Strong Asymptotic Assertions for Discrete MDL in Regression and Classification

Jan Poland JAN@IDSIA.CH
Marcus Hutter MARCUS@IDSIA.CH
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland WWW.IDSIA.CH

Abstract

We study the properties of the MDL (or maximum penalized complexity) estimator for Regression and Classification, where the underlying model class is countable. We show in particular a finite bound on the Hellinger losses under the only assumption that there is a "true" model contained in the class. This implies almost sure convergence of the predictive distribution to the true one at a fast rate. It corresponds to Solomonoff's central theorem of universal induction, however with a bound that is exponentially larger.

Keywords. Regression, Classification, Sequence Prediction, Machine Learning, Minimum Description Length, Bayes Mixture, Marginalization, Convergence, Discrete Model Classes.

1. Introduction

Bayesian methods are popular in Machine Learning. So it is natural to study their predictive properties: How do they behave asymptotically for increasing sample size? Are loss bounds obtainable, either for certain interesting loss functions or even for more general classes of loss functions?

In this paper, we consider the two maybe most important Bayesian methods for prediction in the context of regression and classification. The first one is *marginalization*: Given some data and a model class, obtain a predictive model by integrating over the model class. This *Bayes mixture* is "ideal" Bayesian prediction in many respects, however in many cases it is computationally untractable. Therefore, a commonly employed method is to compute a *maximum penalized complexity* or *maximum a posteriori (MAP)* or *minimum description length (MDL)* estimator. This predicts according to the "best" model instead of a mixture. The MDL principle is important for its own sake, not only as approximation of the Bayes mixture.

Most work on Bayesian prediction has been carried out for *continuous* model classes, e.g. classes with one free parameter $\vartheta \in \mathbb{R}^d$. While the predictive properties of the Bayes mixture are excellent under mild conditions [CB90, Hut03b, GGvdV00, Hut04], corresponding MAP or MDL results are more difficult to establish. For MDL in the strong sense of *description length*, the parameter space has to be discretized appropriately (and dynamically with increasing sample size) [Ris96, BRY98, BC91]. A MAP estimator on the other hand can be very bad in general. In statistical literature, some important work has been performed on the asymptotical discovery of the true parameter, e.g. [CY00]. This can only hold if each model occurs no more than once in the class. Thus it is violated e.g. in the case of an artificial neural network, where exchanging two hidden units in the same layer does not alter the network behavior.

In the case of *discrete* model classes, both loss bounds and asymptotic assertions for the Bayes mixture are relatively easy to prove, compare Theorem 2. In [PH04a], corresponding results for MDL were shown. The setting is sequence prediction but otherwise very general. The only assumption necessary is that the true distribution is contained in the model class. Assertions are given directly for the predictions, thus there is no problem of possibly undistinguishable models. In order to prove that the MDL estimator (precisely, the *static* MDL estimator in terms of [PH04a]) has good predictive properties, we introduce an intermediate step and show first the predictive properties of *dynamic MDL*, where a new MDL estimator is computed for each possible next observation.

In this paper, we will derive analogous results for regression and classification. While results for classification can be generalized from sequence prediction by conditionalizing everything to the input, regression is technically more difficult. Therefore the next section, which deals with the regression setup, covers the major part of the paper. Instead of the popular Euclidian and Kullback-Leibler distances for measuring prediction quality we need to exploit the Hellinger distance. We show that online MDL converges to the true distribution in mean Hellinger sum, which implies "rapid" convergence with probability one. Classification is briefly discussed in Section 3, followed by a discussion and conclusions in Section 4.

2. Regression

We neglect computational aspects and study the properties of the *optimal* Bayes mixture and MDL predictors. When a new sample is observed, the estimator is updated. Thus, regression is considered in an *online framework*: The first input x_1 is presented, we predict the output y_1 and then observe its true value, the second input x_2 is presented and so on.

Setup. Consider a regression problem with arbitrary domain $\mathcal{X}$ (we need no structural assumptions at all on $\mathcal{X}$) and co-domain $\mathcal{Y} = \mathbb{R}$. The task is to learn/fit/infer a function $f : \mathcal{X} \to \mathcal{Y}$, or more generally a conditional probability density $\nu(y|x)$, from data $\{(x_1, y_1), ..., (x_n, y_n)\}$. Formally, we are given a countable class $\mathcal{C}$ of models that are functions ν from $\mathcal{X}$ to *uniformly bounded probability densities* on $\mathbb{R}$. That is, $\mathcal{C} = \{\nu_i : i \geq 1\}$, and there is some $C > 0$ such that

$$0 \leq \nu_i(y|x) \leq C \text{ and } \int_{-\infty}^{\infty} \nu_i(y|x)dy = 1 \quad (1)$$
$$\text{for all } i \geq 1,\ x \in \mathcal{X},\text{ and } y \in \mathcal{Y}.$$

Each ν induces a probability density on $\mathbb{R}^n$ for n-tuples $x_{1:n} \in \mathcal{X}^n$ by $\nu(y_{1:n}|x_{1:n}) = \prod_{t=1}^n \nu(y_t|x_t)$. The notation $x_{1:n}$ for n-tuples is common in sequence prediction. Each model $\nu \in \mathcal{C}$ is associated with a *prior weight* $w_\nu > 0$. The logarithm $\log_2 w_\nu$ has often an interpretation as model *complexity*. We require $\sum_\nu w_\nu = 1$. Then by the Kraft inequality, one can assign to each model $\nu \in \mathcal{C}$ a prefix-code of length $\lceil \log_2 w_\nu \rceil$.

We assume that an infinite stream of data $(x_{1:\infty}, y_{1:\infty})$ is generated as follows: Each x_t may be produced by an arbitrary mechanism, while y_t is sampled from a *true distribution* μ conditioned on x_t. In order to obtain strong convergence results, we will require that $\mu \in \mathcal{C}$.

Example 1 Take $\mathcal{X} = \mathbb{R}$ and $\mathcal{C}_\sigma^{\text{lin1}} \cong \{ax + b + N(0, \sigma^2) : a, b \in \mathbb{Q}\}$ to be the class of linear regression models with rational coefficients a, b, and independent Gaussian noise of fixed variance $\sigma^2 > 0$. That is, $\mathcal{C}_\sigma^{\text{lin1}} = \{\nu^{a,b,\sigma} : a, b \in \mathbb{Q}\}$, where

$$\nu^{a,b,\sigma}(x, y) = \phi_{\sigma^2}(y - ax - b) = \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-ax-b)^2}.$$

Alternatively, you may consider the class $\mathcal{C}_{\geq\sigma_0}^{\text{lin1}} = \{\nu^{a,b,\sigma} : a, b, \sigma \in \mathbb{Q}, \sigma \geq \sigma_0\}$ for some $\sigma_0 > 0$, where also the noise amplitude is part of the models. In the following, we also discuss how to admit degenerate Gaussians that are point measures such as $\mathcal{C}_{\geq 0}^{\text{lin1}}$.

The setup (1) guarantees that all subsequent MDL estimators [(9) and (10)] exist. However, our results and proofs generalize in several directions. First, for the co-domain $\mathcal{Y}$ we may choose any σ-finite measure space instead of $\mathbb{R}$, since we need only Radon-Nikodym densities below. Second, the uniformly boundedness condition can be relaxed, if the MDL estimators still exist. This holds for example for the class $\mathcal{C}_{\geq 0}^{\text{lin1}}$ (see the preceding example), if the definition of the MDL estimators is adapted appropriately (see footnote 2 on page 4). Third, the results remain valid for semimeasures with $\int \nu \leq 1$ instead of measures and $\sum w_\nu \leq 1$, which is however not very relevant for regression (but for universal sequence prediction). In order to keep things simple, we develop all results on the basis of (1). Note finally that the models in $\mathcal{C}$ may be time-dependent, and we need not even make this explicit, since the time can be incorporated into $\mathcal{X}$ ($x_t = (x'_t, t) \in \mathcal{X}' \times \mathbb{N} = \mathcal{X}$). In this way we may also make the models depend on the actual past outcome, if this is desired ($x_t = (x'_{1:t} y_{1:t-1}) \in \mathcal{X}'^* \times \mathcal{Y}^* = \mathcal{X}$).

The case of independent Gaussian noise as in Example 1 is a particularly important one. We therefore introduce the family

$$\mathcal{F}_{\geq}^{\text{Gauss}} = \Big\{\mathcal{C} = \{\nu_i, \sigma_i\}_{i=1}^\infty : \nu_i(x, y) = \phi_{\sigma_i^2}\big(y - f_i(x)\big), \sigma_i \geq \sigma_0 > 0, f_i : \mathcal{X} \to \mathbb{R}\Big\}. \quad (2)$$

of all countable regression model classes with lower bounded Gaussian noise. Clearly, $\mathcal{C}_\sigma^{\text{lin1}}, \mathcal{C}_{\geq\sigma_0}^{\text{lin1}} \in \mathcal{F}_{\geq}^{\text{Gauss}}$ is satisfied. Similarly $\mathcal{F}^{\text{Gauss}} \supset \mathcal{F}_{\geq}^{\text{Gauss}}$ denotes the corresponding family without lower bound on σ_i. Then $\mathcal{C}_{\geq 0}^{\text{lin1}} \in \mathcal{F}^{\text{Gauss}} \setminus \mathcal{F}_{\geq}^{\text{Gauss}}$.

We define the *Bayes mixture*, which for each $n \geq 1$ maps an n-tuple of inputs $x_{1:n} \in \mathcal{X}^n$ to a probability density on $\mathbb{R}^n$:

$$\xi(y_{1:n}|x_{1:n}) = \sum_{\nu \in \mathcal{C}} w_\nu \nu(y_{1:n}|x_{1:n}) = \sum_{\nu \in \mathcal{C}} w_\nu \prod_{t=1}^n \nu(y_t|x_t) \quad (3)$$

(recall $\sum_\nu w_\nu = 1$). Hence, the Bayes mixture *dominates* each ν by means of $\xi(\cdot|x_{1:n}) \geq w_\nu \nu(\cdot|x_{1:n})$ for all $x_{1:n}$. For $\nu \in \mathcal{C}$ and $x_n \in \mathcal{X}$, the ν-prediction of

$y_n \in \mathbb{R}$, that is the *nu*-probability density of observing y_n, is

$$\nu(y_n|x_{1:n}, y_{<n}) = \nu(y_n|x_n).$$

This is independent of the history $(x_{<n}, y_{<n}) = (x_{1:n-1}, y_{1:n-1})$. In contrast, the *Bayes mixture prediction* or regression, which is also a measure on $\mathbb{R}$, depends on the history:

$$\xi(y_n|x_{1:n}, y_{<n}) = \frac{\xi(y_{1:n}|x_{1:n})}{\xi(y_{<n}|x_{<n})} = \frac{\sum_\nu w_\nu \prod_{t=1}^{n} \nu(y_t|x_t)}{\sum_\nu w_\nu \prod_{t=1}^{n-1} \nu(y_t|x_t)}. \tag{4}$$

This is also known as *marginalization*. Observe that the denominator in (4) vanishes only on a set of μ-measure zero, if the true distribution μ is contained in $\mathcal{C}$. Under condition (1), the Bayes mixture prediction is uniformly bounded. It can be argued intuitively that in case of unknown $\mu \in \mathcal{C}$ the Bayes mixture is the best possible model for μ. Formally, its predictive properties are excellent:

Theorem 2 *Let $\mu \in \mathcal{C}$, $n \geq 1$, and $x_{1:n} \in \mathcal{X}^n$, then*

$$\sum_{t=1}^{n} \mathbf{E} \int \Big(\sqrt{\mu(y_t|x_{1:t}, y_{<t})} - \sqrt{\xi(y_t|x_{1:t}, y_{<t})}\Big)^2 dy_t \tag{5}$$
$$\leq \ \ln w_\mu^{-1}.$$

$\mathbf{E}$ denotes the expectation with respect to the true distribution μ. Hence in this case we have $\mathbf{E}\ldots = \int \ldots \mu(dy_{<t})$. The integral expression is also known as *square Hellinger distance*. It will emerge as a main tool in the subsequent proofs. So the theorem states that on any input sequence $x_{<\infty}$ the expected cumulated Hellinger divergence of μ and the Bayes mixture prediction is bounded by $\ln w_\mu^{-1}$. A closely related result was discovered by Solomonoff ([Sol78]) for universal sequence prediction, a "modern" proof can be found in [Hut04]. This proof can be adapted in our regression framework. Alternatively, it is not difficult to give a proof in a few lines analogous to (14) and (15) by using (12).

We introduce the term *convergence in mean Hellinger sum (i.m.H.s.)* for bounds like (5): For some predictive density ψ, the ψ-predictions converge to the μ-predictions i.m.H.s. on a sequence of inputs $x_{<\infty} \in \mathcal{X}^\infty$, if there is $R > 0$ such that $H^2_{x_{<\infty}}(\mu, \psi) \leq R$, where

$$H^2_{x_{<\infty}}(\mu, \psi) = \sum_{t=1}^{\infty} \mathbf{E}[h_t^2] \text{ with} \tag{6}$$
$$h_t^2 = \int \Big(\sqrt{\mu(y_t|x_{1:t}, y_{<t})} - \sqrt{\psi(y_t|x_{1:t}, y_{<t})}\Big)^2 dy_t.$$

Convergence i.m.H.s. is a very strong convergence criterion. It asserts a finite expected cumulative Hellinger loss in the first place. If the co-domain $\mathcal{Y}$ is finite as for classification (see Section 3), then convergence i.m.H.s. implies almost sure (a.s.) convergence of the (finitely many) posterior probabilities. For regression, the situation is more complex, since the posterior probabilities are densities, i.e. Banach space valued. Here, convergence i.m.H.s. implies that with μ-probability one the square roots of the predictive densities converge to the square roots of the μ-densities in $L^2(\mathbb{R})$ (endowed with the Lebesgue measure). In other words, h_t^2 converges to zero a.s.:

$$\begin{aligned} \mathbf{P}\Big(\exists t \geq n : h_t^2 \geq \varepsilon\Big) &= \mathbf{P}\Big(\bigcup_{t\geq n} \{h_t^2 \geq \varepsilon\}\Big) \\ &\leq \sum_{t \geq n} \mathbf{P}\left(h_t^2 \geq \varepsilon\right) \\ &\leq \frac{1}{\varepsilon}\sum_{t=n}^{\infty} \mathbf{E} h_t^2 \stackrel{n\to\infty}{\longrightarrow} 0 \end{aligned} \tag{7}$$

holds by the union bound, the Markov inequality for all $\varepsilon > 0$, and $H^2_{x_{<\infty}} < \infty$, respectively, where $\mathbf{P}$ is the μ-probability. If the densities are uniformly bounded, then also the differences of the densities (as opposed to the difference of the square roots) converge to zero:

$$\psi(y_t|x_{1:t}, y_{<t}) - \mu(y_t|x_{1:t}, y_{<t}) \stackrel{t\to\infty}{\longrightarrow} 0 \quad \text{in } L^2(\mathbb{R}) \text{ a.s.}$$

Moreover, the finite bound on the cumulative Hellinger distances can be interpreted as a convergence rate. Compare the parallel concept "convergence in mean sum" [Hut03b, PH04a, Hut04].

MDL Predictions. In many cases, the Bayes mixture is not only intractable, but even hard to approximate. So a very common substitute is the (ideal) MDL[1] estimator, also known as maximum a posteriori (MAP) or maximum complexity penalized likelihood estimator. Given a model class $\mathcal{C}$ with weights (w_ν) and a data set $(x_{1:n}, y_{1:n})$, we define the two-part MDL estimator as

$$\begin{aligned} \nu^* = \nu^*_{(x_{1:n}, y_{1:n})} &= \arg\max_{\nu \in \mathcal{C}} \{w_\nu \nu(y_{1:n}|x_{1:n})\} \text{ and} \\ \varrho(y_{1:n}|x_{1:n}) &= \max_{\nu \in \mathcal{C}} \{w_\nu \nu(y_{1:n}|x_{1:n})\} \\ &= w_{\nu^*} \nu^*(y_{1:n}|x_{1:n}). \end{aligned} \tag{8}$$

Note that we define both the model ν^* which is the MDL estimator and its weighted density ρ. In our

[1] There is some disagreement about the exact meaning of the term MDL. Sometimes a specific prior is associated with MDL, while we admit arbitrary priors. More importantly, when coding some data x, one can exploit the fact that once the model ν^* is specified, only data which lead to the maximizing element ν^* need to be considered. This allows for a shorter description than $\log_2 \nu^*(x)$. Nevertheless, the *construction principle* is commonly termed MDL, compare for instance the "ideal MDL" in [VL00].

setup (1), the MDL estimator is well defined, since all maxima exist[2]. Moreover, $\varrho(\cdot|x_{1:n})$ is a density but its integral is less than 1 in general. We have $\varrho(\cdot|x_{1:n}) \geq w_\nu \nu(\cdot|x_{1:n})$, so like ξ, ϱ dominates each $\nu \in \mathcal{C}$. Also, $\varrho(\cdot|x_{1:n}) \leq \xi(\cdot|x_{1:n})$ is clear by definition. If we use ν^* for (sequential online) prediction, this is the *static MDL prediction*:

$$\varrho^{\text{static}}(y_n|x_{1:n}, y_{<n}) = \nu^*_{(x_{<n}, y_{<n})}(y_n|x_n). \tag{9}$$

This is the common way of using MDL for prediction. Clearly, the static MDL predictor is a probability density on $\mathbb{R}$. Alternatively, we may compute the MDL estimator for each possible y_n separately, arriving at the *dynamic* MDL predictor:

$$\varrho(y_n|x_{1:n}, y_{<n}) = \frac{\varrho(y_{1:n}|x_{1:n})}{\varrho(y_{<n}|x_{<n})}. \tag{10}$$

We have $\varrho(y_n|x_{1:n}, y_{<n}) \leq \nu^*_{(x_{1:n}, y_{1:n})}(y_n|x_n)$ for each y_n, which shows that under condition (1) the dynamic MDL predictor is uniformly bounded. On the other hand, $\varrho(y_n|x_{1:n}, y_{<n}) \geq \nu^*_{(x_{<n}, y_{<n})}(y_n|x_n)$ holds, so the dynamic MDL predictor may be a density with mass more than 1. Hence we must usually *normalize* it for predicting:

$$\bar\varrho(y_n|x_{1:n}, y_{<n}) = \frac{\varrho(y_{1:n}|x_{1:n})}{\int \varrho(y_{1:n}|x_{1:n})dy_n}. \tag{11}$$

Both fractions in (10) and (11) are well-defined except for a set of measure zero. Dynamic MDL predictions are in a sense computationally (almost) as expensive as the full Bayes mixture.

Convergence Results. Our principal aim is to prove predictive properties of *static* MDL, since this is the practically most relevant variant. To this end, we first need to establish corresponding results for the dynamic MDL. Precisely, the following holds.

Theorem 3 *Assume the setup (1). If $\mu \in \mathcal{C}$, where μ is the true distribution, and $H^2_{x_{<\infty}}(\cdot,\cdot)$ is defined as in (6), then for all input sequences $x_{<\infty} \in \mathcal{X}^\infty$ we have*

$$\begin{array}{ll} (i) & H^2_{x_{<\infty}}(\mu, \bar\varrho) \leq w_\mu^{-1} + \ln w_\mu^{-1}, \\ (ii) & H^2_{x_{<\infty}}(\bar\varrho, \varrho) \leq 2w_\mu^{-1}, \text{ and} \\ (iii) & H^2_{x_{<\infty}}(\varrho, \varrho^{\text{static}}) \leq 3w_\mu^{-1}. \end{array}$$

[2] For a model class with Gaussian noise $\mathcal{C} \in \mathcal{F}^{\text{Gauss}}$ (2), we may dispose of the uniform boundedness condition and admit e.g. also $\mathcal{C}^{\text{lin1}}_{>0}$. In order to compute the MDL estimator, we must then first check if there is nonzero mass concentrated on $(x_{1:n}, y_{1:n})$, in which case the mass is even one and the corresponding model with the largest weight is chosen. Otherwise, the MDL estimator is chosen according to the maximum penalized density. All results and proofs below generalize to this case.

Since the triangle inequality holds for $\sqrt{H^2_{x_{<\infty}}(\cdot,\cdot)}$, we immediately conclude:

Corollary 4 *Given the setup (1) and $\mu \in \mathcal{C}$, then all three predictors $\bar\varrho$, ϱ, and ϱ^{static} converge to the true density μ in mean Hellinger sum, for any input sequence $x_{<\infty}$. In particular, we have $H^2(\mu, \varrho^{\text{static}}) \leq 21 w_\mu^{-1}$.*

We will only prove (i) of Theorem 3 here. The proofs of (ii) and (iii) can be similarly adapted from [PH04a, Theorems 10 and 11], since the Hellinger distance is bounded by the absolute distance: $\int \big(\sqrt{\mu(y)} - \sqrt{\nu(y)}\big)^2 dy \leq \int \big|\mu(y) - \nu(y)\big| dy$ follows from $(\sqrt{a} - \sqrt{b})^2 \leq |a - b|$ for any $a, b \in \mathbb{R}$ (this shows also that the integral h_t^2 in (6) exists). In order to show (i), we make use of the fact that the squared Hellinger distance is bounded by the *Kullback-Leibler divergence*:

$$\int \Big(\sqrt{\mu(y)} - \sqrt{\nu(y)}\Big)^2 dy \leq \int \mu(y) \ln \frac{\mu(y)}{\nu(y)} dy \tag{12}$$

for any two probability densities μ and ν on $\mathbb{R}$ (see e.g. [BM98, p. 178]). So we only need to establish the corresponding bound for the Kullback-Leibler divergence and show

$$\begin{aligned} D_x(\mu\|\bar\varrho) \;&:= \sum_{t=1}^n \mathbf{E} \int \mu(y_t|x_{1:t}, y_{<t}) \ln \frac{\mu(y_t|x_{1:t}, y_{<t})}{\bar\varrho(y_t|x_{1:t}, y_{<t})} dy_t \\ &\leq \;\; w_\mu^{-1} + \ln w_\mu^{-1} \end{aligned} \tag{13}$$

for all $n \geq 1$. In the following computation, we take $x_{<\infty}$ to be fixed and suppress it in the notation, writing e.g. $\mu(y_t|y_{<t})$ instead of $\mu(y_t|x_{1:t}, y_{<t})$. Then

$$\begin{aligned} D_x(\mu\|\bar\varrho) \;&=\; \sum_t \mathbf{E} \ln \frac{\mu(y_t|y_{<t})}{\bar\varrho(y_t|y_{<t})} \qquad (14) \\ &=\; \sum_t \mathbf{E} \left[\ln \frac{\mu(y_t|y_{<t})}{\varrho(y_t|y_{<t})} + \ln \frac{\int \varrho(y_{1:t}) dy_t}{\varrho(y_{<t})} \right]. \end{aligned}$$

The first part of the last term is bounded by

$$\begin{aligned} \sum_t \mathbf{E} \ln \frac{\mu(y_t|y_{<t})}{\varrho(y_t|y_{<t})} \;&=\; \mathbf{E}\; \ln \prod_{t=1}^n \frac{\mu(y_t|y_{<t})}{\varrho(y_t|y_{<t})} \qquad (15) \\ &=\; \mathbf{E}\; \ln \frac{\mu(y_{1:n}|x_{1:n})}{\varrho(y_{1:n}|x_{1:n})} \\ &\leq\; \ln w_\mu^{-1}, \end{aligned}$$

since always $\frac{\mu}{\varrho} \leq w_\mu^{-1}$. For the second part, use $\ln u \leq u - 1$ to obtain

$$\begin{aligned} &\mathbf{E} \ln \frac{\int \varrho(y_{1:t}) dy_t}{\varrho(y_{<t})} \\ &\quad \leq \sum_t \mathbf{E} \left[\frac{\int \varrho(y_{1:t}) dy_t}{\varrho(y_{<t})} - 1 \right] \end{aligned}$$

$$= \int \frac{\mu(y_{<t})(\int \varrho(y_{1:t})dy_t - \varrho(y_{<t})}{\varrho(y_{<t})} dy_{<t}$$

$$\leq w_\mu^{-1} \left[\int \varrho(y_{1:t})dy_{1:t} - \int \varrho(y_{<t})dy_{<t} \right].$$

If this is summed over $t = 1 \ldots n$, the last term is telescoping. So using $\varrho(\emptyset) = \max_\nu w_\nu \geq 0$ and $\varrho \leq \xi$, we conclude

$$\sum_t \mathbf{E} \ln \frac{\int \varrho(y_{1:t})dy_t}{\varrho(y_{<t})} \leq w_\mu^{-1} \left[\int \varrho(y_{1:n})dy_{1:n} - \varrho(\emptyset) \right]$$

$$\leq w_\mu^{-1} \int \xi(y_{1:n})dy_{1:n} \qquad (16)$$

$$= w_\mu^{-1}.$$

Hence, (14), (15), and (16) show together (13). □

We may for example apply the result for the static predictions in a Gaussian noise class $\mathcal{C} \in \mathcal{F}^{\text{Gauss}}$.

Corollary 5 *Let $\mathcal{C} \in \mathcal{F}_{\geq}^{\text{Gauss}}$ [see (2)] then the mean and the variance of the static MDL predictions converge to their true values almost surely. The same holds for $\mathcal{C} \in \mathcal{F}^{\text{Gauss}}$. In particular, if the variance of all models in $\mathcal{C}$ is the same value σ^2, then $\sum_t 2\big[1 - \exp(-\frac{(g^*(x_t|\ldots) - f(x_t))^2}{8\sigma^2})\big] \leq 21 w_\mu^{-1}$, where $f(x_t)$ is the mean value of the true distribution and $g^* = \arg\min_{f_i} \{\frac{1}{n-1}\sum_{t=1}^{n-1}(y_t - f_i(x_t))^2 + 2\sigma^2 \ln w_i^{-1}\}$ is the mean of the MDL predictor.*

For $\mathcal{C} \in \mathcal{F}_{\geq}^{\text{Gauss}}$, almost sure convergence holds since otherwise the cumulative Hellinger distances would be infinite, see (7). This generalizes to $\mathcal{C} \in \mathcal{F}^{\text{Gauss}}$; compare the footnote 2 on page 4. In the case of constant variance, the cumulative Hellinger distances can be explicitly stated as above. Note that since $1 - \exp\big(-\frac{(g^*(x_t|\ldots) - f(x_t))^2}{8\sigma^2}\big) \approx \frac{(g^*(x_t|\ldots) - f(x_t))^2}{8\sigma^2}$ for small $(g^*(x_t|\ldots) - f(x_t))^2$, this implies convergence of g^* to f faster than $O(\frac{1}{\sqrt{t}})$ if the convergence is monotone. Moreover, deviations of a fixed magnitude can only occur finitely often.

Compared with the bound for the Bayes mixture in Theorem 2, MDL bounds are exponentially larger. The bounds are sharp, as shown in [PH04a, Example 9], this example may be also adapted to the regression framework.

3. Classification

The classification setup is technically easier, since only a finite co-domain $\mathcal{Y}$ has to be considered. Results corresponding to Theorem 3 and Corollary 4 follow analogously. Alternatively, one may conditionalize the results for sequence prediction in [PH04a] with respect to the input sequence $x_{<\infty}$, arriving equally at the assertions for classification. The results in [PH04a] are formulated in terms of mean (square) sum convergence instead of Hellinger sum convergence. On finite co-domain, these two convergence notions induce the same topology.

Theorem 6 *Let $\mathcal{X}$ be arbitrary and $\mathcal{Y}$ be a finite set of class labels. $\mathcal{C} = \{\nu_i : i \geq 1\}$ consists of classification models, i.e. for each $\nu \in \mathcal{C}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we have $\nu(y|x) \geq 0$ and $\sum_y \nu(y|x) = 1$. Each model ν is associated with a prior weight $w_\nu > 0$, and $\sum_\nu w_\nu = 1$ holds. Let the MDL predictions be defined analogously to (8), (9) and (10) (the difference being that here probabilities are maximized instead of densities). Assume that $\mu \in \mathcal{C}$, where μ is the true distribution. Then for each $x_{<\infty} \in \mathcal{X}^\infty$,*

$$\sum_{t=1}^{\infty} \mathbf{E} \sum_{y \in \mathcal{Y}} \left(\sqrt{\mu(y|x_t)} - \sqrt{\varrho^{\text{static}}(y|x_{1:t}, y_{<t})} \right)^2 \leq 21 w_\mu^{-1},$$

$$\sum_{t=1}^{\infty} \mathbf{E} \sum_{y \in \mathcal{Y}} \left(\mu(y|x_t) - \varrho^{\text{static}}(y|x_{1:t}, y_{<t}) \right)^2 \leq 21 w_\mu^{-1}$$

holds. Similar assertions are satisfied for the normalized and the un-normalized dynamic MDL predictor. In particular, the predictive probabilities of all three MDL predictors converge to the true probabilities almost surely.

The second bound on the quadratic differences is shown in [PH04a]. The assertions about almost sure convergence follows as in (7).

4. Discussion and Conclusions

We have seen that discrete MDL has good asymptotic predictive properties. On the other hand, the loss bounds for MDL are exponential compared to the Bayes mixture loss bound. This is no proof artifact, as examples are easily constructed where the bound is sharp [PH04a].

This has an important implication for the practical use of MDL: One need to choose the underlying model class and the prior carefully. Then it can be expected that the predictions are good and converge fast: this is supported by theoretical arguments in [Ris96, PH04b]. The Bayes mixture in contrast, which can be viewed as a very large (infinite) weighted committee, also converges rapidly with unfavorable model classes, but at higher computational expenses.

One might be interested in other loss functions than the Hellinger loss. For the classification case, a bound on the expected error loss (number of classification errors) of MDL may be derived with the techniques from

[Hut04], using the bound on the quadratic distance. [Hut03a] gives also bounds for *arbitrary* loss functions, however this requires a bound on the Kullback-Leibler divergence rather than the quadratic distance. Unfortunately, this does not hold for static MDL [PH04a]. For the regression setup, analysis of other, more general or even arbitrary loss functions is even more demanding and, as far as we know, open.

Considering only discrete model classes is certainly a restriction, since many models arising in science (e.g. physics or biology) are continuous. On the other hand there are arguments in favor of discrete classes. From a computational point of view they are definitely sufficient. Real computers may even treat only finite model classes. The class of all programs on a fixed universal Turing machine is countable. It may be related to discrete classes of stochastic models by the means of semimeasures, this is one of the central issues in Algorithmic Information Theory [LV97].

References

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

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

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

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

[CY00] L. Le Cam and G. Yang. *Asymptotics in Statistics*. Springer, 2nd edition, 2000.

[GGvdV00] S. Ghosal, J. Gosh, and A. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28:500–531, 2000.

[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.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004. 300 pp, www.idsia.ch/~marcus/ai/uaibook.htm.

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

[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.

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

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

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