.

[4] Normalization has to be treated differently in this case

3 Error Bounds

3.1 Deterministic Predictors

We start with a very simple measure: making a wrong prediction counts as one error, making a correct prediction counts as no error. In [Hut01a] error bounds have been proven for the binary alphabet $\mathcal{X}=\{0,1\}$. The following generalization to 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

$$e_t^{\Theta_\rho}(x_{<t}) \;:=\; 1-\mu(x_t^{\Theta_\rho}|x_{<t}) \quad , \quad E_n^{\Theta_\rho} \;:=\; \sum_{t=1}^{n}\mathbf{E}_{<t}e_t^{\Theta_\rho}(x_{<t}). \tag{17}$$

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

$$E_n^{\Theta_\mu} \;\leq\; E_n^{\Theta_\rho} \quad \text{for any} \quad \Theta_\rho, \tag{18}$$

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 proved from $\xi\to\mu$. Indeed, this problem occurs in related prediction schemes, where the predictor has to be regularized so that it is continuous [FMG92]. Fortunately this is not necessary here. We prove the following error bound.

Theorem 3 (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 (17) are bounded in the following way

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

where $S_n \leq D_n \leq \ln w_\mu^{-1}$. S_n is the squared distance (11), D_n is the relative entropy (14), and w_μ is the weight (5) of μ in ξ.

The first bound actually contains $E_n^{\Theta_\xi}$ on the r.h.s., so it is not particularly useful, but this is the major bound we will prove, the others follow easily. Furthermore it has a somewhat nicer structure than the second bound. In Section 6 we show that the second bound is optimal. The last bound, which we discuss in the following, has the same asymptotics as the second bound.

First, we observe that the number of errors $E_\infty^{\Theta_\xi}$ of the universal Θ_ξ predictor is finite if the number of errors $E_{\infty\Theta_\mu}$ of the informed Θ_μ predictor is finite. This is especially the case for deterministic μ, as $E_n^{\Theta_\mu}\equiv 0$ in this case[5], i.e. Θ_ξ makes only a finite number of errors on deterministic environments. This can 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 $e_t^{\Theta_\xi} = 1 \leq -\ln\xi(x_t|x_{<t})/\ln 2 = d_t/\ln 2$. If Θ_ξ makes a correct prediction $e_t^{\Theta_\xi} = 0 \leq d_t/\ln 2$ is obvious. Using (14) proves $E_\infty^{\Theta_\xi} \leq D_\infty/\ln 2 \leq \log_2 w_\mu^{-1}$. A combinatoric argument given in Section 6 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 3 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 $E_n^{\Theta_\mu}\to\infty$. Increasing 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 6 we show that the second bound for $E_n^{\Theta_\xi} - E_n^{\Theta_\mu}$ given in Theorem 3 can in general not be improved, i.e. for every predictor Θ (and especially Θ_ξ) there exist $\mathcal{M}$ and $\mu\in\mathcal{M}$ such that the upper bound is achieved. See [Hut01a] for some further discussion and bounds for binary alphabet.

3.3 Proof of Theorem 3

The first inequality in Theorem 3 has already been proven (18). For the second inequality, let us start more modestly and try to find constants $A>0$ and $B>0$ that satisfy the linear

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

inequality

$$E_n^{\Theta_\xi} - E_n^{\Theta_\mu} \;\leq\; A(E_n^{\Theta_\xi} + E_n^{\Theta_\mu}) + BS_n. \tag{19}$$

If we could show

$$e_t^{\Theta_\xi}(x_{<t}) - e_t^{\Theta_\xi}(x_{<t}) \;\leq\; A[e_t^{\Theta_\xi}(x_{<t}) + e_t^{\Theta_\mu}(x_{<t})] + Bs_t(x_{<t}) \tag{20}$$

for all $t \leq n$ and all $x_{<t}$, (19) would follow immediately by summation and the definition of E_n and S_n. With the abbreviations (9) and the abbreviations $m = x_t^{\Theta_\mu}$ and $s = x_t^{\Theta_\xi}$ the various error functions can then be expressed by $e_t^{\Theta_\xi} = 1 - y_s$, $e_t^{\Theta_\mu} = 1 - y_m$ and $s_t = \sum_i (y_i - z_i)^2$. Inserting this into (20) we get

$$y_m - y_s \;\leq\; A[2 - (y_m + y_s)] + B\sum_{i=1}^{N}(y_i - z_i)^2. \tag{21}$$

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

$$B\sum_{i=1}^{N}(y_i - z_i)^2 + A[2 - (y_m + y_s)] - (y_m - y_s) \;\geq\; ... \tag{22}$$

is positive for suitable $A \geq 0$ and $B \geq 0$, which proves (21). For $m = s$ (22) is obviously positive. So we will assume $m \neq s$ in the following. From the square we keep only contributions from $i = m$ and $i = s$.

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

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

$$... \;\geq\; \tfrac{1}{2}Bx^2 + A - x \;\geq\; ... \tag{23}$$

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

$$... \;\geq\; A - \frac{1}{2B} \;\geq\; 0 \quad \text{for} \quad 2AB \geq 1. \tag{24}$$

Inequality (19) therefore holds for any $A > 0$, provided we insert $B = \frac{1}{2A}$. Thus we might minimize the r.h.s. of (19) w.r.t. A leading to the upper bound

$$E_n^{\Theta_\xi} - E_n^{\Theta_\mu} \;\leq\; \sqrt{2(E_n^{\Theta_\xi} + E_n^{\Theta_\mu})S_n} \qquad \text{for} \qquad A^2 = \frac{S_n}{2(E_n^{\Theta_\xi} + E_n^{\Theta_\mu})} \tag{25}$$

which is the first bound in Theorem 3. For the second bound we have to prove

$$\sqrt{2(E_n^{\Theta_\xi} + E_n^{\Theta_\mu})S_n} - S_n \;\leq\; \sqrt{4E_n^{\Theta_\mu}S_n + S_n^2} \tag{26}$$

If we square both sides of this expressions and simplify we just get (25). Hence, (25) implies (26). The last inequality in Theorem 3 is a simple triangle inequality. This completes the proof of Theorem 3 □.

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.

4 Loss Bounds

4.1 Unit 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 in [Hut01b]. To stay in the framework of (passive) prediction we have to assume that the action itself does not influence the environment. Let $\ell_{x_t y_t} \in \mathbb{R}$ 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 demand ℓ to be normalized, i.e. $0 \leq \ell_{x_t y_t} \leq 1$. 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	0.3	0.1
sunglasses	0.0	1.0

Note the small 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 subsection falls into this class together with a special loss function. It assigns unit loss to an erroneous prediction ($\ell_{x_t y_t}=1$ for $x_t \neq y_t$) and no loss to a correct prediction ($\ell_{x_t x_t}=0$).

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 $\mathbf{E}_t\ell_{x_t y_t}$. The goal is to minimize the expected loss. More generally we define the Λ_ρ prediction scheme

$$y_t^{\Lambda_\rho} \;:=\; \arg\min_{y_t\in\mathcal{Y}} \sum_{x_t} \rho(x_t|x_{<t})\ell_{x_t y_t} \tag{27}$$

which minimizes the ρ-expected loss.[6] As the true distribution is μ, the actual μ-expected loss when Λ_ρ predicts the t^{th} symbol and the total μ-expected loss in the first n predictions are

$$l_t^{\Lambda_\rho}(x_{<t}) \;:=\; \mathbf{E}_t\ell_{x_t y_t^{\Lambda_\rho}} \quad , \quad L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^{n} \mathbf{E}_{<t} l_t^{\Lambda_\rho}(x_{<t}). \tag{28}$$

Let Λ be *any* (causal) prediction scheme (deterministic or probabilistic) 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 (28). If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss

$$L_n^{\Lambda_\mu} \;\le\; L_n^\Lambda \quad \text{for any} \quad \Lambda \tag{29}$$

since

$$l_t^{\Lambda_\mu}(x_{<t}) = \mathbf{E}_t\ell_{x_t y_t^{\Lambda_\mu}} = \min_{y_t} \mathbf{E}_t\ell_{x_t y_t} \le \mathbf{E}_t\ell_{x_t y_t^\Lambda} = l_t^\Lambda(x_{<t})$$

for any Λ. The predictor Λ_ξ, based on the universal distribution ξ, is, again, of special interest. Theorem 3 generalizes to arbitrary loss functions.

Theorem 4 (Unit loss 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. A system taking action (or predicting) $y_t\in\mathcal{Y}$ given $x_{<t}$ receives loss $\ell_{x_t y_t}\in[0,1]$ if x_t is the true t^{th} symbol of the sequence. The Λ_ρ-system (27) acts (or predicts) as to minimize the ρ-expected loss. Λ_ξ is the universal prediction scheme based on the universal prior ξ. Λ_μ is the optimal informed prediction scheme. The total μ-expected losses $L_n^{\Lambda_\xi}$ of Λ_ξ and $L_n^{\Lambda_\mu}$ of Λ_μ as defined in (28) are bounded in the following way*

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

where $D_n\le\ln w_\mu^{-1}$ is the relative entropy (14), and w_μ is the weight (5) of μ in ξ.

The loss bounds have the same form as the error bounds when substituting $S_n\le D_n$ in Theorem 3, so most of the discussion of Theorem 3 also applies here. We were not able to derive loss bounds in terms of S_n as in the error case, and indeed one can show that substituting S_n for D_n in Theorem 4 gives an invalid bound. For convenience we collect the most important consequences of Theorem 4 in the following corollary.

[6] $\text{argmin}_y(\cdot)$ is defined as the y which minimizes the argument. A tie is broken arbitrarily. If $\mathcal{Y}$ is finite, then $y_t^{\Lambda_\rho}$ always exists. For infinite action space $\mathcal{Y}$ we assume that a minimizing $y_t^{\Lambda_\rho}\in\mathcal{Y}$ exists. This is for instance the case if $\mathcal{Y}$ is compact and ℓ_{xy} is continuous in y, or for $\mathcal{Y}=\mathbb{N}$, if $\lim_{y\to\infty}\ell_{xy}$ exists for all x and is larger or equal to ℓ_{xy} for most y.

Corollary 5 (Unit loss bound) *Under the same conditions as in Theorem 4 the following relations hold.*

$$\begin{aligned}
&i) \quad L_\infty^{\Lambda_\xi} \text{ is finite } \iff L_\infty^{\Lambda_\mu} \text{ is finite,} \\
&ii) \quad L_\infty^{\Lambda_\xi} \leq 2D_\infty \leq 2\ln w_\mu^{-1} \text{ for deterministic } \mu \text{ if } \forall x \exists y \ell_{xy} = 0, \\
&iii) \quad L_n^{\Lambda_\xi}/L_n^{\Lambda_\mu} = 1 + O((L_n^{\Lambda_\mu})^{-1/2}) \longrightarrow 1 \quad \text{for} \quad L_n^{\Lambda_\mu} \to \infty, \\
&iv) \quad L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} = O(\sqrt{L_n^{\Lambda_\mu}}),
\end{aligned}$$

Let Λ *be* any *prediction scheme.*

$$\begin{aligned}
&v) \quad L_n^{\Lambda_\mu} \leq L_n^{\Lambda} \quad , \quad l_t^{\Lambda_\mu}(x_{<t}) \leq l_t^{\Lambda}(x_{<t}), \\
&vi) \quad L_n^{\Lambda} \geq L_n^{\Lambda_\xi} - 2\sqrt{L_n^{\Lambda_\xi} D_n}, \\
&vii) \quad L_n^{\Lambda_\xi}/L_n^{\Lambda} \leq 1 + O((L_n^{\Lambda})^{-1/2}).
\end{aligned}$$

4.2 Loss Bound of Merhav & Feder

The first general loss bound with no structural assumptions on μ and ℓ (except boundedness) has been derived in a survey paper by Merhav&Feder in [MF98, Sec.3.1.2]. They showed that the regret $L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu}$ is bounded by $\ell_{max}\sqrt{2nD_n}$ for $\ell \in [0,\ell_{max}]$. Assuming $\ell_{max}=1$ (general ℓ_{max} can be recovered by scaling) their bound reads (in our notation)

$$L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} \leq A_n \leq \sqrt{2nD_n}. \tag{30}$$

In Subsection 4.5 we prove

$$l_t^{\Lambda_\xi}(x_{<t}) - l_t^{\Lambda_\mu}(x_{<t}) \leq a_t(x_{<t}) \leq \sqrt{2d_t(x_{<t})}$$

Taking the the expectation $\mathbf{E}_{<t}$ and the average $\frac{1}{n}\sum_{t=1}^n$ and using Jensen's inequality for the concave square root (similarly to (16)) or directly Theorem 2(vi) shows (30).

Bound (30) and our bound (Theorem 4) are in general incomparable. Since $2D_\infty$ is finite and $L_n^{\Lambda_\mu} \leq n$, bound (30) can be at best a factor $\sqrt{2}$ and an additive constant better than our bound. On the other hand, for large n and for $L_n^{\Lambda_\mu} < \frac{n}{2}$ our bound is tighter. The latter condition is satisfied if the best predictor Λ_μ suffers small instantaneous loss $< \frac{1}{2}$ on average. Significant improvement occurs if $L_n^{\Lambda_\mu}$ does not grow linearly with n, but is for instance finite (see Corollary (5), especially (i) and (ii)).

4.3 Example Loss Functions

The case $\mathcal{X} \equiv \mathcal{Y}$ with unit error assignment $\ell_{xy} = 1 - \delta_{xy}$ ($\delta_{xy} = 1$ for $x = y$ and $\delta_{xy} = 0$ for $x \neq y$) has already been discussed and proven in Section 3.

$$y_t^{\Lambda_\rho} = \arg\min_{y_t} \sum_{x_t} \rho(x_t|x_{<t})(1 - \delta_{x_t y_t}) = \arg\max_{x_t} \rho(x_t|x_{<t}) = x_t^{\Theta_\rho}$$

In this case $L_n^{\Lambda_\rho} \equiv E_n^{\Theta_\rho}$ is the total expected number of prediction errors. For $\mathcal{X} = \mathcal{Y} = \{0,1\}$, like in the weather example above, Λ_ρ is a threshold strategy with $y_t^{\Lambda_\rho} = \text{argmin}_{y\in\{0,1\}}\{\rho_1\ell_{1y}+\rho_0\ell_{0y}\} = 0/1$ for $\rho_1 \gtrless \gamma$, where $\gamma := \frac{\ell_{01}-\ell_{00}}{\ell_{01}-\ell_{00}+\ell_{10}-\ell_{11}}$ and $\rho_i = \rho(i|x_{<t})$. In the special error case $\ell_{xy} = 1-\delta_{xy}$, the bit with the highest ρ probability is predicted ($\gamma = \frac{1}{2}$). In the following we consider some standard loss functions for binary outcome $\mathcal{X}=\{0,1\}$ and continuous action y in the unit interval $\mathcal{Y}=[0,1]$. The *absolute loss* is defined as $\ell_{xy}=|x-y|\in[0,1]$. The Λ_ρ scheme predicts $y_t^{\Lambda_\rho}=\text{argmin}_{y\in[0,1]}\{\rho_1(1-y)+\rho_0 y\}=0/1$ for $\rho_0 \gtrless \rho_1$. Since all predictions y lie in the subset $\{0,1\}\subset[0,1]$ and $|x-y|=1-\delta_{xy}$ for $y\in\{0,1\}$ this case coincides with the binary error case above. The same holds for the α-loss $|x-y|^\alpha$ with $0<\alpha\leq 1$. The μ-expected loss is $l_t^{\Lambda_\rho}=\mu(i|x_{<t})$ for the i with $\rho_i>\frac{1}{2}$. For the *quadratic loss* $\ell_{xy}=(x-y)^2\in[0,1]$ the action/prediction $y_t^{\Lambda_\rho}=\text{argmin}_{y\in[0,1]}\{\rho_1(1-y)^2+\rho_0 y^2\}=\rho_1$ is proportional to the ρ-probability of $x_t=1$ and $l_t^{\Lambda_\rho}=\mathbf{E}_t(1-\rho(x_t|x_{<t}))^2$. For the α-loss $|x-y|^\alpha$ with $\alpha>1$ we get $y_t^{\Lambda_\rho}=(1+\sqrt[\alpha-1]{\rho_0/\rho_1})^{-1}$. For arbitrary finite alphabet $\mathcal{X}$ and vector-valued predictions $\mathbf{y}$ the quadratic loss may be generalized to $\ell_{x\mathbf{y}}=\frac{1}{2}\mathbf{y}^T\mathbf{A}_x\mathbf{y}+\mathbf{b}_x^T\mathbf{y}+c_x$. The *Hellinger loss* can be written for binary outcome in the form $\ell_{xy}=1-\sqrt{|1-x-y|}\in[0,1]$ with $y_t^{\Lambda_\rho}=\rho_1^2/(\rho_0^2+\rho_1^2)$ and $l_t^{\Lambda_\rho}=1-(\mu_0\rho_0+\mu_1\rho_1)/\sqrt{\rho_0^2+\rho_1^2}$. The *logarithmic loss* $\ell_{xy}=-\ln|1-x-y|\in[0,\infty]$ is unbounded. But since the corresponding action is $y_t^{\Lambda_\rho}=\rho_1$ the expected loss is $l_t^{\Lambda_\rho}=-\mathbf{E}_t\ln\rho(x_t|x_{<t})$. Hence $l_t^{\Lambda_\xi}-l_t^{\Lambda_\mu}=h_t$ and the total loss regret $L_n^{\Lambda_\xi}-L_n^{\Lambda_\mu}=D_n\leq\ln w_\mu^{-1}$ is finitely bounded anyway and Theorem 4 is not needed. Continuous outcome spaces $\mathcal{X}$ are briefly discussed in Section 10.

4.4 Proof of Theorem 4

The first inequality in Theorem 4 has already been proven (29). For the second and last inequality, we start, as in Theorem 3, by looking for constants $A>0$ and $B>0$, which satisfy the linear inequality

$$L_n^{\Lambda_\xi} \;\leq\; (A+1)L_n^{\Lambda_\mu} + (B+1)D_n. \tag{31}$$

If we could show

$$l_t^{\Lambda_\xi}(x_{<t}) \;\leq\; A' l_t^{\Lambda_\mu}(x_{<t}) + B' d_t(x_{<t}), \quad A' := A+1, \quad B' := B+1 \tag{32}$$

for all $t\leq n$ and all $x_{<t}$, (31) would follow immediately by summation and the definition of L_n and D_n. With the abbreviations (9) and the abbreviations $m=y_t^{\Lambda_\mu}$ and $s=y_t^{\Lambda_\xi}$ the loss and entropy can then be expressed by $l_t^{\Lambda_\xi}=\sum_i y_i\ell_{is}$, $l_t^{\Lambda_\mu}=\sum_i y_i\ell_{im}$ and $d_t=\sum_i y_i\ln\frac{y_i}{z_i}$. Inserting this into (32) we get

$$\sum_{i=1}^N y_i\ell_{is} \;\leq\; A'\sum_{i=1}^N y_i\ell_{im} + B'\sum_{i=1}^N y_i\ln\frac{y_i}{z_i} \tag{33}$$

By definition (27) of $y_t^{\Lambda_\mu}$ and $y_t^{\Lambda_\xi}$ we have

$$\sum_i y_i\ell_{im}\leq\sum_i y_i\ell_{ij} \quad\text{and}\quad \sum_i z_i\ell_{is}\leq\sum_i z_i\ell_{ij} \quad\text{for all}\quad j. \tag{34}$$

Actually, we need the first constraint only for $j=s$ and the second for $j=m$. In Appendix D we reduce the problem to the binary $N=2$ case, which we will consider in the following. We take $\sum_{i=0}^{1}$ instead of $\sum_{i=1}^{2}$ for convenience.

$$B'\sum_{i=0}^{1} y_i \ln\frac{y_i}{z_i} + \sum_{i=0}^{1} y_i(A'\ell_{im}-\ell_{is}) \stackrel{?}{\geq} 0 \tag{35}$$

The cases $\ell_{im}>\ell_{is}\forall i$ and $\ell_{is}>\ell_{im}\forall i$ contradict the first/second inequality (34). Hence we can assume $\ell_{0m}\geq\ell_{0s}$ and $\ell_{1m}\leq\ell_{1s}$. The symmetric case $\ell_{0m}\leq\ell_{0s}$ and $\ell_{1m}\geq\ell_{1s}$ is proven analogously or can be reduced to the first case by renumbering the indices $(0\leftrightarrow 1)$. Using the abbreviations $a:=\ell_{0m}-\ell_{0s}$, $b:=\ell_{1s}-\ell_{1m}$, $c:=y_1\ell_{1m}+y_0\ell_{0s}$, $y=y_1=1-y_0$ and $z=z_1=1-z_0$ we can write (35) as

$$f(y,z) \;:=\; B'[y\ln\tfrac{y}{z}+(1-y)\ln\tfrac{1-y}{1-z}]+A'(1-y)a-yb+Ac \;\stackrel{?}{\geq}\; 0 \tag{36}$$

for $zb\leq(1-z)a$ and $0\leq a,b,c,y,z\leq 1$. The constraint (34) on y has been dropped since (36) will turn out to be true for all y. Furthermore, we can assume that $d:=A'(1-y)a-yb\leq 0$ since for $d>0$, f is trivially positive. Multiplying d with a constant ≥ 1 will decrease f. Let us first consider the case $z\leq\frac{1}{2}$. We multiply the d term by $1/b\geq 1$, i.e. replace it with $A'(1-y)\frac{a}{b}-y$. From the constraint on z we known that $\frac{a}{b}\geq\frac{z}{1-z}$. We can decrease f further by replacing $\frac{a}{b}$ by $\frac{z}{1-z}$ and by dropping Ac. Hence, (36) is proven for $z\leq\frac{1}{2}$ if we can prove

$$B'[y\ln\tfrac{y}{z}+(1-y)\ln\tfrac{1-y}{1-z}]+A'(1-y)\tfrac{z}{1-z}-y \;\stackrel{?}{\geq}\; 0 \quad\text{for}\quad z\leq\tfrac{1}{2}. \tag{37}$$

In Appendix B we prove that it holds for $B\geq\frac{1}{A}+1$. The case $z\geq\frac{1}{2}$ is treated similarly. We scale d with $1/a\geq 1$, i.e. replace it with $A'(1-y)-y\frac{b}{a}$. From the constraint on z we know that $\frac{b}{a}\leq\frac{1-z}{z}$. We decrease f further by replacing $\frac{b}{a}$ by $\frac{1-z}{z}$ and by dropping Ac. Hence (36) is proven for $z\geq\frac{1}{2}$ if we can prove

$$B'[y\ln\tfrac{y}{z}+(1-y)\ln\tfrac{1-y}{1-z}]+A'(1-y)-y\tfrac{1-z}{z} \;\stackrel{?}{\geq}\; 0 \quad\text{for}\quad z\geq\tfrac{1}{2}. \tag{38}$$

In Appendix C we prove that it holds for $B\geq\frac{1}{A}+1$. So in summary we proved that (31) holds for $B\geq\frac{1}{A}+1$. Inserting $B=\frac{1}{A}+1$ into (31) and minimizing the r.h.s. w.r.t. A leads to the last bound of Theorem 4 with $A=\sqrt{D_n/L_n^{\Lambda_\mu}}$. Actually inequalities (37) and (38) also hold for $B\geq\frac{1}{4}A+\frac{1}{A}$, which, by the same minimization argument, proves the slightly tighter second bound in Theorem 4. Unfortunately, the current proof is very long and complex, and involves some numerical or graphical analysis for determining intersection properties of some higher order polynomials. This or a hopefully simplified proof will be postponed. The cautious reader may check the inequalities (37) and (38) numerically for $B=\frac{1}{4}A+\frac{1}{A}$ □.

4.5 Convergence of Instantaneous Losses

Since $L_n^{\Lambda_\xi}-L_n^{\Lambda_\mu}$ is not finitely bounded by Theorem 4 it cannot be used directly to conclude $l_t^{\Lambda_\xi}-l_t^{\Lambda_\mu}\to 0$. It would follow from $\xi\to\mu$ by continuity if $l_t^{\Lambda_\xi}$ and $l_t^{\Lambda_\mu}$ would be continuous

functions of ξ and μ. $l_t^{\Lambda_\mu}$ is a continuous piecewise linear concave function, but $l_t^{\Lambda_\xi}$ is an, in general, discontinuous function of ξ (and μ). Fortunately it is continuous at the one necessary point $\xi=\mu$. This allows to bound $l_t^{\Lambda_\xi}-l_t^{\Lambda_\mu}$ in terms of $\xi(x_t|x_{<t})-\mu(x_t|x_{<t})$.

Theorem 6 (Instantaneous Loss Bound) *Under the same conditions as in Theorem 4 the following relations hold for the instantaneous losses $l_t^{\Lambda_\mu}(x_{<t})$ and $l_t^{\Lambda_\xi}(x_{<t})$ at time t of the informed and universal prediction schemes Λ_μ and Λ_ξ:*

$$
\begin{aligned}
i)\quad & \sum_{t=1}^{n}\mathbf{E}_{<t}(l_t^{\Lambda_\xi}(x_{<t})-l_t^{\Lambda_\mu}(x_{<t}))^2 \;\leq\; 2D_n \;\leq\; 2\ln w_\mu^{-1} \;<\; \infty \\
ii)\quad & 0 \;\leq\; l_t^{\Lambda_\xi}(x_{<t})-l_t^{\Lambda_\mu}(x_{<t}) \;\leq\; \sum_{x_t}|\xi(x_t|x_{<t})-\mu(x_t|x_{<t})| \;\leq\; \sqrt{2d_t(x_{<t})} \overset{t\to\infty}{\underset{w.\mu.p.1}{\longrightarrow}} 0. \\
iii)\quad & 0 \;\leq\; l_t^{\Lambda_\xi}(x_{<t})-l_t^{\Lambda_\mu}(x_{<t}) \;\leq\; 2d_t(x_{<t})+2\sqrt{l_t^{\Lambda_\mu}(x_{<t})\,d_t(x_{<t})} \overset{t\to\infty}{\underset{w.\mu.p.1}{\longrightarrow}} 0.
\end{aligned}
$$

Proof: (ii) follows from

$$
\begin{aligned}
l_t^{\Lambda_\xi}(x_{<t})-l_t^{\Lambda_\mu}(x_{<t}) \;&\equiv\; \sum_i y_i\ell_{is}-\sum_i y_i\ell_{im} \;\leq\; \sum_i(y_i-z_i)(\ell_{is}-\ell_{im}) \;\leq \\
\leq\; \sum_i|y_i-z_i|\cdot|\ell_{is}-\ell_{im}| \;&\leq\; \sum_i|y_i-z_i| \;\leq\; \sqrt{2\sum_i y_i\ln\frac{y_i}{z_i}} \;\equiv\; \sqrt{2d_t(x_{<t})}
\end{aligned}
$$

To arrive at the first inequality we added $\sum_i z_i(\ell_{im}-\ell_{is})$ which is positive due to (34). $|\ell_{is}-\ell_{im}|\leq 1$ since $\ell\in[0,1]$. The last inequality follows from Lemma 1a. $d_t\to 0$ has been proven in Theorem 2(ii). (i) follows by inserting (ii) and using (14). (iii) follows from the proof of Theorem 4 by inserting $B=\frac{1}{A}+1=\sqrt{l_t^{\Lambda_\mu}/d_t}+1$ into (32). Convergence to zero holds for μ random sequences, i.e. with μ probability 1, since $l_t^{\Lambda_\mu}\leq 1$ is bounded. The losses $l_t^{\Lambda_\rho}(x_{<t})$ itself need not to converge. □

Note, that the inequalities in (ii) and (iii) hold for all individual sequences. The sum/average is only taken over the current outcome x_t, but the history $x_{<t}$ is fixed. Bound (ii) and (iii) are in general incomparable, but for large t and for $l_t^{\Lambda_\mu}<\frac{1}{2}$ (especially if $l_t^{\Lambda_\mu}\to 0$) bound (iii) is tighter than bound (ii).

4.6 General Loss

There are only very few restrictions imposed on the loss $\ell_{x_ty_t}$ in Theorem 4, namely that it is static and in the unit interval $[0,1]$. If we look at the proof of Theorem 4, we see that the time-independence has not been used at all. The proof is still valid for an individual loss function $\ell^t_{x_ty_t}\in[0,1]$ for each step t. The loss might even depend on the actual history $x_{<t}$. The case of a loss $\ell^t_{x_ty_t}(x_{<t})$ bounded to a general interval $[\ell_{min},\ell_{max}]$ can be reduced to the unit interval case by rescaling ℓ. We introduce a scaled loss ℓ'

$$
0 \;\leq\; \ell'^t_{x_ty_t}(x_{<t}) := \frac{\ell^t_{x_ty_t}(x_{<t})-\ell_{min}}{\ell_\Delta} \;\leq\; 1, \quad \text{where} \quad \ell_\Delta := \ell_{max}-\ell_{min}.
$$

The prediction scheme Λ'_ρ based on ℓ' is identical to the original prediction scheme Λ_ρ based on ℓ, since minarg in (27) is not affected by linear transformation of its argument. From $y_t^{\Lambda'_\rho} = y_t^{\Lambda_\rho}$ it follows that $l_t'^{\Lambda_\rho} = (l_t^{\Lambda_\rho} - \ell_{min})/\ell_\Delta$ and $L_n'^{\Lambda_\rho} = (L_n^{\Lambda_\rho} - \ell_{min})/\ell_\Delta$ ($D'_n \equiv D_n$, since ℓ is not involved). Theorem 4 is valid for the primed quantities, since $\ell' \in [0,1]$. Inserting $L'_{n\Lambda_{\mu/\xi}}$ and rearranging terms we get

Theorem 7 (General loss 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. A system taking action (or predicting) $y_t \in \mathcal{Y}$ given $x_{<t}$ receives loss $\ell^t_{x_t y_t}(x_{<t}) \in [\ell_{min}, \ell_{min}+\ell_\Delta]$ if x_t is the true t^{th} symbol of the sequence. The Λ_ρ-system (27) acts (or predicts) as to minimize the ρ-expected loss. Λ_ξ is the universal prediction scheme based on the universal prior ξ. Λ_μ is the optimal informed prediction scheme. The total μ-expected losses $L_n^{\Lambda_\xi}$ and $L_n^{\Lambda_\mu}$ of Λ_ξ and Λ_μ as defined in (28) are bounded in the following way*

$$0 \;\leq\; L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} \;\leq\; \ell_\Delta D_n + \sqrt{4(L_n^{\Lambda_\mu} - n\ell_{min})\ell_\Delta D_n + \ell_\Delta^2 D_n^2}$$

where $D_n \leq \ln w_\mu^{-1}$ is the relative entropy (14), and w_μ is the weight (5) of μ in ξ.

5 Application to Games of Chance

5.1 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 behaviour 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 2 we know that Solomonoff's universal prior $\xi(x_t|x_{<t})$ 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 [LV97]. 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.

5.2 Games of Chance

We use Theorem 7 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 $p_t = r_t - s_t$. The loss, which we want to minimize, can be defined as the negative profit, $\ell_{x_t y_t} = -p_t$. 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 $P_n^{\Lambda_\rho} = -L_n^{\Lambda_\rho}$. 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 $P_n^{\Lambda_\mu} \geq P_n^{\Lambda_\rho}\ \forall\rho$). 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 Theorem 7 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. Theorem 7 allows us to lower bound the universal profit $P_n^{\Lambda_\xi}$

$$P_n^{\Lambda_\xi} \;\geq\; P_n^{\Lambda_\mu} - p_\Delta D_n - \sqrt{4(np_{max} - P_n^{\Lambda_\mu}) p_\Delta D_n + p_\Delta^2 D_n^2} \tag{39}$$

where p_{max} is the maximal profit per round and p_Δ the profit range. 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 (39) is positive if, and only if

$$n \;>\; \frac{2p_\Delta(2p_{max} - \bar{p}_n^{\Lambda_\mu})}{(\bar{p}_n^{\Lambda_\mu})^2} \cdot D_n. \tag{40}$$

Theorem 8 (Time to Win) *Let there be sequences $x_1 x_2 ...$ 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 (27) 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:*

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

where $w_\mu = e^{-b_\mu}$ is the weight (5) of μ in ξ.

By dividing (39) by n and using $D_n \leq b_\mu$ (14) 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 (40). $P_n^{\Lambda_\xi}$ 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 (40), since $p_\Delta > p_{max}$, and (40) implies positive (39), i.e. $P_n^{\Lambda_\xi} > 0$, which proves (ii).

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

5.3 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 ($\gamma = \frac{1}{2}$). 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 Solomonoff'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 Subsection 2.7). Then we know that betting on the outcome with higher ξ probability leads asymptotically to the same profit (Theorem 8(i)) and Λ_ξ reaches the winning threshold no later than $n_{thresh} = 900\ln 2\cdot K(\mu)$ (Theorem 8(ii)) or sharper $n_{thresh} = 330\ln 2\cdot K(\mu)$ from (40), 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 t_{thresh}. The true expected turnaround t might be smaller. However, every game for which there exists a computable winning strategy with $\bar{p}_{n\rho} > \varepsilon > 0$, Λ_ξ is guaranteed to get into the winning zone for some $t \sim K(\mu)$.

5.4 Information-theoretic Interpretation

We try to give an intuitive explanation of Theorem 8(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 $\ln 2\cdot D_\infty \leq b_\mu\cdot\ln 2$. One might think of a Shannon-Fano prefix code of $\nu \in \mathcal{M}$ of length $\lceil b_\nu\cdot\ln 2\rceil$, which exists since the Kraft inequality $\sum_\nu 2^{-\lceil b_\nu\cdot\ln 2\rceil} \leq \sum_\nu w_\nu \leq 1$ is satisfied. $b_\mu\cdot\ln 2$ 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 Theorems 4 and 7 with their rather complex proofs.

6 Optimality Properties

6.1 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 3 is sharp. In the general probabilistic case we can show by a similar argument that the upper bound of Theorem 3 is sharp.

Theorem 9 (Lower Error Bound) *Let Θ be* any deterministic *predictor not knowing from which distribution $\mu\in\mathcal{M}$ the observed sequence $x_1x_2...$ is sampled from. Θ knows (depends on) $\mathcal{M}$ and has at time t access to the previous outcomes $x_{<t}$. Then there is for every n an $\mathcal{M}$ and $\mu\in\mathcal{M}$ and weights w_ν such that*

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

where E_n^Θ and $E_n^{\Theta_\mu}$ are the total expected number of errors of Θ and Θ_μ, and s_t and S_n are defined in (11). The equalities especially hold for the universal predictor Θ_ξ.

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

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

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

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

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

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

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

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

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

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

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

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

which proves the total regret formula of Theorem 9. □

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 3 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 in Theorem 4 can also not be improved. Furthermore, $E_n^\Theta - E_n^{\Theta_\mu} = n\varepsilon = n\sqrt{\frac{2S_n}{n}} \to \sqrt{2nD_n}$, which shows that the bound (30) of Merhav&Feder is also tight.

An n independent set $\mathcal{M}$ leading to a good (but not tight) lower bound is $\mathcal{M}=\{\mu_1,\mu_2\}$ with $\mu_{1/2}(1|x_{<t})=\frac{1}{2}\pm\varepsilon_t$ with $\varepsilon_t=\min\{\frac{1}{2},\sqrt{\ln w_{\mu_1}^{-1}}/\sqrt{t}\ln t\}$. For $w_{\mu_1}\ll w_{\mu_2}$ and $n\to\infty$ one can show that $E_n^{\Theta_\xi} - E_n^{\Theta_{\mu_1}} \sim \frac{1}{\ln n}\sqrt{E_n^{\Theta_\mu}\ln w_{\mu_1}^{-1}}$.

6.2 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 all performance measures studied in this work.

Definition 10 (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 11 (Pareto Optimality) *The universal prior ξ is Pareto-optimal w.r.t. the instantaneous and total squared distances s_t and S_n (11), entropy distances d_t and D_n (13), errors e_t and E_n (17), and losses l_t and L_n (28).*

Proof: We first proof Theorem 11 for the instantaneous expected loss l_t. We need the more general ρ expected instantaneous losses

$$l^\Lambda_{t\rho}(x_{<t}) \;:=\; \sum_{x_t}\rho(x_t|x_{<t})\ell_{x_t y^\Lambda_t} \tag{41}$$

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

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

The two equalities follow from inserting (7) into (41). 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 (27) the ξ-expected loss (similarly to (29)). The contradiction $l^\Lambda_{t\xi}<l^\Lambda_{t\xi}$ proves Pareto-optimality of ξ w.r.t. l_t.

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

$$L^\Lambda_{n\rho} \;:=\; \sum_{t=1}^n\sum_{x_{<t}}\rho(x_{<t})l^\Lambda_{t\rho}(x_{<t}) \;=\; \sum_{t=1}^n\sum_{x_{1:t}}\rho(x_{1:t})\ell_{x_t y^\Lambda_t} \tag{42}$$

for a predictor Λ, and by assuming $L^\Lambda_{n\nu}\leq L^{\Lambda_\xi}_{n\nu}$ for all ν and strict inequality for some ν, from which we get the contradiction $L^\Lambda_{n\xi}=\sum_\nu w_\nu L^\Lambda_{n\nu}<\sum_\nu w_\nu L^{\Lambda_\xi}_{n\nu}=L^{\Lambda_\xi}_{n\xi}\leq L^\Lambda_{n\xi}$ with the help

[7] According to definition 10 we should look for a ρ, but for each deterministic predictor Λ there exists a ρ with $\Lambda=\Lambda_\rho$.

of (5). 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

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

where we have used $\sum_\nu w_\nu=1$ and $\sum_\nu w_\nu y_{\nu i}=z_i$ (7). $\sum_i(r_i-z_i)^2\leq 0$ implies $\mathbf{r}=\mathbf{z}$ proving unique Pareto-optimality of ξ w.r.t. s_t. Similarly for d_t the assumption $\sum_i y_{\nu i}\ln\frac{y_{\nu i}}{r_i}\leq\sum_i y_{\nu i}\ln\frac{y_{\nu i}}{z_i}\ \forall\nu$ implies

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

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 what counts.

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

Theorem 12 ((Non)Pareto-optimality) *ξ is Pareto-optimal w.r.t.*

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

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

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

Intuition on this problem can be gained by considering probability vectors $\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}\in\Delta\subset\mathbb{R}^3$, where Δ is the 2d probability triangle, and $\mathbf{z}=w\mathbf{x}+(1-w)\mathbf{y}$ is a mixture of $\mathbf{x}$ and $\mathbf{y}$. Consider the sets $M_\mathbf{x}:=\{\mathbf{r}:\mathcal{F}(\mathbf{x},\mathbf{r})\leq\mathcal{F}(\mathbf{x},\mathbf{z})\}$ and analogously $M_\mathbf{y}$. $M_\mathbf{x}\cap M_\mathbf{y}$ is not empty; it contains $\mathbf{z}$. If $M_\mathbf{x}\cap M_\mathbf{y}$ has an interior, then $\mathbf{z}$ is not Pareto-optimal. Visualize the 1d boundaries of the 2d areas $M_\mathbf{x}$ and $M_\mathbf{y}$ qualitatively for the various performance measures $\mathcal{F}$. This gives some intuition of how to prove Pareto-optimality and to construct counter-examples. A proof of Theorem 12 will be given elsewhere.

6.3 Balanced Pareto Optimality of ξ

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

Theorem 13 (Balanced Pareto Optimality w.r.t. L)

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

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

This means that a weighted loss decrease $\Delta_\mathcal{H}$ by using $\tilde{\Lambda}$ instead of Λ_ξ is compensated by an at least as large weighted increase $\Delta_\mathcal{L}$ on other environments. If the increase is small, the decrease can also only be small. In the special case of only a single environment with increased loss Δ_λ, the decrease is bound by $\Delta_\eta \leq \frac{w_\lambda}{w_\eta}|\Delta_\lambda|$, i.e. an increase by an amount Δ_λ can only cause a decrease by at most the same amount times a factor $\frac{w_\lambda}{w_\eta}$. A increase can only cause a smaller decrease in simpler environments, but a scaled decrease in more complex environments. Finally note that pure Pareto-optimality (11) follows from balanced Pareto-optimality in the special case of no increase $\Delta_\mathcal{L} \equiv 0$.

Proof: $\Delta \geq 0$ follows from $\Delta = \sum_\nu w_\nu [L_\nu^{\tilde{\Lambda}} - L_\nu^{\Lambda_\xi}] = L_\xi^{\tilde{\Lambda}} - L_\xi^{\Lambda_\xi} \geq 0$, where we have used linearity of L_ρ in ρ and $L_\xi^{\Lambda_\xi} \leq L_\xi^{\Lambda}$. The remainder of Theorem 13 is obvious from $0 \leq \Delta = \Delta_\mathcal{L} - \Delta_\mathcal{H}$ and by bounding the weighted average Δ_η by its maximum.□

6.4 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 Subsection 2.7 we have discussed the set $\mathcal{M}$ of all enumerable semimeasures which we regarded as sufficiently large from a computational point of view (see [Sch00] for even larger sets, but which are still in the computational realm). Agreeing on this $\mathcal{M}$ still leaves open the question of how to choose the weights (prior believes) w_ν, since every ξ_w with $w_\nu > 0 \forall \nu$ is Pareto-optimal and leads asymptotically to optimal predictions.

We have derived bounds for the mean squared sum $S_{n\nu}^{\xi_w} \leq \ln w_\nu^{-1}$ 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 $w_\nu = 2^{-K(\nu)}$ which gives Solomonoff's prior.

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

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

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

Theorem 14 (Optimality of universal weights) *Within the set $\mathcal{V}$ of enumerable weight functions with short program, the universal weights $w_\nu = 2^{-K(\nu)}$ lead to the smallest performance bounds within an additive (to $\ln w_\mu^{-1}$) constant in all enumerable environments.*

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

6.5 Occam's razor versus No Free Lunches

We do not regard Theorem 13 as a "No Free Lunch" (NFL) theorem [WM97]. Since most environments are completely random, a small concession on the loss in each of these completely uninteresting environments provides enough margin $\Delta_{\mathcal{H}}$ to yield distinguished performance on the few non-random (interesting) environments. Indeed, we would interpret

[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 (2), the error bound (Theorem 3), the loss bound (4), 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 can be shown by a more easy and direct argument [Sch02].

the NFL theorems for optimization and search in [WM97] as balanced Pareto-optimality results. Interestingly, whereas for prediction only Bayes-mixes are Pareto-optimal, for search and optimization every algorithm is Pareto-optimal. There is an ongoing battle between believers in Occam's razor and believers in "no free lunches" that cannot be dealt with here [Sto01].

7 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 Subsection 2.7. 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

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

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) then Theorem 7 applies. 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

$$\xi(x_{1:n}) \;:=\; \int_\Theta d\theta\, w(\theta)\cdot\mu_\theta(x_{1:n}), \qquad \int_\Theta d\theta\, w(\theta) \;=\; 1, \qquad w(\theta) \;\geq\; 0. \tag{43}$$

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 (5) 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 $\ln\mu_\theta$ is approximately flat on average. The averaged instantaneous, mean, and total curvature matrices of $\ln\mu$ are

$$\begin{aligned} j_t(x_{<t}) \;&:=\; \mathbf{E}_t\nabla_\theta\ln\mu_\theta(x_t|x_{<t})\nabla_\theta^T\ln\mu_\theta(x_t|x_{<t})_{|\theta=\theta_0}, \qquad \bar{J}_n \;:=\; \tfrac{1}{n}J_n \\ J_n \;&:=\; \sum_{t=1}^n \mathbf{E}_{<t}j_t(x_{<t}) \;=\; \mathbf{E}_{1:n}\nabla_\theta\ln\mu_\theta(x_{1:n})\nabla_\theta^T\ln\mu_\theta(x_{1:n})_{|\theta=\theta_0} \end{aligned} \tag{44}$$

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 line of reasoning as in (14) for D_n.

Theorem 15 (Continuous Entropy Bound) *Let μ_θ be twice continuously differentiable at $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ 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 (43)) 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 (43) of μ in ξ and $o(1)$ tends to zero for $n \to \infty$.

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 in [CB90, Theorem 2.3]. In this case $J^{[CB90]}(\theta_0) \equiv \bar{\jmath}_n \equiv j_n$ 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 [CB90, 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 [CB90, (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 15. See [CB90, 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 [Ris96].

So Theorems 2...7 are also applicable to the case of continuously parameterized probability classes. Theorem 15 is also valid for a mixture of the discrete and continuous case $\xi = \sum_a \int d\theta \, w^a(\theta) \, \mu_\theta^a$ with $\sum_a \int d\theta \, w^a(\theta) = 1$.

8 Further Applications

8.1 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 need 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 $\ell^t_{zy} \equiv 0$ at the z positions. Any dummy prediction y is then consistent with (27). 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 $\rho(x_{1:n})$ ($\rho = \mu, \nu, \xi$) 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 2...15 also hold in this case for all individual z's.

8.2 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 $x \in \mathcal{X}$.

9 Comparison to Weighted Majority

There are two schools of universal sequence prediction: We considered expected performance bounds for Bayesian prediction based on mixtures. The other approach are weighted majority (WM) algorithms with worst case loss bounds in the spirit of Littlestone, Warmuth, Vovk and others. The two schools usually do not refer to each other much. We briefly describe WM and compare both approaches. For a more comprehensive comparison see [MF98]. In the following we focus on topics not covered in [MF98]. WM was invented in [LW89, LW94] and [Vov92] and further developed in [Ces97, HKW98, KW99] and by others. Many variations known by many names (weighted average, aggregating strategy, learning with expert advice, boosting, hedge algorithm, ...) have meanwhile been invented. Early works in this direction are [Daw84, Ris89]. See [Vov99] for a review and further references. We describe the setting and basic idea of

WM 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 (17). The WM 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 WM 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 in [Ces97]:

$$L_p(\mathbf{x}) \;\leq\; L_\varepsilon(\mathbf{x}) + 2.8 \ln |\mathcal{E}| + 4\sqrt{L_\varepsilon(\mathbf{x}) \ln |\mathcal{E}|}. \tag{45}$$

The constants 2.8 and 4 have been improved in [AG00, YE01]. The last bound in Theorem 3 with $D_n \leq \ln|\mathcal{M}|$ for uniform weights and with $L_n^{\Lambda_\mu}$ increased to L_n^Λ reads

$$L_n^{\Lambda_\xi} \;\leq\; L_n^\Lambda + 2 \ln |\mathcal{M}| + 2\sqrt{L_n^\Lambda \ln |\mathcal{M}|}.$$

It has a quite similar structure as (45), although the algorithms, the settings, the proofs, and the interpretation are quite different. Whereas WM performs well in any environment, but only relative to a given set of experts $\mathcal{E}$, our Λ_ξ predictor competes with the best possible Λ_μ predictor (and hence with any other Λ predictor), but only in expectation and for a given set of environments $\mathcal{M}$. WM depends on the set of expert, Λ_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 ($|\mathcal{E}| \hookrightarrow \ln \frac{1}{w_\nu}$) [LW89, Vov92], more general loss functions [HKW98], continuous valued outcomes [HKW98], and multi-dimensional predictions [KW99] (but not yet for the absolute loss). The work of [Yam98] lies somewhat in between WM and this work; "WM" 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 WM 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 WM 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 is taken over the probabilistic prediction, whereas for the deterministic Λ_ξ algorithm the expectation is taken over the environmental distribution μ. The multi-dimensional case [KW99] 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 [FS97] 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 WM and bound (45) to arbitrary alphabet and to general loss functions with probabilistic interpretation.

[10] The original WM version [LW89] had discrete prediction $x_t^p \in \{0,1\}$ with (necessarily) double as many errors as the best expert and is only of historical interest any more.

10 Outlook

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

10.1 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) alphabet. 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 $y_t^{\Lambda_\rho}\in\mathcal{Y}$ (27). 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.

10.2 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 in [Hut01a] can also be generalized to arbitrary alphabet.

10.3 More Active Systems

Prediction means guessing the future, but not influencing it. A small step in the direction to 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 (27) 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 in [Hut01b], but loss bounds have yet to be proven.

10.4 Miscellaneous

Another direction is to investigate the learning aspect of universal prediction. Many prediction schemes explicitly learn and exploit a model of the environment. Learning and exploitation are melted together in the framework of universal Bayesian prediction. A separation of these two aspects in the spirit of hypothesis learning with MDL [VL00] could lead to new insights. Also, the separation of noise from useful data, usually an important issue [GTV01], did not play a role here. The attempt at an information theoretic interpretation of Theorem 8 may be made more rigorous in this or another way. In the end, this may lead to a simpler proof of Theorem 8 and maybe even for the loss bounds. A unified picture of the loss bounds obtained here and the loss bounds for the weighted majority (WM) algorithm could also be fruitful. Yamanishi [Yam98] used WM 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 WM. Maximum-likelihood predictors may also be studied. Finally, the system should be applied to specific induction problems for specific $\mathcal{M}$ with computable ξ.

11 Summary

We compared universal predictions based on Bayes-mixtures ξ to the infeasible informed predictor based on the unknown true generating distribution μ. We have shown that the universal posterior ξ converges to μ and that $\xi/\mu \to 1$. 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, and this is true for any μ independent predictor. We have also shown Pareto-optimality of ξ in the sense that there is no other predictor which performs better or equal in all environments $\nu \in \mathcal{M}$ and strictly better in at least one. Optimal predictors can (in most cases) be based on a mixture distributions ξ. Finally we gave an Occam's razor argument that Solomonoff's 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 [CBL01]. 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 and worst case (WM) approaches and results, classification tasks, games of chances, infinite alphabet, active systems influencing the environment, and others.

Acknowledgements I want to thank Ray Solomonoff and Jürgen Schmidhuber for many valuable discussions and for encouraging me to generalize the error bounds obtained

in [Hut01a] in various ways. Furthermore I want to thank the anonymous referees of preliminary versions for pointing out relevant literature and for their valuable comments which helped improving the work. This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

A Entropy Inequalities (Lemma 1)

[11]We show that

$$\frac{1}{2}\sum_{i=1}^{N} f(y_i-z_i) \;\leq\; f(\sqrt{\frac{1}{2}\sum_{i=1}^{N} y_i \ln\frac{y_i}{z_i}}) \quad \text{for} \quad y_i\geq 0, \quad z_i\geq 0, \quad \sum_{i=1}^{N} y_i = 1 = \sum_{i=1}^{N} z_i. \tag{46}$$

for any convex and even ($f(x)=f(-x)$) function with $f(0)\leq 0$. For $f(x)=x^2$ we get inequality (Lemma 1s), for $f(x)=|x|$ we get inequality (15). To prove (46) we partition $i\in\{1,...,N\}=G^+\cup G^-$, $G^+\cap G^-=\{\}$, and define $y^\pm:=\sum_{i\in G^\pm} y_i$ and $z^\pm:=\sum_{i\in G^\pm} z_i$. It is well known that the relative Entropy is positive, i.e.

$$\sum_{i\in G^\pm} p_i \ln\frac{p_i}{q_i} \;\geq\; 0 \quad \text{for} \quad p_i\geq 0, \quad q_i\geq 0, \quad \sum_{i\in G^\pm} p_i = 1 = \sum_{i\in G^\pm} q_i. \tag{47}$$

Note that there are 4 probability distributions (p_i and q_i for $i\in G^+$ and $i\in G^-$). For $i\in G^\pm$, $p_i:=y_i/y^\pm$ and $q_i:=z_i/z^\pm$ satisfy the conditions on p and q. Inserting this into (47) and rearranging the terms we get

$$\sum_{i\in G^\pm} y_i \ln\frac{y_i}{z_i} \;\geq\; y^\pm \ln\frac{y^\pm}{z^\pm}.$$

If we sum over $\pm$ and define $y\equiv y^+=1-y^-$ and $z\equiv z^+=1-z^-$ we get

$$\sum_{i=1}^{N} y_i \ln\frac{y_i}{z_i} \;\geq\; \sum_{\pm} y^\pm \ln\frac{y^\pm}{z^\pm} \;=\; y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z} \;\geq\; 2(y-z)^2 \tag{48}$$

The last inequality is elementary and well known. For the special choice $G^\pm:=\{i: y_i \gtrless z_i\}$, we can upper bound $\sum_i f(y_i-z_i)$ as follows

$$\begin{aligned}\sum_{i\in G^\pm} f(y_i-z_i) &\overset{(a)}{=} \sum_{i\in G^\pm} f(|y_i-z_i|) \overset{(b)}{\leq} f(\sum_{i\in G^\pm}|y_i-z_i|) \overset{(c)}{=} f(|\sum_{i\in G^\pm} y_i-z_i|) \overset{(d)}{=} \\ &\overset{(d)}{=} f(|y^\pm-z^\pm|) \overset{(e)}{=} f(|y-z|) \overset{(f)}{=} f(\sqrt{(y-z)^2}) \overset{(g)}{\leq} f(\sqrt{\frac{1}{2}\sum_{i=1}^{N} y_i\ln\frac{y_i}{z_i}})\end{aligned} \tag{49}$$

[11]We will not explicate every subtlety and only sketch the proofs. Subtleties regarding $y,z=0/1$ have been checked but will be passed over. $0\ln\frac{0}{z_i}:=0$ even for $z_i=0$. Positive means ≥ 0. The probability constraints in (46) on y and z apply to all appendices. $z>0$ if $y>0$.

(a) follows from the symmetry of f. (b) follows from the convexity[12] of f and from $f(0)\leq 0$. (c) is true, since all $y_i - z_i$ are positive/negative for $i \in G^{\pm}$ due to the special choice of $G^{\pm}$. (d) and (e) follow from the definition of $y^{(\pm)}$ and $z^{(\pm)}$, (f) is obvious. (g) follows from (48) and the monotonicity[13] of $\sqrt{\ }$ and f for positive arguments. Inequality (46) follows by summation of (49) over $\pm$ and noting that $f(\sqrt{\ })$ is independent of $\pm$.

This proves (Lemma 1f). Inserting $f(x) = x^2$ yields (Lemma 1s), inserting $f(x) = |x|$ yields (Lemma 1a). (Lemma 1h) is proven differently. For arbitrary $y \geq 0$ and $z \geq 0$ we define

$$f(y,z) := y \ln \frac{y}{z} - (\sqrt{y} - \sqrt{z})^2 + z - y = 2yg(\sqrt{z/y}) \quad \text{with} \quad g(t) := -\ln t + t - 1 \geq 0.$$

This shows $f \geq 0$, and hence $\sum_i f(y_i,z_i) \geq 0$, which implies

$$\sum_i y_i \ln \frac{y_i}{z_i} - \sum_i (\sqrt{y_i} - \sqrt{z_i})^2 \geq \sum_i y_i - \sum_i z_i = 1 - 1 = 0.$$

This proves (Lemma 1h). □

B Binary Loss Inequality for $z \leq \frac{1}{2}$ (37)

With the definition

$$f(y,z) \;:=\; B' \cdot \left[y \ln \frac{y}{z} + (1-y) \ln \frac{1-y}{1-z} \right] + A' \cdot (1-y) \frac{z}{1-z} - y \quad , \quad z \leq \tfrac{1}{2} \qquad (50)$$

we show $f(y,z) \geq 0$ for suitable $A' \equiv A+1$ and $B' \equiv B+1$. We do this by showing that $f \geq 0$ at all extremal values and "at" boundaries. $f \to +\infty$ for $z \to 0$, if we choose $B' > 0$. For the boundary $z = \frac{1}{2}$ we lower bound the relative entropy by the sum over squares (Lemma 1s)

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

The r.h.s. is quadratic in y with minimum at $y^* = \frac{A'+2B'+1}{4B'}$, which implies

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

Furthermore, for $A \geq 4$ and $B \geq 1$ we have $f(y,\frac{1}{2}) \geq 2(1-y)(3-2y) \geq 0$. Hence $f(y,\frac{1}{2}) \geq 0$ for $B \geq \frac{1}{A}+1$, since for $A \geq 4$ it implies $B \geq 1$ and for $A \leq 4$ it implies $B \geq \frac{1}{4}A + \frac{1}{A}$.

The extremal condition $\partial f / \partial z = 0$ (keeping y fixed) leads to

$$y \;=\; y^* \;:=\; z \cdot \frac{B'(1-z) + A'}{B'(1-z) + A'z}.$$

[12] Inserting $y = 0$ and $x = a+b$ in the convexity definition $\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y)$ leads to $\alpha f(a+b) + (1-\alpha) f(0) \geq f(\alpha(a+b))$. Inserting $\alpha = \frac{a}{a+b}$ and $\alpha = \frac{b}{a+b}$ and adding both inequalities gives $f(a+b) + f(0) \geq f(a) + f(b)$ for $a,b \geq 0$. Using $f(0) \leq 0$ we get $f(\sum_i x_i) \geq \sum_i f(x_i)$ for $x_i \geq 0$ by induction.

[13] Inserting $b = y = -x$ and $\alpha = \frac{1}{2}$ into the convexity definition and using the symmetry of f we get $f(b) \geq f(0)$. Inserting this into $f(a+b) + f(0) \geq f(a) + f(b)$ we get $f(a+b) \geq f(a)$ which proves that f is monotonically increasing for positive arguments $(a,b \geq 0)$.

Inserting y^* into the definition of f and, again, replacing the relative entropy by the sum over squares (Lemma 1s), we get

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

$$g(z) \;:=\; 2B'A'^2z(1-z) + [(A'-1)B'(1-z) - A'](B' + A'\tfrac{z}{1-z}).$$

We have reduced the problem to showing $g\geq 0$. If the bracket [...] is positive, then g is positive. If the bracket is negative, we can decrease g by increasing $\frac{z}{1-z}\leq 1$ in $(B'+A'\frac{z}{1-z})$ to 1. The resulting expression is now quadratic in z with minima at the boundary values $z=0$ and $z=\frac{1}{2}$. It is therefore sufficient to check

$$g(0) \geq (AB-1)(A+B+2) \geq 0 \quad \text{and} \quad g(\tfrac{1}{2}) \geq \tfrac{1}{2}(AB-1)(2A+B+3) \geq 0$$

which is true for $B\geq\frac{1}{A}$. In summary we have proved (50) for $B\geq\frac{1}{A}+1$ and $A>0$ □.

C Binary Loss Inequality for $z\geq\frac{1}{2}$ (38)

With the definition

$$f(y,z) \;:=\; B'\cdot\left[y\ln\frac{y}{z} + (1-y)\ln\frac{1-y}{1-z}\right] + A'\cdot(1-y) - y\frac{1-z}{z} \quad , \quad z\geq\tfrac{1}{2} \qquad (51)$$

we show $f(y,z)\geq 0$ for suitable $A'\equiv A+1>1$ and $B'\equiv B+1>2$ similarly to Appendix B by proving that $f\geq 0$ at all extremal values and "at" boundaries. $f\to+\infty$ for $z\to 1$. The boundary $z=\frac{1}{2}$ has already been checked in Appendix B. The extremal condition $\partial f/\partial z=0$ (keeping y fixed) leads to

$$y \;=\; y^* \;:=\; z\cdot\frac{B'z}{(B'+1)z-1}.$$

Inserting y^* into the definition of f and replacing the relative entropy by the sum over squares (Lemma 1s), we get

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

$$g(z) \;:=\; [(A'-1)B'z - A' + 2z(1-z)](B'+1-\tfrac{1}{z}) + 2(1-z)^2.$$

We have reduced the problem to showing $g\geq 0$. Since $(B'+1-\frac{1}{z})\geq 0$ it is sufficient to show that the bracket is positive. We solve $[...]\geq 0$ w.r.t. B and get

$$B \geq \frac{1-2z(1-z)}{z}\cdot\frac{1}{A} + \frac{1-z}{z}.$$

For $B\geq\frac{1}{A}+1$ this is satisfied for all $\frac{1}{2}\leq z\leq 1$. In summary we have proved (51) for $B\geq\frac{1}{A}+1$ and $A>0$ □.

D General Loss Inequality (33)

We reduce

$$f(\mathbf{y},\mathbf{z}) := B'\sum_{i=1}^{N} y_i \ln\frac{y_i}{z_i} + A'\sum_{i=1}^{N} y_i\ell_{im} - \sum_{i=1}^{N} y_i\ell_{is} \;\geq\; 0 \tag{52}$$

$$\text{for} \quad \textstyle\sum_{i=1}^{N} z_i d_i \geq 0, \quad d_i := \ell_{im} - \ell_{is} \tag{53}$$

to the binary $N=2$ case. We do this by keeping $\mathbf{y}$ fixed and showing that f as a function of $\mathbf{z}$ is positive at all extrema in the interior of the simplex $\Delta := \{\mathbf{z}: \sum_i z_i = 1, z_i \geq 0\}$ of the domain of $\mathbf{z}$ and "at" all boundaries. First, the boundaries $z_i \to 0$ are safe as $f \to \infty$ for $B' > 0$. Variation of f w.r.t. to $\mathbf{z}$ leads to a minimum at $\mathbf{z}=\mathbf{y}$. If $\sum_i z_i d_i \geq 0$, we have

$$f(\mathbf{y},\mathbf{y}) \;=\; \sum_i y_i(A'\ell_{im} - \ell_{is}) \;\geq\; \sum_i y_i(\ell_{im} - \ell_{is}) \;=\; \sum_i z_i d_i \;\geq\; 0.$$

In the first inequality we used $A' > 1$. If $\sum_i z_i d_i < 0$, $\mathbf{z}=\mathbf{y}$ is outside the valid domain due to the constraint (53) and the valid minima are attained at the boundary $\Delta \cap P$, $P := \{\mathbf{z}: \sum_i z_i d_i = 0\}$. We implement the constraints with the help of Lagrange multipliers and extremize

$$L(\mathbf{y},\mathbf{z}) := f(\mathbf{y},\mathbf{z}) + \lambda \sum z_i + \mu \sum z_i d_i.$$

$\partial L/\partial z_i = 0$ leads to $y_i = y_i^* := z_i(\lambda + \mu d_i)$. Summing this equation over i we obtain $\lambda = 1$. μ is a function of $\mathbf{y}$ for which a formal expression might be given. If we eliminate y_i in favor of z_i, we get

$$f(\mathbf{y}^*,\mathbf{z}) \;=\; \sum_i c_i z_i \quad , \qquad c_i := (1+\mu d_i)(B'\ln(1+\mu d_i) + A'\ell_{im} - \ell_{is}).$$

In principle μ is a function of $\mathbf{y}$ but we can treat μ directly as an independent variable, since $\mathbf{y}$ has been eliminated.

The next step is to determine the extrema of the function $f = \sum c_i z_i$ for $\mathbf{z} \in \Delta \cap P$. For clearness we state the line of reasoning for $N = 3$. In this case Δ is a triangle. As f is linear in $\mathbf{z}$ it assumes its extrema at the vertices of the triangle, where all $z_i = 0$ except one. But we have to take into account a further constraint $\mathbf{z} \in P$. The plane P intersects triangle Δ in a finite line (for $\Delta \cap P = \{\}$ the only boundaries are $z_i \to 0$ which have already been treated). Again, as f is linear, it assumes its extrema at the ends of the line, i.e. at edges of the triangle Δ on which all but two z_i are zero. With a similar line of arguments for $N > 3$ we conclude that a necessary condition for a minimum of f at the boundary is that at most two z_i are non-zero. But this implies that all but two y_i are zero. If we had eliminated $\mathbf{z}$ in favor of $\mathbf{y}$, we could not have made the analogous conclusion because $y_i = 0$ does not necessarily imply $z_i = 0$. We have effectively reduced the problem of showing $f(\mathbf{y}^*,\mathbf{z}) \geq 0$ to the case $N = 2$. We can go back one step further and prove (52) for $N = 2$, which implies $f(\mathbf{y}^*,\mathbf{z}) \geq 0$ for $N = 2$. A proof of (52) for $N = 2$ implies, by the arguments given above, that it holds for all N. This is what we set out to show here $\Box$.

The $N = 2$ case is proven in the main text and in Appendices B and C .

References

[AG00] 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.

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

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

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–471, 1990.

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

[Ces97] N. Cesa-Bianchi et al. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[Cha75] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.

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

[Daw84] A. P. Dawid. Statistical theory. The prequential approach. *J.R. Statist. Soc. A*, 147:278–292, 1984.

[Doo53] J. L. Doob. *Stochastic Processes*. John Wiley & Sons, New York, 1953.

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

[FS97] 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.

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

[GTV01] P. Gács, J. Tromp, and M. B. Vitányi. Algorithmic statistics. *IEEE Transactions on Information Theory*, 47(6):2443–2463, 2001.

[HKW98] Haussler, Kivinen, and Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

[Hut01a] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

[Hut01b] M. Hutter. Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. *Proceedings of the 12th Eurpean Conference on Machine Learning (ECML-2001)*, pages 226–238, 2001.

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

[KW99] 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.

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

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

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

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

[LW89] 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.

[LW94] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[MF98] M. Merhav and N. Feder. Universal prediction. *IEEETIT: IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[Ris89] J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publ. Co., 1989.

[Ris96] J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans on Information Theory*, 42(1):40–47, January 1996.

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

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

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

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

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

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

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

[Vov87] Vovk. On a randomness criterion. *DOKLADY: Russian Academy of Sciences Doklady. Mathematics (formerly Soviet Mathematics–Doklady)*, 35, 1987.

[Vov92] V. G. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992.

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

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

[Yam98] K. Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*, 44:1424–1439, 1998.

[YE01] R. Yaroshinsky and R. El-Yaniv. Smooth online learning of expert advice. *Submitted for publication*, 2001.

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