Convergence and Loss Bounds for Bayesian Sequence Prediction

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland*

marcus@idsia.ch, http://www.idsia.ch/~marcus

Abstract. The probability of observing x_t at time t, given past observations $x_1...x_{t-1}$ can be computed with Bayes' rule if the true generating distribution μ of the sequences $x_1x_2x_3...$ is known. If μ is unknown, but known to belong to a class $\mathcal{M}$ one can base ones prediction on the Bayes mix ξ defined as a weighted sum of distributions $\nu\in\mathcal{M}$. Various convergence results of the mixture posterior ξ_t to the true posterior μ_t are presented. In particular a new (elementary) derivation of the convergence $\xi_t/\mu_t\to 1$ is provided, which additionally gives the rate of convergence. A general sequence predictor is allowed to choose an action y_t based on $x_1...x_{t-1}$ and receives loss $\ell_{x_t y_t}$ if x_t is the next symbol of the sequence. No assumptions are made on the structure of ℓ (apart from being bounded) and $\mathcal{M}$. The Bayes-optimal prediction scheme Λ_ξ based on mixture ξ and the Bayes-optimal informed prediction scheme Λ_μ are defined and the total loss L_ξ of Λ_ξ is bounded in terms of the total loss L_μ of Λ_μ. It is shown that L_ξ is bounded for bounded L_μ and $L_\xi/L_\mu\to 1$ for $L_\mu\to\infty$. Convergence of the instantaneous losses are also proven.

Keywords. Bayesian sequence prediction; general loss function and bounds; convergence; mixture distributions

1 Introduction

Setup. We consider inductive inference problems in the following form: Given a string $x_1x_2...x_{t-1}$, we want to predict its continuation x_t. We assume that the strings which have to be continued are drawn from a probability distribution μ. The maximal prior information a prediction algorithm can possess is the exact knowledge of μ, but in many cases the true generating distribution is not known. In order to overcome this problem a mixture distribution ξ is defined as a w_ν weighted sum over distributions $\nu\in\mathcal{M}$, where $\mathcal{M}$ is any discrete (hypothesis) set including μ. We assume that $\mathcal{M}$ is known and contains the true distribution, i.e. $\mu\in\mathcal{M}$. Since the posterior ξ_t can be shown to converge rapidly to the true posterior μ_t, making decisions based on ξ is often nearly as good as the infeasible optimal decision based on the unknown μ [MF98]. In this work we compare the expected loss of predictors based on mixture ξ to the expected loss of informed predictors based on μ.

Contents. Section 2 introduces concepts and notation needed later, including strings, probability distributions, mixture distributions, expectations, and various types of convergence and distance measures. Section 3 summarizes various convergence results of the mixture distribution ξ to the true distribution μ. We provide a new (elementary) derivation of the posterior convergence in ratio, which is not based on Martingales, but on the Hellinger distance, and compare it to related known results [Doo53, LV97, Vov87, VL00a]. Section 4 introduces the decision theoretic setup, where an action/prediction y_t results in a loss $\ell_{x_t y_t}$ if x_t is the next symbol of the sequence. Improving upon previous results in [MF98, Hut01a, Hut01b], the expected total (or cumulative) loss L_ξ made by the Bayes-optimal prediction scheme based on mixture ξ minus the expected total loss L_μ of the optimal informed prediction scheme based on μ is bounded by $O(\sqrt{L_\mu})$. Some popular loss functions, including the absolute, square, logarithmic, Hellinger, and error loss are discussed. A Proof of the loss bound is given in Section 5. Convergence of the instantaneous losses are briefly studied in Section 6. Section 7 recapitulates the assumptions made in this work and possible relaxations, mentions some optimality properties of ξ proven in [Hut02a], and provides an outlook to future work.

2 Preliminaries

Strings and Probability Distributions. We denote strings over a finite alphabet $\mathcal{X}$ by $x_1x_2...x_n$ with $x_t\in\mathcal{X}$. We abbreviate $x_{n:m}:=x_nx_{n+1}...x_{m-1}x_m$ and $x_{<n}:=x_1...x_{n-1}$. We use Greek letters for probability distributions/measures, especially ρ for arbitrary ones, $\mu\in\mathcal{M}$ for the true (generating) one, $\nu\in\mathcal{M}$ for arbitrary ones in $\mathcal{M}$, and ξ for the mixture (1). Let $\rho(x_{1:t})$ be the probability that an (infinite) sequence starts with $x_1...x_t$. The

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

conditional ρ probability that a given string $x_1...x_{t-1}$ is continued by x_t is $\rho_t := \rho(x_t|x_{<t}) = \rho(x_{1:t})/\rho(x_{<t})$. The considered prediction schemes will be based on these posteriors.

Mixture distributions. Let $\mathcal{M} := \{\mu_1,\mu_2,...\}$ be a finite or countable set of candidate probability distributions on strings. We define a weighted average on $\mathcal{M}$

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

ξ is called a Bayes-mixture. The weights w_ν may be interpreted as the prior belief in environment $\nu\in\mathcal{M}$. The most interesting property the mixture distribution ξ is that it multiplicatively dominates all distributions in $\mathcal{M}$:

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

In the following, we assume that $\mathcal{M}$ is known and contains the true distribution, i.e. $\mu\in\mathcal{M}$. If $\mathcal{M}$ is chosen sufficiently large, then $\mu\in\mathcal{M}$ is not a serious constraint. Generic classes, especially where $\mathcal{M}$ contains *all* (semi)computable probability distributions are discussed in [Sol78, LV97, Hut01a, Hut02a]. Generalizations to the case where $\mathcal{M}$ does not contain μ are briefly discussed in [Hut02a] and more intensively in a related context in [Grü98].

Expectations and convergence measures. We use $\mathbf{E}[..]$ to denote expectations w.r.t. the "true" distribution μ and abbreviate $\mathbf{E}_t[..] := \mathbf{E}[..|x_{<t}]$. If $[..]$ depends on $x_{1:t}$ only, i.e. is independent of $x_{t+1:\infty}$, we have

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

where $\sum'$ sums over all x_t or $x_{1:t}$ for which $\mu(x_{1:t})\neq 0$. Similarly we use $\mathbf{P}[..]$ to denote the μ probability of event $[..]$. We need the following kinds of convergence of a random sequence $z_1,z_2,...$ to (a random variable) z_*:

$$\begin{array}{rcl}
\text{with probability 1} & \text{(w.p.1)} & \mathbf{P}[z_t \stackrel{t\to\infty}{\longrightarrow} z_*] = 1 \\
\text{in probability} & \text{(i.p.)} & \forall\varepsilon : \mathbf{P}[|z_t - z_*| \geq \varepsilon] \stackrel{t\to\infty}{\longrightarrow} 0 \\
\text{in mean sum} & \text{(i.m.s.)} & \sum_{t=1}^\infty \mathbf{E}[(z_t - z_*)^2] < \infty \\
\text{in the mean} & \text{(i.m.)} & \mathbf{E}[(z_t - z_*)^2] \stackrel{t\to\infty}{\longrightarrow} 0
\end{array}$$

Convergence in one sense may imply convergence in another sense. The following implications are valid, strict, and complete:

$$i.m.s. \begin{array}{c} \nearrow\; w.p.1 \;\searrow \\ \searrow\; i.m. \;\nearrow \end{array} i.p.$$

Convergence i.m.s. is very strong: it provides a rate of convergence in the sense that the expected number of times t in which z_t deviates more than ε from z_* is finite and bounded by $\sum_{t=1}^\infty \mathbf{E}[(z_t - z_*)^2]/\varepsilon^2$.

Distance Measures. We need several distance measures between probability distributions $y_i\geq 0$, $z_i\geq 0$, $\sum_i y_i = \sum_i z_i = 1$, $i = \{1,...,N\}$, namely the

$$\begin{array}{rl}
\text{absolute distance:} & a = \sum_i |y_i - z_i| \qquad (3)\\
\text{square or Euclidian distance:} & s = \sum_i (y_i - z_i)^2 \\
\text{Hellinger distance:} & h = \sum_i (\sqrt{y_i} - \sqrt{z_i})^2 \\
\text{relative entropy or KL divergence:} & d = \sum_i y_i \ln\frac{y_i}{z_i} \\
\text{absolute divergence:} & b = \sum_i y_i |\ln\frac{y_i}{z_i}|
\end{array}$$

All bounds we prove in this work heavily rely on the following inequalities:

$$s \;\leq\; d, \qquad h \;\leq\; d, \qquad b-d \;\leq\; a \;\leq\; \sqrt{2d}. \qquad (4)$$

See [Hut01a], [CT91, Lem.12.6.1], and [BM98, p178] for proofs of $s\leq d$, $a\leq\sqrt{2d}$, and $h\leq d$, respectively. $b-d\leq a$ is elementary and follows from $-\ln x \leq \frac{1}{x} - 1$. Inequality $s\leq d$ is a generalization of the binary $N=2$ case used in [Sol78, Hut01c, LV97]. If we insert

$$\mathcal{X} = \{1,...,N\}, \quad N = |\mathcal{X}|, \quad i = x_t, \qquad (5)$$

$$y_i = \mu_t := \mu(x_t|x_{<t}), \quad z_i = \xi_t := \xi(x_t|x_{<t}) \qquad (6)$$

into (3) we get various *instantaneous distances* (at time t) between μ and ξ. If we take the expectation (over $x_{<t}$) and sum over $t = 1..n$, $(\sum_{t=1}^n \mathbf{E}[...])$ we get various *total distances* between μ and ξ:

$$\begin{array}{ll}
a_t(x_{<t}) := \sum_{x_t} |\mu_t - \xi_t|, & A_n := \sum_{t=1}^n \mathbf{E}[a_t] \\
s_t(x_{<t}) := \sum_{x_t} (\mu_t - \xi_t)^2, & S_n := \sum_{t=1}^n \mathbf{E}[s_t] \\
h_t(x_{<t}) := \sum_{x_t} (\sqrt{\mu_t} - \sqrt{\xi_t})^2, & H_n := \sum_{t=1}^n \mathbf{E}[h_t] \qquad (7)\\
d_t(x_{<t}) := \sum_{x_t} \mu_t \ln\frac{\mu_t}{\xi_t}, & D_n := \sum_{t=1}^n \mathbf{E}[d_t] \\
b_t(x_{<t}) := \sum_{x_t} \mu_t |\ln\frac{\mu_t}{\xi_t}|, & B_n := \sum_{t=1}^n \mathbf{E}[b_t]
\end{array}$$

3 Convergence of ξ to μ

For D_n the following representation and bound is well known and crucial [Sol78, LV97, Hut01a]

$$D_n \;\equiv\; \sum_{t=1}^n \mathbf{E}[d_t(x_{<t})] \;=\; \mathbf{E}[\ln\frac{\mu(x_{1:n})}{\xi(x_{1:n})}] \;\leq\; \ln w_\mu^{-1} \;<\; \infty \qquad (8)$$

The inequality follows from (2). The following theorem summarizes various bounds and convergence results needed later. The major new part is Theorem 1(iv) which allows for an elementary proof of $\xi_t/\mu_t \to 1$ w.p.1 based on the Hellinger distance.

Theorem 1 (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 mixture conditional probability $\xi'_t := \xi(x'_t|x_{<t})$ of the next symbol x'_t*

given $x_{<t}$ is related to the true conditional probability $\mu'_t := \xi(x'_t|x_{<t})$ in the following way:

i) $\sum_{t=1}^n \mathbf{E}[\sum_{x'_t}(\mu'_t - \xi'_t)^2] \equiv S_n \leq D_n \leq \ln w_\mu^{-1} < \infty$

ii) $\sum_{x'_t}(\mu'_t - \xi'_t)^2 \equiv s_t(x_{<t}) \leq d_t(x_{<t}) \stackrel{t\to\infty}{\longrightarrow} 0$ *w.p.1*

iii) $\xi'_t - \mu'_t \to 0$ *for* $t\to\infty$ *w.p.1 (and i.m.s) for any* x'_t

iv) $\sum_{t=1}^n \mathbf{E}[(\sqrt{\frac{\xi_t}{\mu_t}}-1)^2] \leq H_n \leq D_n \leq \ln w_\mu^{-1} < \infty$

v) $\sqrt{\frac{\xi_t}{\mu_t}} \to 1$ *i.m.s and* $\frac{\xi_t}{\mu_t} \to 1$ *w.p.1 for* $t\to\infty$

vi) $b_t - d_t \leq a_t \leq \sqrt{2d_t}, \quad B_n - D_n \leq A_n \leq \sqrt{2nD_n},$

where μ_t, ξ_t are defined in (6), d_t, D_n are the relative entropies (7), and w_μ is the weight (1) of μ in ξ.

Proof. The inequality in (ii) follows from the definitions (7) and from the entropy inequality $s \leq d$ (4). From the definition and finiteness of D_∞ (8) and from $d_t(x_{<t}) \geq 0$ one sees that $\sqrt{d_t(x_{<t})} \stackrel{i.m.s.}{\longrightarrow} 0$ for $t\to\infty$, which implies $d_t(x_{<t}) \stackrel{w.p.1}{\longrightarrow} 0$. The (first) inequality in (i) follows from (ii) by taking the $\mathbf{E}$ expectation and the $\sum_{t=1}^n$ sum. (iii) follows from (i) by dropping $\sum_{x'_t}$. (iv) and (v) are related to (i) and (iii), but are incomparable convergence results. (iv) is proven as follows:

$$\mathbf{E}_t[(\sqrt{\tfrac{\xi_t}{\mu_t}}-1)^2] = \textstyle\sum'_{x_t}\mu_t(\sqrt{\tfrac{\xi_t}{\mu_t}}-1)^2 = \tag{9}$$
$$= \textstyle\sum'_{x_t}(\sqrt{\xi_t}-\sqrt{\mu_t})^2 \leq h_t(x_{<t}) \leq d_t(x_{<t}).$$

The inequalities follow from (7) and $h \leq d$ (4). (iv) now follows by taking the $\mathbf{E}$ expectation and the $\sum_{t=1}^n$ sum. (v) follows from (iv) by the definition of convergence i.m.s., which implies convergence w.p.1. The first two inequalities in (vi) immediately follow from inequalities (4) and definitions (7). The third inequality of (vi) follows from the first by linearity of $\mathbf{E}$ and $\sum$. The last inequality follows from

$$\tfrac{1}{n}A_n \equiv \tfrac{1}{n}\textstyle\sum_{t=1}^n \mathbf{E}[a_t] \leq \tfrac{1}{n}\sum_{t=1}^n \mathbf{E}[\sqrt{2d_t}] \leq \tag{10}$$
$$\leq \tfrac{1}{n}\textstyle\sum_{t=1}^n \sqrt{\mathbf{E}[2d_t]} \leq \sqrt{\tfrac{1}{n}\sum_{t=1}^n \mathbf{E}[2d_t]} \equiv \sqrt{\tfrac{2}{n}D_n}$$

where we have used Jensen's inequality for exchanging the averages ($\frac{1}{n}\sum_{t=1}^n$ and $\mathbf{E}$) with the concave function $\sqrt{\cdot}$. $\square$

Since the conditional probabilities are the basis of the prediction algorithms considered in the next section and ξ'_t converges rapidly to μ'_t, we expect a good prediction performance if we use ξ as a guess of μ. Performance measures are defined in the next section.

Without the use of the Hellinger distance, a somewhat weaker statement than (v) can be derived from (vi):

$$\mathbf{E}|\ln\tfrac{\mu_t}{\xi_t}| = \mathbf{E}[b_t] \leq \mathbf{E}[d_t] + \mathbf{E}[\sqrt{2d_t}] \leq \mathbf{E}[d_t] + \sqrt{2\mathbf{E}[d_t]} \stackrel{t\to\infty}{\longrightarrow} 0,$$

since $\mathbf{E}[d_t]\to 0$. I.e. $\sqrt{|\ln\frac{\mu_t}{\xi_t}|} \stackrel{i.m.}{\longrightarrow} 0$, which implies $\frac{\xi_t}{\mu_t} \stackrel{i.p.}{\longrightarrow} 1$. The explicit appearance of n in the last expression of (vi) prevents proving stronger convergence of ξ_t/μ_t w.p.1 from (vi). Similarly [Bar00, Th.2] shows (in our notation) convergence of $\ln\frac{\mu(x_{1:t})}{\xi(x_{1:t})}$ in L_1-norm, which implies $\frac{\xi_t}{\mu_t} \stackrel{i.p.}{\longrightarrow} 1$, but is also not strong enough to derive (v).

The elementary proof for (v) w.p.1 given here does not rely on the semi-martingale convergence Theorem [Doo53, pp. 324–325] as the proof of Gács in [LV97, Th.5.2.2]. Furthermore, (iv) (and (i)) give a "rate" of convergence in the sense that the number of times ξ_t can depart from μ_t by more than ε in the sense of $|\sqrt{\xi_t/\mu_t}-1|>\varepsilon$ (or $|\xi'_t-\mu'_t|>\varepsilon$) is bounded by $\varepsilon^{-2}\ln w_\mu^{-1}$. Note also 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 $\mu'_t - \xi'_t$ converges to zero, but no statement is possible for ξ'_t/μ'_t, since $\liminf \mu'_t$ could be zero. On the other hand, if we stay *on* the μ-random sequence ($x'_{1:\infty} = x_{1:\infty}$), (v) shows that $\xi_t/\mu_t \to 1$ (whether $\inf\mu_t$ tends to zero or not does not matter). Indeed, it is easy to see that $\xi(1|0_{<t})/\mu(1|0_{<t}) \propto t \to \infty$ diverges for $\mathcal{M} = \{\mu,\nu\}$, $\mu(1|x_{<t}) := \frac{1}{2}t^{-3}$ and $\nu(1|x_{<t}) := \frac{1}{2}t^{-2}$, although $0_{1:\infty}$ is μ-random [Hut02a].

An interesting open question is whether ξ converges to μ (in difference (iii) or ratio (v)) individually for all Martin-Löf (M.L.) random sequences. Convergence M.L. implies convergence *w.p.*1, but the converse may fail on a set of sequences with μ-measure zero. A convergence M.L. result would be particularly interesting for $\mathcal{M}$ being the set of all enumerable semimeasures and ξ being Solomonoff's universal prior. Vovk's interesting results [Vov87] are not strong enough to settle this point, and the proof given in [VL00a] is incomplete. See [Hut02a] for further discussions.

4 Loss Bounds

Setup. A prediction is very often the basis for some decision. The decision results in an action, which itself leads to some reward or loss. We assume that the action itself does not influence the environment. Let $\ell_{x_t y_t} \in I\!R$ be the received loss when acting $y_t \in \mathcal{Y}$, and $x_t \in \mathcal{X}$ is the actual outcome. In many cases the prediction of x_t can be identified or is already the action y_t. $\mathcal{X} \equiv \mathcal{Y}$ in these cases. 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

$$y_t^{\Lambda_\rho} \;:=\; \arg\min_{y_t\in\mathcal{Y}} \sum_{x_t} \rho(x_t|x_{<t})\ell_{x_t y_t} \tag{11}$$

which minimizes the ρ-expected loss.[1] 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

$$l_t^{\Lambda_\rho}(x_{<t}) \;:=\; \mathbf{E}_t[\ell_{x_t y_t^{\Lambda_\rho}}], \quad L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^n \mathbf{E}[l_t^{\Lambda_\rho}(x_{<t})]. \quad (12)$$

Let Λ be *any* (causal) prediction scheme (deterministic or probabilistic does not matter) with no constraint at all, predicting *any* $y_t^\Lambda \in \mathcal{Y}$ with losses l_t^Λ and L_n^Λ similarly defined as (12). If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss

$$L_n^{\Lambda_\mu} \;\leq\; L_n^\Lambda \quad \text{for any} \quad \Lambda. \quad (13)$$

We prove the following loss bound for the Λ_ξ predictor based on mixture ξ:

Theorem 2 (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 (11) acts (or predicts) as to minimize the ρ-expected loss. Λ_ξ is the prediction scheme based on the mixture 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 (12) are bounded in the following way*

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

where the relative entropy D_n (8) is bounded by $\ln w_\mu^{-1} < \infty$.

The implications of Theorem 2 can best be read off from the following corollary.

Corollary 3 (Loss bound) *Under the same conditions as in Theorem 2 the following relations hold*

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

Let Λ be any *prediction scheme.*

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

The Corollary is a trivial consequence of Theorem 2 and (13). (vi) follows from Theorem 2 by replacing $L_n^{\Lambda_\mu}$ with L_n^Λ and solving the quadratic inequality w.r.t. L_n^Λ. The main message is that the total loss $L_\infty^{\Lambda_\xi}$ of the mixture Λ_ξ predictor is finite if the total loss $L_\infty^{\Lambda_\mu}$ of the informed Λ_μ predictor is finite, and that $L_n^{\Lambda_\xi}/L_n^{\Lambda_\mu} \to 1$ if $L_\infty^{\Lambda_\mu}$ is not finite. (vi) shows that *no* (causal) predictor Λ whatsoever achieves significantly less (expected) loss than Λ_ξ. Worst case bounds for aggregating strategies, especially the one derived in [CB97], explicitly depend on the comparison class. There are always predictors which perform significantly better than the aggregating strategy. On the other hand these algorithms have the remarkable property that the bounds hold for *any* sequence, whereas our bounds only hold in an expected sense and depend on the environment $\mu \in \mathcal{M}$. See [Hut01b] for a more detailed discussion of the bounds in general and this duality in particular.

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]. (The special case of the error-loss has earlier been considered in [BCH93]). 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)

$$L_n^{\Lambda_\xi} - L_n^{\Lambda_\mu} \;\leq\; A_n \;\leq\; \sqrt{2nD_n}. \quad (14)$$

In Section 6 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 $\mathbf{E}$ and the average $\frac{1}{n}\sum_{t=1}^n$ and using Theorem 1 shows (14).

Bound (14) and our bound (Theorem 2) are in general incomparable. Since $2D_\infty$ is finite and $L_n^{\Lambda_\mu} \leq n$, bound (14) 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 3, especially (i) and (ii)).

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 [Hut01a]. In this case $L_n^{\Lambda_\rho} \equiv E_n^{\Theta_\rho}$ is the total expected number of prediction errors. For $\mathcal{X} = \mathcal{Y} = \{0,1\}$, 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

[1] $\text{argmin}_y(\cdot)$ 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, although even this assumption may be removed.

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 $l_t^{\Lambda_\xi}-l_t^{\Lambda_\mu}=d_t$ and the total loss excess $L_n^{\Lambda_\xi}-L_n^{\Lambda_\mu}=D_n\leq\ln w_\mu^{-1}$ is finitely bounded anyway and Theorem 2 is not needed.

5 Loss Bound Proof

Main steps. The first inequality in Theorem 2 has already been proven (13). For the second and last inequality, we start looking for constants $A>0$ and $B>0$, which satisfy the linear inequality

$$L_n^{\Lambda_\xi} \;\leq\; (A+1)L_n^{\Lambda_\mu}+(B+1)D_n. \tag{15}$$

If we could show

$$l_t^{\Lambda_\xi}(x_{<t}) \;\leq\; A' l_t^{\Lambda_\mu}(x_{<t})+B' d_t(x_{<t}) \tag{16}$$

with $A':=A+1$ and $B':=B+1$ for all $t\leq n$ and all $x_{<t}$, (15) would follow immediately by summation and the definition of L_n and D_n. With the abbreviations the $m=y_t^{\Lambda_\mu}$ and $s=y_t^{\Lambda_\xi}$ and the abbreviations (5) and (6) 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 (16) we get

$$\sum_{i=1}^N y_i\ell_{is} \;\leq\; A'\sum_{i=1}^N y_i\ell_{im}+B'\sum_{i=1}^N y_i\ln\frac{y_i}{z_i} \tag{17}$$

By definition (11) of $y_t^{\Lambda_\mu}$ and $y_t^{\Lambda_\xi}$ we have

$$\sum_i y_i\ell_{im}\leq\sum_i y_i\ell_{ij} \quad\text{and}\quad \sum_i z_i\ell_{is}\leq\sum_i z_i\ell_{ij} \tag{18}$$

for all j. Actually, we need the first constraint only for $j=s$ and the second for $j=m$. In the final paragraph of this section 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.

$$B'\sum_{i=0}^1 y_i\ln\frac{y_i}{z_i}+\sum_{i=0}^1 y_i(A'\ell_{im}-\ell_{is}) \;\overset{?}{\geq}\; 0 \tag{19}$$

The cases $\ell_{im}>\ell_{is}\forall i$ and $\ell_{is}>\ell_{im}\forall i$ contradict the first/second inequality (18). 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 (19) as

$$f(y,z) \;:= \tag{20}$$

$$B'[y\ln\tfrac{y}{z}+(1-y)\ln\tfrac{1-y}{1-z}]+A'(1-y)a-yb+Ac \;\overset{?}{\geq}\; 0$$

for $zb\leq(1-z)a$ and $0\leq a,b,c,y,z\leq 1$. The constraint (18) on y has been dropped since (20) 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, (20) is proven for $z\leq\frac{1}{2}$ if we can prove

$$f_1(y,z):=B'[...]+A'(1-y)\tfrac{z}{1-z}-y \;\overset{?}{\geq}\; 0 \text{ for } z\leq\tfrac{1}{2}. \tag{21}$$

In the next paragraph of this section 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 (20) is proven for $z\geq\frac{1}{2}$ if we can prove

$$f_2(y,z):=B'[...]+A'(1-y)-y\tfrac{1-z}{z} \;\overset{?}{\geq}\; 0 \text{ for } z\geq\tfrac{1}{2}. \tag{22}$$

In the second next paragraph of this section we prove that it holds for $B\geq\frac{1}{A}+1$. So in summary we proved that (15) holds for $B\geq\frac{1}{A}+1$. Inserting $B=\frac{1}{A}+1$ into (15) and minimizing the r.h.s. w.r.t. A leads to the last bound of Theorem 2 with $A=\sqrt{D_n/L_n^{\Lambda_\mu}}$. Actually inequalities (21) and (22) 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 2. 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 (21) and (22) numerically for $B=\frac{1}{4}A+\frac{1}{A}$. $\square$

Binary loss inequality for $z\leq\frac{1}{2}$ (21). We now prove $f_1(y,z)\geq 0$ for $z\leq\frac{1}{2}$ and suitable $A'\equiv A+1$ and $B'\equiv B+1$. We do this by showing that $f_1\geq 0$ at all extremal values and "at" boundaries. $f_1\to+\infty$ 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 $s\leq d$ (4)

$$f_1(y,\tfrac{1}{2})\geq 2B'(y-\tfrac{1}{2})^2+A'(1-y)-y\geq 0 \quad \text{for} \quad B\geq\tfrac{1}{4}A+\tfrac{1}{A}$$

as can be shown by minimizing the r.h.s. w.r.t. y. Furthermore for $A\geq 4$ and $B\geq 1$ we have $f_1(y,\frac{1}{2})\geq 2(1-y)(3-2y)\geq 0$. Hence $f_1(y,\frac{1}{2})\geq 0$ 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

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

Inserting y^* into the definition of f_1 and, again, replacing the relative entropy by the sum over squares ($y\ln\frac{y}{z}+(1-y)\ln\frac{1-y}{1-z}\geq 2(y-z)^2$), which is a special case of $s\leq d$ (4), we get

$$f_1(y^*,z)\geq 2B'(y^*-z)^2+A'(1-y^*)\tfrac{z}{1-z}-y^*=\tfrac{z(1-z)\cdot g_1(z)}{[B'(1-z)+A'z]^2},$$

$$g_1(z):=2B'A'^2z(1-z)+[(A'-1)B'(1-z)-A'](B'+A'\tfrac{z}{1-z}).$$

We have reduced the problem to showing $g_1\geq 0$. If the bracket [...] is positive, then g_1 is positive. If the bracket is negative, we can decrease g_1 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

$$g_1(0)\;\geq\;(AB-1)(A+B+2)\geq 0 \quad \text{and}$$

$$g_1(\tfrac{1}{2})\;\geq\;\tfrac{1}{2}(AB-1)(2A+B+3)\geq 0$$

which is true for $B\geq\frac{1}{A}$. In summary we have proved (21) for $B\geq\frac{1}{A}+1$ and $A>0$. $\square$

Binary loss inequality for $z\geq\frac{1}{2}$ (22). We now prove we show $f_2(y,z)\geq 0$ for $z\geq\frac{1}{2}$ and suitable $A'\equiv A+1>1$ and $B'\equiv B+1>2$ similarly as in the last paragraph by proving that $f_2\geq 0$ at all extremal values and "at" boundaries. $f_2\to+\infty$ for $z\to 1$. The boundary $z=\frac{1}{2}$ has already been checked in in the last paragraph. 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_2 and replacing the relative entropy by the sum over squares $s\leq d$ (4), we get

$$f_2(y^*,z)\geq 2B'(y^*-z)^2+A'(1-y^*)-y^*\tfrac{1-z}{z}=\tfrac{z(1-z)\cdot g_2(z)}{[(B'+1)z-1]^2},$$

$$g_2(z)\;:=\;[(A'-1)B'z-A'+2z(1-z)](B'+1-\tfrac{1}{z})+2(1-z)^2.$$

We have reduced the problem to showing $g_2\geq 0$. 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 (22) for $B\geq\frac{1}{A}+1$ and $A>0$. $\square$

General loss inequality (17). We reduce

$$f(\mathbf{y},\mathbf{z}):=B'\sum_{i=1}^N y_i\ln\frac{y_i}{z_i}+A'\sum_{i=1}^N y_i\ell_{im}-\sum_{i=1}^N y_i\ell_{is}\;\geq\;0 \quad (23)$$

$$\text{for} \quad \textstyle\sum_{i=1}^N z_id_i\geq 0, \quad d_i:=\ell_{im}-\ell_{is} \quad (24)$$

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 $\sum_i z_id_i\geq 0$, we have

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

In the first inequality we used $A'>1$. If $\sum_i z_id_i<0$, $\mathbf{z}=\mathbf{y}$ is outside the valid domain due to the constraint (24) and the valid minima are attained at the boundary $\Delta\cap P$, $P:=\{\mathbf{z}:\sum_i z_id_i=0\}$. We implement the constraints with the help of Lagrange multipliers and extremize

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

$\partial L/\partial z_i=0$ 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

$$f(\mathbf{y}^*,\mathbf{z})\;=\;\textstyle\sum_i c_iz_i \quad \text{with}$$

$$c_i:=(1+\mu d_i)(B'\ln(1+\mu d_i)+A'\ell_{im}-\ell_{is}).$$

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 $f=\sum c_iz_i$ 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 (23) for $N=2$, which implies $f(\mathbf{y}^*,\mathbf{z})\geq 0$ for $N=2$. A proof of (23) for $N=2$ implies, by the arguments given above, that it holds for all N. This is what we set out to show here. $\square$

The $N=2$ case has been proven in the previous paragraphs. This completes the proof of Theorem 2. $\square$

6 Instantaneous Losses

Since $L_n^{\Lambda_\xi}-L_n^{\Lambda_\mu}$ is not finitely bounded by Theorem 2 it cannot be used directly to conclude analogously $l_t^{\Lambda_\xi}-l_t^{\Lambda_\mu}\to 0$. It would follow from $\xi_t\to\mu_t$ by continuity if $l_t^{\Lambda_\xi}$ and $l_t^{\Lambda_\mu}$ were continuous functions of ξ_t and μ_t. $l_t^{\Lambda_\mu}$ is a continuous piecewise linear concave function, but $l_t^{\Lambda_\xi}$ is an, in general, discontinuous function of ξ_t (and μ_t). Fortunately it is continuous at the one necessary point $\xi_t=\mu_t$. This allows to bound $l_t^{\Lambda_\xi}-l_t^{\Lambda_\mu}$ in terms of $\xi_t-\mu_t$.

Theorem 4 (Instantaneous Loss Bound) *Under the same conditions as in Theorem 2, for discrete $\mathcal{M}$ 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 mixture prediction schemes Λ_μ and Λ_ξ:*

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

Proof. (ii) follows from

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

To arrive at the first inequality we added $\sum_i z_i(\ell_{im}-\ell_{is})$ which is positive due to (18). $|\ell_{is}-\ell_{im}|\leq 1$ since $\ell\in[0,1]$. The last inequality follows from $a\leq\sqrt{2d}$ (4). (i) follows by inserting (ii) and using (8). (iii) follows from the proof of Theorem 2 by inserting $B=\frac{1}{A}+1=\sqrt{l_t^{\Lambda_\mu}/d_t}+1$ into (16). Convergence to zero holds for μ random sequences, i.e. w.p.1, since $l_t^{\Lambda_\mu}\leq 1$ is bounded. The losses $l_t^{\Lambda_\rho}(x_{<t})$ itself need not to converge. $\square$

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).

7 Conclusions

Generalization. The only assumptions we made in this work were that $\mu\in\mathcal{M}$, the loss ℓ is bounded to [0,1], and that the decision y_t does not influence the environment, i.e. μ is independent y_t. No other structural assumptions on $\mathcal{M}$ and ℓ have been made. The case $\mu\notin\mathcal{M}$ is briefly discussed in [Hut02a] and more intensively in [Grü98] in a related context. Simple scaling allows loss functions in arbitrary bounded interval [Hut01b]. Asymptotic loss/value bounds for an acting agent influencing the environment can be found in [Hut02b].

Optimality properties. In [Hut02a] we show that there are $\mathcal{M}$ and $\mu\in\mathcal{M}$ and weights w_ν such that the derived loss bounds are tight. This shows that the loss bounds cannot be improved in general, i.e. without making extra assumptions on ℓ, $\mathcal{M}$, or w_ν. 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 (in a decision theoretic sense) can always be based on a mixture distribution ξ. This still leaves open how to choose the weights. We give an Occam's razor argument that the choice $w_\nu\sim 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program describing ν, is optimal.

Outlook. The presented Theorems and proofs are independent of the size of $\mathcal{X}$ and hence should generalize to countably infinite and continuous alphabets under (minor) technical conditions. An infinite prediction space $\mathcal{Y}$ was no problem at all as long as we assumed the existence of $y_t^{\Lambda_\rho}\in\mathcal{Y}$ (11), but even this is not essential. The Λ_ρ schemes and theorems may be generalized to delayed sequence prediction, where the true symbol x_t is given only in cycle $t+d$. Another direction is to investigate the learning aspect of mixture 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 [VL00b] could lead to new insights. A unified picture of the loss bounds obtained here and the loss bounds for predictors based on expert advice (PEA) could also be fruitful. Also, bounds which say that the actual (not expected) loss suffered by Λ_ξ divided by the loss suffered by Λ_μ is with high probability close to 1 for sufficiently large n, would be interesting. Maximum-likelihood predictors may also be studied. See [Hut02a] for further references and discussions on the relation Bayes and PEA approaches and results, classification tasks, games of chances, infinite alphabet, continuous classes $\mathcal{M}$, universal mixtures, and others.

Summary. We compared mixture predictions based on Bayes-mixes ξ to the infeasible informed predictor based on the unknown true generating distribution μ. Convergence results of the mixture posterior ξ_t to the true posterior μ_t have been derived. A new (elementary) derivation of the convergence in ratio has been presented, including a rate of convergence. The 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, improving on various previous results.

References

[Bar00] A. R. Barron. Limits of information, markov chains, and projection. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 25–25, Sorrento, Italy, 2000.

[BCH93] A. R. Barron, B. S. Clarke, and D. Haussler. Information bounds for the risk of bayesian predictions and the redundancy of universal codes. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 54–54, 1993.

[BM98] A. A. Borovkov and A. Moullagaliev. *Mathematical Statistics.* Gordon & Breach, 1998.

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

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

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

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

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

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

[Hut01c] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

[Hut02a] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. Technical Report IDSIA-02-02, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Manno(Lugano), Switzerland, 2002.

[Hut02b] M. Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, Lecture Notes in Artificial Intelligence, pages 364–379, Sydney, Australia, 2002. Springer.

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

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

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

[VL00a] P. M. Vitányi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. on Information Theory*, 46(2):446–464, 2000.

[VL00b] 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] V. G. Vovk. On a randomness criterion. *DOKLADY: Russian Academy of Sciences Doklady. Mathematics (formerly Soviet Mathematics–Doklady)*, 35, 1987.