Asymptotics of Discrete MDL for Online Prediction*

Jan Poland and **Marcus Hutter**
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
{jan,marcus}@idsia.ch http://www.idsia.ch

6 June 2005

Abstract

Minimum Description Length (MDL) is an important principle for induction and prediction, with strong relations to optimal Bayesian learning. This paper deals with learning non-i.i.d. processes by means of two-part MDL, where the underlying model class is countable. We consider the online learning framework, i.e. observations come in one by one, and the predictor is allowed to update his state of mind after each time step. We identify two ways of predicting by MDL for this setup, namely a *static* and a *dynamic* one. (A third variant, hybrid MDL, will turn out inferior.) We will prove that under the only assumption that the data is generated by a distribution contained in the model class, the MDL predictions converge to the true values almost surely. This is accomplished by proving finite bounds on the quadratic, the Hellinger, and the Kullback-Leibler loss of the MDL learner, which are however exponentially worse than for Bayesian prediction. We demonstrate that these bounds are sharp, even for model classes containing only Bernoulli distributions. We show how these bounds imply regret bounds for arbitrary loss functions. Our results apply to a wide range of setups, namely sequence prediction, pattern classification, regression, and universal induction in the sense of Algorithmic Information Theory among others.

Keywords

Minimum Description Length, Sequence Prediction, Consistency, Discrete Model Class, Universal Induction, Stabilization, Algorithmic Information Theory, Loss Bounds, Classification, Regression.

*A shorter version of this paper [PH04a] appeared in COLT 2004.

1 Introduction

"Always prefer the *simplest* explanation for your observation," says Occam's razor. In Learning and Information Theory, simplicity is often quantified in terms of description length, giving immediate rise to the Minimum Description Length (MDL) principle [WB68, Ris78, Grü98]. Thus MDL can be seen as a strategy against overfitting. An alternative way to think of MDL is Bayesian. The explanations for the observations (the *models*) are endowed with a prior. Then the model having maximum a posteriori (MAP) probability is also a two-part MDL estimate, where the correspondence between probabilities and description lengths is simply by a negative logarithm.

How does two-part MDL perform for prediction? Some very accurate answers to this question have been already given. If the data is generated by an independently identically distributed (i.i.d.) process, then the MDL estimates are consistent [BC91]. In this case, an important quantity to consider is the *index of resolvability*, which depends on the complexity of the data generating process. This quantity is a tight bound on the regret in terms of coding (i.e. the excess code length). Moreover, the index of resolvability also bounds the predictive regret, namely the rate of convergence of the predictive distribution to the true one. These results apply to both discrete and continuously parameterized model classes, where in the latter case the MDL estimator must be discretized with an appropriate precision.

Under the relaxed assumption that the data generating process obeys a central limit theorem and some additional conditions, Rissanen [Ris96, BRY98] proves an asymptotic bound on the regret of MDL codes. Here, he also removes the coding redundancy arising if two-part codes are defined in the straightforward way. The resulting bound is very similar to that in [CB90] for Bayes mixture codes and i.i.d. processes, where the i.i.d. assumption may also be relaxed [Hut03b]. Other similar and related results can be found in [GV01, GV04].

In this work, we develop new methods in order to arrive at very general consistency theorems for MDL on *countable model classes*. Our setup is *online sequence prediction*, that is, the symbols $x_1, x_2, \ldots$ of an infinite sequence are revealed successively by the environment, where our task is to predict the next symbol in each time step. Consistency is established by proving *finite cumulative bounds* on the differences of the predictive to the true distribution. Differences will be measured in terms of the relative entropy, the quadratic distance, and the Hellinger distance. Most of our results are based on the only assumption that the data generating process is *contained in the models class*. (The discussion of how strong this assumption is, will be postponed to the last section.) Our results imply regret bounds with *arbitrary* loss functions. Moreover, they can be directly applied to important general setups such as pattern classification, regression, and universal induction.

As many scientific models (e.g. in physics or biology) are smooth, much statistical work is focussed on continuous model classes. On the other hand, the largest relevant classes from a computational point of view are at most countable. In particular,

the field of Algorithmic Information Theory (also known as Kolmogorov Complexity, e.g. [ZL70, LV97, Cal02, Hut04]) studies the class of *all lower-semicomputable semimeasures*. Then there is a one-to-one correspondence of models and programs on a fixed universal Turing Machine. (Since programs need not halt on each input, models are semimeasures instead of measures, see e.g. [LV97] for details). This model class can be considered the largest one which can be in the limit processed under standard computational restrictions. We will develop all our results for semimeasures, so that they can be applied in this context, which we refer to as *universal* sequence prediction.

In the universal setup, the Bayes mixture is also termed Solomonoff-Levin prior and has been intensely studied first by Solomonoff [Sol64, Sol78]. Its predictive properties are excellent [Hut01, Hut04]. Precisely one can bound the cumulative loss by the complexity of the data generating process. This is the reference performance we compare MDL to. It turns out that the predictive properties of MDL can be exponentially worse, even in the case that the model class contains only Bernoulli distributions. Another related quantity in the universal setup is *one-part MDL*, which has been studied in [Hut03c]. We will briefly encounter it in Section 8.4.

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 Sections 6 and 7, we study the stabilization properties of the MDL estimator. Section 8 presents more general loss bounds as well as three important applications: pattern classification, regression, and universal induction. Finally, Section 9 contains the conclusions.

2 Prerequisites and Notation

We build on the notation of [LV97] and [Hut04]. 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*. Intuitively, the quantity $\nu(x)$ can be understood as the probability that a data generating process yields a string starting with x. Then, for a measure, the probabilities of all joint continuations of x add up to $\nu(x)$, while for a semimeasure, there may be a "probability leak" (1). Recall that we are interested in semimeasures (and not only in measures) because of their correspondence to programs on a fixed universal Turing machine in the universal setup and our inability to decide the halting problem.

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 every $\nu_i \in \mathcal{C}$ there is a constant $c_i > 0$ such that $\nu(x) \geq c_i \cdot \nu_i(x)$ holds for all $x \in \mathcal{X}^*$. A dominant semimeasure $\tilde{\nu}$ need not be contained in $\mathcal{C}$.

Each (semi)measure $\nu \in \mathcal{C}$ is associated with a weight $w_\nu > 0$, and we require $\sum_\nu w_\nu \leq 1$. We may interpret the weights as a *prior* on $\mathcal{C}$. Then it is obvious that the Bayes mixture

$$\xi(x) \equiv \xi_{[\mathcal{C}]}(x) := \sum_{\nu \in \mathcal{C}} w_\nu \nu(x) \quad (\text{for } x \in \mathcal{X}^*) \tag{2}$$

dominates $\mathcal{C}$. Assume that there is some measure $\mu \in \mathcal{C}$, the *true distribution*, generating sequences $x_{<\infty} \in \mathcal{X}^\infty$. Typically μ is unknown. (Note that we require μ to be a measure: The data stream always continues, there are no "probability leaks".) 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

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

and, for well-definedness, $\mu(x_t|x_{<t}) = 0$ if $\mu(x_{<t}) = 0$ (this case has probability zero). Note that $\mu(x_t|x_{<t})$ can depend on the complete history $x_{<t}$. 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

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

More general, the expectation may be defined as an integral over infinite sequences. But since we won't need it, we can keep things simple. The following is a central result about prediction with Bayes mixtures in a form independent of Algorithmic Information Theory.

Theorem 1 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, which is a measure, we have*

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

This was found by Solomonoff ([Sol78]) for universal sequence prediction. A proof is also given in [LV97] (only for binary alphabet) or [Hut04] (arbitrary alphabet). It is surprisingly simple once Lemma 6 is known. A few lines analogous to (14) and (15) exploiting the dominance of ξ are sufficient.

One should be aware that the condition $\mu \in \mathcal{C}$ is essential in general, for both Bayes and MDL predictions [GL04]. On the other hand, one can show that if there is an element in $\mathcal{C}$ which is sufficiently close to μ in an appropriate sense, then still good predictive properties hold [Hut03b].

Note that although w_ν can be interpreted as a prior on the model class, we do not assume any probabilistic structure for $\mathcal{C}$ (e.g. a sampling mechanism). The theorem rather states that the cumulative loss is bounded by a quantity depending on the complexity $\ln w_\mu^{-1}$ of the true distribution. The same kind of assertion will be proven for MDL predictions later.

The bound (5) implies that the ξ-predictions converge to the μ-predictions almost surely (i.e. with μ-probability one). This is not hard to see, since with the abbreviation $s_t = \sum_a \big(\mu(a|x_{<t}) - \xi(a|x_{<t})\big)^2$ and for each $\varepsilon > 0$, we have

$$\begin{aligned}\mathbf{P}\Big(\exists t \geq n : s_t \geq \varepsilon\Big) &= \mathbf{P}\Big(\bigcup_{t\geq n}\{s_t \geq \varepsilon\}\Big) \\ &\leq \sum_{t\geq n}\mathbf{P}(s_t \geq \varepsilon) \leq \frac{1}{\varepsilon}\sum_{t=n}^{\infty}\mathbf{E}s_t \stackrel{n\to\infty}{\longrightarrow} 0. \qquad (6)\end{aligned}$$

Actually, (5) yields an even stronger assertion, since it characterizes the *speed of convergence* by the quantity on the right hand side. Precisely, the expected number of times t in which $\xi(a|x_{<t})$ deviates by more than ε from $\mu(a|x_{<t})$ is finite and bounded by $\ln w_\mu^{-1}/\varepsilon^2$, and the probability that the number of ε-deviations exceeds $\frac{\ln w_\mu^{-1}}{\varepsilon^2\delta}$ is smaller than δ. (However, we *cannot* conclude a convergence rate of $s_t = o(\frac{1}{t})$ from (5), since the quadratic differences generally do not decrease monotonically.)

Since we will encounter this type of convergence (5) frequently in the following, we call it *convergence in mean sum (i.m.s)*:

$$\varphi \stackrel{i.m.s.}{\longrightarrow} \mu \quad\Longleftrightarrow\quad \exists\, C > 0 : \sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\mu(a|x_{<t}) - \varphi(a|x_{<t})\Big)^2 < \infty. \qquad (7)$$

Then Theorem 1 states that the ξ predictions converge to the μ predictions i.m.s., or "ξ converges to μ i.m.s." for short. By (6), convergence i.m.s. implies almost sure convergence (with respect to the true distribution μ). Note that in contrast, convergence in the mean, i.e. $\mathbf{E}[\sum_a(\mu(a|x_{<t}) - \varphi(a|x_{<t}))^2] \stackrel{t\to\infty}{\longrightarrow} 0$, only implies convergence in probability.

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. We introduce the abbreviation

$$K\ldots = -\log_2\ldots,\ \text{e.g. } K\nu(x) = -\log_2\nu(x) \qquad (8)$$

for a semimeasure ν and $x \in \mathcal{X}^*$ and $K\xi(x) = -\log_2\xi(x)$ for the Bayes mixture ξ. It is common to ignore the somewhat irrelevant restriction that code lengths must be

integer. In particular, given a class of semimeasures $\mathcal{C}$ together with weights, each weight w_ν corresponds to a description length or complexity

$$Kw(\nu) = -\log_2 w_\nu. \tag{9}$$

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 (maximum a posteriori) estimator, is defined as

$$\nu^x = \nu^x_{[\mathcal{C}]} = \arg\max_{\nu\in\mathcal{C}}\{w_\nu \nu(x)\}. \tag{10}$$

In case of a tie, we need not specify the further choice at this point, just pick any of the maximizing elements. But for concreteness, you may think that ties are broken in favor of larger prior weights. The maximum is always attained in (10) since for each $\varepsilon > 0$ at most a finite number of elements fulfil $w_\nu \nu(x) > \varepsilon$. Observe immediately the correspondence in terms of *description lengths* rather than *probabilities*:

$$\nu^x = \arg\min_{\nu\in\mathcal{C}} \{Kw(\nu) + K\nu(x)\}.$$

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 (8) and (9)). As explained before, we stick to the product notation.

For notational simplicity, let $\nu^*(x) = \nu^x(x)$. 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. Note that the use of the term "estimator" is non-standard,

[1]The term MAP estimator is more precise. 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 (compare the discussion section at the end of this paper). Second, when coding some data x, one can exploit the fact that once the distribution ν^x is specified, only data which leads to this ν^x needs to be considered. This allows for a description shorter than $Kw(\nu^x)$. Nevertheless, the *construction principle* is commonly termed MDL, compare e.g. the "ideal MDL" in [VL00].

since ϱ is a product of the estimator ν^* (this use is standard) and its prior weight. There will be no confusion between these two meanings of "estimator" in the following.

For each $x, y \in \mathcal{X}^*$,

$$\xi(x) \geq \varrho(x) \geq \varrho^y(x) \tag{11}$$

is immediate. If $\mathcal{C}$ contains only measures, we have $\sum_a \varrho(xa) \geq \sum_a \varrho^x(xa) = \varrho^x(x) = \varrho(x)$ for all $x \in \mathcal{X}^*$, so ϱ has some "anti-semimeasure" property. If $\mathcal{C}$ contains semimeasures, no semimeasure or anti-semimeasure property can be established for ϱ.

We can define MDL predictors according to (3). There are basically *two* possible ways to use MDL for prediction.

Definition 2 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 Bayes mixture ξ-predictor.

Definition 3 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. This is usually more efficient in practice.

We will define another MDL predictor, the *hybrid* one, in Section 6. It can be paraphrased as "do dynamic MDL but drop weights". We will see that its predictive performance is weaker.

The range of the static MDL predictor is obviously contained in $[0, 1]$. For the dynamic MDL predictor, this holds by

$$\varrho^x(x) \geq \varrho^{xa}(x) \geq \varrho^{xa}(xa). \tag{12}$$

Static MDL is omnipresent in machine learning and applications, see also Section 8. 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[2] 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

[2]There are however regularization methods which cannot be interpreted in this way but build on a different theoretical foundation, such as structural risk minimization.

MDL is applied more rarely due to its larger computational effort. For example, the similarity metric proposed in [LCL+03] can be interpreted as (a deterministic variant of) dynamic MDL.

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

$$\varphi_{\rm norm}(a|x) \;:=\; \frac{\varphi(a|x)}{\sum_{b\in\mathcal{X}}\varphi(b|x)} \;=\; \frac{\varphi(xa)}{\sum_{b\in\mathcal{X}}\varphi(xb)}$$

is a measure (assume that the denominator is different from zero, which is always true with probability 1 (w.p.1) if φ is an MDL predictor). This procedure is known as *Solomonoff normalization* [Sol78, LV97] and results in

$$\varphi_{\rm norm}(x_{1:n}) = \frac{\varphi(x_{1:n})}{\varphi(\epsilon)}\prod_{t=1}^{n}\frac{\varphi(x_{<t})}{\sum_{a\in\mathcal{X}}\varphi(x_{<t}a)} = \frac{\varphi(x_{1:n})}{\varphi(\epsilon)N_\varphi(x_{<n})},$$

where

$$N_\varphi(x) = \prod_{t=1}^{\ell(x)+1}\frac{\sum_{a\in\mathcal{X}}\varphi(x_{<t}a)}{\varphi(x_{<t})} \tag{13}$$

is the normalizer.

We conclude this section with a simple example.

Bernoulli and i.i.d. classes. Let $n\in\mathbb{N}$, $\mathcal{X}=\{1,\dots,n\}$, 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 $\mathbf{P}(x_t=i)=\vartheta_i$ 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 $\alpha_i=\frac{1}{t}|\{s\leq t : x_s=i\}|$, then it is easy to see that

$$\nu^{x_{1:t}} = \arg\min_{\vartheta\in\Theta}\{Kw(\vartheta)\cdot\ln 2 + t\cdot D(\alpha\|\vartheta)\},$$

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 $\mathcal{X}=\{0,1\}$ in this case.

4 Dynamic MDL

We may now develop convergence results, beginning with the dynamic MDL predictor from Definition 2. The following simple lemma is crucial for all subsequent proofs.

Lemma 4 *For an arbitrary class of (semi)measures $\mathcal{C}$, we have*

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

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. □

The following proposition demonstrates how simple it can be to obtain a convergence result, however a weak one. Various similar results have been already obtained in the past, e.g. in [BD62, Bar85].

Proposition 5 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

$$\frac{\varrho(x_t|x_{<t})}{\mu(x_t|x_{<t})} \to 1 \quad w.\mu.p.1$$

Proof. Since $\xi - \varrho$ is a positive semimeasure by Lemma 4, $\frac{\xi-\varrho}{\mu}$ is a positive super-martingale. By Doob's martingale convergence theorem (see e.g. [Doo53] or [CT88] or any textbook on advanced probability theory), it therefore converges on a set of μ-measure one. Moreover, $\frac{\xi}{\mu}$ converges on a set of measure one, being a positive super-martingale as well [LV97, Thm.5.2.2]. Thus $\frac{\varrho}{\mu}$ must converge on a set of measure one. We denote this limit by f and observe that $f \geq w_\mu$ since $\frac{\varrho}{\mu} \geq w_\mu$ everywhere. On this set of measure one, the denominator $\varrho(x_{<t})/\mu(x_{<t})$ of

$$\frac{\varrho(x_{1:t})/\mu(x_{1:t})}{\varrho(x_{<t})/\mu(x_{<t})} = \frac{\varrho(x_t|x_{<t})}{\mu(x_t|x_{<t})}$$

converges to $f > 0$, and so does the numerator. The whole fraction thus converges to one, which was to be shown. □

Proposition 5 gives only a statement about "on-sequence" ($\varrho(x_t|x_{<t})$) convergence of the ϱ-predictions. Indeed, no conclusion about "off-sequence" convergence, i.e. $\varrho(a|x_{<t})$ for arbitrary $a \in \mathcal{X}$, can be drawn from the proposition, not even in the

deterministic case. There, the true measure μ is concentrated on the particular sequence $x_{<\infty}$. So for $a \neq x_t$, we have $\mu(x_{<t}a) = 0$, and thus no assertion for $\varrho(a|x_{<t})$ can be made. On the other hand, an off-sequence result is essential for prediction: Even if on the *correct* next symbol the predictive probability is very close to the true value, we must be sure that this is so also for all *alternatives*. This is particularly important if we base some decision on the prediction; compare Section 8.1.

The following theorem closes this gap. In addition, it provides a statement about the speed of convergence. In order to prove it, we need a lemma establishing a relation between the square distance and the Kullback-Leibler distance, which is proven for instance in [Hut04, Sec.3.9.2].

Lemma 6 *Let μ and ρ be measures on $\mathcal{X}$, then*

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

Theorem 7 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ (which is a measure), we have*

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

That is, $\varrho_{\text{norm}}(a|x_{<t}) \stackrel{i.m.s.}{\longrightarrow} \mu(a|x_{<t})$ (see (7)), which implies $\varrho_{\text{norm}}(a|x_{<t}) \to \mu(a|x_{<t})$ with μ-probability one.

Proof. Let $n \in \mathbb{N}$. From Lemma 6, we know

$$\begin{aligned}\sum_{t=1}^{n} \mathbf{E} \sum_{a\in\mathcal{X}} \big(\mu(a|x_{<t}) - \varrho_{\text{norm}}(a|x_{<t})\big)^2 &\leq \sum_{t=1}^{n} \mathbf{E} \sum_{a\in\mathcal{X}} \mu(a|x_{<t}) \ln \frac{\mu(a|x_{<t})}{\varrho_{\text{norm}}(a|x_{<t})} \\ = \sum_{t=1}^{n} \mathbf{E} \ln \frac{\mu(x_t|x_{<t})}{\varrho_{\text{norm}}(x_t|x_{<t})} &= \sum_{t=1}^{n} \mathbf{E} \left[\ln \frac{\mu(x_t|x_{<t})}{\varrho(x_t|x_{<t})} + \ln \frac{\sum_{a\in\mathcal{X}} \varrho(x_{<t}a)}{\varrho(x_{<t})}\right]. \quad (14)\end{aligned}$$

Then we can estimate

$$\sum_{t=1}^{n} \mathbf{E} \ln \frac{\mu(x_t|x_{<t})}{\varrho(x_t|x_{<t})} \;=\; \mathbf{E} \ln \prod_{t=1}^{n} \frac{\mu(x_t|x_{<t})}{\varrho(x_t|x_{<t})} \;=\; \mathbf{E} \ln \frac{\mu(x_{1:n})}{\varrho(x_{1:n})} \;\leq\; \ln w_\mu^{-1}, \quad (15)$$

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

$$\mathbf{E} \ln \frac{\sum_a \varrho(x_{<t}a)}{\varrho(x_{<t})} \leq \mathbf{E} \left[\frac{\sum_a \varrho(xa)}{\varrho(x)} - 1\right] = \sum_{\ell(x)=t-1} \frac{\mu(x)\Big[\big(\sum_a \varrho(xa)\big) - \varrho(x)\Big]}{\varrho(x)}$$

$$\leq \sum_{\ell(x)=t-1} \frac{\mu(x)\Big[\big(\sum_{a\in\mathcal{X}}\varrho(xa)\big)-\varrho(x)+\max\big\{0,\varrho(x)-\sum_{a\in\mathcal{X}}\varrho(xa)\big\}\Big]}{\varrho(x)}$$

$$\leq w_\mu^{-1}\sum_{\ell(x)=t-1}\left[\Big(\sum_{a\in\mathcal{X}}\varrho(xa)\Big)-\varrho(x)+\max\Big\{0,\varrho(x)-\sum_{a\in\mathcal{X}}\varrho(xa)\Big\}\right]. \quad (16)$$

If $\mathcal{C}$ contains only measures, the max-term is not necessary, since ϱ is an "anti-semimeasure" in this case. We proceed by observing

$$\sum_{t=1}^{n}\sum_{\ell(x)=t-1}\Big[\Big(\sum_{a\in\mathcal{X}}\varrho(xa)\Big)-\varrho(x)\Big]=\sum_{t=1}^{n}\Big[\sum_{\ell(x)=t}\varrho(x)-\sum_{\ell(x)=t-1}\varrho(x)\Big]=\Big[\sum_{\ell(x)=n}\varrho(x)\Big]-\varrho(\epsilon) \quad (17)$$

which is true since for successive t the positive and negative terms cancel. From Lemma 4 we know $\varrho(x)-\sum_{a\in\mathcal{X}}\varrho(xa) \leq \xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)$ and therefore

$$\begin{aligned}\sum_{t=1}^{n}\sum_{\ell(x)=t-1}\max\Big\{0,\varrho(x)-\sum_{a\in\mathcal{X}}\varrho(xa)\Big\} &\leq \sum_{t=1}^{n}\sum_{\ell(x)=t-1}\max\Big\{0,\xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)\Big\}\\ =\sum_{t=1}^{n}\sum_{\ell(x)=t-1}\Big[\xi(x)-\sum_{a\in\mathcal{X}}\xi(xa)\Big] &= \xi(\epsilon)-\sum_{\ell(x)=n}\xi(x). \quad (18)\end{aligned}$$

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 (16), (17) and (18), and observing $\varrho\leq\xi\leq 1$, we obtain

$$\sum_{t=1}^{n}\mathbf{E}\ln\frac{\sum_a\varrho(x_{<t}a)}{\varrho(x_{<t})}\leq w_\mu^{-1}\left[\xi(\epsilon)-\varrho(\epsilon)+\sum_{\ell(x)=n}(\varrho(x)-\xi(x))\right]\leq w_\mu^{-1}\xi(\epsilon)\leq w_\mu^{-1}. \quad (19)$$

Therefore, (14), (15) and (19) finally prove the assertion. □

We point out again that the proof gets a bit simpler if $\mathcal{C}$ contains only measures, since then (18) becomes irrelevant. However, this case doesn't give a tighter bound.

This is the first convergence result "in mean sum", see (7). 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 $w_\mu^{-1}=2^{Kw(\mu)}$. The following example will show that this bound is sharp (save for a constant factor). Observe that in the corresponding result for mixtures, Theorem 1, the bound is much smaller, namely $\ln w_\mu^{-1}=Kw(\mu)\ln 2$.

Example 8 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 $z_{<\infty}^{(i)}=1^{i-1}0^\infty$, 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 $\varrho_{\rm norm}(0|x_{<t}) = \varrho_{\rm norm}(1|x_{<t}) = \frac{1}{2}$. Hence, $\sum_t \mathbf{E} \sum_a (\mu(a|x_{<t}) - \varrho_{\rm norm}(a|x_{<t}))^2 = \frac{1}{2}(N-1) \stackrel{\times}{=} w_\mu^{-1}$. In Example 15 we will even see a case where the model class contains only Bernoulli distributions and nevertheless the exponential bound is sharp.

The next result implies that convergence holds also for the un-normalized dynamic MDL predictor.

Theorem 9 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

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

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

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

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

Here, $|u| = u^+ + (-u)^+ = -u + 2u^+$, $\ln u \leq u-1$, and $\varrho \geq w_\mu \mu$ have been used, the latter implies also $\sum_a \varrho(xa) \geq w_\mu \sum_a \mu(xa) = w_\mu \mu(x)$. 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 (16) - (19) as in the proof of Theorem 7.

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

$$\begin{aligned}
\sum_a \Big| \varrho_{\rm norm}(a|x) - \varrho(a|x) \Big| &= \sum_a \frac{\varrho(a|x)}{\sum_b \varrho(b|x)} \Big| 1 - \sum_b \varrho(b|x) \Big| = \Big| 1 - \sum_b \varrho(b|x) \Big| \qquad (20) \\
&= \frac{(\sum_a \varrho(xa) - \varrho(x))^+}{\varrho(x)} + \frac{(\varrho(x) - \sum_a \varrho(xa))^+}{\varrho(x)}.
\end{aligned}$$

Then take the expectation $\mathbf{E}$ and the sum $\sum_{t=1}^{\infty}$ and proceed as in (i). $\square$

Corollary 10 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

$$\sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\mu(a|x_{<t})-\varrho(a|x_{<t})\Big)^2 \;\leq\; 8w_\mu^{-1}.$$

That is, $\varrho(a|x_{<t})\stackrel{i.m.s.}{\longrightarrow}\mu(a|x_{<t})$ (see (7)).

Proof. For two functions $\varphi_1,\varphi_2:\mathcal{X}^*\to[0,1]$, let

$$\Delta(\varphi_1,\varphi_2)=\left(\sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\varphi_1(a|x_{<t})-\varphi_2(a|x_{<t})\Big)^2\right)^{\frac{1}{2}}. \tag{21}$$

Then the triangle inequality holds for $\Delta(\cdot,\cdot)$, since Δ is (proportional to) an Euclidian distance (2-norm). Moreover, $\Delta(\mu,\varrho_{\rm norm})\leq\sqrt{2w_\mu^{-1}}$ by Theorem 7 and $\ln w_\mu^{-1}\leq w_\mu^{-1}-1\leq w_\mu^{-1}$. We also have $\Delta(\varrho_{\rm norm},\varrho)\leq\sqrt{2w_\mu^{-1}}$ by multiplying $|\varrho_{\rm norm}-\varrho|$ in Theorem 9(ii) with another $|\varrho_{\rm norm}-\varrho|$. Note $|\varrho_{\rm norm}-\varrho|\leq 1$, since both $\varrho(a|x),\varrho_{\rm norm}(a|x)\in[0,1]$, for ϱ this holds by (12). This implies $\Delta(\mu,\varrho)\leq\Delta(\mu,\varrho_{\rm norm})+\Delta(\varrho_{\rm norm},\varrho)\leq 2\sqrt{2w_\mu^{-1}}$. □

Corollary 11 *For almost all $x_{<\infty}\in\mathcal{X}^\infty$, the normalizer N_ϱ defined in (13) converges to a number which is finite and greater than zero, i.e. $0<N_\varrho(x_{<\infty})<\infty$. Moreover, the sum of the MDL posterior estimates converges to one almost surely,*

$$\sum_{a\in\mathcal{X}}\varrho(a|x_{<t})=\frac{\sum_{a\in\mathcal{X}}\varrho(x_{<t}a)}{\varrho(x_{<t})}\to 1\quad\text{as}\quad t\to\infty\quad w.\mu.p.1. \tag{22}$$

Proof. Theorem 9 implies that with probability one, the sum $\sum_1^n|\ln\frac{\sum_a\varrho(x_{<t}a)}{\varrho(x_{<t})}|$ is bounded in n, hence converges absolutely, hence also the limit

$$\ln N_\varrho(x_{<\infty})\;=\;\sum_{t=1}^{\infty}\ln\frac{\sum_{a\in\mathcal{X}}\varrho(x_{<t}a)}{\varrho(x_{<t})}$$

exists and is finite. For these sequences, $0<N_\varrho(x_{<\infty})<\infty$ and (22) follows. □

5 Static MDL

Static MDL as introduced in Definition 3 is usually more efficient and thus preferred in practice, since only one MDL estimator has to be computed. The following technical result will allow to conclude that the static MDL predictions converge in mean sum like the dynamic ones.

Theorem 12 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

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

Proof. We proceed in a similar way as in the proof of Theorem 7, (16) - (18). From Lemma 4, we know $\varrho(x)-\sum_a\varrho^x(xa) \;\leq\; \xi(x)-\sum_a\xi(xa)$. Then

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

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. $\Box$

Corollary 13 *For any class of (semi)measures $\mathcal{C}$ containing the true distribution μ, we have*

$$\begin{aligned}
\sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\mu(a|x_{<t})-\varrho^{x_{<t}}(a|x_{<t})\Big)^2 &\leq 21w_\mu^{-1} \quad \textit{and}\\
\sum_{t=1}^{\infty}\mathbf{E}\sum_{a\in\mathcal{X}}\Big(\mu(a|x_{<t})-\varrho^{x_{<t}}_{\rm norm}(a|x_{<t})\Big)^2 &\leq 32w_\mu^{-1}.
\end{aligned}$$

That is, $\varrho^{x_{<t}}(a|x_{<t})\stackrel{i.m.s.}{\longrightarrow}\mu(a|x_{<t})$ and $\varrho^{x_{<t}}_{\rm norm}(a|x_{<t})\stackrel{i.m.s.}{\longrightarrow}\mu(a|x_{<t})$.

Proof. Using $\varrho(xa)\geq\varrho^x(xa)$ and the triangle inequality, we see

$$\sum_a\Big|\varrho(a|x)-\varrho^x(a|x)\Big|=\Big|\sum_a\varrho(a|x)-\sum_a\varrho^x(a|x)\Big|\leq\Big|\sum_a\varrho(a|x)-1\Big|+\Big|1-\sum_a\varrho^x(a|x)\Big|$$

With $\Delta(\cdot,\cdot)$ as in (21), using $|\varrho - \varrho^x| \leq 1$ we therefore have

$$\Delta^2(\varrho, \varrho^{\text{static}}) \leq \sum_{t=1}^{\infty} \mathbf{E} \sum_a \Big| \varrho(a|x) - \varrho^x(a|x) \Big| \leq 3 w_\mu^{-1}$$

according to Theorem 9 (ii) and Theorem 12. Since $\Delta(\mu, \varrho) \leq 2\sqrt{2 w_\mu^{-1}}$ holds by Corollary 10, we obtain $\Delta(\mu, \varrho^{\text{static}}) \leq \Delta(\mu, \varrho) + \Delta(\varrho, \varrho^{\text{static}}) \leq \sqrt{21 w_\mu^{-1}}$. Theorem 12 also asserts $\Delta(\varrho^{\text{static}}, \varrho^{\text{static}}_{\text{norm}}) \leq \sqrt{w_\mu^{-1}}$, hence $\Delta(\mu, \varrho^{\text{static}}_{\text{norm}}) \leq \sqrt{32 w_\mu^{-1}}$ follows. $\Box$

Distance measures. The total expected square error is not the only possible choice for measuring distance of distributions and speed of convergence. In fact, looking at the proof of Theorem 7, the expected Kullback-Leibler distance may seem more natural at a first glance. However this quantity behaves well only under dynamic MDL, not static MDL. To see this, let $\mathcal{C} \cong \{0, \frac{1}{2}\}$ contain two Bernoulli distributions, both with prior weight $\frac{1}{2}$, and let $\mu \cong \frac{1}{2}$ be the uniform measure. If the first symbol happens to be 0, which occurs with probability $\frac{1}{2}$, then the static MDL estimate is $\nu^0 \cong 0$. Then $D(\mu\|\nu^0) = \infty$, hence the expectation is ∞, too. The quadratic distance behaves locally like the Kullback-Leibler distance (Lemma 6), but otherwise is bounded and thus more convenient.

Another possible choice is the *Hellinger distance*

$$h_t(\mu, \varphi)|_{x_{<t}} = \sum_{a \in \mathcal{X}} \Big(\sqrt{\mu(a|x_{<t})} - \sqrt{\varphi(a|x_{<t})} \Big)^2 \text{ and} \tag{23}$$

$$H_{1:n}(\mu, \varphi) = \sum_{t=1}^{n} \mathbf{E} h_t(\mu, \varphi). \tag{24}$$

Like the square distance, the Hellinger distance is bounded by both the relative entropy and the absolute distance:

$$h_t(\mu, \varphi) \leq \sum_{a \in \mathcal{X}} \mu(a|x_{<t}) \ln \frac{\mu(a|x_{<t})}{\varphi(a|x_{<t})} \text{ and} \tag{25}$$

$$h_t(\mu, \varphi) \leq \sum_{a \in \mathcal{X}} \Big| \mu(a|x_{<t}) - \varphi(a|x_{<t}) \Big|. \tag{26}$$

The former is e.g. shown in [Hut04, Lem.3.11, p.114], the latter follows from $(\sqrt{u} - \sqrt{v})^2 \leq |u - v|$ for any $u, v \in \mathbb{R}$. Therefore, the same bounds we have proven for the square distance also hold for the Hellinger distance; they are subsumed in Corollary 14 below. Although for simplicity of notation we have preferred the square distance over the Hellinger distance in the presentation so far, in Sections 8.1 and 8.3 we will meet situations where the quadratic distance is not sufficient. Then the Hellinger distance will be useful.

The following corollary recapitulates our results and states convergence i.m.s (and therefore also w.μ-p.1) for all combinations of un-normalized/normalized and dynamic/static MDL predictions.

Corollary 14 *Let $\mathcal{C}$ contain the true distribution μ, then*

$$\begin{aligned} S_{<\infty}(\mu,\varrho_{\text{norm}}) &\leq 2w_\mu^{-1}, & H_{<\infty}(\mu,\varrho_{\text{norm}}) &\leq 2w_\mu^{-1}, \\ S_{<\infty}(\mu,\varrho) &\leq 8w_\mu^{-1}, & H_{<\infty}(\mu,\varrho) &\leq 8w_\mu^{-1}, \\ S_{<\infty}(\mu,\varrho^{\text{static}}) &\leq 21w_\mu^{-1}, & H_{<\infty}(\mu,\varrho^{\text{static}}) &\leq 21w_\mu^{-1}, \\ S_{<\infty}(\mu,\varrho^{\text{static}}_{\text{norm}}) &\leq 32w_\mu^{-1}, & H_{<\infty}(\mu,\varrho^{\text{static}}_{\text{norm}}) &\leq 32w_\mu^{-1}, \end{aligned}$$

where $S_{<\infty}(\mu,\varphi)=\sum_t \mathbf{E}\sum_a\left(\mu(a|x_{<t})-\varphi(a|x_{<t})\right)^2$ and $H_{<\infty}$ is as in (24).

The following example shows that the exponential bound is sharp (except for a multiplicative constant), even if the model class contains only Bernoulli distributions. It is stated in terms of static MDL, however it equally holds for dynamic MDL.

Example 15 Let $N \geq 1$ and $\mathcal{C} \cong \Theta = \{\frac{1}{2}\} \cup \{\frac{1}{2}+2^{-k-1} : 1 \leq k \leq N\}$ be a Bernoulli class as discussed at the end of Section 3. Let μ be Bernoulli with parameter $\frac{1}{2}$, i.e. the distribution generating fair coin flips. Assume that all weights are equally $\frac{1}{N+1}$. Then it is shown in [PH04b, Prop. 5] that

$$\sum_{t=1}^{\infty} \mathbf{E}(\tfrac{1}{2} - \varrho^{x_{<t}}(1|x_{<t}))^2 \geq \tfrac{1}{84}(N-4).$$

So the bound equals w_μ^{-1} within a multiplicative constant.

This shows that in general there is no hope to improve the bounds, even for very simple model classes. But the situation is not as bad as it might seem. First, the bounds may be exponentially smaller under certain regularity conditions on the class and the weights, as [Ris96] and the positive assertions in [PH04b] show. It is open to define such conditions for more general model classes. Second, the example just given behaves differently than Example 8. There, the error remains at a significant level for $O(w_\mu^{-1})$ time steps, which must be regarded critical. Here in contrast, the error drops to zero as $\frac{1}{n}$ for a very long time, namely $O(2^{w_\mu^{-1}})$ steps, and decreases more rapidly only afterwards. This behavior is tolerable in practice. Recently, [Li99, Zha04] have proven that this favorable case always occurs for i.i.d., if the weights satisfy the *light tails* condition $\sum w_\nu^\alpha \leq 1$ for some $\alpha < 1$ [BC91]. Precisely, they give a rapidly decaying bound on the instantaneous error. It is open if similar results also hold in more general setups than i.i.d. Example 8 shows that at least some additional assumption is necessary.

6 Hybrid MDL

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, as we will see with the hybrid MDL predictor in the subsequent example.

Definition 16 The *hybrid* MDL predictor is given by

$$\varrho^{\mathrm{hyb}}(a|x) = \frac{\nu^*(xa)}{\nu^*(x)}$$

(compare (10)). This can be paraphrased as "do dynamic MDL and drop the weights". It is somewhat in-between static and dynamic MDL.

Example 17 Let $\mathcal{X} = \{0,1\}$ 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 predict probabilities of constantly

$$\tfrac{1}{2} = \lambda(a|x_{<t}) = \nu(a|x_{<t}) = \frac{w_\lambda\lambda(x_{<t}a)}{w_\nu\nu(x_{<t})} = \frac{w_\nu\nu(x_{<t}a)}{w_\lambda\lambda(x_{<t})}$$

for all $a \in \{0,1\}$. 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. in favor of larger weights, then the hybrid MDL predictor does converge for this example. We replace (10) by this rule:

$$\nu^x = \arg\max\{w_\nu : \nu \in \{\nu = \arg\max_{\nu\in\mathcal{C}} w_\nu\nu(x)\}\}.$$

Then, do the hybrid MDL predictions always converge? This is equivalent to asking if the process of selecting a maximizing element eventually *stabilizes*. If stabilization does not occur, 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. We show that this can indeed happen, even for different reasons.

Example 18 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 $\frac{\nu(111\ldots)}{\mu(111\ldots)}$ converges and oscillates. In fact, each sequence having positive measure under μ and ν contains eventually only ones, and the quotient oscillates.

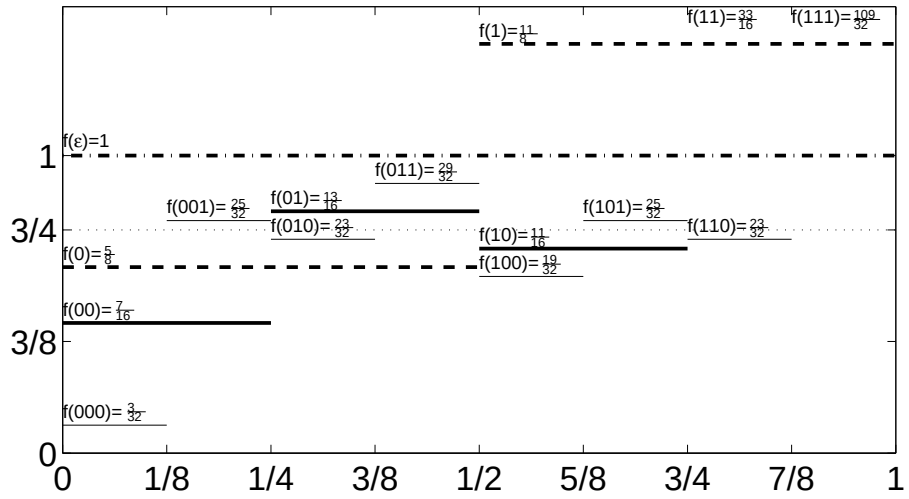

Figure 1: Construction of a martingale that with high probability converges to $\frac{3}{4}$ oscillating infinitely often.

Example 19 This example is a little more complex. We assume the uniform distribution λ to be the true distribution. We now construct a positive martingale $f(\cdot)$ that converges to $\frac{3}{4}$ with high probability and thereby oscillates infinitely often.

The martingale is defined on strings x of successively increasing length. Of course, $f(\epsilon) := 1$. If $f(x)$ is already defined for strings of length $n-1$, we extend the definition on strings of length n in the following way: If $f(x) > \frac{3}{4}$, we set

$$\begin{aligned} f(x0) &:= \frac{3}{4} - 2^{-n-2} \quad \text{and} \\ f(x1) &:= 2f(x) - \Big(\frac{3}{4} - 2^{-n-2}\Big). \end{aligned}$$

This guarantees the martingale property $f(x) = \frac{1}{2}(f(x0) + f(x1))$. If $f(x) \leq \frac{3}{4}$ and $f(x) \geq \frac{3}{8} + 2^{-n-3}$, then we can similarly define

$$\begin{aligned} f(x0) &:= 2f(x) - \Big(\frac{3}{4} + 2^{-n-2}\Big) \quad \text{and} \\ f(x1) &:= \frac{3}{4} + 2^{-n-2}. \end{aligned}$$

However, if $f(x) < \frac{3}{8} + 2^{-n-3}$, we cannot proceed in this way, since f must be positive. Therefore, we set $f(x0) := f(x1) := f(x)$ in this case and call those x "dead" strings. Strings that are not dead will be called "alive". A few steps of the construction are shown in Figure 1. For example, it can be observed that the string 000 is dead, all other strings in the figure are alive.

It is obvious from the construction that $f(x_{1:t})$ is a martingale, it oscillates and converges to $\frac{3}{4}$ as $t \to \infty$ for all sequences $x_{<\infty}$ that always stay alive. The only thing we must show is that many sequences in fact stay alive.

Claim 20 *We have* $\lambda(\{x_{<\infty} : \exists t \text{ such that } x_{1:t} \text{ is dead}\}) \leq \frac{1}{4}$.

Proof. After the nth step, i.e. when f has been defined for strings of length n, $f(x)$ assumes the value

$$a_0^n = \frac{3}{4} - 2^{-n-2}$$

on a set of measure at most $\frac{1}{2}$. In the next step $n+1$, f is defined to

$$a_1^n = \frac{3}{4} - 2^{-n-1}\Big(1 + \frac{1}{4}\Big)$$

on half of the extended strings. Generally, in the kth next step, f is defined to

$$a_k^n = \frac{3}{4} - 2^{-n+k-2}\Big(\sum_{j=0}^{k} 2^{-2j}\Big)$$

on a 2^{-k} fraction of the extended strings.

The extended strings stay alive as long as $a_k^n \geq \frac{3}{8} + 2^{-n-k-3}$ holds. Some elementary calculations show that this is equivalent to $k \leq n$. So precisely after $n+1$ additional steps, a fraction of 2^{-n-1} of the extended strings die.

We already noted that for $A_n = \{x : \ell(x) = n \wedge f(x) = a_0^n\}$, we have $\lambda(A_n) \leq \frac{1}{2}$. Thus,

$$\lambda(\{x_{<\infty} : x_{1:n} \in A_n \text{ and } x_{1:2n+1} \text{ is dead}\}) \leq 2^{-n-2}.$$

Hence, one can conclude

$$\lambda(\{x_{<\infty} : \exists t \text{ such that } x_{1:t} \text{ is dead}\}) \leq \sum_{n=1}^{\infty} 2^{-n-2} = \frac{1}{4},$$

which proves the claim. □

We now define the measure ν by

$$\nu(x) = f(x) \cdot \lambda(x) = f(x) \cdot 2^{-\ell(x)},$$

and set the weights to $w_\lambda = \frac{3}{7}$ and $w_\nu = \frac{4}{7}$. Then this provides an example where the maximizing element never stops oscillating with probability at least $\frac{3}{4}$.

Both examples point out different possible reasons for failure of stabilizing. Example 18 works since the measure μ and ν are asymptotically very similar and close to deterministic. In contrast, in Example 19 stabilizing fails because of lack of independence: The quantity $\nu(a|x)$ strongly depends on x. In particular, one can note that even Markovian dependence may spoil the stabilization, since $\nu(a|x)$ only depends on the last symbol of x.

7 Stabilization

In the light of the previous section, it is therefore natural to ask when the maximizing element stabilizes (almost surely). Barron [Bar85, BRY98] has shown that this happens if all distributions in $\mathcal{C}$ are *asymptotically mutually singular*. Under this condition, the true distribution is even eventually identified almost surely.[3]

The condition of asymptotic mutual singularity holds in many important cases, e.g. if the distributions are i.i.d. However, one cannot always build on it.[4] Therefore, in this section we give a different approach: In order to prevent stabilization, it is necessary that the ratio of two predictive distributions oscillates around the inverse ratio of the respective weights. Therefore, stabilization must occur almost surely if the ratio of two predictive distributions converges almost surely but is not *concentrated* in the limit. This is satisfied under appropriate conditions, as we will prove. We start with a general theorem which allows to conclude almost sure stabilization in a countable model class, if for any *pair* of models we have almost sure stabilization.

Theorem 21 *Let $\mathcal{C}$ be a countable class of (semi)measures containing the true measure μ. Assume that for each two $\nu_1, \nu_2 \in \mathcal{C}$ the maximizing element chosen from $\{\nu_1, \nu_2\}$ eventually stabilizes almost surely. Then also the maximizing element chosen from all of $\mathcal{C}$ stabilizes almost surely.*

Proof. It is immediate that the maximizing element chosen from any finite subset of $\mathcal{C}$ stabilizes almost surely. Now, for all $\nu \in \mathcal{C}$ and $c > 0$, define the set A_c^ν by

$$A_c^\nu = \left\{ x_{<\infty} : \exists\, t \geq 1 \ \text{ such that } \ \frac{\nu(x_{1:t})}{\mu(x_{1:t})} \geq c \right\}.$$

Then we have

$$\begin{aligned}
\mu(A_c^\nu) &= \mu\left(\bigcup \left\{ \Gamma_x : \frac{\nu(x)}{\mu(x)} \geq c \wedge \frac{\nu(x_{1:s})}{\mu(x_{1:s})} < c \ \forall\ s < \ell(x) \right\} \right) \\
&= \sum \left\{ \mu(x) : \frac{\nu(x)}{\mu(x)} \geq c \wedge \frac{\nu(x_{1:s})}{\mu(x_{1:s})} < c \ \forall\ s < \ell(x) \right\} \\
&\leq \sum \left\{ \frac{\nu(x)}{c} : \frac{\nu(x)}{\mu(x)} \geq c \wedge \frac{\nu(x_{1:s})}{\mu(x_{1:s})} < c \ \forall\ s < \ell(x) \right\} \\
&= \frac{1}{c} \sum \{ \nu(x) : \dots \} \ \leq \ \frac{1}{c},
\end{aligned}$$

[3] In general, stabilization does not imply that the true distribution is identified. Consider for instance a model class containing two measures: the true measure is concentrated on 0^∞ and has prior weight $\frac{1}{8}$, the other one assigns probability $\nu(x_t = 1) = 2^{-t}$ independently of the past $x_{<t}$. Then the maximizing element will remain the incorrect distribution ν, however with predictions rapidly converging to the truth.

[4] Here is a simple example: let the true measure be Bernoulli($\frac{1}{2}$) and another measure be a product of Bernoullis with parameter rapidly converging to $\frac{1}{2}$. These distributions are not asymptotically mutually singular, nevertheless a.s. stabilization holds, as we will see.

since ν is a (semi)measure and the set $\left\{x \in \mathcal{X}^* : \frac{\nu(x)}{\mu(x)} \geq c \wedge \frac{\nu(x_{1:s})}{\mu(x_{1:s})} < c \;\forall\; s < \ell(x)\right\}$ is prefix-free. Let

$$B^\nu = \left\{x_{<\infty} : \exists\, t \geq 1 \;\text{ such that }\; \frac{w_\nu \nu(x_{1:t})}{w_\mu \mu(x_{1:t})} \geq 1\right\} = A^\nu_{(w_\mu/w_\nu)},$$

then $\mu(B^\nu) \leq \frac{w_\nu}{w_\mu}$ holds. We arrange the (semi)measures $\nu \in \mathcal{C}$ in an order $\nu_1, \nu_2, \ldots$ such that the weights $w_{\nu_1}, w_{\nu_2}, \ldots$ are descending. For each $c \geq 1$, we can now find an index k and a set

$$\mathcal{N}_c = \{\nu_i : i \geq k\} \;\text{ such that }\; \sum_{\nu \in \mathcal{N}_c} w_\nu \leq \frac{w_\mu}{c}.$$

Defining $B_c = \bigcup_{\nu \in \mathcal{N}^c} B^\nu$, we get

$$\mu(B_c) \leq \sum_{\nu \in \mathcal{N}^c} \frac{w_\nu}{w_\mu} \leq \frac{1}{c}.$$

For all $x_{<\infty} \notin B^\nu$, ν can never be the maximizing element. Therefore, for all $x_{<\infty} \notin B_c$, there are only finitely many $\nu \notin \mathcal{N}_c$ having the chance of becoming the maximizing element at any time. By assumption, the maximizing element chosen from the finite set $\mathcal{C} \setminus \mathcal{N}_c$ stabilizes a.s. Thus, we conclude almost sure stabilization on the sequences in $\mathcal{X}^\infty \setminus B_c$. Since this holds for all B_c and $\mu(\mathcal{X}^\infty \setminus B_c) \to 1$ as $c \to \infty$, the maximizing element stabilizes with μ-probability one. $\Box$

For the rest of this section, we assume that the model class $\mathcal{C}$ contains only proper measures. A measure μ is called *factorizable* if there are measures μ_i on $\mathcal{X}$ such that

$$\mu(x) = \prod_{i=1}^{\ell(x)} \mu_i(x_i)$$

for all $x \in \mathcal{X}^*$. That is, the symbols of sequences $x_{<\infty}$ generated by μ are independent. A factorizable 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

$$\mu_i(a) > 0 \Rightarrow \mu_i(a) \geq \delta \;\text{ for all }\; a \in \mathcal{X} \;\text{ and }\; i \geq 1. \tag{27}$$

In particular, all deterministic measures and all i.i.d. distributions 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 $\pi = 3.1415\ldots$

Lemma 22 *Let μ, ν, and $\tilde{\nu}$ be factorizable and uniformly stochastic measures, where μ is the true distribution.*
(i) The maximizing element chosen from μ and ν stabilizes almost surely.
(ii) If μ is not eventually always preferred over ν or $\tilde{\nu}$ (in which case we the maximizing element stabilizes a.s. by (i)), then the maximizing element chosen from ν and $\tilde{\nu}$ stabilizes almost surely.

Proof. We will show only (ii), as the proof of (i) is similar but simpler. So we assume that both ν and $\tilde{\nu}$ remain competitive in the process of choosing the maximizing element, and show that then maximizing element chosen from ν and $\tilde{\nu}$ stabilizes almost surely.

Let $\nu = \prod_i \nu_i$, $\tilde{\nu} = \prod_i \tilde{\nu}_i$, and $X_i = \frac{\tilde{\nu}_i(x_i)}{\nu_i(x_i)}$. The X_i are independent random variables depending on the event $x_{<\infty}$. Moreover, both fractions $\frac{\nu(x_{1:t})}{\mu(x_{1:t})}$ and $\frac{\tilde{\nu}(x_{1:t})}{\mu(x_{1:t})}$ are martingales (with respect to μ) and thus converge almost surely for $t \to \infty$. We are interested only in the events in

$$A_\nu = \left\{ x_{<\infty} \in \mathcal{X}^\infty : \tfrac{\nu(x_{1:t})}{\mu(x_{1:t})} \text{ converges to a value } > 0 \right\},$$

since otherwise ν eventually is no longer competitive. So we assume that $\mu(A_\nu) > 0$, which implies $\mu(A_\nu) = 1$ by the Kolmogorov zero-one-law (see e.g. [CT88]). Similarly, $\mu(A_{\tilde{\nu}}) = 1$ for the analogously defined set $A_{\tilde{\nu}}$. That is,

$$\prod_{i=1}^{t} X_i = \frac{\tilde{\nu}(x_{1:t})}{\nu(x_{1:t})} = \tfrac{\tilde{\nu}(x_{1:t})}{\mu(x_{1:t})} \Big/ \tfrac{\nu(x_{1:t})}{\mu(x_{1:t})}$$

converges to a value > 0 almost surely, and in particular $0 < X_i < \infty$ a.s.

Now we will use the *concentration function* of a real valued random variable U,

$$Q(U, \eta) = \sup_{u \in \mathbb{R}} \mu(u \le U \le u + \eta), \ \eta \ge 0. \tag{28}$$

This quantity was introduced by Lèvy, see e.g. [Pet95]. The concentration function is non-decreasing in η. Moreover, when two independent random variables U and V are added, we have [Pet95, Lemma 1.11]

$$Q(U + V, \eta) \le \min \{Q(U, \eta), Q(V, \eta)\} \ \forall \ \eta \ge 0. \tag{29}$$

We first assume that the following set is unbounded:

$$B = \left\{ \sum_{i=1}^{n} \Big(1 - Q(X_i, \eta) \Big) : n \in \mathbb{N}, \eta > 0 \right\} \subset \mathbb{R}^+, \qquad \text{that is} \tag{30}$$

$$\sup(B) = +\infty, \tag{31}$$

We show that then $\frac{\tilde{\nu}(x_{1:t})}{\nu(x_{1:t})}$ (which converges a.s.) is not concentrated in the limit. That is, it converges to some given $c > 0$, in particular to $c = \frac{w_\nu}{w_{\tilde{\nu}}}$, with μ-probability zero. This shows that almost surely it does not oscillate around $\frac{w_\nu}{w_{\tilde{\nu}}}$.

Define independent random variables $Y_i = \ln(X_i)$. Let $S_n := \sum_1^n Y_i$ and denote its almost everywhere existing limit by $S = \sum_1^\infty Y_i$. The assertion is verified under condition (31), if we can show that the distribution of S is not concentrated to any point since then also $\prod_1^\infty X_i = \exp(S)$ is not concentrated to any point. In terms of the concentration function defined in (28), this reads $Q(S, 0) = 0$. According to

(31), for each $R > 0$, we find $\eta > 0$ and $n \in \mathbb{N}$ such that $\sum_{i=1}^{n} (1 - Q(X_n, \eta)) > R$. Then, because of $X_i < \infty$ (ignoring the measure-zero set where this may fail),

$$W = \max_{1 \le i \le n} X_i = \max \left\{ \tfrac{\tilde{\nu}_i(x_i)}{\nu_i(x_i)} : 1 \le i \le n \text{ and } \mu(x_i) > 0 \right\}$$

is finite. The mapping

$$(0, W] \ni w \ \mapsto \ \ln(w) \in (-\infty, \ \ln W]$$

is bijective and has derivative at least W^{-1}. Let $\tilde{\eta} = \frac{\eta}{W}$. Then by definition of Y_i, we have $Q(Y_i, \tilde{\eta}) \le Q(X_i, \eta)$ for $1 \le i \le n$ and consequently

$$\sum_{i=1}^{n} \Big(1 - Q(Y_i, \tilde{\eta})\Big) > R.$$

By the Kolmogorov-Rogozin inequality (see [Pet95, Theorem 2.15]), there is a constant C such that

$$Q(S_n, \tilde{\eta}) \le C \left(\sum_{i=1}^{n} \Big(1 - Q(Y_i, \tilde{\eta})\Big) \right)^{-\frac{1}{2}}.$$

Thus, for each $\varepsilon > 0$, we can choose R sufficiently large to guarantee $C \cdot R^{-\frac{1}{2}} < \varepsilon$. Then $Q(S_n, \tilde{\eta}) < \varepsilon$ for n and $\tilde{\eta}$ as before. By (29) we conclude

$$Q(S, \tilde{\eta}) = Q\left(S_n + \Big(\sum_{i=n+1}^{\infty} Y_i \Big), \tilde{\eta} \right) \le Q(S_n, \tilde{\eta}) < \varepsilon$$

and consequently $Q(S, 0) = 0$ since Q is non-decreasing. This proves the assertion under assumption (31).

Now assume that B is bounded, i.e. (31) does not hold. Then there is $R > 0$ such that $\sum_{1}^{n} (1 - Q(X_i, \eta)) \le R$ for all $\eta > 0$ and $n \in \mathbb{N}$. Since the distribution of X_i is a finite convex combination of point measures, for each i there is an $\eta > 0$ such that $Q(X_i, \eta) = Q(X_i, 0)$ and thus $\sum_{i=1}^{n} (1 - Q(X_i, 0)) \le R$ for all $n \in \mathbb{N}$. Therefore, also $\sum_{1}^{\infty} (1 - Q(X_i, 0)) \le R$ holds. Since $\tilde{\nu}_i(x_i) = c_i \nu_i(x_i)$ is equivalent to $X_i = c_i$, this implies that there are constants $c_i \ge 0$ such that

$$\sum_{i=1}^{\infty} \mu_i\{a : \tilde{\nu}_i(a) \ne c_i \nu_i(a)\} \le R. \tag{32}$$

Next we argue that if $c_i \ne 1$ for infinitely many i, then either ν or $\tilde{\nu}$ is eventually not competitive. To verify this claim, let $N_i = \{a : \tilde{\nu}_i(a) \ne c_i \nu_i(a)\}$ and $M_i = \mathcal{X} \setminus N_i$ and observe that $\mu_i(N_i) < \delta$ holds for sufficiently large i, since the sum (32) is bounded. On the other hand μ is uniformly stochastic, so there are no events of probability $\mu_i(a) \in (0, \delta)$, hence $\mu_i(N_i) = 0$ and $\mu_i(M_i) = 1$ for sufficiently large

i. Now for these i, $c_i > 1$ together with $\nu_i(M_i) = 1$ implies the contradiction $\tilde{\nu}_i(M_i) = c_i > 1$. So $c_i > 1$ necessarily requires $\nu_i(M_i) < 1$, hence $\nu_i(M_i) \leq 1 - \delta$, since ν is uniformly stochastic. If this happens infinitely often, then ν is eventually not competitive. A symmetric argument with $\tilde{\nu}$ holds for $c_i < 1$.

The last paragraph shows that, if both ν and $\tilde{\nu}$ stay competitive, eventually $\tilde{\nu}_i = \nu_i$ holds a.s. In this case, $\frac{\tilde{\nu}(x_{1:t})}{\nu(x_{1:t})}$ is eventually constant, which completes the proof. □

Corollary 23 *Let $\mathcal{C}$ be a countable class of factorizable and uniformly stochastic measures, then the maximizing element stabilizes almost surely.*

Proof. This follows from Theorem 21 and Lemma 22. □

Lemma 22 and Corollary 23 are certainly not the only or the strongest assertions obtainable for stabilization. They rather give a flavor how a proof can look like, even if the distributions are not asymptotically mutually singular. On the other hand, the given result is optimal at least in some sense, as shown by the previous Examples 18 and 19. In the former example, μ is not uniformly stochastic but both μ and ν are factorizable, while in the latter one, μ is uniformly stochastic but ν is not factorizable.

The proof of Lemma 22 crucially relies on the independence assumption, which is necessary in order to use the Kolmogorov-Rogozin inequality. It is possible to relax this and require independent sampling only "every so often". It is however not clear how to remove this condition completely.

8 Applications

In the following, we present some applications of the theory developed so far. We begin by stating general loss bounds. After that, three very general applications are discussed.

8.1 Loss bounds

So far we have only considered special loss functions, like the square loss, the Hellinger loss, or the relative entropy. We now show how these results, in particular the bounds for the Hellinger loss, imply regret bounds for *arbitrary loss functions*. (As we will see, square distance is not sufficient.) This parallels the bounds in [Hut03a, Hut03b]. The proofs are simplified, in particular Lemma 24 facilitates the analysis considerably. The reader should compare the results to the bounds for "prediction with expert advice", e.g. [CB97, HP05].

In order to keep things simple, we restrict to binary alphabet $\mathcal{X} = \{0, 1\}$ in this section. Our results extend to general alphabet by the techniques used in [Hut03a]. Consider a binary predictor having access to a belief probability φ depending on the

current history, e.g. $\varphi(x_t = 1|x_{<t}) = \frac{1}{3}$. Which actual prediction should he output, 0 or 1? We can answer this question if we know the *loss function*, according to which losses are assigned to the (wrong) predictions. Consider for example the 0/1 loss (also known as classification error loss), i.e. a wrong prediction gives loss of 1 and a right prediction gives no loss. Then we should predict 1 if our belief is $\varphi > \frac{1}{2}$. This may be different under other loss functions. In general, we should predict in a *Bayes optimal* way: We should output the symbol with the least expected loss,

$$x^\varphi := \arg\min_{\tilde{x}\in\{0,1\}}\{(1-\varphi)\ell(0,\tilde{x}) + \varphi\ell(1,\tilde{x})\},$$

where $\ell(x,\tilde{x})$ is the loss incurred by prediction $\tilde{x}$ if the true symbol is x. In the following, we will restrict to *bounded* loss functions $\ell(x,\tilde{x}) \in [0,1]$. Breaking ties in the above expression in an arbitrary deterministic way, the resulting prediction is *deterministic* for given φ and loss function ℓ. If μ is the true distribution as usual, then let $l_t^\varphi := \sum_a \mu(a|x_{<t})\ell(a,x_t^\varphi)$ be the μ-expected loss of the φ-predictor. Then, by

$$L_{1:n}^\varphi = \mathbf{E}[l_1^\varphi + ... + l_n^\varphi] = \sum_{t=1}^n \mu(x_{<t}) l_t^\varphi(x_{<t})$$

we denote the cumulative μ-expected loss of the φ-predictor. With φ being the variants of the MDL predictor, we will bound the quantity $\Delta_{1:n} = L_{1:n}^\varphi - L_{1:n}^\mu$, i.e. the cumulative *regret*, by an expression depending on $L_{1:n}^\mu$ and w_μ^{-1}.

We admit arbitrary non-stationary loss functions $\ell_{x_{<t}}$ which may depend on the history. Our analysis considers the worst possible choice of loss functions and consists of three steps. First the cumulative regret bound is reduced to an instantaneous regret bound (Lemma 24). Then the instantaneous bound is reduced to a bound in terms of special functions of μ and φ (Lemma 25). Finally, the bound for the special functions is given (Lemma 26).

Lemma 24 *Assume that some φ-predictor satisfies the instantaneous regret bound $\delta_t = l_t^\varphi - l_t^\mu \leq 2h_t + 2\sqrt{2h_t l_t^\mu}$, where $h_t = h_t(\mu,\varphi)$ is the Hellinger distance of the instantaneous predictive probabilities (23). Then the cumulative φ-regret is bounded in the same way:*

$$\Delta_{1:n} = L_{1:n}^\varphi - L_{1:n}^\mu \leq 2H_{1:n}(\mu,\varphi) + 2\sqrt{2H_{1:n}(\mu,\varphi)L_{1:n}^\mu}.$$

This and the following lemma hold with arbitrary constants, the choices 2 and $2\sqrt{2}$ are the smallest ones for which Lemma 26 is true. Note that if the Hellinger distance is replaced by the relative entropy, then $2\sqrt{2}$ may be replaced by 2. Thus, normalized dynamic MDL and Bayes mixture admit smaller bounds, compare [Hut03a]. However, this is not true for the other MDL variants, as we have no relative entropy bound there.

Proof. The key property is the *super-additivity* of the bound. A function $f : [0,\infty)^2 \to [0,\infty)$ is said to be super-additive if

$$f(x_1+x_2, y_1+y_2) \geq f(x_1,y_1) + f(x_2,y_2).$$

The function $(H, L) \mapsto \sqrt{HL}$ satisfies this condition. We now use an inductive argument. Assume $\Delta^0_{2:n} \le 2H^0_{2:n} + 2\sqrt{2H^0_{2:n}L^{\mu,0}_{2:n}}$, where the summation starts at $t = 2$ and the superscript 0 indicates that the first symbol of the sequence was 0. Let the same hold for the first symbol 1. Writing $\mu_1 = \mu(1|\epsilon)$ and using $\delta_1 \le 2h_1 + 2\sqrt{2h_1 l^\mu_1}$, we obtain

$$\begin{aligned}
\Delta_{1:n} &= \delta_1 + (1-\mu_1)\Delta^0_{2:n} + \mu_1\Delta^1_{2:n} \\
&\le 2\left[h_1 + \sqrt{2h_1\ell_1} + (1-\mu_1)\Big(H^0_{2:n} + \sqrt{2H^0_{2:n}L^{\mu,0}_{2:n}}\Big) + \mu_1\Big(H^1_{2:n} + \sqrt{2H^1_{2:n}L^{\mu,1}_{2:n}}\Big)\right] \\
&\le 2\left[H_{1:n} + \sqrt{2h_1\ell_1} + \sqrt{2\Big((1-\mu_1)H^0_{2:n} + \mu_1 H^1_{2:n}\Big)\Big((1-\mu_1)L^{\mu,0}_{2:n} + \mu_1 L^{\mu,1}_{2:n}\Big)}\right] \\
&\le 2H_{1:n} + 2\sqrt{2H_{1:n}L^\mu_{1:n}}.
\end{aligned}$$

Here, the first inequality is the induction hypothesis together with the instantaneous bound, the second bound is Cauchy-Schwarz's inequality, and the last estimate is the super-additivity. □

Lemma 25 *Assume that some φ-predictor satisfies $\tilde\delta \le 2h + 2\sqrt{2h\tilde\ell}$ for all $\mu, \varphi \in [0,1]$, with the Hellinger distance $h = h(\mu,\varphi)$ and the special functions $\tilde\delta(\mu,\varphi)$ and $\tilde\ell(\mu,\varphi)$ defined in the following way, where we slightly abuse notation and abbreviate $\mu = \mu(1|\ldots)$ and $\varphi = \varphi(1|\ldots)$:*

$$\tilde\delta = \frac{|\varphi-\mu|}{\max\{\varphi, 1-\varphi\}} \qquad \text{and} \qquad \tilde\ell = \begin{cases} \mu & \text{if } \mu \le \varphi \le \frac12, \\ \mu(1-\varphi)/\varphi & \text{if } \mu \le \varphi \wedge \frac12 \le \varphi, \\ 1-\mu & \text{if } \frac12 \le \varphi \le \mu, \\ (1-\mu)\varphi/(1-\varphi) & \text{if } \varphi \le \mu \wedge \varphi \le \frac12. \end{cases}$$

Then for arbitrary bounded loss function $\ell : \{0,1\}^2 \to [0,1]$, we have

$$\delta \le 2h + 2\sqrt{2hl^\mu}. \tag{33}$$

Proof. First we show that we may assume $\ell(0,0) = \ell(1,1) = 0$, i.e. we do not incur loss for correct predictions. To this end, consider the modified loss function $\ell'(x,\tilde x) = \ell(x,\tilde x) - \ell(x,x)$ and assume w.l.o.g $\ell'(x,\tilde x) \in [0,1]$. Then it is not hard to see that the regrets under the original and the modified loss functions coincide, while the expected loss of the μ-predictor clearly decreases with the modified loss function. Thus, (33) holds for ℓ if it holds for ℓ'. Hence we may assume $\ell(0,0) = \ell(1,1) = 0$. For each possible outcome $x \in \{0,1\}$, we abbreviate $\ell^x = \ell(x, 1-x)$.

Now assume w.l.o.g. $\mu \le \varphi$. In order to show the assertion, we need to consider the cases in the definition of $\tilde\ell$ separately. We show this only for the first case, i.e. $\mu \le \varphi \le \frac12$. Then $l^\mu = \mu\ell^1$, $l^\varphi = (1-\mu)\ell^0$. We assume that the μ-predictor outputs 0 and the φ-predictor 1, otherwise they give the same prediction and the φ-predictor

has no regret at all. This condition is equivalent to $\ell^0 = \ell^1 \frac{u}{1-u}$ for some $u \in [\mu, \varphi]$. We consider the worst case by maximizing l^φ, i.e. choosing u as large as possible. For this $u = \varphi$, we obtain $\ell^0 = \ell^1 \frac{\varphi}{1-\varphi}$ and

$$\delta = \ell^1[\tfrac{(1-\mu)\varphi}{1-\varphi} - \mu] = \ell^1\tilde{\delta} \leq \ell^1[2h + 2\sqrt{2h\tilde{\ell}}] \leq 2h + 2\sqrt{2h\ell^1\tilde{\ell}} \leq 2h + 2\sqrt{2hl^\mu},$$

showing (33) provided that $\mu \leq \varphi \leq \frac{1}{2}$. The other cases are shown similarly. □

Lemma 26 *The bound $\tilde{\delta} \leq 2h + 2\sqrt{2h\tilde{\ell}}$ holds for all $\mu, \varphi \in [0,1]$, with the functions $\tilde{\delta}, \tilde{\ell} : [0,1]^1 \to [0,1]$ as defined in Lemma 25.*

The technical and not very interesting proof of this lemma is omitted. The careful reader may check the assertion numerically or graphically, as it is just the boundedness of some function on the unit square. We remark that the bound does *not* hold if the Hellinger distance is replaced by the quadratic distance, not even with larger constants.

Theorem 27 *For arbitrary non-stationary loss function which is bounded in $[0,1]$ and known to the MDL predictors, their respective losses are bounded by*

$$L_{1:n}^{\varrho_{\rm norm}}, L_{1:n}^{\varrho}, L_{1:n}^{\varrho^{\rm static}}, L_{1:n}^{\varrho_{\rm norm}^{\rm static}} \quad \leq \quad L_{1:n}^{\mu_{\rm norm}} + 2\sqrt{2cL_{1:n}^{\mu_{\rm norm}} \cdot w_\mu^{-1}} + 2cw_\mu^{-1},$$

where the constant $c = 2, 8, 21,$ or 32, according to which MDL predictor is used (compare Corollary 14).

Proof. This follows from the above three lemmas and from $H_{1:n} \leq c \cdot w_\mu^{-1}$ (Corollary 14). □

This shows in particular that, regardless of the loss function, the average expected per-round regret tends to zero. Again, the direct practical relevance of the bounds is limited because of the potentially huge w_μ^{-1}.

8.2 Classification

Transferring our results to pattern classification is very easy. All we have to do is to add *inputs* to our models. That is, we consider an arbitrary input space $\mathcal{U}$ and (as before) a finite observation or output space $\mathcal{X}$. A model is now a *measure*

$$\nu(x|u) \in [0,1], \ x \in \mathcal{X}, \ u \in \mathcal{U}, \ \text{where} \ \sum_{x \in \mathcal{X}} \nu(x|u) = 1 \ \text{ for all } \ u \in \mathcal{U}.$$

That is, we have a distribution which is conditionalized to the input. We restrict our discussion to measures, since there is no motivation to consider semimeasures

for classification. The definition of a model does not include history dependence. There is no loss of generality: We may include the history in the arbitrary input space.

Transferring the proofs in the previous sections to the present setup is straightforward. We therefore obtain immediately the following corollaries.

Corollary 28 *Let $\mathcal{C}$ be a countable set of classification models containing the true distribution μ. Then for any sequence of inputs $u_{<\infty} \in \mathcal{U}$, we have*

$$\begin{aligned}
\textstyle\sum_t \mathbf{E} \sum_a \left(\mu(a|u_t) - \varrho_{\mathrm{norm}}(a|u_t, u_{<t}, x_{<t})\right)^2 &\leq 2w_\mu^{-1}, \\
\textstyle\sum_t \mathbf{E} \sum_a \left(\mu(a|u_t) - \varrho(a|u_t, u_{<t}, x_{<t})\right)^2 &\leq 8w_\mu^{-1}, \\
\textstyle\sum_t \mathbf{E} \sum_a \left(\mu(a|u_t) - \varrho^{\mathrm{static}}(a|u_t, u_{<t}, x_{<t})\right)^2 &\leq 21w_\mu^{-1}.
\end{aligned}$$

(Note that although each single model formally does not depend on the history, the MDL estimators necessarily do.)

We need not consider the normalized static variant here, since all models are measures anyway. If there is a distribution over $\mathcal{U}$, the result therefore also holds in expectation over the inputs. An analogue of Corollary 23 is obtained as easily. If the inputs are i.i.d., which is usually assumed for classification, then the two conditions of factorizability and uniform stochasticity are trivially satisfied. Therefore, the true distribution μ is eventually discovered by MDL almost surely. Note that in this case, the distributions are also asymptotically mutually singular, so that the assertion also follows from Barron's [Bar85] earlier result.

Note that again, the assumption $\mu \in \mathcal{C}$ is essential. In practical applications, if this is not clear, it may be therefore favorable to choose a different method having guarantees without this condition, compare [GL04].

8.3 Regression

We may also apply our results in the regression setup, that is for predicting continuous densities. Our use of the term regression is a bit non-standard here, since it normally refers to just estimating the mean of some prediction, where the distribution is often assumed to be Gaussian. Again the assumption $\mu \in \mathcal{C}$ is essential, so that in practice some other method not relying on it might be preferred.

Continuous densities cause some additional difficulties. The observation space is now $\mathbb{R}$. This implies in particular that, like for the loss bounds, the square distance is no longer appropriate for our purpose[5] (note that our use of the squared error

[5]To see this, define a distribution f by its density $f_n = \frac{n}{3}\chi_{[-\frac{1}{n},0]} + \frac{2n}{3}\chi_{(0,\frac{1}{n}]}$, where χ is the characteristic function of an interval. Let $\tilde{f}(x) = f(-x)$, then the quadratic distance is $\int (f-\tilde{f})^2 dx = \frac{2n}{9} \stackrel{n\to\infty}{\longrightarrow} \infty$, whereas the relative entropy $\int f \ln(f/\tilde{f}) dx = \frac{\ln 2}{3}$ is constant.

is completely different from the standard use in regression). So we will use the Hellinger distance instead, defined similarly to (23) by

$$h(f,\tilde{f}) = \int \left(\sqrt{f(x)} - \sqrt{\tilde{f}(x)}\right)^2 dx \quad \text{for integrable } f,\tilde{f}:\mathbb{R}\to[0,\infty). \tag{34}$$

Accordingly, $H_{1:n}(\mu,\varphi) = \sum_t \mathbf{E}h(\mu(\cdot|u_t),\varphi(\cdot|u_t,u_{<t},x_{<t}))$ is the cumulative Hellinger distance of two predictive distributions μ and φ. Similarly as in (25) and (26), the Hellinger distance is bounded by the (continuous) relative entropy and absolute distance. This shows in particular that the integral (34) exists.

We now consider a countable class $\mathcal{C}$ of models that are functions ν from $\mathcal{U}$ to *uniformly bounded probability densities* on $\mathcal{X} = \mathbb{R}$. That is, there is some $C > 0$ such that

$$0 \le \nu_i(x|u) \le C \;\text{ and }\; \int_{-\infty}^{\infty} \nu_i(x|u)dx = 1 \;\text{ for all }\; i \ge 1,\; u \in \mathcal{U}, \;\text{ and }\; x \in \mathbb{R}. \tag{35}$$

for all $i \ge 1$, $u \in \mathcal{U}$, and $x \in \mathbb{R}$. The MDL estimator is then defined as the element which maximizes the *density*, $\nu^* = \arg\max_{\nu\in\mathcal{C}}\{w_\nu \nu(x_{1:n}|u_{1:n})\}$. The uniform boundedness condition asserts that the MDL estimator exists. It may be relaxed, provided that the MDL estimator remains well-defined, such as for a family of Gaussian densities which tend to the point measure.

With these definitions, the proofs of the theorems for static and dynamic MDL can be adapted. Since the triangle inequality holds for the Hellinger distance $\sqrt{H^2}$, we obtain the following.

Corollary 29 *Let $\mathcal{C}$ be a countable model class according to (35), containing the true distribution μ. Then for any sequence of inputs $u_{<\infty} \in \mathcal{U}$, we have $H_{1:n}(\mu,\varrho_{\rm norm}) \le 2w_\mu^{-1}$, $H_{1:n}(\mu,\varrho) \le 8w_\mu^{-1}$, and $H_{1:n}(\mu,\varrho^{\rm static}) \le 21w_\mu^{-1}$.*

We may apply this for example to model classes with Gaussian noise, concluding that the mean and the variances converge to the true values, see [PH05] for an example. It is not immediately clear how to obtain an analogue of Corollary 23 for continuous densities.

8.4 Universal Induction

Since the assertions on static and dynamic MDL have been proven generally for *semimeasures*, we may apply them to the universal setup. Here $\mathcal{C} = \mathcal{M}$ is the countable set of all lower semicomputable (= enumerable) semimeasures on $\mathcal{X}^*$. So $\mathcal{M}$ contains stochastic models in general, and in particular all models for computable deterministic sequences. There is a one-to-one correspondence of $\mathcal{M}$ to the class of all programs on some fixed universal monotone Turing machine U, see e.g. [LV97]. We will assume programs to be *binary*, in contrast to outputs, which are strings

$x \in \mathcal{X}^*$. This relation defines in particular the complexities and weights of each ν by

$$Kw(\nu) = \text{length of the program for } \nu \text{ on } U, \quad \text{and} \quad w_\nu = 2^{Kw(\nu)}. \tag{36}$$

We call these weights the *canonical weights*. They satisfy $w_\nu > 0$ for all ν and $\sum_\nu w_\nu \leq 1$.

An enumerable semimeasure which dominates all other enumerable semimeasures is called universal. The Bayes mixture ξ defined in (2) has this property. One can show that ξ is equal within a multiplicative constant to Solomonoff's prior [Sol64, Eq. (7)], which is the a priori probability that (some extension of) a string x is generated provided that the input of U consists of fair coin flips. That is

$$\xi(x) \quad \stackrel{\times}{=} \quad M(x) \quad = \sum_{p \text{ minimal: } U(p)=x*} 2^{-\ell(p)} \text{ for all } x \in \mathcal{X}^*.$$

Here, we use the notations

$$f \stackrel{+}{\leq} g :\Leftrightarrow f \leq g + O(1), \qquad f \stackrel{+}{=} g :\Leftrightarrow f \stackrel{+}{\leq} g \wedge g \stackrel{+}{\leq} f,$$
$$f \stackrel{\times}{\leq} g :\Leftrightarrow f \leq g \cdot O(1), \qquad f \stackrel{\times}{=} g :\Leftrightarrow f \stackrel{\times}{\leq} g \wedge g \stackrel{\times}{\leq} f.$$

The MDL definitions in Section 3 directly transfer to this setup. All bounds on the cumulative square loss (subsumed in Corollary 14) therefore apply to $\varrho = \varrho_{[\mathcal{M}]}$. The necessary assumption now reads that μ must be a recursive (= computable) measure. Also, Theorem 1 implies Solomonoff's important universal induction theorem.

In addition to $\mathcal{M}$, we also consider the set of all recursive measures $\tilde{\mathcal{M}}$ together with the same canonical weights (36). We define $\tilde{\xi} = \xi_{[\tilde{\mathcal{M}}]}$ and $\tilde{\varrho} = \varrho_{[\tilde{\mathcal{M}}]}$. Then $\tilde{\varrho}(x) \leq \tilde{\xi}(x) \leq \xi(x)$ and $\varrho(x) \leq \xi(x)$ for all $x \in \mathcal{X}^*$ is immediate. It is straightforward that $\xi(x) \stackrel{\times}{\leq} \varrho(x)$ since $\xi \in \mathcal{M}$. Moreover, for any string $x \in \mathcal{X}^*$, define the *monotone complexity* $Km(x) = \min\{\ell(p) : U(p) = x*\}$ as the length of the shortest program such that U's output starts with x. The following assertion holds.

Proposition 30 *We have* $K\tilde{\varrho} \stackrel{+}{\geq} Km$.

Proof. We must 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 $Kw(\nu) + K\nu(x) + c$ that outputs a string starting with x, where constant c is independent of x and ν.

Consider all strings $y_i \in \mathcal{X}^n$ $(1 \leq i \leq |\mathcal{X}|^n)$ of length $n = \ell(x)$ arranged in lexicographical order. Each y_i has measure $P_i = \mu(y_i)$. Let S_i be the cumulated measures: $S_0 = 0$ and $S_i = \sum_{k=1}^{i} P_k$. Let j be the index of x, i.e. $x = y_j$. Then, the interval $[S_{j-1}, S_j) \subset [0,1)$ has measure P_j and therefore contains exactly one $\lceil -\log_2 P_j \rceil$-bit number $z \in [S_{j-1}, S_j)$. We describe x with the number z, this is known

as *arithmetic encoding* (see e.g. [CT91]). The coding is injective since $[S_{i-1}, S_i)$ and $[S_{k-1}, S_k)$ are disjoint for $i \neq k$.

In order to decode z, we may descend the $|\mathcal{X}|$-ary tree of all possible strings y, first considering strings of length one, then of length two, etc. For each possible string y, we can determine its binary code by approximating $\nu(x)$ sufficiently accurately. Eventually we will find z, then we print the current y. At this stage, y might be only a prefix of x, since an extension of y might have a measure very close to y and thus map to the same code z. Therefore we continue the procedure until all codes starting with z are proper extensions of z (which may never be the case, then the algorithm runs forever). In each step, the appropriate additional symbol is written on the output tape. The resulting output will be x or some extension of x.

This algorithm can be specified in a constant c' number of bits. The description of ν needs another $Kw(\nu)$ bits. Finally, z has length $\lceil -\log_2 P_j \rceil \leq -\log_2 \nu(x) + 1$. Thus, the overall description has length $Kw(\nu) + K\nu(x) + c$ as required. □

It is also possible to prove the proposition indirectly using [LV97, Thm.4.5.4]. This implies that $Km(x) \stackrel{+}{\leq} Kw(\nu) + K\nu(x)$ for all $x \in \mathcal{X}^*$ and all recursive measures $\nu \in \tilde{\mathcal{M}}$. Then, also $Km(x) \stackrel{+}{\leq} \min\{Kw(\nu) + K\nu(x)\} = K\tilde{\varrho}(x)$ holds.

So together with the above observations, we have

$$Km(x) \stackrel{+}{\leq} K\tilde{\varrho}(x) \stackrel{+}{\geq} K\tilde{\xi}(x) \stackrel{+}{\geq} K\varrho(x) \stackrel{\pm}{=} KM(x). \tag{37}$$

On the other hand, there is a deep result in Algorithmic information theory which states that an exact coding theorem does *not* hold on continuous sample space, $Km(x) \stackrel{+}{\not\geq} KM(x)$ [Gác83]. Therefore, at least one of the above $\stackrel{+}{\geq}$ must be proper.

Problem 31 *Which of the two inequalities* $K\tilde{\varrho}(x) \stackrel{\times}{\geq} K\tilde{\xi}(x)$ *and* $K\tilde{\xi}(x) \stackrel{\times}{\geq} K\varrho(x)$ *is proper (or are both)?*

The proof in [Gác83] is very subtle, and the phenomenon is still not completely understood. There is some hope that by answering Problem 31, one arrives at a better understanding of the continuous coding theorem and even at a simpler proof for its failure.

9 Discussion

In this last section, we recapitulate the main achievements of this work and discuss their philosophical and practical consequences. In the first place, we have shown that if two-part MDL is used for predicting a stochastic sequence, then the predictive probabilities converge to the true ones in mean sum, provided that the distribution generating the sequence is contained in the model class. The two most important implications are almost sure convergence and loss bounds for arbitrary loss functions.

The guaranteed convergence is slow in general: All bounds depend linearly on w_μ^{-1}, the inverse of the prior weight of the true distribution. For large model classes, this number must be regarded too huge to be relevant for practical applications. Examples show that this bound is sharp. This is in contrast to the exponentially smaller corresponding bound for the Bayes mixture. The latter predictor however is often computationally more expensive to approximate in practice. We believe that this principally indicates that with MDL, some care has to be taken when choosing the model class and the prior. Conditions which are sufficient for fast convergence have been given for instance in [Ris96, BRY98, PH04b]. It remains a major challenge to generalize these results in order to obtain fast convergence under assumptions that are as weak as possible. In particular for universal induction, this question is interesting and possibly difficult. Even when considering only computable Bernoulli distributions endowed with a universal prior, fast convergence possibly holds for many environments, but maybe not for all [PH04b]. We also need to distinguish how the large error cumulates. Either the instantaneous error remains significant for a long time, which is critical, or the instantaneous error drops just too slowly to be summable, e.g. as $O(\frac{1}{n})$, which is tolerable. We have seen instances for both cases; compare the discussion after Example 15. In this light, the cumulative error might not be the right quantity to assess convergence speed.

The main results have been shown under the only assumption that the data generating process is contained in the model class. This condition is essential in general, as [GL04] shows that in its absence MDL can fail dramatically. In the universal setup, the assumption merely requires that the data is generated in some (probabilistically) computable way. This is a very weak condition. Laplace, Zuse [Zus67] and successors argue that nature operates in a computable way, and consequently *all* thinkable data satisfies the assumption. On the other hand, predicting with a universal model is computationally very expensive. In particular it is provably infeasible if the thesis of computable nature holds. Despite these practical problems, the theory of universal prediction is valuable since it explores the limits of computational induction.

Acknowledgements. Thanks to Peter Grünwald and an anonymous reviewer for their very valuable comments and suggestions. This work was supported by SNF grant 2100-67712.02.

References

[Bar85] A. R. Barron. *Logically smooth density estimation.* PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

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

[BD62] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

[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.

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

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

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

[CT88] Y. S. Chow and H. Teicher. *Probability Theory*. Springer-Verlag, New York, 2nd edition, 1988.

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

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

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

[GL04] P. Grünwald and J. Langford. Suboptimal behaviour of Bayes and MDL in classification under misspecification. In *17th Annual Conference on Learning Theory (COLT)*, pages 331–347, 2004.

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

[GV01] S. Ghosal and A. van der Vaart. Entropies and rates of convergence for Bayes and maximum likelihood estimation for mixture of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001.

[GV04] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. preprint, 2004.

[HP05] M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005.

[Hut01] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, June 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. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003.

[Hut03c] 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. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[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.

[Li99] J. Q. Li. *Estimation of Mixture Models*. PhD thesis, Dept. of Statistics. Yale University, 1999.

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

[Pet95] V. V. Petrov. *Limit Theorems of Probability Theory*. Clarandon Press, Oxford, 1995.

[PH04a] J. Poland and M. Hutter. Convergence of discrete MDL for sequential prediction. In *17th Annual Conference on Learning Theory (COLT)*, pages 300–314, 2004.

[PH04b] J. Poland and M. Hutter. On the convergence speed of MDL predictions for Bernoulli sequences. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 294–308, 2004.

[PH05] J. Poland and M. Hutter. Strong asymptotic assertions for discrete MDL in regression and classification. In *Benelearn 2005 (Ann. Machine Learning Conf. of Belgium and the Netherlands)*, 2005.

[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.

[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. 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.

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

[Zha04] T. Zhang. On the convergence of MDL density estimation. In *Proc. 17th Annual Conference on Learning Theory (COLT)*, pages 315–330, 2004.

[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.

[Zus67] K. Zuse. Rechnender Raum. *Elektronische Datenverarbeitung*, 8:336–344, 1967. English translation: Calculating Space, MIT Technical Translation AZT-70-164-GEMIT, 1970.