Technical Report IDSIA-13-04

On the Convergence Speed of MDL Predictions for Bernoulli Sequences

Jan Poland and **Marcus Hutter**

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

15 July 2004

Abstract

We consider the Minimum Description Length principle for online sequence prediction. If the underlying model class is discrete, then the total expected square loss is a particularly interesting performance measure: (a) this quantity is bounded, implying convergence with probability one, and (b) it additionally specifies a *rate of convergence*. Generally, for MDL only exponential loss bounds hold, as opposed to the linear bounds for a Bayes mixture. We show that this is even the case if the model class contains only Bernoulli distributions. We derive a new upper bound on the prediction error for countable Bernoulli classes. This implies a small bound (comparable to the one for Bayes mixtures) for certain important model classes. The results apply to many Machine Learning tasks including classification and hypothesis testing. We provide arguments that our theorems generalize to countable classes of i.i.d. models.

Keywords

MDL, Minimum Description Length, Convergence Rate, Prediction, Bernoulli, Discrete Model Class.

*This work was supported by SNF grant 2100-67712.02.

1 Introduction

"Bayes mixture", "Solomonoff induction", "marginalization", all these terms refer to a central induction principle: Obtain a predictive distribution by integrating the product of prior and evidence over the model class. In many cases however, the Bayes mixture cannot be computed, and even a sophisticated approximation is expensive. The MDL or MAP (maximum a posteriori) estimator is both a common approximation for the Bayes mixture and interesting for its own sake: Use the model with the largest product of prior and evidence. (In practice, the MDL estimator is usually being approximated too, in particular when only a local maximum is determined.)

How good are the predictions by Bayes mixtures and MDL? This question has attracted much attention. In the context of prediction, arguably the most important quality measure is the *total* or cumulative *expected loss* of a predictor. A very common choice of loss function is the square loss. Throughout this paper, we will study this quantity in an *online setup*.

Assume that the outcome space is finite, and the model class is continuously parameterized. Then for Bayes mixture prediction, the cumulative expected square loss is usually small but unbounded, growing with $\log n$, where n is the sample size [CB90]. This corresponds to an *instantaneous* loss bound of $\frac{1}{n}$. For the MDL predictor, the losses behave similarly [Ris96, BRY98] under appropriate conditions, in particular with a specific prior. (Note that in order to do MDL for continuous model classes, one needs to *discretize* the parameter space, e.g. [BC91].)

On the other hand, if the model class is discrete, then Solomonoff's theorem [Sol78, Hut01] bounds the cumulative expected square loss for the Bayes mixture predictions finitely, namely by $\ln w_\mu^{-1}$, where w_μ is the prior weight of the "true" model μ. The only necessary assumption is that the true distribution μ is contained in the model class. For the corresponding MDL predictions, we have shown [PH04] that a bound of w_μ^{-1} holds. This is exponentially larger than the Solomonoff bound, and it is sharp in general. A finite bound on the total expected square loss is particularly interesting:

1. It implies convergence of the predictive to the true probabilities with probability one. In contrast, an instantaneous loss bound which tends to zero implies only convergence in probability.

2. Additionally, it gives a *convergence speed*, in the sense that errors of a certain magnitude cannot occur too often.

So for both, Bayes mixtures and MDL, convergence with probability one holds, while the convergence rate is exponentially worse for MDL compared to the Bayes mixture.

It is therefore natural to ask if there are model classes where the cumulative loss of MDL is comparable to that of Bayes mixture predictions. Here we will concentrate

on the simplest possible stochastic case, namely discrete Bernoulli classes (compare also [Vov97]). It might be surprising to discover that in general the cumulative loss is still exponential. On the other hand, we will give mild conditions on the prior guaranteeing a small bound. We will provide arguments that these results generalize to arbitrary i.i.d. classes. Moreover, we will see that the instantaneous (as opposed to the cumulative) bounds are always small ($\approx \frac{1}{n}$). This corresponds to the well-known fact that the instantaneous square loss of the Maximum Likelihood estimator decays as $\frac{1}{n}$ in the Bernoulli case.

A particular motivation to consider discrete model classes arises in Algorithmic Information Theory. From a computational point of view, the largest relevant model class is the countable class of all computable models (isomorphic to programs) on some fixed universal Turing machine. We may study the corresponding Bernoulli case and consider the countable set of computable reals in $[0,1]$. We call this the *universal setup*. The description length $K(\vartheta)$ of a parameter $\vartheta \in [0,1]$ is then given by the length of the shortest program that outputs ϑ, and a prior weight may be defined by $2^{K(\vartheta)}$.

Many Machine Learning tasks are or can be reduced to sequence prediction tasks. An important example is classification. The task of classifying a new instance z_n after having seen (instance,class) pairs $(z_1,c_1),...,(z_{n-1},c_{n-1})$ can be phrased as to predict the continuation of the sequence $z_1c_1...z_{n-1}c_{n-1}z_n$. Typically the (instance,class) pairs are i.i.d.

Our main tool for obtaining results is the Kullback-Leibler divergence. Lemmata for this quantity are stated in Section 2. Section 3 shows that the exponential error bound obtained in [PH04] is sharp in general. In Section 4, we give an upper bound on the instantaneous and the cumulative losses. The latter bound is small e.g. under certain conditions on the distribution of the weights, this is the subject of Section 5. Section 6 treats the universal setup. Finally, in Section 7 we discuss the results and give conclusions.

2 Kullback-Leibler Divergence

Let $\mathbb{B} = \{0,1\}$ and consider finite strings $x \in \mathbb{B}^*$ as well as infinite sequences $x_{<\infty} \in \mathbb{B}^\infty$, with the first n bits denoted by $x_{1:n}$. If we know that x is generated by an i.i.d random variable, then $P(x_i = 1) = \vartheta_0$ for all $1 \leq i \leq \ell(x)$ where $\ell(x)$ is the length of x. Then x is called a Bernoulli sequence, and $\vartheta_0 \in \Theta \subset [0,1]$ the *true parameter*. In the following we will consider only countable Θ, e.g. the set of all computable numbers in $[0,1]$.

Associated with each $\vartheta \in \Theta$, there is a *complexity* or description length $Kw(\vartheta)$ and a *weight* or (semi)probability $w_\vartheta = 2^{-Kw(\vartheta)}$. The complexity will often but need not be a natural number. Typically, one assumes that the weights sum up to at most one, $\sum_{\vartheta\in\Theta} w_\vartheta \leq 1$. Then, by the Kraft inequality, for all $\vartheta \in \Theta$ there exists a prefix-code of length $Kw(\vartheta)$. Because of this correspondence, it is only a matter of

convenience if results are developed in terms of description lengths or probabilities. We will choose the former way. We won't even need the condition $\sum_\vartheta w_\vartheta \leq 1$ for most of the following results. This only means that Kw cannot be interpreted as a prefix code length, but does not cause other problems.

Given a set of distributions $\Theta \subset [0,1]$, complexities $(Kw(\vartheta))_{\vartheta\in\Theta}$, a true distribution $\vartheta_0 \in \Theta$, and some observed string $x \in \mathbb{B}^*$, we define an *MDL estimator*[1]:

$$\vartheta^x = \arg\max_{\vartheta\in\Theta}\{w_\vartheta P(x|\vartheta_0 = \vartheta)\}.$$

Here, $P(x|\vartheta_0 = \vartheta)$ is the probability of observing x if ϑ is the true parameter. Clearly, $P(x|\vartheta_0 = \vartheta) = \vartheta^{\mathbb{1}(x)}(1-\vartheta)^{\ell(x)-\mathbb{1}(x)}$, where $\mathbb{1}(x)$ is the number of ones in x. Hence $P(x|\vartheta_0 = \vartheta)$ depends only on $\ell(x)$ and $\mathbb{1}(x)$. We therefore see

$$\begin{aligned} \vartheta^x \;=\; \vartheta^{(\alpha,n)} \;&=\; \arg\max_{\vartheta\in\Theta}\{w_\vartheta \left(\vartheta^\alpha(1-\vartheta)^{1-\alpha}\right)^n\} \\ &=\; \arg\min_{\vartheta\in\Theta}\{n\cdot D(\alpha\|\vartheta) + Kw(\vartheta)\cdot\ln 2\}, \end{aligned} \tag{1}$$

where $n = \ell(x)$ and $\alpha := \frac{\mathbb{1}(x)}{\ell(x)}$ is the *observed fraction* of ones and

$$D(\alpha\|\vartheta) = \alpha\ln\tfrac{\alpha}{\vartheta} + (1-\alpha)\ln\tfrac{1-\alpha}{1-\vartheta}$$

is the Kullback-Leibler divergence. Let $\vartheta, \tilde\vartheta \in \Theta$ be two parameters, then it follows from (1) that in the process of choosing the MDL estimator, ϑ is being preferred to $\tilde\vartheta$ iff

$$n(D(\alpha\|\tilde\vartheta) - D(\alpha\|\vartheta)) \geq \ln 2\cdot(Kw(\vartheta) - Kw(\tilde\vartheta)). \tag{2}$$

In this case, we say that ϑ *beats* $\tilde\vartheta$. It is immediate that for increasing n the influence of the complexities on the selection of the maximizing element decreases. We are now interested in the *total expected square prediction error* (or cumulative square loss) of the MDL estimator $\sum_{n=1}^\infty \mathbf{E}(\vartheta^{x_{1:n}} - \vartheta_0)^2$. In terms of [PH04], this is the *static MDL prediction* loss, which means that a predictor/estimator ϑ^x is chosen according to the current observation x. The *dynamic* method on the other hand would consider both possible continuations $x0$ and $x1$ and predict according to ϑ^{x0} *and* ϑ^{x1}. In the following, we concentrate on static predictions. They are also preferred in practice, since computing only one model is more efficient.

Let $A_n = \{\frac{k}{n} : 0 \leq k \leq n\}$. Given the true parameter ϑ_0 and some $n \in \mathbb{N}$, the *expectation* of a function $f^{(n)} : \{0,\ldots,n\} \to \mathbb{R}$ is given by

$$\mathbf{E}f^{(n)} = \sum_{\alpha\in A_n} p(\alpha|n)f(\alpha n), \text{ where } p(\alpha|n) = \binom{n}{k}\left(\vartheta_0^\alpha(1-\vartheta_0)^{1-\alpha}\right)^n. \tag{3}$$

[1]Precisely, we define a MAP (maximum a posteriori) estimator. For two reasons, our definition might not be considered as MDL in the strict sense. First, MDL is often associated with a specific prior, while we admit arbitrary priors. Second and more importantly, when coding some data x, one can exploit the fact that once the parameter ϑ^x is specified, only data which leads to this ϑ^x needs to be considered. This allows for a description shorter than $Kw(\vartheta^x)$. Nevertheless, the *construction principle* is commonly termed MDL, compare e.g. the "ideal MDL" in [VL00].

(Note that the probability $p(\alpha|n)$ depends on ϑ_0, which we do not make explicit in our notation.) Therefore,

$$\sum_{n=1}^{\infty}\mathbf{E}(\vartheta^{x_{1:n}}-\vartheta_0)^2=\sum_{n=1}^{\infty}\sum_{\alpha\in A_n}p(\alpha|n)(\vartheta^{(\alpha,n)}-\vartheta_0)^2. \tag{4}$$

Denote the relation $f=O(g)$ by $f\stackrel{\times}{\leq}g$. Analogously define "$\stackrel{\times}{\geq}$" and "$\stackrel{\times}{=}$". From [PH04, Corollary 12], we immediately obtain the following result.

Theorem 1 *The cumulative loss bound* $\sum_n\mathbf{E}(\vartheta^{x_{1:n}}-\vartheta_0)^2\stackrel{\times}{\leq}2^{Kw(\vartheta_0)}$ *holds.*

This is the "slow" convergence result mentioned in the introduction. In contrast, for a Bayes mixture, the total expected error is bounded by $Kw(\vartheta_0)$ rather than $2^{Kw(\vartheta_0)}$ (see [Sol78] or [Hut01, Th.1]). An upper bound on $\sum_n\mathbf{E}(\vartheta^{x_{1:n}}-\vartheta_0)^2$ is termed as *convergence in mean sum* and implies convergence $\vartheta^{x_{1:n}}\to\vartheta_0$ with probability 1 (since otherwise the sum would be infinite).

We now establish relations between the Kullback-Leibler divergence and the quadratic distance. We call bounds of this type *entropy inequalities*.

Lemma 2 *Let* $\vartheta,\tilde\vartheta\in(0,1)$ *and* $\vartheta^*=\arg\min\{|\vartheta-\frac12|,|\tilde\vartheta-\frac12|\}$, *i.e.* ϑ^* *is the element from* $\{\vartheta,\tilde\vartheta\}$ *which is closer to* $\frac12$. *Then*

$$\begin{aligned}2\cdot(\vartheta-\tilde\vartheta)^2&\stackrel{(i)}{\leq}D(\vartheta\|\tilde\vartheta)\stackrel{(ii)}{\leq}\tfrac83(\vartheta-\tilde\vartheta)^2\ \textit{and}\\ \frac{(\vartheta-\tilde\vartheta)^2}{2\vartheta^*(1-\vartheta^*)}&\stackrel{(iii)}{\leq}D(\vartheta\|\tilde\vartheta)\stackrel{(iv)}{\leq}\frac{3(\vartheta-\tilde\vartheta)^2}{2\vartheta^*(1-\vartheta^*)}.\end{aligned}$$

Thereby, (ii) *requires* $\vartheta,\tilde\vartheta\in[\frac14,\frac34]$, (iii) *requires* $\vartheta,\tilde\vartheta\leq\frac12$, *and* (iv) *requires* $\vartheta\leq\frac14$ *and* $\tilde\vartheta\in[\frac{\vartheta}{3},3\vartheta]$. *Statements* (iii) *and* (iv) *have symmetric counterparts for* $\vartheta\geq\frac12$.

Proof. The lower bound (i), is standard, see e.g. [LV97, p. 329]. In order to verify the upper bound (ii), let $f(\eta)=D(\vartheta\|\eta)-\frac83(\eta-\vartheta)^2$. Then (ii) follows from $f(\eta)\leq0$ for $\eta\in[\frac14,\frac34]$. We have that $f(\vartheta)=0$ and $f'(\eta)=\frac{\eta-\vartheta}{\eta(1-\eta)}-\frac{16}{3}(\eta-\vartheta)$. This difference is nonnegative if and only $\eta-\vartheta\leq0$ since $\eta(1-\eta)\geq\frac{3}{16}$. This implies $f(\eta)\leq0$. Statements (iii) and (iv) giving bounds if ϑ is close to the boundary are proven similarly. □

Lemma 2 (ii) is sufficient to prove the lower bound on the error in Proposition 5. The bounds (iii) and (iv) are only needed in the technical proof of the upper bound in Theorem 8, which will be omitted. It requires also similar upper and lower bounds for the absolute distance, and if the second argument of $D(\cdot\|\cdot)$ tends to the boundary. The lemma remains valid for the extreme cases $\vartheta,\tilde\vartheta\in\{0,1\}$ if the

fraction $\frac{0}{0}$ is properly defined. It is likely to generalize to arbitrary alphabet, for (i) this is shown in [Hut01].

It is a well-known fact that the binomial distribution may be approximated by a Gaussian. Our next goal is to establish upper and lower bounds for the binomial distribution. Again we leave out the extreme cases.

Lemma 3 *Let $\vartheta_0 \in (0,1)$ be the true parameter, $n \geq 2$ and $1 \leq k \leq n-1$, and $\alpha = \frac{k}{n}$. Then the following assertions hold.*

$$(i) \qquad p(\alpha|n) \leq \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} \exp\big(-nD(\alpha\|\vartheta_0)\big),$$

$$(ii) \qquad p(\alpha|n) \geq \frac{1}{\sqrt{8\alpha(1-\alpha)n}} \exp\big(-nD(\alpha\|\vartheta_0)\big).$$

The lemma is verified using Stirling's formula. The upper bound is sharp for $n \to \infty$ and fixed α. Lemma 3 can be easily combined with Lemma 2, yielding Gaussian estimates for the Binomial distribution. The following lemma is proved by simply estimating the sums by appropriate integrals.

Lemma 4 *Let $z \in \mathbb{R}^+$, then*

$$(i) \qquad \frac{\sqrt{\pi}}{2z^3} - \frac{1}{z\sqrt{2e}} \leq \sum_{n=1}^{\infty} \sqrt{n} \cdot \exp(-z^2 n) \leq \frac{\sqrt{\pi}}{2z^3} + \frac{1}{z\sqrt{2e}} \text{ and}$$

$$(ii) \qquad \sum_{n=1}^{\infty} n^{-\frac{1}{2}} \exp(-z^2 n) \leq \sqrt{\pi}/z.$$

3 Lower Bound

We are now in the position to prove that even for Bernoulli classes the upper bound from Theorem 1 is sharp in general.

Proposition 5 *Let $\vartheta_0 = \frac{1}{2}$ be the true parameter generating sequences of fair coin flips. Assume there are $2^N - 1$ other parameters $\vartheta_1, \ldots, \vartheta_{2^N-1}$ with $\vartheta_k = \frac{1}{2} + 2^{-k-1}$. Let all complexities be equal, i.e. $Kw(\vartheta_0) = Kw(\vartheta_1) = \ldots = Kw(\vartheta_{2^N-1}) = N$. Then*

$$\sum_{n=1}^{\infty} \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \geq \tfrac{1}{84}(2^N - 5) \stackrel{\times}{=} 2^{Kw(\vartheta_0)}.$$

Proof. Recall that $\vartheta^x = \vartheta^{(\alpha,n)}$ the maximizing element for some observed sequence x only depends on the length n and the observed fraction of ones α. In order to obtain an estimate for the total prediction error $\sum_n \mathbf{E}(\vartheta_0 - \vartheta^x)^2$, partition the

interval $[0,1]$ into 2^N disjoint intervals I_k, such that $\bigcup_{k=0}^{2^N-1} I_k = [0,1]$. Then consider the contributions for the observed fraction α falling in I_k *separately*:

$$C(k) = \sum_{n=1}^{\infty} \sum_{\alpha \in A_n \cap I_k} p(\alpha|n)(\vartheta^{(\alpha,n)} - \vartheta_0)^2 \tag{5}$$

(compare (3)). Clearly, $\sum_n \mathbf{E}(\vartheta_0 - \vartheta^x)^2 = \sum_k C(k)$ holds. We define the partitioning (I_k) as $I_0 = [0, \frac{1}{2} + 2^{-2^N}) = [0, \vartheta_{2^N-1})$, $I_1 = [\frac{3}{4}, 1] = [\vartheta_1, 1]$, and

$$I_k = [\vartheta_k, \vartheta_{k-1}) \text{ for all } 2 \le k \le 2^N - 1.$$

Fix $k \in \{2, \ldots, 2^N - 1\}$ and assume $\alpha \in I_k$. Then

$$\vartheta^{(\alpha,n)} = \arg\min_{\vartheta} \{nD(\alpha\|\vartheta) + Kw(\vartheta)\ln 2\} = \arg\min_{\vartheta} \{nD(\alpha\|\vartheta)\} \in \{\vartheta_k, \vartheta_{k-1}\}$$

according to (1). So clearly $(\vartheta^{(\alpha,n)} - \vartheta_0)^2 \ge (\vartheta_k - \vartheta_0)^2 = 2^{-2k-2}$ holds. Since $p(\alpha|n)$ decreases for increasing $|\alpha - \vartheta_0|$, we have $p(\alpha|n) \ge p(\vartheta_{k-1}|n)$. The interval I_k has length 2^{-k-1}, so there are at least $\lfloor n2^{-k-1} \rfloor \ge n2^{-k-1} - 1$ observed fractions α falling in the interval. From (5), the total contribution of $\alpha \in I_k$ can be estimated by

$$C(k) \ge \sum_{n=1}^{\infty} 2^{-2k-2}(n2^{-k-1} - 1)p(\vartheta_{k-1}|n).$$

Note that the terms in the sum even become negative for small n, which does not cause any problems. We proceed with

$$p(\vartheta_{k-1}|n) \ge \frac{1}{\sqrt{8 \cdot 2^{-2}n}} \exp\left[-nD(\tfrac{1}{2} + 2^{-k}\|\tfrac{1}{2})\right] \ge \frac{1}{\sqrt{2n}} \exp\left[-n\tfrac{8}{3}2^{-2k}\right]$$

according to Lemma 3 and Lemma 2 (ii). By Lemma 4 (i) and (ii), we have

$$\begin{aligned} \sum_{n=1}^{\infty} \sqrt{n} \exp\left[-n\tfrac{8}{3}2^{-2k}\right] &\ge \frac{\sqrt{\pi}}{2}\left(\frac{3}{8}\right)^{\frac{3}{2}} 2^{3k} - \frac{1}{\sqrt{2e}}\sqrt{\frac{3}{8}}2^k \text{ and} \\ -\sum_{n=1}^{\infty} n^{-\frac{1}{2}} \exp\left[-n\tfrac{8}{3}2^{-2k}\right] &\ge -\sqrt{\pi}\sqrt{\frac{3}{8}}2^k. \end{aligned}$$

Considering only $k \ge 5$, we thus obtain

$$\begin{aligned} C(k) &\ge \frac{1}{\sqrt{2}}\sqrt{\frac{3}{8}}2^{-2k-2}\left[\frac{3\sqrt{\pi}}{16}2^{2k-1} - \frac{1}{\sqrt{2e}}2^{-1} - \sqrt{\pi}2^k\right] \\ &\ge \frac{\sqrt{3}}{16}\left[3\sqrt{\pi}2^{-5} - \frac{1}{\sqrt{2e}}2^{-2k-1} - \sqrt{\pi}2^{-k}\right] \ge \frac{\sqrt{3\pi}}{8}2^{-5} - \frac{\sqrt{3}}{16\sqrt{2e}}2^{-11} > \frac{1}{84}. \end{aligned}$$

Ignoring the contributions for $k \le 4$, this implies the assertion. □

This result shows that if the parameters and their weights are chosen in an appropriate way, then the total expected error is of order w_0^{-1} instead of $\ln w_0^{-1}$. Interestingly, this outcome seems to depend on the arrangement and the weights of the *false* parameters rather than on the weight of the *true* one. One can check with moderate effort that the proposition still remains valid if e.g. w_0 is twice as large as the other weights. Actually, the proof of Proposition 5 shows even a slightly more general result, namely the same bound holds when there are additional arbitrary parameters with larger complexities. This will be used for Example 14. Other and more general assertions can be proven similarly.

4 Upper Bounds

Although the cumulative error may be large, as seen in the previous section, the instantaneous error is always small.

Proposition 6 *For $n \geq 3$, the expected instantaneous square loss is bounded:*

$$\mathbf{E}(\vartheta_0 - \hat{\vartheta}^{x_{1:n}})^2 \leq \frac{(\ln 2)Kw(\vartheta_0)}{2n} + \frac{\sqrt{2(\ln 2)Kw(\vartheta_0)\ln n}}{n} + \frac{6\ln n}{n}.$$

Proof. We give an elementary proof for the case $\vartheta_0 \in (\frac{1}{4}, \frac{3}{4})$ only. Like in the proof of Proposition 5, we consider the contributions of different α separately. By Hoeffding's inequality, $\mathbf{P}(|\alpha - \vartheta_0| \geq \frac{c}{\sqrt{n}}) \leq 2e^{-2c^2}$ for any $c > 0$. Letting $c = \sqrt{\ln n}$, the contributions by these α are thus bounded by $\frac{2}{n^2} \leq \frac{\ln n}{n}$.

On the other hand, for $|\alpha - \vartheta_0| \leq \frac{c}{\sqrt{n}}$, recall that ϑ_0 *beats* any ϑ iff (2) holds. According to $Kw(\vartheta) \leq 1$, $|\alpha - \vartheta_0| \leq \frac{c}{\sqrt{n}}$, and Lemma 2 (i) and (ii), (2) is already implied by $|\alpha - \vartheta| \geq \sqrt{\frac{\frac{1}{2}(\ln 2)Kw(\vartheta_0) + \frac{4}{3}c^2}{n}}$. Clearly, a contribution only occurs if ϑ beats ϑ_0, therefore if the opposite inequality holds. Using $|\alpha - \vartheta_0| \leq \frac{c}{\sqrt{n}}$ again and the triangle inequality, we obtain that

$$(\vartheta - \vartheta_0)^2 \leq \frac{5c^2 + \frac{1}{2}(\ln 2)Kw(\vartheta_0) + \sqrt{2(\ln 2)Kw(\vartheta_0)c^2}}{n}$$

in this case. Since we have chosen $c = \sqrt{\ln n}$, this implies the assertion. □

One can improve the bound in Proposition 6 to $\mathbf{E}(\vartheta_0 - \hat{\vartheta}^{x_{1:n}})^2 \stackrel{\times}{\leq} \frac{Kw(\vartheta_0)}{n}$ by a refined argument, compare [BC91]. But the high-level assertion is the same: Even if the cumulative upper bound may tend to infinity, the instantaneous error converges rapidly to 0. Moreover, the convergence speed depends on $Kw(\vartheta_0)$ as opposed to $2^{Kw(\vartheta_0)}$. Thus $\hat{\vartheta}$ tends to ϑ_0 rapidly in probability (recall that the assertion is not strong enough to conclude almost sure convergence). The proof does not exploit $\sum w_\vartheta \leq 1$, but only $w_\vartheta \leq 1$, hence the assertion even holds for a maximum likelihood

estimator (i.e. $w_\vartheta = 1$ for all $\vartheta \in \Theta$). The theorem generalizes to i.i.d. classes. For the example in Proposition 5, the instantaneous bound implies that the bulk of losses occurs very late. This does *not* hold for general (non-i.i.d.) model classes: The losses in [PH04, Example 9] grow linearly in the first n steps.

We will now state our main positive result that upper bounds the cumulative loss in terms of the negative logarithm of the true weight and the *arrangement* of the false parameters. We will only give the proof idea – which is similar to that of Proposition 5 – and omit the lengthy and tedious technical details.

Consider the cumulated sum square error $\sum_n \mathbf{E}(\vartheta^{(\alpha,n)} - \vartheta_0)^2$. In order to upper bound this quantity, we will partition the open unit interval $(0,1)$ into a sequence of intervals $(I_k)_{k=1}^\infty$, each of measure 2^{-k}. (More precisely: Each I_k is either an interval or a union of two intervals.) Then we will estimate the contribution of each interval to the cumulated square error,

$$C(k) = \sum_{n=1}^{\infty} \sum_{\alpha \in A_n, \vartheta^{(\alpha,n)} \in I_k} p(\alpha|n)(\vartheta^{(\alpha,n)} - \vartheta_0)^2$$

(compare (3) and (5)). Note that $\vartheta^{(\alpha,n)} \in I_k$ precisely reads $\vartheta^{(\alpha,n)} \in I_k \cap \Theta$, but for convenience we generally assume $\vartheta \in \Theta$ for all ϑ being considered. This partitioning is also used for α, i.e. define the contribution $C(k,j)$ of $\vartheta \in I_k$ where $\alpha \in I_j$ as

$$C(k,j) = \sum_{n=1}^{\infty} \sum_{\alpha \in A_n \cap I_j, \vartheta^{(\alpha,n)} \in I_k} p(\alpha|n)(\vartheta^{(\alpha,n)} - \vartheta_0)^2.$$

We need to distinguish between α that are located close to ϑ_0 and α that are located far from ϑ_0. "Close" will be roughly equivalent to $j > k$, "far" will be approximately $j \leq k$. So we get $\sum_n \mathbf{E}(\vartheta^{(\alpha,n)} - \vartheta_0)^2 = \sum_k^\infty C(k) = \sum_k \sum_j C(k,j)$. In the proof,

$$p(\alpha|n) \stackrel{\times}{\leq} [n\alpha(1-\alpha)]^{-\frac{1}{2}} \exp[-nD(\alpha\|\vartheta_0)]$$

is often applied, which holds by Lemma 3 (recall that $f \stackrel{\times}{\leq} g$ stands for $f = O(g)$). Terms like $D(\alpha\|\vartheta_0)$, arising in this context and others, can be further estimated using Lemma 2. We now give the constructions of intervals I_k and complementary intervals J_k.

Definition 7 Let $\vartheta_0 \in \Theta$ be given. Start with $J_0 = [0,1)$. Let $J_{k-1} = [\vartheta_k^l, \vartheta_k^r)$ and define $d_k = \vartheta_k^r - \vartheta_k^l = 2^{-k+1}$. Then $I_k, J_k \subset J_{k-1}$ are constructed from J_{k-1} according to the following rules.

$$\vartheta_0 \in [\vartheta_k^l, \vartheta_k^l + \tfrac{3}{8}d_k) \;\Rightarrow\; J_k = [\vartheta_k^l, \vartheta_k^l + \tfrac{1}{2}d_k),\ I_k = [\vartheta_k^l + \tfrac{1}{2}d_k, \vartheta_k^r), \quad (6)$$

$$\vartheta_0 \in [\vartheta_k^l + \tfrac{3}{8}d_k, \vartheta_k^l + \tfrac{5}{8}d_k) \;\Rightarrow\; J_k = [\vartheta_k^l + \tfrac{1}{4}d_k, \vartheta_k^l + \tfrac{3}{4}d_k), \quad (7)$$

$$I_k = [\vartheta_k^l, \vartheta_k^l + \tfrac{1}{4}d_k) \cup [\vartheta_k^l + \tfrac{3}{4}d_k, \vartheta_k^r),$$

$$\vartheta_0 \in [\vartheta_k^l + \tfrac{5}{8}d_k, \vartheta_k^r) \;\Rightarrow\; J_k = [\vartheta_k^l + \tfrac{1}{2}d_k, \vartheta_k^r),\ I_k = [\vartheta_k^l, \vartheta_k^l + \tfrac{1}{2}d_k). \quad (8)$$

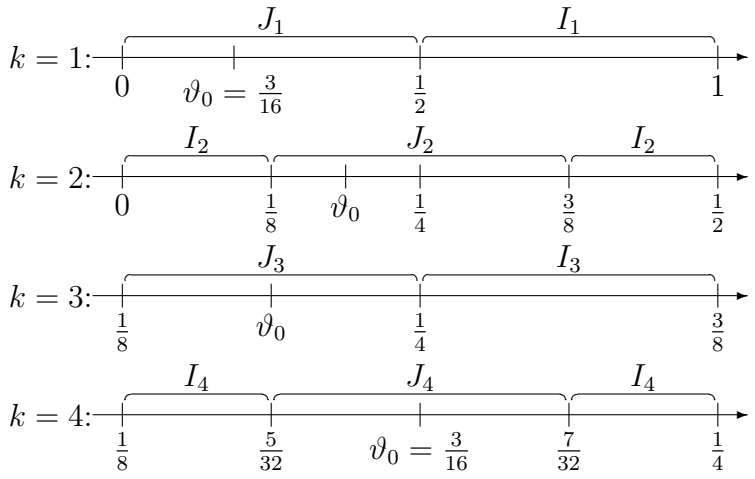

Figure 1: Example of the first four intervals for $\vartheta_0 = \frac{3}{16}$. We have an l-step, a c-step, an l-step and another c-step. All following steps will be also c-steps.

We call the kth step of the interval construction an *l-step* if (6) applies, a *c-step* if (7) applies, and an *r-step* if (8) applies, respectively. Fig. 1 shows an example for the interval construction.

Clearly, this is not the only possible way to define an interval construction. Maybe the reader wonders why we did not center the intervals around ϑ_0. In fact, this construction would equally work for the proof. However, its definition would not be easier, since one still has to treat the case where ϑ_0 is located close to the boundary. Moreover, our construction has the nice property that the interval bounds are finite binary fractions. Given the interval construction, we can identify the $\vartheta \in I_k$ with lowest complexity:

$$\begin{aligned} \vartheta_k^I &= \arg\min\{Kw(\vartheta) : \vartheta \in I_k \cap \Theta\}, \\ \vartheta_k^J &= \arg\min\{Kw(\vartheta) : \vartheta \in J_k \cap \Theta\}, \text{ and} \\ \Delta(k) &= \max\{Kw(\vartheta_k^I) - Kw(\vartheta_k^J), 0\}. \end{aligned}$$

If there is no $\vartheta \in I_k \cap \Theta$, we set $\Delta(k) = Kw(\vartheta_k^I) = \infty$.

Theorem 8 *Let $\Theta \subset [0,1]$ be countable, $\vartheta_0 \in \Theta$, and $w_\vartheta = 2^{-Kw(\vartheta)}$, where $Kw(\vartheta)$ is some complexity measure on Θ. Let $\Delta(k)$ be as introduced in the last paragraph, then*

$$\sum_{n=0}^{\infty} \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} Kw(\vartheta_0) + \sum_{k=1}^{\infty} 2^{-\Delta(k)} \sqrt{\Delta(k)}.$$

The proof is omitted. But we briefly discuss the assertion of this theorem. It states an error bound in terms of the arrangement of the false parameters which directly depends on the interval construction. As already indicated, a different interval construction would do as well, provided that it exponentially contracts to the true parameter. For a reasonable distribution of parameters, we might expect that $\Delta(k)$ increases linearly for k large enough, and thus $\sum 2^{-\Delta(k)}\sqrt{\Delta(k)}$ remains bounded. In the next section, we identify cases where this holds.

5 Uniformly Distributed Weights

We are now able to state some positive results following from Theorem 8.

Theorem 9 *Let $\Theta \subset [0,1]$ be a countable class of parameters and $\vartheta_0 \in \Theta$ the true parameter. Assume that there are constants $a \geq 1$ and $b \geq 0$ such that*

$$\min\{Kw(\vartheta) : \vartheta \in [\vartheta_0 - 2^{-k}, \vartheta_0 + 2^{-k}] \cap \Theta, \vartheta \neq \vartheta_0\} \geq \frac{k-b}{a} \tag{9}$$

holds for all $k > aKw(\vartheta_0) + b$. Then we have

$$\sum_{n=0}^{\infty} \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} aKw(\vartheta_0) + b \stackrel{\times}{\leq} Kw(\vartheta_0).$$

Proof. We have to show that

$$\sum_{k=1}^{\infty} 2^{-\Delta(k)} \sqrt{\Delta(k)} \stackrel{\times}{\leq} aKw(\vartheta_0) + b,$$

then the assertion follows from Theorem 8. Let $k_1 = \lceil aKw(\vartheta_0) + b + 1 \rceil$ and $k' = k - k_1$. It is not hard to see that $\max_{\vartheta \in I_k} |\vartheta - \vartheta_0| \leq 2^{-k+1}$ holds. Together with (9), this implies

$$\begin{aligned}
\sum_{k=1}^{\infty} 2^{-\Delta(k)} \sqrt{\Delta(k)} &\leq \sum_{k=1}^{k_1} 1 + \sum_{k=k_1+1}^{\infty} 2^{-Kw(\vartheta_k^I) + Kw(\vartheta_0)} \sqrt{Kw(\vartheta_k^I) - Kw(\vartheta_0)} \\
&\leq k_1 + 2^{Kw(\vartheta_0)} \sum_{k=k_1+1}^{\infty} 2^{-\frac{k-b}{a}} \sqrt{\frac{k-b}{a}} \\
&\leq k_1 + 2^{Kw(\vartheta_0)} \sum_{k'=1}^{\infty} 2^{-\frac{k'+k_1-b}{a}} \sqrt{\frac{k'+k_1-b}{a}} \\
&\leq aKw(\vartheta_0) + b + 2 + \sum_{k'=1}^{\infty} 2^{-\frac{k'}{a}} \sqrt{\frac{k'}{a} + Kw(\vartheta_0)}.
\end{aligned}$$

Observe $\sqrt{\frac{k'}{a} + Kw(\vartheta_0)} \leq \sqrt{\frac{k'}{a}} + \sqrt{Kw(\vartheta_0)}$, $\sum_{k'} 2^{-\frac{k'}{a}} \stackrel{\times}{\leq} a$, and by Lemma 4 (i), $\sum_{k'} 2^{-\frac{k'}{a}} \sqrt{\frac{k'}{a}} \stackrel{\times}{\leq} a$. Then the assertion follows. □

Letting $j = \frac{k-b}{a}$, (9) asserts that parameters ϑ with complexity $Kw(\vartheta) = j$ must have a minimum distance of 2^{-ja-b} from ϑ_0. That is, if parameters with equal weights are (approximately) uniformly distributed in the neighborhood of ϑ_0, in the sense that they are not too close to each other, then fast convergence holds. The next two results are special cases based on the set of all finite binary fractions,

$$\mathbb{Q}_{\mathbb{B}^*} = \{\vartheta = 0.\beta_1\beta_2\ldots\beta_{n-1}1 : n \in \mathbb{N}, \beta_i \in \mathbb{B}\} \cup \{0,1\}.$$

If $\vartheta = 0.\beta_1\beta_2\ldots\beta_{n-1}1 \in \mathbb{Q}_{\mathbb{B}^*}$, its length is $\ell(\vartheta) = n$. Moreover, there is a binary code $\beta'_1\ldots\beta'_{n'}$ for n, having at most $n' \leq \lfloor \log_2(n+1) \rfloor$ bits. Then $0\beta'_1 0\beta'_2\ldots 0\beta'_{n'}1\beta_1\ldots\beta_{n-1}$ is a prefix-code for ϑ. For completeness, we can define the codes for $\vartheta = 0, 1$ to be 10 and 11, respectively. So we may define a complexity measure on $\mathbb{Q}_{\mathbb{B}^*}$ by

$$Kw(0) = 2,\ Kw(1) = 2,\ \text{and}\ Kw(\vartheta) = \ell(\vartheta) + 2\lfloor \log_2(\ell(\vartheta)+1) \rfloor \text{ for } \vartheta \neq 0, 1. \quad (10)$$

There are other similar simple prefix codes on $\mathbb{Q}_{\mathbb{B}^*}$ such that $Kw(\vartheta) \geq \ell(\vartheta)$.

Corollary 10 *Let $\Theta = \mathbb{Q}_{\mathbb{B}^*}$, $\vartheta_0 \in \Theta$ and $Kw(\vartheta) \geq \ell(\vartheta)$, then $\sum_n \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} Kw(\vartheta_0)$ holds.*

The proof is trivial, since Condition (9) holds with $a = 1$ and $b = 0$. This is a special case of a uniform distribution of parameters with equal complexities. The next corollary is more general, it proves fast convergence if the uniform distribution is distorted by some function φ.

Corollary 11 *Let $\varphi : [0,1] \to [0,1]$ be an injective, N times continuously differentiable function. Let $\Theta = \varphi(\mathbb{Q}_{\mathbb{B}^*})$, $Kw(\varphi(t)) \geq \ell(t)$ for all $t \in \mathbb{Q}_{\mathbb{B}^*}$, and $\vartheta_0 = \varphi(t_0)$ for a $t_0 \in \mathbb{Q}_{\mathbb{B}^*}$. Assume that there is $n \leq N$ and $\varepsilon > 0$ such that*

$$\left|\frac{d^n\varphi}{dt^n}(t)\right| \geq c > 0 \quad \text{for all } t \in [t_0 - \varepsilon, t_0 + \varepsilon] \text{ and}$$
$$\frac{d^m\varphi}{dt^m}(t_0) = 0 \quad \text{for all } 1 \leq m < n.$$

Then we have

$$\sum \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} nKw(\vartheta_0) + 2\log_2(n!) - 2\log_2 c + n\log_2\varepsilon \stackrel{\times}{\leq} nKw(\vartheta_0).$$

Proof. Fix $j > Kw(\vartheta_0)$, then

$$Kw(\varphi(t)) \geq j \text{ for all } t \in [t_0 - 2^{-j}, t_0 + 2^{-j}] \cap \mathbb{Q}_{\mathbb{B}^*}. \quad (11)$$

Moreover, for all $t \in [t_0 - 2^{-j}, t_0 + 2^{-j}]$, Taylor's theorem asserts that

$$\varphi(t) = \varphi(t_0) + \frac{\frac{d^n\varphi}{dt^n}(\tilde{t})}{n!}(t - t_0)^n \quad (12)$$

for some $\tilde{t}$ in (t_0, t) (or (t, t_0) if $t < t_0$). We request in addition $2^{-j} \leq \varepsilon$, then $|\frac{d^n\varphi}{dt^n}| \geq c$ by assumption. Apply (12) to $t = t_0 + 2^{-j}$ and $t = t_0 - 2^{-j}$ and define $k = \lceil jn + \log_2(n!) - \log_2 c \rceil$ in order to obtain $|\varphi(t_0 + 2^{-j}) - \vartheta_0| \geq 2^{-k}$ and $|\varphi(t_0 -$

$2^{-j}) - \vartheta_0| \geq 2^{-k}$. By injectivity of φ, we see that $\varphi(t) \notin [\vartheta_0 - 2^{-k}, \vartheta_0 + 2^{-k}]$ if $t \notin [t_0 - 2^{-j}, t_0 + 2^{-j}]$. Together with (11), this implies

$$Kw(\vartheta) \geq j \geq \frac{k - \log_2(n!) + \log_2 c - 1}{n} \text{ for all } \vartheta \in [\vartheta_0 - 2^{-k}, \vartheta_0 + 2^{-k}] \cap \Theta.$$

This is condition (9) with $a = n$ and $b = \log_2(n!) - \log_2 c + 1$. Finally, the assumption $2^{-j} \leq \varepsilon$ holds if $k \geq k_1 = n\log_2\varepsilon + \log_2(n!) - \log_2 c + 1$. This gives an additional contribution to the error of at most k_1. □

Corollary 11 shows an implication of Theorem 8 for *parameter identification*: A class of models is given by a set of parameters $\mathbb{Q}_{\mathbb{B}^*}$ and a mapping $\varphi : \mathbb{Q}_{\mathbb{B}^*} \to \Theta$. The task is to identify the true parameter t_0 or its image $\vartheta_0 = \varphi(t_0)$. The injectivity of φ is not necessary for fast convergence, but it facilitates the proof. The assumptions of Corollary 11 are satisfied if φ is for example a polynomial. In fact, it should be possible to prove fast convergence of MDL for many common parameter identification problems. For sets of parameters other than $\mathbb{Q}_{\mathbb{B}^*}$, e.g. the set of all rational numbers $\mathbb{Q}$, similar corollaries can easily be proven.

How large is the constant hidden in "$\stackrel{\times}{\leq}$"? When examining carefully the proof of Theorem 8, the resulting constant is quite large. This is mainly due to the frequent "wasting" of small constants. Supposably a smaller bound holds as well, perhaps 16. On the other hand, for the actual *true* expectation (as opposed to its upper bound) and complexities as in (10), numerical simulations indicate that $\sum_n \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \leq \frac{1}{2} Kw(\vartheta_0)$.

Finally, we state an implication which almost trivially follows from Theorem 8, since there $\sum_k 2^{-\Delta(k)}\sqrt{\Delta(k)} \leq N$ is obvious. However, it may be very useful for practical purposes, e.g. for hypothesis testing.

Corollary 12 *Let Θ contain N elements, $Kw(\cdot)$ be any complexity function on Θ, and $\vartheta_0 \in \Theta$. Then we have*

$$\sum_{n=1}^{\infty} \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} N + Kw(\vartheta_0).$$

6 The Universal Case

We briefly discuss the important universal setup, where $Kw(\cdot)$ is (up to an additive constant) equal to the prefix Kolmogorov complexity K (that is the length of the shortest self-delimiting program printing ϑ on some universal Turing machine). Since $\sum_k 2^{-K(k)}\sqrt{K(k)} = \infty$ no matter how late the sum starts (otherwise there would be a shorter code for large k), we cannot apply Theorem 8. This means in particular that we do not even obtain our previous result, Theorem 1. But probably the following strengthening of the theorem holds under the same conditions, which then easily implies Theorem 1 up to a constant.

Conjecture 13 $\sum_n \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} K(\vartheta_0) + \sum_k 2^{-\Delta(k)}$.

Then, take an incompressible finite binary fraction $\vartheta_0 \in \mathbb{Q}_{\mathbb{B}^*}$, i.e. $K(\vartheta_0) \stackrel{+}{=} \ell(\vartheta_0) + K(\ell(\vartheta_0))$. For $k > \ell(\vartheta_0)$, we can reconstruct ϑ_0 and k from ϑ_k^I and $\ell(\vartheta_0)$ by just truncating ϑ_k^I after $\ell(\vartheta_0)$ bits. Thus $K(\vartheta_k^I)+K(\ell(\vartheta_0)) \stackrel{\times}{\geq} K(\vartheta_0)+K(k|\vartheta_0, K(\vartheta_0))$ holds. Using Conjecture 13, we obtain

$$\sum_n \mathbf{E}(\vartheta_0 - \vartheta^x)^2 \stackrel{\times}{\leq} K(\vartheta_0) + 2^{K(\ell(\vartheta_0))} \stackrel{\times}{\leq} \ell(\vartheta_0)(\log_2 \ell(\vartheta_0))^2, \tag{13}$$

where the last inequality follows from the example coding given in (10). So, under Conjecture 13, we obtain a bound which slightly exceeds the complexity $K(\vartheta_0)$ if ϑ_0 has a certain structure. It is not obvious if the same holds for all computable ϑ_0. In order to answer this question positive, one could try to use something like [Gác83, Eq.(2.1)]. This statement implies that as soon as $K(k) \geq K_1$ for all $k \geq k_1$, we have $\sum_{k \geq k_1} 2^{-K(k)} \stackrel{\times}{\leq} 2^{-K_1} K_1 (\log_2 K_1)^2$. It is possible to prove an analogous result for ϑ_k^I instead of k, however we have not found an appropriate coding that does without knowing ϑ_0. Since the resulting bound is exponential in the code length, we therefore have not gained anything.

Another problem concerns the size of the multiplicative constant that is hidden in the upper bound. Unlike in the case of uniformly distributed weights, it is now of exponential size, i.e. $2^{O(1)}$. This is no artifact of the proof, as the following example shows.

Example 14 Let U be some universal Turing machine. We construct a second universal Turing machine U' from U as follows: Let $N \geq 1$. If the input of U' is $1^N p$, where 1^N is the string consisting of N ones and p is some program, then U will be executed on p. If the input of U' is 0^N, then U' outputs $\frac{1}{2}$. Otherwise, if the input of U' is x with $x \in \mathbb{B}^N \setminus \{0^N, 1^N\}$, then U' outputs $\frac{1}{2} + 2^{-x-1}$. For $\vartheta_0 = \frac{1}{2}$, the conditions of a slight generalization of Proposition 5 are satisfied (where the complexity is relative to U'), thus $\sum_n \mathbf{E}(\vartheta^x - \vartheta_0)^2 \stackrel{\times}{\geq} 2^N$.

Can this also happen if the underlying universal Turing machine is not "strange" in some sense, like U', but "natural"? Again this is not obvious. One would have to define first a "natural" universal Turing machine which rules out cases like U'. If N is not too large, then one can even argue that U' *is* natural in the sense that its compiler constant relative to U is small.

There is a relation to the class of all *deterministic* (generally non-i.i.d.) measures. For this setup, MDL predicts the next symbol just according to the *monotone complexity* Km, see [Hut03b]. According to [Hut03b, Theorem 5], 2^{-Km} is very close to the universal semimeasure M (this is due to [ZL70]). Then the total prediction error (which is defined slightly differently in this case) can be shown to be bounded by $2^{O(1)} Km(x_{<\infty})^3$ [Hut04]. The similarity to the (unproven) bound (13) "huge constant $\times$ polynomial" for the universal Bernoulli case is evident.

7 Discussion and Conclusions

We have discovered the fact that the instantaneous and the cumulative loss bounds can be *incompatible*. On the one hand, the cumulative loss for MDL predictions may be exponential, i.e. $2^{Kw(\vartheta_0)}$. Thus it implies almost sure convergence at a slow rate, even for arbitrary discrete model classes [PH04]. On the other hand, the instantaneous loss is always of order $\frac{1}{n}Kw(\vartheta_0)$, implying fast convergence in probability and a cumulative loss bound of $Kw(\vartheta_0)\ln n$. Similar logarithmic loss bounds can be found in the literature for continuous model classes [Ris96].

A different approach to assess convergence speed is presented in [BC91]. There in index of resolvability is introduced, which can be interpreted as the difference of the expected MDL code length and the expected code length under the true model. For discrete model classes, they show that the index of resolvability converges to zero as $\frac{1}{n}Kw(\vartheta_0)$ [BC91, Equation (6.2)]. Moreover, they give a convergence of the predictive distributions in terms of the Hellinger distance [BC91, Theorem 4]. This implies a cumulative (Hellinger) loss bound of $Kw(\vartheta_0)\ln n$ and therefore fast convergence in probability.

If the prior weights are arranged nicely, we have proven a small finite loss bound $Kw(\vartheta_0)$ for MDL (Theorem 8). If parameters of equal complexity are uniformly distributed or not too strongly distorted (Theorem 9 and Corollaries), then the error is within a small multiplicative constant of the complexity $Kw(\vartheta_0)$. This may be applied e.g. for the case of parameter identification (Corollary 11). A similar result holds if Θ is finite and contains only few parameters (Corollary 12), which may be e.g. satisfied for hypothesis testing. In these cases and many others, one can interpret the conditions for fast convergence as the presence of prior knowledge. One can show that if a predictor converges to the correct model, then it performs also well under arbitrarily chosen bounded loss-functions [Hut03a, Theorem 4]. Moreover, we can then conclude good properties for other machine learning tasks such as classification, as discussed in the introduction. From an information theoretic viewpoint one may interpret the conditions for a small bound in Theorem 8 as "good codes".

The main restriction of our positive result is the fact that we have proved it only for the Bernoulli case. We therefore argue that it generalizes to arbitrary i.i.d settings. Let $\vartheta_0 \in [0,1]^N$, $\sum_i \vartheta_0^{(i)} = 1$ be a probability vector that generates sequences of i.i.d. samples in $\{1,\dots,N\}^\infty$. Assume that ϑ_0 stays away from the boundary (the other case is treated similarly). Then we can define a sequence of nested sets in dimension $N-1$ in analogy to the interval construction. The main points of the proof are now the following two: First, for an observed parameter α far from ϑ_0, the probability of α decays exponentially, and second, for α close to ϑ_0, some ϑ far from ϑ_0 can contribute at most for short time. These facts hold in the general i.i.d case like in the Bernoulli case. However, the rigorous proof of it is yet more complicated and technical than for the Bernoulli case. (Compare the proof of the main result in [Ris96].)

We conclude with an open question. In abstract terms, we have proven a con-

vergence result for the Bernoulli (or i.i.d) case by mainly exploiting the *geometry* of the space of distributions. This is in principle very easy, since for Bernoulli this space is just the unit interval, for i.i.d it is the space of probability vectors. It is not obvious how (or if at all) this approach can be transferred to general (computable) measures.

References

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. on Information Theory*, 37(4):1034–1054, 1991.

[BRY98] A. R. Barron, J. J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, 44(6):2743–2760, 1998.

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

[Gác83] P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theoretical Computer Science*, 22:71–93, 1983.

[Hut01] M. Hutter. Convergence and error bounds for universal prediction of nonbinary sequences. *Proc. 12th Eurpean Conference on Machine Learning (ECML-2001)*, pages 239–250, December 2001.

[Hut03a] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Trans. on Information Theory*, 49(8):2061–2067, 2003.

[Hut03b] M. Hutter. Sequence prediction based on monotone complexity. In *Proc. 16th Annual Conference on Learning Theory (COLT-2003)*, Lecture Notes in Artificial Intelligence, pages 506–521, Berlin, 2003. Springer.

[Hut04] M. Hutter. Sequential predictions based on algorithmic complexity. Technical report, 2004. IDSIA-16-04.

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

[PH04] J. Poland and M. Hutter. Convergence of discrete MDL for sequential prediction. In *17th Annual Conference on Learning Theory (COLT)*, pages 300–314, 2004.

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

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

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

[Vov97] V. G. Vovk. Learning about the parameter of the bernoulli model. *Journal of Computer and System Sciences*, 55:96–104, 1997.

[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. *Russian Mathematical Surveys*, 25(6):83–124, 1970.