General Loss Bounds for Universal Sequence Prediction

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch http://www.idsia.ch/~marcus
Technical Report IDSIA-03-01, 10 April 2001

Keywords

Bayesian and deterministic prediction; general loss function; Solomonoff induction; Kolmogorov complexity; leaning; universal probability; loss bounds; games of chance; partial and delayed prediction; 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 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 estimate of the true distribution. In order to overcome this problem a universal distribution ξ is defined as a weighted sum of distributions $\mu_i \in M$, where M is any countable set of distributions including μ. This is a generalization of Solomonoff induction, in which M is the set of all enumerable semi-measures. Systems which predict y_t, given $x_1...x_{t-1}$ and which receive loss $l_{x_t y_t}$ if x_t is the true next symbol of the sequence are considered. It is proven that using the universal ξ as a prior is nearly as good as using the unknown true distribution μ. Furthermore, games of chance, defined as a sequence of bets, observations, and rewards are studied. The time needed to reach the winning zone is bounded in terms of the relative entropy of μ and ξ. Extensions to arbitrary alphabets, partial and delayed prediction, and more active systems are discussed.

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 of observation sequences $x_1x_2x_3....$ 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

Solomonoff [Sol64] had the idea to define a universal probability distribution[1] ξ as a weighted average over all possible computable probability distributions. Lower weights were assigned to more complex distributions. He unified Epicurus' principle of multiple explanations, Occams' razor, and Bayes' rule into an elegant formal theory. For a binary alphabet, the universal conditional probability used for predicting x_t converges to the true conditional probability for $t \to \infty$ with probability 1. The convergence serves as a justification of using ξ as a substitution

[1] We use the term *distribution* slightly unprecisely for a *probability measure*.

for the usually unknown μ. The framework can easily be generalized to other probability classes and weights [Sol78].

1.3 Contents

The main aim of this work is to prove expected loss bounds for general loss functions which measure the performance of ξ relative to μ, and to apply the results to games of chance. Details and proofs can be found in [Hut01]. There are good introductions and surveys of Solomonoff sequence prediction [LV97], inductive inference [AS83, Sol97], reasoning under uncertainty [Grü98], and competitive online statistics [Vov99] with interesting relations to this work. See [Hut01] and subsection 5.4 for details.

Section 2 explains notation and defines the generalized universal distribution ξ as the w_{μ_i} weighted sum of probability distributions μ_i of a set M, which must include the true distribution μ. This generalization is straightforward and causes no problems. ξ multiplicatively dominates all $\mu_i \in M$, and the relative entropy between μ and ξ is bounded by $\ln \frac{1}{w_\mu}$. Convergence of ξ to μ is shown in Theorem 1.

Section 3 considers the case where a prediction or action $y_t \in \mathcal{Y}$ results in a loss $l_{x_t y_t}$ if x_t is the next symbol of the sequence. Optimal universal Λ_ξ and optimal informed Λ_μ prediction schemes are defined for this case and loss bounds are proved. Theorems 2 and 3 bound the total loss L_ξ of Λ_ξ by the total loss L_μ of Λ_μ *plus* $O(\sqrt{L_\mu})$ terms.

Section 4 applies Theorem 3 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, asymptotically the universal Λ_ξ scheme will also become profitable. Theorem 4 lower bounds the time needed to reach the winning zone in terms of the relative entropy of μ and ξ. An attempt is made to give an information theoretic interpretation of the result.

Section 5 outlines possible extensions of the presented theory and results. They include arbitrary alphabets, partial, delayed and probabilistic prediction, classification, even more general loss functions, active systems influencing the environment, learning aspects, and a comparison to the weighted majority algorithm(s) and loss bounds.

2 Setup and Convergence

2.1 Strings and Probability Distributions

We denote binary strings by $x_1 x_2 ... x_n$ with $x_t \in \{0,1\}$. 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. Let $\rho(x_{1:t})$ be the probability that an (infinite) sequence starts with $x_1 ... x_t$. The conditional probability

$$\rho(x_t | x_{<t}) \;=\; \frac{\rho(x_{1:t})}{\rho(x_{<t})} \tag{1}$$

that a given string $x_1 ... x_{t-1}$ is continued by x_t is obtained by using Bayes' rule. The prediction schemes will be based on these posteriors.

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

$$\begin{aligned} \xi(x_{1:n}) \;&:=\; \sum_{\mu_i \in M} w_{\mu_i} \cdot \mu_i(x_{1:n}), \\ \sum_{\mu_i \in M} w_{\mu_i} &= 1, \quad w_{\mu_i} > 0. \end{aligned} \tag{2}$$

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

$$\xi(x_{1:n}) \;\geq\; w_{\mu_i} \cdot \mu_i(x_{1:n}) \quad \text{for all} \quad \mu_i \in M. \tag{3}$$

In the following, we assume that M is known and contains the true distribution, i.e. $\mu \in M$. This is not a serious

[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*.

constraint if we include *all* computable probability distributions in M with a high weight assigned to simple μ_i. Solomonoff's universal semi-measure is obtained if we include all enumerable semi-measures in M with weights $w_{\mu_i} \sim 2^{-K(\mu_i)}$, where $K(\mu_i)$ is the length of the shortest program for μ_i [Sol64, Sol78, LV97]. A detailed discussion of various general purpose choices for M is given in [Hut01].

Furthermore, we need the relative entropy between μ and ξ:

$$h_t(x_{<t}) \;:=\; \sum_{x_t\in\{0,1\}} \mu(x_t|x_{<t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} \qquad (4)$$

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

$$H_n := \sum_{t=1}^{n} \sum_{x_{<t}\in\{0,1\}^{t-1}} \mu(x_{<t})\cdot h_t(x_{<t}) \;\leq\; \ln \frac{1}{w_\mu} =: d_\mu \qquad (5)$$

The following theorem shows the important property of ξ converging to the true distribution μ, in a sense.

Theorem 1 (Convergence) *Let there be binary sequences $x_1x_2...$ 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:*

$$i) \quad \sum_{t=1}^{n} \sum_{x_{1:t}\in\{0,1\}^t} \mu(x_{<t})\Big(\mu(x_t|x_{<t}) - \xi(x_t|x_{<t})\Big)^2 \;\leq\; \leq\; H_n \;\leq\; d_\mu \;=\; \ln\tfrac{1}{w_\mu} \;<\; \infty$$

$$ii) \quad \xi(x_t|x_{<t}) \to \mu(x_t|x_{<t}) \quad \text{for } t\to\infty \text{ with } \mu \text{ probability 1}$$

where H_n is the relative entropy (5), and w_μ is the weight (2) of μ in ξ.

(i) and (5) are easy generalizations of [Sol78] to arbitrary weights w_μ and an arbitrary probability set M. For $n\to\infty$ the l.h.s. of (i) is an infinite t-sum over positive arguments, which is bounded by the finite constant d_μ on the r.h.s. Hence the arguments must converge to zero for $t\to\infty$. Since the arguments are μ expectations of the squared difference of ξ and μ, this means that $\xi(x_t|x_{<t})$ converges[3] to $\mu(x_t|x_{<t})$ with μ probability 1. This proves (ii). 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 next section.

[3]More precisely $\xi(x_t|x_{<t})-\mu(x_t|x_{<t})$ converges to zero for $t\to\infty$ with μ probability 1 or, more stringent, in a mean squared sense.

3 Loss Bounds

3.1 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 [Hut00]. To stay in the framework of (passive) prediction we have to assume that the action itself does not influence the environment. Let $l_{x_ty_t}\in\mathbb{R}$ be the received loss when taking action $y_t\in\mathcal{Y}$ and $x_t\in\{0,1\}$ is the t^{th} symbol of the sequence. We demand l to be normalized, i.e. $0\leq l_{x_ty_t}\leq 1$. For instance, if we make a sequence of weather forecasts $\{0,1\}=\{$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). Reasonable losses may be

Loss	sunny	rainy
umbrella	0.3	0.1
sunglasses	0.0	1.0

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*. In the following, we assume "predictive" actions of this kind, i.e. $\mathcal{Y}=\{0,1\}$. General action spaces $\mathcal{Y}$ and general alphabets $\mathcal{A}$ are considered in [Hut01].

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 $\mu(1|x_{<t})l_{1y_t}+\mu(0|x_{<t})l_{0y_t}$. The goal is to minimize the expected loss. More generally we define the Λ_ρ prediction scheme

$$y_t^{\Lambda_\rho} \;:=\; \arg\min_{y_t} \sum_{x_t\in\{0,1\}} \rho(x_t|x_{<t}) l_{x_ty_t} \qquad (6)$$

which minimizes the ρ-expected loss. This is a threshold strategy with $y_t^{\Lambda_\rho}=0/1$ for $\rho(1|x_{<t})\gtrless\gamma$, where $\gamma := \frac{l_{01}-l_{00}}{l_{01}-l_{00}+l_{10}-l_{11}}$. 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

$$\begin{aligned} l_{t\Lambda_\rho}(x_{<t}) \;&:=\; \sum_{x_t} \mu(x_t|x_{<t}) l_{x_t y_t^{\Lambda_\rho}}, \\ L_{n\Lambda_\rho} \;&:=\; \sum_{t=1}^{n}\sum_{x_{<t}} \mu(x_{<t})\cdot l_{t\Lambda_\rho}(x_{<t}). \end{aligned} \qquad (7)$$

In the special case $l_{01}=l_{10}=1$ and $l_{00}=l_{11}=0$, the bit with the highest ρ probability is predicted ($\gamma=\frac{1}{2}$), and $L_{n\Lambda_\rho}$ is the total expected number of prediction errors.

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

$$L_{n\Lambda_\mu} \;\leq\; L_{n\Lambda_\rho} \quad \text{for any} \quad \Lambda_\rho \tag{8}$$

The predictor Λ_ξ, based on the universal distribution ξ, is of special interest.

Theorem 2 (Unit loss bound) *Let there be binary sequences $x_1x_2...$ drawn with probability $\mu(x_{1:n})$ for the first n symbols. A system predicting $y_t \in \{0,1\}$ given $x_{<t}$ receives loss $l_{x_t y_t} \in [0,1]$ if x_t is the true t^{th} symbol of the sequence. The Λ_ρ-system (6) 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 (7) are bounded in the following way*

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

where $H_n \leq \ln \frac{1}{w_\mu}$ is the relative entropy (5), and w_μ is the weight (2) of μ in ξ.

First, we observe that the total loss $L_{\infty\Lambda_\xi}$ of the universal Λ_ξ predictor is finite if the total loss $L_{\infty\Lambda_\mu}$ of the informed Λ_μ predictor is finite. This is especially the case for deterministic μ and $l_{00} = l_{11} = 0$, as $L_{n\Lambda_\mu} \equiv 0$ in this case[4], i.e. Λ_ξ receives a finite loss on deterministic environments if a correct prediction results in zero loss. More precisely, $L_{\infty\Lambda_\xi} \leq 2H_\infty \leq 2\ln\frac{1}{w_\mu}$. A combinatoric argument shows that there are M and $\mu \in M$ with $L_{\infty\Lambda_\xi} \geq \log_2 |M|$. This shows that the upper bound $L_{\infty\Lambda_\xi} \leq 2\ln|M|$ for uniform w is rather tight. For more complicated probabilistic environments, where even the ideal informed system makes an infinite number of errors, the theorem ensures that the loss excess $L_{n\Lambda_\xi} - L_{n\Lambda_\mu}$ is only of order $\sqrt{L_{n\Lambda_\mu}}$. The excess is quantified in terms of the information content H_n of μ (relative to ξ), or the weight w_μ of μ in ξ. This ensures that the loss densities L_n/n of both systems converge to each other for $n \to \infty$. Actually, the theorem ensures more, namely that the quotient converges to 1, and also gives the speed of convergence $L_{n\Lambda_\xi}/L_{n\Lambda_\mu} = 1 + O(L_{n\Lambda_\mu}^{-1/2}) \longrightarrow 1$ for $L_{n\Lambda_\mu} \to \infty$.

3.2 Proof Sketch of Theorem 2

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

$$L_{n\Lambda_\xi} \;\leq\; (A+1)L_{n\Lambda_\mu} + (B+1)H_n. \tag{9}$$

If we could show

$$l_{t\Lambda_\xi}(x_{<t}) \;\leq\; A' l_{t\Lambda_\mu}(x_{<t}) + B' h_t(x_{<t}) \tag{10}$$

with $A' := A+1$ and $B' := B+1$ for all $t \leq n$ and all $x_{<t}$, (9) would follow immediately by summation and the definition of L_n and H_n. With the abbreviations

$$i = x_t, \quad y_i = \mu(x_t|x_{<t}), \quad z_i = \xi(x_t|x_{<t})$$

$$m = y_t^{\Lambda_\mu}, \quad s = y_t^{\Lambda_\xi}$$

the loss and entropy can be expressed by $l_{t\Lambda_\xi} = \sum_i y_i l_{is}$, $l_{t\Lambda_\mu} = \sum_i y_i l_{im}$ and $h_t = \sum_i y_i \ln \frac{y_i}{z_i}$. Inserting this into (10) and rearranging terms we have to prove

$$B' \sum_{i=0}^{1} y_i \ln \frac{y_i}{z_i} + \sum_{i=0}^{1} y_i (A' l_{im} - l_{is}) \;\stackrel{?}{\geq}\; 0. \tag{11}$$

By definition (6) of $y_t^{\Lambda_\mu}$ and $y_t^{\Lambda_\xi}$ we have

$$\sum_i y_i l_{im} \leq \sum_i y_i l_{ij} \quad \text{and} \quad \sum_i z_i l_{is} \leq \sum_i z_i l_{ij} \tag{12}$$

for all j. Actually, we need the first constraint only for $j=s$ and the second for $j=m$. The cases $l_{im} > l_{is} \forall i$ and $l_{is} > l_{im} \forall i$ contradict the first/second inequality (12). Hence we can assume $l_{0m} \geq l_{0s}$ and $l_{1m} \leq l_{1s}$. The symmetric case $l_{0m} \leq l_{0s}$ and $l_{1m} \geq l_{1s}$ is proved analogously or can be reduced to the first case by renumbering the indices $(0 \leftrightarrow 1)$. Using the abbreviations $a := l_{0m} - l_{0s}$, $b := l_{1s} - l_{1m}$, $c := y_1 l_{1m} + y_0 l_{0s}$, $y = y_1 = 1 - y_0$ and $z = z_1 = 1 - z_0$ we can write (11) as

$$f(y,z) \;:= \tag{13}$$

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

for $zb \leq (1-z)a$ and $0 \leq a,b,c,y,z \leq 1$. The constraint (12) on y has been dropped since (13) 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 ($h_t \geq 0$). 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, (13) is proved for $z \leq \frac{1}{2}$ if we can prove

$$B'[...] + A'(1-y)\tfrac{z}{1-z} - y \;\stackrel{?}{\geq}\; 0 \quad \text{for} \quad z \leq \tfrac{1}{2}. \tag{14}$$

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

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 (13) is proved for $z \geq \frac{1}{2}$ if we can prove

$$B'[...] + A'(1-y) - y\tfrac{1-z}{z} \stackrel{?}{\geq} 0 \quad \text{for} \quad z \geq \tfrac{1}{2}. \qquad (15)$$

In [Hut01] we prove that (14) and (15) indeed hold for $B \geq \frac{1}{4}A + \frac{1}{A}$. The cautious reader may check the inequalities numerically. So in summary we proved that (9) holds for $B \geq \frac{1}{4}A + \frac{1}{A}$. Inserting $B = \frac{1}{4}A + \frac{1}{A}$ into (9) and minimizing the r.h.s. with respect to A leads to the bound of Theorem 2 (with $A^2 = H_n/(L_{n\Lambda_\mu} + \frac{1}{4}H_n)$) $\Box$.

3.3 General Loss

There are only very few restrictions imposed on the loss $l_{x_t y_t}$ in Theorem 2, namely that it is static and in the unit interval $[0,1]$. If we look at the proof of Theorem 2, we see that the time-independence has not been used at all. The proof is still valid for an individual loss function $l^t_{x_t y_t} \in [0,1]$ for each step t. The loss might even depend on the actual history $x_{<t}$. The case of a loss $l^t_{x_t y_t}(x_{<t})$ bounded to a general interval $[l_{min}, l_{max}]$ can be reduced to the unit interval case by rescaling l. We introduce a scaled loss l'

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

The prediction scheme Λ'_ρ based on l' is identical to the original prediction scheme Λ_ρ based on l, since argmin in (6) is not affected by a constant scaling and a shift of its argument. From $y_t^{\Lambda'_\rho} = y_t^{\Lambda_\rho}$ it follows that $l'_{t\Lambda_\rho} = (l_{t\Lambda_\rho} - l_{min})/l_\Delta$ and $L'_{n\Lambda_\rho} = (L_{n\Lambda_\rho} - l_{min})/l_\Delta$ ($H'_n \equiv H_n$, since l is not involved). Theorem 2 is valid for the primed quantities, since $l' \in [0,1]$. Inserting $L'_{n\Lambda_{\mu/\xi}}$ and rearranging terms we get

Theorem 3 (General loss bound) *Let there be binary sequences $x_1 x_2 ...$ 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 $l^t_{x_t y_t}(x_{<t}) \in [l_{min}, l_{min} + l_\Delta]$ if x_t is the true t^{th} symbol of the sequence. The Λ_ρ-system (6) 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 (7) are bounded in the following way*

$$0 \;\leq\; L_{n\Lambda_\xi} - L_{n\Lambda_\mu} \;\leq$$
$$\leq\; l_\Delta H_n + \sqrt{4(L_{n\Lambda_\mu} - n l_{min}) l_\Delta H_n + l_\Delta^2 H_n^2}$$

where $H_n \leq \ln \frac{1}{w_\mu}$ is the relative entropy (5), and w_μ is the weight (2) of μ in ξ.

4 Application to Games of Chance

4.1 Introduction/Example

Think of 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 1 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 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.

4.2 Games of Chance

We use Theorem 3 (or its generalization to arbitrary action and alphabet, proved in [Hut01]) 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, $l_{x_t y_t} = -p_t$. 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 $\mu \in M$ for some M, for instance that μ is a computable probability distribution. From Theorem 3 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 3 allows us to lower bound the universal profit $P_{n\Lambda_\xi}$

$$P_{n\Lambda_\xi} \geq P_{n\Lambda_\mu} - p_\Delta H_n - \sqrt{4(np_{max} - P_{n\Lambda_\mu}) p_\Delta H_n + p_\Delta^2 H_n^2} \tag{16}$$

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 (16) is positive if, and only if

$$n \;>\; \frac{2p_\Delta(2p_{max} - \bar{p}_{n\Lambda_\mu})}{\bar{p}_{n\Lambda_\mu}^2} \cdot H_n. \tag{17}$$

Theorem 4 (Time to Win) *Let there be binary sequences $x_1 x_2$... 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 (6) 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:*

$$i) \quad \bar{p}_{n\Lambda_\xi} = \bar{p}_{n\Lambda_\mu} - O(n^{-1/2}) \longrightarrow \bar{p}_{n\Lambda_\mu} \quad for \quad n \to \infty$$

$$ii) \quad if \quad n > \left(\frac{2p_\Delta}{\bar{p}_{n\Lambda_\mu}}\right)^2 \cdot d_\mu \quad and \quad \bar{p}_{n\Lambda_\mu} > 0 \implies \bar{p}_{n\Lambda_\xi} > 0$$

where $w_\mu = e^{-d_\mu}$ is the weight (2) of μ in ξ.

By dividing (16) by n and using $H_n \leq d_\mu$ (5) we see that the leading order of $\bar{p}_{n\Lambda_\xi} - \bar{p}_{n\Lambda_\mu}$ is bounded by $\sqrt{4p_\Delta p_{max} d_\mu / n}$, which proves (i). The condition in (ii) is actually a weakening of (17). $P_{n\Lambda_\xi}$ 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 (17), since $p_\Delta > p_{max}$, and (17) implies positive (16), 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.

4.3 Information-Theoretic Interpretation

We try to give an intuitive explanation of Theorem 4(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 $\ln 2 \cdot H_\infty \leq d_\mu \cdot \ln 2$. One might think of a Shannon-Fano prefix code of $\mu_i \in M$ of length $\lceil d_{\mu_i} \cdot \ln 2 \rceil$, which exists since the Kraft inequality $\sum_i 2^{-\lceil d_{\mu_i} \ln 2 \rceil} \leq \sum_i w_{\mu_i} \leq 1$ is satisfied. $d_\mu \cdot \ln 2$ bits have to be learned before Λ_ξ can be as good as Λ_μ. In the worst case, the only information contained in 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, $d_\mu \ln 2$ bits have to be transmitted, which requires $n \geq O((p_\Delta / \bar{p}_{n\Lambda_\mu})^2) \cdot d_\mu \ln 2$ 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 2 and 3.

5 Outlook

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

5.1 General Alphabet

In many, cases the prediction unit is not a bit, but a letter from a finite alphabet $\mathcal{A}$. Non-binary prediction cannot be (easily) reduced to the binary case. One might think of a binary coding of the symbols $x_t \in \mathcal{A}$ in the sequence $x_1 x_2$.... 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, considered here and in [Sol78]! Fortunately, all theorems (1-4) take over to general alphabet [Hut01]. Unfortunately, the proofs are rather complex. In many cases the basic prediction unit is not even a letter from a finite alphabet, but a number (for inducing number sequences), or a word (for completing sentences), a real number or vector (for physical measurements). The prediction may either be generalized to a block by block prediction of symbols or, more suitably, the finite alphabet $\mathcal{A}$ could be generalized to countable

(numbers, words) or continuous (real or vector) alphabet. The theorems should generalize to countably infinite alphabets by appropriately taking the limit $|\mathcal{A}| \to \infty$ and to continuous alphabets by a denseness or separability argument.

5.2 Partial Prediction, Delayed Prediction, Classification

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

5.3 More Active Systems

Prediction means guessing the future, but not influencing it. We mentioned the possibility of interpreting $y_t \in \mathcal{Y}$ as an action with $\mathcal{Y} \neq \mathcal{A}$. This tiny step towards a more active system is described in more detail in [Hut01]. 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 (6) 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 [Hut00], but loss bounds have yet to be proven.

5.4 The Weighted Majority Algorithm(s)

The Weighted Majority (WM) algorithm is a related universal forecasting algorithm. It was invented by Littlestone and Warmuth [LW89, LW94] and Vovk [Vov92] and further developed in [Ces97, HKW98, KW99] and others. Many variations known by many names have meanwhile been invented. Early works in this direction are [Daw84, Ris89]. See [Vov99] for a review and further references. The setting and basic idea of WM are the following. 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 defined in [Hut01]. The WM algorithm $p_{\beta n}$ combines the predictions of all experts. It forms its own prediction x_t^p 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 most general bound of this kind is [Ces97]

$$L_p(\mathbf{x}) \leq L_\varepsilon(\mathbf{x}) + 2.8 \ln|\mathcal{E}| + 4\sqrt{L_\varepsilon(\mathbf{x}) \ln|\mathcal{E}|}. \qquad (18)$$

It is interesting that our bound in Theorem 2 (with $H_n \leq \ln|M|$ for uniform weights) has a quite similar structure as this bound, 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 ρ predictor), but only for a given set of environments M. WM depends on the set of expert, Λ_ξ depends on the set of environments M. The basic $p_{\beta n}$ algorithm has been extended in different directions: incorporation of different initial weights ($|\mathcal{E}| \hookrightarrow \ln \frac{1}{w_i}$) [LW89, Vov92], more general loss functions [HKW98], continuous valued outcomes [HKW98], and multi-dimensional predictions [KW99] (but not yet for the absolute loss). The works of Yamanishi [Yam97] and [Yam98] lie 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 different classes of loss functions). Finally, note that the predictions of WM are continuous. In a sense it is more natural to predict 0 or 1 on a binary sequence, rather than some real number. On the other hand it is possible to convert the continuous prediction of WM into a probabilistic binary prediction by interpreting $x_t^p \in [0,1]$ as the probability of predicting 1, and $|x_t - x_t^p|$ as 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 $\mathcal{A} = \{0,1\}^d$, but error bounds for the absolute loss have yet to be proven. It would be interesting to generalize WM and bound (18) to arbitrary alphabet and to general loss functions with probabilistic interpretation.

5.5 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. The attempt at an information theoretic interpretation of Theorem 4 may be made more rigorous in this or another way. In the end, this may lead to a simpler proof of Theorem 4 and maybe even for the loss bounds. Finally, the system should be implemented and tested on specific induction problems for specific finite M with computable ξ.

6 Summary

Solomonoff's universal probability measure has been generalized to arbitrary probability classes and weights. A wise choice of M widens the applicability by reducing the computational burden for ξ. A framework, where predictions result in losses of arbitrary, but known form, has been considered. Loss bounds for general loss functions have been proved, which show that the universal prediction scheme Λ_ξ can compete with the best possible informed scheme Λ_μ. The results show that universal prediction is ideally suited for games of chance with a sequence of bets, observations, and rewards. Extensions in various directions have been suggested.

Acknowledgements

I want to thank Ray Solomonoff for many valuable discussions and for encouraging me to derive the general loss bounds presented here.

References

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

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

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

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

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

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

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

[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.

[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.

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

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

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

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

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

[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.

[Yam97] K. Yamanishi. On-line maximum likelihood prediction with respect to general loss functions. *Journal of Computer and System Sciences*, 55(1):105–118, 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.