TECHNICAL REPORT IDSIA-03-04

CONVERGENCE OF DISCRETE MDL FOR SEQUENTIAL PREDICTION

Jan Poland and **Marcus Hutter**

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

{jan,marcus}@idsia.ch, http://www.idsia.ch/~{jan,marcus}

April 28, 2004

Abstract

We study the properties of the Minimum Description Length principle for sequence prediction, considering a two-part MDL estimator which is chosen from a countable class of models. This applies in particular to the important case of *universal sequence prediction*, where the model class corresponds to all algorithms for some fixed universal Turing machine (this correspondence is by enumerable semimeasures, hence the resulting models are stochastic). We prove convergence theorems similar to Solomonoff's theorem of universal induction, which also holds for general Bayes mixtures. The bound characterizing the convergence speed for MDL predictions is exponentially larger as compared to Bayes mixtures. We observe that there are at least *three* different ways of using MDL for prediction. One of these has worse prediction properties, for which predictions only converge if the MDL estimator stabilizes. We establish sufficient conditions for this to occur. Finally, some immediate consequences for complexity relations and randomness criteria are proven.

Keywords

Minimum Description Length, Sequence Prediction, Convergence, Discrete Model Classes, Universal Induction, Stabilization, Algorithmic Information Theory.

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

1 Introduction

The Minimum Description Length (MDL) principle is one of the most important concepts in Machine Learning, and serves as a scientific guide, in general. In particular, the process of building a model for any kind of given data is governed by the MDL principle in the majority of cases. The following illustrating example is probably familiar to many readers: A Bayesian net (or neural network) is constructed from (trained with) some data. We may just determine (train) the net in order to fit the data as closely as possible, then we are describing the data very precisely, but disregard the description of the net itself. The resulting net is a maximum likelihood estimator. Alternatively, we may *simultaneously* minimize the "residual" description length of the data given the net *and* the description length of the net. This corresponds to minimizing a *regularized* error term, and the result is a maximum a posteriori or MDL estimator. The latter way of modelling is not only superior to the former in most applications, it is also conceptually appealing since it implements the simplicity principle, Occam's razor.

The MDL method has been studied on all possible levels from very concrete and highly tuned practical applications up to general theoretical assertions (see e.g. [WB68, Ris78, Grü98]). The aim of this work is to contribute to the theory of MDL. We regard Bayesian or neural nets or other models as just some particular class of models. We identify (probabilistic) models with *(semi)measures*, *data* with the initial part of a *sequence* $x_1, x_2, \ldots, x_{t-1}$, and the task of learning with the problem of *predicting* the next symbol x_t (or more symbols). The sequence $x_1, x_2, \ldots$ itself is generated by some *true* but unknown *distribution* μ.

An two-part MDL estimator for some string $x = x_1, \ldots, x_{t-1}$ is then some short description of the semimeasure, while simultaneously the probability of the data under the related semimeasure is large. Surprisingly little work has been done on this general setting of sequence *prediction* with MDL. In contrast, most work addresses MDL for *coding and modeling*, or others, see e.g. [BRY98, Ris96, BC91, Ris99]. Moreover, there are some results for the prediction of independently identically distributed (i.i.d.) sequences, see e.g. [BC91]. There, discrete model classes are considered, while most of the material available focusses on continuous model classes. In our work we will study countable classes of *arbitrary* semimeasures.

There is a strong motivation for considering both countable classes and semimeasures: In order to derive performance guarantees one has to assume that the model class contains the true model. So the larger we choose this class, the less restrictive is this assumption. From a computational point of view the largest relevant class is the class of *all* lower-semicomputable semimeasures. We call this setup *universal sequence prediction*. This class is at the foundations of and has been intensely studied in Algorithmic Information Theory [ZL70, LV97, Cal02]. Since algorithms do not necessarily halt on each string, one is forced to consider the more general class of semimeasures, rather than measures. Solomonoff [Sol64, Sol78] defined a universal induction system, essentially based on a Bayes mixture over this class (see

[Hut01b, Hut03a] for recent developments). There seems to be no work on MDL for this class, which this paper intends to change. What has been studied intensely in [Hut03b] is the so called one-part MDL over the class of deterministic computable models (see also Section 7).

The paper is structured as follows. Section 2 establishes basic definitions. In Section 3, we introduce the MDL estimator and show how it can be used for sequence prediction in at least three ways. Sections 4 and 5 are devoted to convergence theorems. In Section 6, we study the stabilization properties of the MDL estimator. The setting of universal sequence prediction is treated in Section 7. Finally, Section 8 contains the conclusions.

2 Prerequisites and Notation

We build on the notation of [LV97] and [Hut03b]. Let the alphabet $\mathcal{X}$ be a finite set of symbols. We consider the spaces $\mathcal{X}^*$ and $\mathcal{X}^\infty$ of finite strings and infinite sequences over $\mathcal{X}$. The initial part of a sequence up to a time $t \in \mathbb{N}$ or $t-1 \in \mathbb{N}$ is denoted by $x_{1:t}$ or $x_{<t}$, respectively. The empty string is denoted by ϵ.

A *semimeasure* is a function $\nu : \mathcal{X}^* \to [0,1]$ such that

$$\nu(\epsilon) \leq 1 \text{ and } \nu(x) \geq \sum_{a \in \mathcal{X}} \nu(xa) \text{ for all } x \in \mathcal{X}^* \tag{1}$$

holds. If equality holds in both inequalities of (1), then we have a *measure*. Let $\mathcal{C}$ be a countable class of (semi)measures, i.e. $\mathcal{C} = \{\nu_i : i \in I\}$ with finite or infinite index set $I \subseteq \mathbb{N}$. A (semi)measure $\tilde{\nu}$ *dominates* the class $\mathcal{C}$ iff for all $\nu_i \in \mathcal{C}$ there is a constant $c(\nu_i) > 0$ such that $\nu(x) \geq c(\nu_i) \cdot \nu_i(x)$ holds for all $x \in \mathcal{X}^*$. The dominant semimeasure $\tilde{\nu}$ need not be contained in $\mathcal{C}$, but if it is, we call it a *universal* element of $\mathcal{C}$.

Let $\mathcal{C}$ be a countable class of (semi)measures, where each $\nu \in \mathcal{C}$ is associated with a weight $w_\nu > 0$ and $\sum_\nu w_\nu \leq 1$. We may interpret the weights as a *prior* on $\mathcal{C}$. Then it is obvious that the Bayes mixture

$$\xi(x) = \xi_{[\mathcal{C}]}(x) = \sum_{\nu \in \mathcal{C}} w_\nu \nu(x), \; x \in \mathcal{X}^*, \tag{2}$$

dominates $\mathcal{C}$. Assume that there is some measure $\mu \in \mathcal{C}$, the *true distribution*, generating sequences $x_{<\infty} \in \mathcal{X}^\infty$. Normally μ is unknown. (Note that we require μ to be a measure, while $\mathcal{C}$ may contain also semimeasures in general. This is motivated by the setting of universal sequence prediction as already indicated.) If some initial part $x_{<t}$ of a sequence is given, the probability of observing $x_t \in \mathcal{X}$ as a next symbol is given by

$$\mu(x_t|x_{<t}) = \frac{\mu(x_{<t}x_t)}{\mu(x_{<t})} \text{ if } \mu(x_{<t}) > 0 \;\text{ and }\; \mu(x_t|x_{<t}) = 0 \text{ if } \mu(x_{<t}) = 0. \tag{3}$$

The case $\mu(x_{<t})=0$ is stated only for well-definedness, it has probability zero. Note that $\mu(x_t|x_{<t})$ can depend on $x_{<t}$. We may generally define the quantity (3) for *any* function $\varphi:\mathcal{X}^*\to[0,1]$, we call $\varphi(x_t|x_{<t})=\frac{\varphi(x_{1:t})}{\varphi(x_{<t})}$ the *φ-prediction*. Clearly, this is not necessarily a probability on $\mathcal{X}$ for general φ. For a semimeasure ν in particular, the ν-prediction $\nu(\cdot|x_{<t})$ is a semimeasure on $\mathcal{X}$.

We define the *expectation* with respect to the true probability μ: Let $n\geq 0$ and $f:\mathcal{X}^n\to\mathbb{R}$ be a function, then

$$\mathbf{E}\ f=\mathbf{E}\ f(x_{1:n})=\sum_{x_{1:n}\in\mathcal{X}^n}\mu(x_{1:n})f(x_{1:n}). \tag{4}$$

Generally, we may also define the expectation as an integral over infinite sequences. But since we won't need it, we can keep things simple. We can now state a central result about prediction with Bayes mixtures in a form independent of Algorithmic Information Theory.

2.1 Theorem. *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ and any $n\geq 1$, we have*

$$\sum_{t=1}^{n}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\mu(a|x_{<t})-\xi(a|x_{<t})\Big)^2\ \leq\ \ln w_\mu^{-1}. \tag{5}$$

This was found by Solomonoff ([Sol78]) for universal sequence prediction. A proof is also given in [LV97] (only for binary alphabet) or [Hut01a] (arbitrary alphabet). It is surprisingly simple once Lemma 4.2 is known. A few lines analogous to (8) and (9) exploiting the dominance of ξ are sufficient.

The bound (5) asserts convergence of the ξ-predictions to the μ-predictions *in mean sum (i.m.s.)*, since we define

$$\varphi\xrightarrow{i.m.s.}\mu\quad\Longleftrightarrow\quad\exists\ C>0:\ \sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\mu(a|x_{<t})-\varphi(a|x_{<t})\Big)^2\leq C. \tag{6}$$

Convergence i.m.s. implies convergence with μ-probability one (w.μ-p.1), since otherwise the sum would be infinite. Moreover, convergence i.m.s. provides a rate or speed of convergence in the sense that the expected number of times t in which $\varphi(a|x_{<t})$ deviates more than ε from $\mu(a|x_{<t})$ is finite and bounded by C/ε^2 and the probability that the number of ε-deviations exceeds $\frac{C}{\varepsilon^2\delta}$ is smaller than δ. If the quadratic differences were monotonically decreasing (which is usually not the case), we could even conclude convergence faster than $\frac{1}{t}$.

2.2 Probabilities vs. Description Lengths. By the Kraft inequality, each (semi)measure can be associated with a code length or *complexity* by means of the negative logarithm, where all (binary) codewords form a prefix-free set. The converse holds as well. E.g. for the weights w_ν with $\sum w_\nu\leq 1$, codes of lengths $\lceil-\log_2 w_\nu\rceil$

can be found. It is often only a matter of notational convenience if description lengths or probabilities are used, but description lengths are generally preferred in Algorithmic Information Theory. Keeping the equivalence in mind, we will develop the general theory in terms of probabilities, but formulate parts of the results in universal sequence prediction rather in terms of complexities.

3 MDL Estimator and Predictions

Assume that $\mathcal{C}$ is a countable class of semimeasures together with weights $(w_\nu)_{\nu\in\mathcal{C}}$, and $x \in \mathcal{X}^*$ is some string. Then the *maximizing element* ν^x, often called MAP estimator, is defined as

$$\nu^x = \nu^x_{[\mathcal{C}]} = \arg\max_{\nu\in\mathcal{C}}\{w_\nu \nu(x)\}.$$

In fact the maximum is attained since for each $\varepsilon \in (0,1)$ only a finite number of elements fulfil $w_\nu \nu(x) > \varepsilon$. Observe immediately the correspondence in terms of *description lengths* rather than *probabilities*: $\nu^x = \arg\min_{\nu\in\mathcal{C}}\{-\log_2 w(\nu) - \log_2 \nu(x)\}$. Then the *minimum description length principle* is obvious: ν^x minimizes the joint description length of the model plus the data given the model[1] (see the last paragraph of the previous section). As explained before, we stick to the product notation.

For notational simplicity we set $\nu^*(x) = \nu^x(x)$. The *two-part MDL estimator* is defined by

$$\varrho(x) = \varrho_{[\mathcal{C}]}(x) = w_{\nu^x}\nu^x(x) = \max_{\nu\in\mathcal{C}}\{w_\nu \nu(x)\}.$$

So ϱ chooses the maximizing element with respect to its argument. We may also use the version $\varrho^y(x) := w_{\nu^y}\nu^y(x)$ for which the choice depends on the superscript instead of the argument. For each $x, y \in \mathcal{X}^*$, $\xi(x) \geq \varrho(x) \geq \varrho^y(x)$ is immediate.

We can define MDL predictors according to (3). There are *at least three* possible ways to use MDL for prediction.

3.1 Definition. The *dynamic* MDL predictor is defined as

$$\varrho(a|x) = \frac{\varrho(xa)}{\varrho(x)} = \frac{\varrho^{xa}(xa)}{\varrho^x(x)}.$$

That is, we look for a short description of xa and relate it to a short description of $x = x_{<t}$. We call this dynamic since for each possible a we have to find a new MDL estimator. This is the closest correspondence to the ξ-predictor.

[1] Precisely, we define a MAP (maximum a posteriori) estimator. For two reasons, information theorists and statisticians would not consider our definition as MDL in the strong sense. First, MDL is often associated with a specific prior. Second, when coding some data x, one can exploit the fact that once the model ν^x is specified, only data which leads to the maximizing element ν^x needs to be considered. This allows for a description shorter than $\log_2 \nu^x(x)$. Since however most authors refer to MDL, we will keep using this general term instead of MAP, too.

3.2 Definition. The *static* MDL predictor is given by

$$\varrho^{\text{static}}(a|x) = \varrho^x(a|x) = \frac{\varrho^x(xa)}{\varrho(x)} = \frac{\varrho^x(xa)}{\varrho^x(x)} = \frac{\nu^x(xa)}{\nu^x(x)}.$$

Here obviously only *one* MDL estimator ϱ^x has to be identified, which may be more efficient in practice.

3.3 Definition. The *hybrid* MDL predictor is given by $\varrho^{\text{hyb}}(a|x) = \frac{\nu^*(xa)}{\nu^*(x)}$. This can be paraphrased as "do dynamic MDL and drop the weights". It is somewhat in-between static and dynamic MDL.

The range of the static MDL predictor is obviously contained in $[0,1]$. For the dynamic MDL predictor, this holds by $\varrho^x(x) \geq \varrho^{xa}(x) \geq \varrho^{xa}(xa)$, while for the hybrid MDL predictor it is generally false.

Static MDL is omnipresent in machine learning and applications. In fact, many common prediction algorithms can be abstractly understood as static MDL, or rather as approximations. Namely, if a prediction task is accomplished by building a *model* such as a neural network with a suitable regularization to prevent "overfitting", this is just searching an MDL estimator within a certain class of distributions. After that, only this model is used for prediction. Dynamic and hybrid MDL are applied more rarely due to their larger computational effort. For example, the similarity metric proposed in [LCL$^+$03] can be interpreted as (a deterministic variant of) dynamic MDL. For hybrid MDL, we will see that the prediction properties are worse than for dynamic and static MDL.

We will need to convert our MDL predictors to *measures* on $\mathcal{X}$ by means of *normalization*. If $\varphi : \mathcal{X}^* \to [0,1]$ is any function, then

$$\varphi_{norm}(a|x_{<t}) \;=\; \frac{\varphi(a|x_{<t})}{\sum_{a'\in\mathcal{X}} \varphi(a'|x_{<t})} \;=\; \frac{\varphi(x_{<t}a)}{\sum_{a'\in\mathcal{X}} \varphi(x_{<t}a')}$$

(assume that the denominator is different from zero, which is always true with probability 1 if φ is an MDL predictor). This procedure is known as *Solomonoff normalization* ([Sol78, LV97]) and results in $\nu_{norm}(x_{1:n}) = \nu(x_{1:n})/[\nu(\epsilon)N_\nu(x_{<n})]$, where

$$N_\nu(x) = \prod_{t=1}^{\ell(x)+1} \frac{\sum_{a\in\mathcal{X}} \nu(x_{<t}a)}{\nu(x_{<t})} \tag{7}$$

is the normalizer. Before proceeding with the theory, an example is in order.

3.4 Example. Let $n \in \mathbb{N}$, $\mathcal{X} = \{1, \ldots, n\}$, and

$$\mathcal{C} = \Big\{\nu_\vartheta(x_{1:t}) = \vartheta_{x_1} \cdot \ldots \cdot \vartheta_{x_t} : \vartheta \in \Theta\Big\} \quad \text{with} \quad \Theta = \Big\{\vartheta \in ([0,1] \cap \mathbb{Q})^n : \sum_{i=1}^n \vartheta_i = 1\Big\}$$

be the set of all rational probability vectors with any prior $(w_\vartheta)_{\vartheta\in\Theta}$. Each $\vartheta \in \Theta$ generates sequences $x_{<\infty}$ of *independently identically distributed (i.i.d)* random variables such that $P(x_t = i) = \vartheta_i$ for all $t \geq 1$ and $1 \leq i \leq n$. If $x_{1:t}$ is the initial part of a sequence and $\alpha \in \Theta$ is defined by $\alpha_i = |\{s \leq t : x_s = i\}|$, then it is easy to see that

$$\nu^{x_{1:t}} = \arg\max_{\vartheta\in\Theta} \{w(\vartheta) \cdot \exp[-t \cdot D(\alpha\|\vartheta)]\},$$

where $D(\alpha\|\vartheta) = \sum_{i=1}^n \alpha_i \ln \frac{\alpha_i}{\vartheta_i}$ is the *Kullback-Leibler divergence*. If $|\mathcal{X}| = 2$, then Θ is also called a *Bernoulli class*, and one usually takes the binary alphabet $\mathcal{X} = \mathbb{B} = \{0, 1\}$ in this case.

4 Dynamic MDL

We can start to develop results. It is surprisingly easy to give a convergence proof w.p.1 of the non-normalized dynamic MDL predictions based on martingales. However we omit it, since it does not include a convergence speed assertion as i.m.s. results do, nor does it yield an off-sequence statement about $\varrho(a|x_{<t})$ for $a \neq x_t$ which is necessary for prediction.

4.1 Lemma. *For an arbitrary class of (semi)measures $\mathcal{C}$, we have*

$$\begin{aligned} (i) \quad & \varrho(x) - \sum_{a\in\mathcal{X}} \varrho(xa) \;\leq\; \xi(x) - \sum_{a\in\mathcal{X}} \xi(xa) \text{ and} \\ (ii) \quad & \varrho^x(x) - \sum_{a\in\mathcal{X}} \varrho^x(xa) \;\leq\; \xi(x) - \sum_{a\in\mathcal{X}} \xi(xa) \end{aligned}$$

for all $x \in \mathcal{X}^*$. In particular, $\xi - \varrho$ is a semimeasure.

Proof. For all $x \in \mathcal{X}^*$, with $f := \xi - \varrho$ we have

$$\begin{aligned} \sum_{a\in\mathcal{X}} f(xa) &= \sum_{a\in\mathcal{X}} \Big(\xi(xa) - \varrho(xa)\Big) \leq \sum_{a\in\mathcal{X}} \Big(\xi(xa) - \varrho^x(xa)\Big) \\ &= \sum_{\nu\in\mathcal{M}\setminus\{\nu^x\}} \sum_{a\in\mathcal{X}} w_\nu \nu(xa) \leq \sum_{\nu\in\mathcal{M}\setminus\{\nu^x\}} w_\nu \nu(x) = \xi(x) - \varrho(x) = f(x). \end{aligned}$$

The first inequality follows from $\varrho^x(xa) \leq \varrho(xa)$, and the second one holds since all ν are semimeasures. Finally, $f(x) = \xi(x) - \varrho(x) = \sum_{\nu\in\mathcal{M}\setminus\{\nu^x\}} w_\nu \nu(x) \geq 0$ and $f(\epsilon) = \xi(\epsilon) - \varrho(\epsilon) \leq 1$. Hence f is a semimeasure. □

4.2 Lemma. *Let μ and $\tilde{\mu}$ be measures on $\mathcal{X}$, then*

$$\sum_{a\in\mathcal{X}} (\mu(a) - \tilde{\mu}(a))^2 \leq \sum_{a\in\mathcal{X}} \mu(a) \ln \frac{\mu(a)}{\tilde{\mu}(a)}.$$

See e.g. [Hut01a, Sec.3.2] for a proof.

4.3 Theorem. *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ and for all $n \in \mathbb{N}$, we have*

$$\sum_{t=1}^{n} \mathbf{E} \sum_{a \in \mathcal{X}} \big(\mu(a|x_{<t}) - \varrho_{norm}(a|x_{<t})\big)^2 \leq w_\mu^{-1} + \ln w_\mu^{-1}.$$

That is, $\varrho_{norm}(a|x_{<t}) \stackrel{i.m.s.}{\longrightarrow} \mu(a|x_{<t})$ (see (6)), which implies $\varrho_{norm}(a|x_{<t}) \to \mu(a|x_{<t})$ with μ-probability one.

Proof. From Lemma 4.2, we know

$$\sum_{t=1}^{n} \mathbf{E} \sum_{a \in \mathcal{X}} \big(\mu(a|x_{<t}) - \varrho_{norm}(a|x_{<t})\big)^2 \leq \sum_{t=1}^{n} \mathbf{E} \sum_{a \in \mathcal{X}} \mu(a|x_{<t}) \ln \frac{\mu(a|x_{<t})}{\varrho_{norm}(a|x_{<t})}$$
$$= \sum_{t=1}^{n} \mathbf{E} \ln \frac{\mu(x_t|x_{<t})}{\varrho_{norm}(x_t|x_{<t})} = \sum_{t=1}^{n} \mathbf{E} \left[\ln \frac{\mu(x_t|x_{<t})}{\varrho(x_t|x_{<t})} + \ln \frac{\sum_{a \in \mathcal{X}} \varrho(x_{<t}a)}{\varrho(x_{<t})} \right]. \tag{8}$$

Then we can estimate

$$\sum_{t=1}^{n} \mathbf{E} \ln \frac{\mu(x_t|x_{<t})}{\varrho(x_t|x_{<t})} \;=\; \mathbf{E} \;\ln \prod_{t=1}^{n} \frac{\mu(x_t|x_{<t})}{\varrho(x_t|x_{<t})} \;=\; \mathbf{E} \;\ln \frac{\mu(x_{1:n})}{\varrho(x_{1:n})} \;\leq\; \ln w_\mu^{-1}, \tag{9}$$

since always $\frac{\mu}{\varrho} \leq w_\mu^{-1}$. Moreover, by setting $x = x_{<t}$, using $\ln u \leq u - 1$, adding an always positive max-term, and finally using $\frac{\mu}{\varrho} \leq w_\mu^{-1}$ again, we obtain

$$\mathbf{E} \;\ln \frac{\sum_a \varrho(x_{<t}a)}{\varrho(x_{<t})} \leq \mathbf{E} \left[\frac{\sum_a \varrho(xa)}{\varrho(x)} - 1 \right] = \sum_{\ell(x)=t-1} \frac{\mu(x) \Big[\big(\sum_a \varrho(xa) \big) - \varrho(x) \Big]}{\varrho(x)}$$
$$\leq \sum_{\ell(x)=t-1} \frac{\mu(x) \Big[\big(\sum_{a \in \mathcal{X}} \varrho(xa) \big) - \varrho(x) + \max \big\{ 0, \varrho(x) - \sum_{a \in \mathcal{X}} \varrho(xa) \big\} \Big]}{\varrho(x)}$$
$$\leq w_\mu^{-1} \sum_{\ell(x)=t-1} \left[\Big(\sum_{a \in \mathcal{X}} \varrho(xa) \Big) - \varrho(x) + \max \Big\{ 0, \varrho(x) - \sum_{a \in \mathcal{X}} \varrho(xa) \Big\} \right]. \tag{10}$$

We proceed by observing

$$\sum_{t=1}^{n} \sum_{\ell(x)=t-1} \left[\Big(\sum_{a \in \mathcal{X}} \varrho(xa) \Big) - \varrho(x) \right] = \sum_{t=1}^{n} \left[\sum_{\ell(x)=t} \varrho(x) - \sum_{\ell(x)=t-1} \varrho(x) \right] = \left[\sum_{\ell(x)=n} \varrho(x) \right] - \varrho(\epsilon) \tag{11}$$

which is true since for successive t the positive and negative terms cancel. From Lemma 4.1 we know $\varrho(x)-\sum_{a\in\mathcal{X}}\varrho(xa) \;\leq\; \xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)$ and therefore

$$\sum_{t=1}^{n}\sum_{\ell(x)=t-1}\max\Big\{0,\varrho(x)-\sum_{a\in\mathcal{X}}\varrho(xa)\Big\} \;\leq\; \sum_{t=1}^{n}\sum_{\ell(x)=t-1}\max\Big\{0,\xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)\Big\}$$
$$=\sum_{t=1}^{n}\sum_{\ell(x)=t-1}\Big[\xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)\Big] \;=\; \xi(\epsilon)-\sum_{\ell(x)=n}\xi(x). \tag{12}$$

Here we have again used the fact that positive and negative terms cancel for successive t, and moreover the fact that ξ is a semimeasure. Combining (10), (11) and (12), and observing $\varrho\leq\xi\leq 1$, we obtain

$$\sum_{t=1}^{n}\mathbf{E}\ln\frac{\sum_a\varrho(x_{<t}a)}{\varrho(x_{<t})}\leq w_\mu^{-1}\left[\xi(\epsilon)-\varrho(\epsilon)+\sum_{\ell(x)=n}(\varrho(x)-\xi(x))\right]\leq w_\mu^{-1}\xi(\epsilon)\leq w_\mu^{-1}. \tag{13}$$

Therefore, (8), (9) and (13) finally prove the assertion. □

This is the first convergence result in mean sum, see (6). It implies both on-sequence and off-sequence convergence. Moreover, it asserts the convergence is "fast" in the sense that the sum of the total expected deviations is bounded by $w_\mu^{-1}+\ln w_\mu^{-1}$. Of course, w_μ^{-1} can be very large, namely 2 to the power of complexity of μ. The following example will show that this bound is sharp (save for a constant factor). Observe that in the corresponding result for mixtures, Theorem 2.1, the bound is much smaller, namely $\ln w_\mu^{-1}$ = complexity of μ.

4.4 Example. Let $\mathcal{X}=\{0,1\}$, $N\geq 1$ and $\mathcal{C}=\{\nu_1,\ldots,\nu_{N-1},\mu\}$. Each ν_i is a deterministic measure concentrated on the sequence $1^{i-1}0^\infty$, while the true distribution μ is deterministic and concentrated on $x_{<\infty}=1^\infty$. Let $w_{\nu_i}=w_\mu=\frac{1}{N}$ for all i. Then μ generates $x_{<\infty}$, and for each $t\leq N-1$ we have $\varrho_{norm}(0|x_{<t})=\varrho_{norm}(1|x_{<t})=\frac{1}{2}$. Hence, $\sum_t\mathbf{E}\sum_a(\mu(a|x_{<t})-\varrho_{norm}(a|x_{<t}))^2=\frac{1}{2}(N-1)\approx\frac{1}{2}w_\mu^{-1}$ for large N. Here, μ is Bernoulli, while the ν_i are not. It might be surprising at a first glance that there are even classes $\mathcal{C}$ containing *only* Bernoulli distributions, where the exponential bound is sharp [PH04].

4.5 Theorem. *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

$$(i)\quad \sum_{t=1}^{\infty}\mathbf{E}\left|\ln\sum_{a\in\mathcal{X}}\varrho(a|x_{<t})\right| \;\leq\; 2w_\mu^{-1}\quad\text{and}$$

$$(ii)\quad \sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big|\varrho_{norm}(a|x_{<t})-\varrho(a|x_{<t})\Big| \;=\; \sum_{t=1}^{\infty}\mathbf{E}\Big|1-\sum_{a\in\mathcal{X}}\varrho(a|x_{<t})\Big| \;\leq\; 2w_\mu^{-1}.$$

Consequently, $\varrho(a|x_{<t}) \stackrel{i.m.s.}{\longrightarrow} \mu(a|x_{<t})$*, and for almost all* $x_{<\infty} \in \mathcal{X}^\infty$*, the normalizer* N_ϱ *defined in (7) converges to a number which is finite and greater than zero, i.e.* $0 < N_\varrho(x_{<\infty}) < \infty$*.*

Proof. (i) Define $u^+ = \max\{0, u\}$ for $u \in \mathbb{R}$, then for $x := x_{<t} \in \mathcal{X}^{t-1}$ we have

$$
\begin{aligned}
\mathbf{E}\Big|\ln\sum_{a\in\mathcal{X}}\varrho(a|x)\Big| = \mathbf{E}\Big|\ln\frac{\sum_a \varrho(xa)}{\varrho(x)}\Big| &= \mathbf{E}\Big[\Big(\ln\frac{\sum_a \varrho(xa)}{\varrho(x)}\Big)^+ + \Big(\ln\frac{\varrho(x)}{\sum_a \varrho(xa)}\Big)^+\Big] \\
&\le \mathbf{E}\frac{(\sum_a \varrho(xa) - \varrho(x))^+}{\varrho(x)} + \mathbf{E}\frac{(\varrho(x) - \sum_a \varrho(xa))^+}{\sum_a \varrho(xa)} \\
&= \sum_{\ell(x)=t-1}\frac{\mu(x)(\sum_a \varrho(xa) - \varrho(x))^+}{\varrho(x)} + \sum_{\ell(x)=t-1}\frac{\mu(x)(\varrho(x) - \sum_a \varrho(xa))^+}{\sum_a \varrho(xa)} \\
&\le w_\mu^{-1}\sum_{\ell(x)=t-1}(\textstyle\sum_a \varrho(xa) - \varrho(x))^+ + w_\mu^{-1}\displaystyle\sum_{\ell(x)=t-1}(\varrho(x) - \textstyle\sum_a \varrho(xa))^+ \\
&= w_\mu^{-1}\sum_{\ell(x)=t-1}\big[\textstyle\sum_a \varrho(xa) - \varrho(x) + 2(\varrho(x) - \sum_a \varrho(xa))^+\big].
\end{aligned}
$$

Here, $|u| = u^+ + (-u)^+ = -u + 2u^+$, $\ln u \le u - 1$, and $\varrho \ge w_\mu \mu$ have been used, the latter implies also $\sum_a \varrho(xa) \ge w_\mu \sum_a \mu(xa) = w_\mu \mu(x)$. The last expression in this (in)equality chain, when summed over $t = 1...\infty$ is bounded by $2w_\mu^{-1}$ by essentially the same arguments (10) - (13) as in the proof of Theorem 4.3.

(ii) Let again $x := x_{<t}$ and use $\varrho_{norm}(a|x) = \varrho(a|x)/\sum_b \varrho(b|x)$ to obtain

$$
\begin{aligned}
\sum_a \Big|\varrho_{norm}(a|x) - \varrho(a|x)\Big| &= \sum_a \frac{\varrho(a|x)}{\sum_b \varrho(b|x)}\Big|1 - \sum_b \varrho(b|x)\Big| = \Big|1 - \sum_b \varrho(b|x)\Big| \\
&= \frac{(\sum_a \varrho(xa) - \varrho(x))^+}{\varrho(x)} + \frac{(\varrho(x) - \sum_a \varrho(xa))^+}{\varrho(x)}
\end{aligned}
$$

Then take the expectation $\mathbf{E}$ and the sum $\sum_{t=1}^\infty$ and proceed as in (i). Finally, $\varrho(a|x_{<t}) \stackrel{i.m.s.}{\longrightarrow} \mu(a|x_{<t})$ follows by combining (ii) with Theorem 4.3, and by (i), $\sum_1^n |\ln\frac{\sum_a \varrho(x_{<t}a)}{\varrho(x_{<t})}|$ is bounded in n with μ-probability 1, thus the same is true for $\ln N_\varrho(x_{<\infty}) = \sum_1^\infty \ln\frac{\sum_{a\in\mathcal{X}} \varrho(x_{<t}a)}{\varrho(x_{<t})}$. $\square$

5 Static MDL

So far, we have considered dynamic MDL from Definition 3.1. We turn now to the static variant (Definition 3.2), which is usually more efficient and thus preferred in practice.

5.1 Theorem. *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

$$\sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\left|\varrho_{norm}^{x_{<t}}(a|x_{<t})-\varrho^{x_{<t}}(a|x_{<t})\right| \;=\; \sum_{t=1}^{\infty}\mathbf{E}\left|1-\sum_{a\in\mathcal{X}}\varrho^{x_{<t}}(a|x_{<t})\right| \;\leq\; w_\mu^{-1}.$$

Proof. We proceed in a similar way as in the proof of Theorem 4.3, (10) - (12). From Lemma 4.1, we know $\varrho(x)-\sum_a\varrho^x(xa) \;\leq\; \xi(x)-\sum_a\xi(xa)$. Then

$$\sum_{t=1}^{n}\mathbf{E}\left|1-\sum_{a\in\mathcal{X}}\varrho^{x_{<t}}(a|x_{<t})\right| = \sum_{t=1}^{n}\;\mathbf{E}\frac{\varrho(x_{<t})-\sum_{a\in\mathcal{X}}\varrho^{x_{<t}}(x_{<t}a)}{\varrho(x_{<t})}$$
$$= \;\sum_{t=1}^{n}\;\sum_{\ell(x)=t-1}\mu(x)\frac{\varrho(x)-\sum_{a\in\mathcal{X}}\varrho^x(xa)}{\varrho(x)} \leq w_\mu^{-1}\;\sum_{t=1}^{n}\;\sum_{\ell(x)=t-1}\Big[\varrho(x)-\sum_{a\in\mathcal{X}}\varrho^x(xa)\Big]$$
$$\leq\; w_\mu^{-1}\sum_{t=1}^{n}\;\sum_{\ell(x)=t-1}\left[\xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)\right] \leq w_\mu^{-1}\left[\xi(\epsilon)-\sum_{\ell(x)=n}\xi(x)\right] \;\leq\; w_\mu^{-1}.$$

for all $n\in\mathbb{N}$. This implies the assertion. Again we have used $\frac{\mu}{\varrho}\leq w_\mu^{-1}$ and the fact that positive and negative terms cancel for successive t. □

5.2 Corollary. *Let $\mathcal{C}$ contain the true distribution μ, then*

$$\begin{array}{lcl}
\sum_t\mathbf{E}\sum_a\;(\mu(a|x_{<t})-\varrho_{norm}(a|x_{<t}))^2 & \leq & 2w_\mu^{-1},\\
\sum_t\mathbf{E}\sum_a\;(\mu(a|x_{<t})-\varrho(a|x_{<t}))^2 & \leq & 8w_\mu^{-1},\\
\sum_t\mathbf{E}\sum_a\;(\mu(a|x_{<t})-\varrho^{x_{<t}}(a|x_{<t}))^2 & \leq & 21w_\mu^{-1},\\
\sum_t\mathbf{E}\sum_a\;(\mu(a|x_{<t})-\varrho_{norm}^{x_{<t}}(a|x_{<t}))^2 & \leq & 32w_\mu^{-1}.
\end{array}$$

Proof. This follows by combining the assertions of Theorems 4.3 - 5.1 with the triangle inequality. For static MDL, use in addition $\sum_a|\varrho(a|x)-\varrho^x(a|x)| = |\sum_a\varrho(a|x)-\sum_a\varrho^x(a|x)| \leq |\sum_a\varrho(a|x)-1|+|1-\sum_a\varrho^x(a|x)|$ which follows from $\varrho(xa)\geq\varrho^x(xa)$. □

This corollary recapitulates our results and states convergence i.m.s (and therefore also with μ-probability 1) for all combinations of un-normalized/normalized and dynamic/static MDL predictions.[2]

[2] We briefly discuss the choice of the total expected square error for measuring speed of convergence. The expected Kullback-Leibler distance may seem more natural in the light of our proofs. However, this quantity behaves well only under dynamic MDL, not static MDL. To see this, let $\mathcal{C}$ be the class of all computable Bernoulli distributions and μ the measure having $\mu(0)=\mu(1)=\frac{1}{2}$. Then the sequence $x=0^n$ has nonzero probability. For sufficiently large n, $\nu^x=\nu_0$ holds (typically already for small n), where ν_0 is the distribution generating only 0. Then $D(\mu\|\nu^x)=\infty$, and the expectation is ∞, too. The quadratic distance behaves locally like the Kullback-Leibler distance (Lemma 4.2), but otherwise is bounded and thus more convenient.

6 Hybrid MDL and Stabilization

We now turn to the hybrid MDL variant (see Definition 3.3). So far we have not cared about what happens if two or more (semi)measures obtain the same value $w_\nu \nu(x)$ for some string x. In fact, for the previous results, the *tie-breaking strategy* can be completely arbitrary. This need not be so for all thinkable prediction methods other than static and dynamic MDL, as the following example shows.

6.1 Example. Let $\mathcal{X} = \mathbb{B}$ and $\mathcal{C}$ contain only two measures, the uniform measure λ which is defined by $\lambda(x) = 2^{-\ell(x)}$, and another measure ν having $\nu(1x) = 2^{-\ell(x)}$ and $\nu(0x) = 0$. The respective weights are $w_\lambda = \frac{2}{3}$ and $w_\nu = \frac{1}{3}$. Then, for each x starting with 1, we have $w_\nu \nu(x) = w_\lambda \lambda(x) = \frac{1}{3} 2^{-\ell(x)+1}$. Therefore, for all $x_{<\infty}$ starting with 1 (a set which has uniform measure $\frac{1}{2}$), we have a tie. If the maximizing element ν^* is chosen to be λ for even t and ν for odd t, then both static and dynamic MDL constantly predict probabilities of $\frac{1}{2}$ for all $a \in \mathbb{B}$. However, the hybrid MDL predictor values $\frac{\nu^*(x_{<t}a)}{\nu^*(x_{<t})}$ oscillate between $\frac{1}{4}$ and 1.

If the ambiguity in the tie-breaking process is removed, e.g. if always the measure with the larger weight w_ν is been chosen, then the hybrid MDL predictor *does* converge for this example. If there are more (semi)measures in the class and there remains still a tie of shortest programs, an arbitrary program can be selected, since then the respective measures are equal, too. In the following, we assume that this tie-breaking rule is applied.

Do the hybrid MDL predictions always converge then? This is equivalent to asking if the process of selecting a maximizing element eventually *stabilizes*. If there is no stabilization, then hybrid MDL will necessarily fail as soon as the weights are not equal. A possible counterexample could consist of two measures the fraction of which oscillates perpetually around a certain value. This can indeed happen.

6.2 Example. Let $\mathcal{X}$ be binary, $\mu(x) = \prod_{i=1}^{\ell(x)} \mu_i(x_i)$ and $\nu(x) = \prod_{i=1}^{\ell(x)} \nu_i(x_i)$ with

$$\mu_i(1) = 1 - 2^{-2\lceil \frac{i}{2} \rceil} \text{ and } \nu_i(1) = 1 - 2^{-2\lceil \frac{i+1}{2} \rceil + 1}.$$

Then one can easily see that $\mu(111\ldots) = \prod_1^\infty \mu_i(1) > 0$, $\nu(111\ldots) = \prod_1^\infty \nu_i(1) > 0$, and $\frac{\nu(111\ldots)}{\mu(111\ldots)}$ is convergent but oscillates around its limit. Therefore, we can set w_μ and w_ν appropriately to prevent the maximizing element from stabilizing on $x_{<\infty} = 111\ldots$ (Moreover, each sequence having positive measure under μ and ν contains eventually only ones, and the quotient oscillates.)

The reason for the oscillation in this example is the fact that measures μ and ν are asymptotically very similar. One can also achieve a similar effect by constructing a measure which is *dependent* on the past. This shows in particular that we need both parts of the following definition which states properties sufficient for a positive result.

6.3 Definition. (i) A (semi)measure ν on $\mathcal{X}^\infty$ is called *factorizable* if there are (semi)measures ν_i on $\mathcal{X}$ such that $\nu(x) = \prod_{i=1}^{\ell(x)} \nu_i(x_i)$ for all $x \in \mathcal{X}^*$. That is, the symbols of sequences $x_{<\infty}$ generated by ν are independent.
(ii) A factorizable (semi)measure $\mu = \prod \mu_i$ is called *uniformly stochastic*, if there is some $\delta > 0$ such that at each time i the probability of all symbols $a \in \mathcal{X}$ is either 0 or at least δ. That is, $\mu_i(a) > 0 \Rightarrow \mu_i(a) \geq \delta$ for all $a \in \mathcal{X}$ and $i \geq 1$.

In particular, all deterministic measures are uniformly stochastic. Another simple example of a uniformly stochastic measure is a probability distribution which generates alternately random bits by fair coin flips and the digits of the binary representation of π.

6.4 Theorem. *Let $\mathcal{C}$ be a countable class of factorizable (semi)measures and μ be uniformly stochastic. Then the maximizing element stabilizes almost surely.*

We omit the proof. So in particular, under the conditions of Theorem 6.4, the hybrid MDL predictions converge almost surely. No statement about the convergence speed can be made.

7 Complexities and Randomness

In this section, we concentrate on universal sequence prediction. It was mentioned already in the introduction that this is one interesting application of the theory developed so far. So $\mathcal{C} = \mathcal{M}$ is the countable set of all enumerable (i.e. lower semi-computable) semimeasures on $\mathcal{X}^*$. (Algorithms are identified with semimeasures rather than measures since they need not terminate.) $\mathcal{M}$ contains stochastic models in general, and in particular all models for computable deterministic sequences. One can show that this class $\mathcal{M}$ is determined by *all* algorithms on some fixed universal monotone Turing machine U [LV97, Th. 4.5.2]. By this correspondence, each semimeasure $\nu \in \mathcal{M}$ is assigned a *canonical weight* $w_\nu = 2^{-K(\nu)}$ (where $K(\nu)$ is the Kolmogorov complexity of ν, see [LV97, Eq. 4.11]), and $\sum w_\nu \leq 1$ holds. We will assume programs to be *binary*, i.e. $p \in \mathbb{B}^*$, in contrast to outputs, which are strings $x \in \mathcal{X}^*$.

The MDL definitions in Section 3 directly transfer to this setup. All our results (Theorems 4.3 - 5.1) therefore apply to $\varrho = \varrho_{[\mathcal{M}]}$ if the true distribution μ is a measure, which is not very restrictive. Then μ is necessarily computable. Also, Theorem 2.1 implies Solomonoff's important *universal induction* theorem: ξ converges to the true distribution i.m.s., if the latter is computable. Note that the Bayes mixture ξ is within a multiplicative constant of the *Solomonoff-Levin prior* $M(x)$, which is the algorithmic probability that U produces an output starting with x if its input is random.

In addition to $\mathcal{M}$, we also consider the set of all recursive measures $\tilde{\mathcal{M}}$ together with the same canonical weights, and the mixture $\tilde{\xi}(x) = \sum_{\nu \in \tilde{\mathcal{M}}} w_\nu \nu(x)$. Likewise, define $\tilde{\varrho} = \varrho_{[\tilde{\mathcal{M}}]}$. Then we obviously have $\tilde{\varrho}(x) \leq \tilde{\xi}(x) \leq \xi(x)$ and $\varrho(x) \leq \xi(x)$ for all $x \in \mathcal{X}^*$. It is even immediate that $\xi(x) \stackrel{\times}{\leq} \varrho(x)$ since $\xi \in \mathcal{M}$. Here, by $f \stackrel{\times}{\leq} g$ we mean $f \leq g \cdot O(1)$, "$\stackrel{\times}{\geq}$" and "$\stackrel{\times}{=}$" are defined analogously.

Moreover, for any string $x \in \mathcal{X}^*$, there is also a *universal one-part MDL estimator* $m(x) = 2^{-Km(x)}$ derived from the monotone complexity $Km(x) = \min\{\ell(p) : U(p) = x*\}$. (I.e. the monotone complexity is the length of the shortest program such that U's output starts with x.) The minimal program p defines a measure ν with $\nu(x) = 1$ and $w_\nu \geq 2^{-\ell(p)} \cdot O(1)$ (recall that programs are binary). Therefore, $m(x) \stackrel{\times}{\leq} \tilde{\varrho}(x)$ for all $x \in \mathcal{X}^*$. Together with the following proposition, we thus obtain

$$m(x) \;\stackrel{\times}{=}\; \tilde{\varrho}(x) \;\stackrel{\times}{\leq}\; \tilde{\xi}(x) \;\stackrel{\times}{\leq}\; \varrho(x) \;\stackrel{\times}{=}\; \xi(x) \text{ for all } x \in \mathcal{X}^*. \tag{14}$$

7.1 Proposition. *We have $\tilde{\varrho}(x) \stackrel{\times}{\leq} m(x)$ for all $x \in \mathcal{X}^*$.*

Proof. (Sketch only.) It is not hard to show that given a string $x \in \mathcal{X}^*$ and a recursive measure ν (which in particular may be the MDL descriptor $\nu^*(x)$) it is possible to specify a program p of length at most $-\log_2 w_\nu - \log_2 \nu(x) + c$ that outputs a string starting with x, where constant c is independent of x and ν. This is done via arithmetic encoding. Alternatively, it is also possible to prove the proposition indirectly using [LV97, Th.4.5.4]. This implies that $m(x) \stackrel{\times}{\geq} w_\nu \nu(x)$ for all $x \in \mathcal{X}^*$ and all recursive measures ν. Then, also $m(x) \stackrel{\times}{\geq} \max\{w_\nu \nu(x)\}$ holds. □

On the other hand, we know from [Gác83] that $m \stackrel{\times}{\neq} \xi$. Therefore, at least one of the two inequalities in (14) must be proper.

7.2 Problem. *Which of the inequalities $\tilde{\varrho} \stackrel{\times}{\leq} \tilde{\xi}$ and $\tilde{\xi} \stackrel{\times}{\leq} \varrho$ is proper (or are both)?*

Equation (14) also has an easy consequence in terms of randomness criteria.

7.3 Proposition. *A sequence $x_{<\infty} \in \mathcal{X}^\infty$ is Martin-Löf random with respect to some computable measure μ iff for any $f \in \{m, \tilde{\varrho}, \tilde{\xi}, r, M\}$ there is a constant $C > 0$ such that $f(x_{1:n}) \leq C\mu(x_{1:n})$ for all $n \in \mathbb{N}$ holds.*

Proof. It is a standard result that if $x_{<\infty}$ is random then $M(x_{1:n}) \leq C\mu(x_{1:n})$ for some C [Lev73, Th.3]. Then by (14), $f(x_{1:n}) \stackrel{\times}{\leq} \mu(x_{1:n})$ for all f. Conversely, if $f(x_{1:n}) \stackrel{\times}{\leq} \mu(x_{1:n})$ for some f, then there is C such that $m(x_{1:n}) \leq C\mu(x_{1:n})$. This implies μ-randomness of $x_{<\infty}$ ([Lev73, Th.2] or [LV97, p295]). □

Interestingly, these randomness criteria partly depend on the weights. The criteria for $\tilde{\xi}$ and $\tilde{\varrho}$ are not equivalent any more if weights other than the canonical weights are used, as the following example will show. In contrast, for ξ and ϱ there is no weight dependency as long as the weights are strictly greater than zero, since $\xi \in \mathcal{M}$.

7.4 Example. There are other randomness criteria than Martin-Löf randomness, e.g. rec-randomness. A rec-random sequence $x_{<\infty}$ (with respect to the uniform distribution) satisfies $\nu(x_{1:n}) \leq c(\nu)2^{-n}$ for each computable measure ν and for all n. It is obvious that Martin-Löf random sequences are also rec-random. The converse does not hold, there are sequences $x_{<\infty}$ that are rec-random but not Martin-Löf random, as shown e.g. in [Sch71, Wan96].

Let $x_{<\infty}$ be such a sequence, i.e. $\nu(x_{1:n}) \leq c(\nu)2^{-n}$ for all computable measures ν and for all n, but where $x_{<\infty}$ is not Martin-Löf random. Let $\nu_1, \nu_2, \ldots$ be a (non-effective) enumeration of all computable measures. Define $w'_i = 2^{-i}c(\nu_i)^{-1}$. Then

$$\tilde{M}'(x_{1:n}) = \sum_{i=1}^{\infty} w'_i \nu_i(x_{1:n}) \leq \sum_{i=1}^{\infty} 2^{-i}c(\nu_i)^{-1}c(\nu_i)2^{-n} = 2^{-n} \text{ for all } n,$$

i.e. $x_{<\infty}$ is $\tilde{M}'$-random. Thus, $x_{<\infty}$ is also $\tilde{r}'$-random with $\tilde{r}' = \max_i\{w'_i\nu_i\}$.

8 Conclusions

We have proven convergence theorems for MDL prediction for arbitrary countable classes of semimeasures, the only requirement being that the true distribution μ is a measure. Our results hold for both static and dynamic MDL and provide a statement about convergence speed in mean sum. This also yields both on-sequence and off-sequence assertions. Our results are to our knowledge the strongest available for the discrete case.

Compared to the bound for Solomonoff prediction in Theorem 2.1, the error bounds for MDL are exponentially worse, namely w_μ^{-1} instead of $\ln w_\mu^{-1}$. Our bounds are sharp in general, as Example 4.4 shows. There are even classes of Bernoulli distributions where the exponential bound is sharp [PH04].

In the case of continuously parameterized model classes, finite error bounds do not hold [BC91, BRY98], but the error grows slowly as $\ln t$. Under additional assumptions (i.i.d. for instance) and with a reasonable prior, one can prove similar behavior of MDL and Bayes mixture predictions [Ris96]. In this sense, MDL converges as fast as Bayes mixture, and this is even true for the "slow" Bernoulli example presented in [PH04]. However in Example 4.4, the error grows as t, which shows that the Bayes mixture may be superior to MDL in general.

References

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions 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 Transactions on Information Theory*, 44(6):2743–2760, 1998.

[Cal02] C. S. Calude. *Information and Randomness*. Springer, Berlin, 2nd edition, 2002.

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

[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, December 2001.

[Hut01b] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, June 2001.

[Hut03a] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003.

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

[LCL+03] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. In *Proc. 14th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2003.

[Lev73] L. A. Levin. On the notion of a random sequence. *Soviet Math. Dokl.*, 14(5):1413–1416, 1973.

[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. On the convergence speed of MDL predictions for Bernoulli sequences. submitted to International Conference on Algorithmic Learning Theory (ALT), 2004.

[Ris78] J. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

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

[Ris99] J. J. Rissanen. Hypothesis selection and testing by the MDL principle. *The Computer Journal*, 42(4):260–269, 1999.

[Sch71] C. P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit*, volume 218 of *Lecture Notes in Mathematics*. Springer, Chichester, England, 1971.

[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.

[Wan96] Y. Wang. *Randomness and Complexity.* PhD thesis, Ruprecht-Karls-Universität Heidelberg, 1996.

[WB68] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Jrnl.*, 11(2):185–194, August 1968.

[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.