Convergence of Discrete MDL for Sequential Prediction

Jan Poland and Marcus Hutter

IDSIA • Lugano • Switzerland

Overview

- Sequential online prediction in a Bayesian framework
- No i.i.d. assumption!!!
- Applies to classification and to regression
- Applies to Universal Prediction in the sense of AIT
- We will obtain strong asymptotic assertions
- ... and also (weak) loss bounds

Rough Problem Setup

- Given an initial part $x = x_{1:t}$ of a sequence, predict the next symbol x_{t+1}
- Examples:
  - $x = 01010101010101$
  - $x = 1100100100001111110110101010001000100001$
  - $x = 0001111001010010001111110110101001001111$

(Semi)Measures

- This is a "binary talk", but everything also works for arbitrary alphabet!
- Let $\mathbb{B} = \{0,1\}$, $\mathbb{B}^\infty = \{$all binary sequences$\}$
- ϵ is the empty string
- A *measure* μ is a function $\mu : \mathbb{B}^* \to [0,1]$ s.t.

$$\mu(\epsilon) = 1 \text{ and } \mu(x) = \mu(x0) + \mu(x1) \text{ for all } x$$

- A *semimeasure* ν has

$$\nu(\epsilon) \leq 1 \text{ and } \nu(x) \geq \nu(x0) + \nu(x1) \text{ for all } x$$

Examples: (Semi)Measures

- $\lambda(x) = 2^{-length(x)}$ is the uniform measure
- $\mu_1(111...1) = 1$ and $\mu_1(x) = 0$ if x contains at least one 0, is a deterministic measure
- $M_U(x) =$ the probability that some universal Turing machine (UTM) U outputs a string starting with x when the input is random coin flips
- The latter is a semimeasure, not a measure, since U does not halt on each input!
- Binary classification: $\mu(1|z)$ is i.i.d. given some input z (conditionalized measure)

Classes of (Semi)Measures

- Let $\mathcal{C}$ be a *countable* class of (semi)measures
- Each $\nu \in \mathcal{C}$ is assigned a *prior weight* $w_\nu > 0$
- Kraft inequality: $\sum_{\nu \in \mathcal{C}} w_\nu \leq 1$
- Universal setup: $\mathcal{C} = \mathcal{M} \cong$ all programs on a UTM U
- $w_\nu = 2^{-K(\nu)}$ where $K(\nu)$ is the *prefix Kolmogorov Complexity* of ν, i.e. the length of the shortest self-delimiting program defining ν

Assumptions

- We make *no probabilistic* assumption on $\mathcal{C}$
- We show bounds for given *true distribution* μ
- which is a *measure* (not a semimeasure)
- *and assumed to be in* $\mathcal{C}$
- Thus, bounds depend on the complexity (or prior weight w_μ) of the true distribution
- Occam's razor
- priors correspond to regularization

Bayes Mixtures

- We denote a Bayes mixture by ξ
- Given observation x and a countable class together with weights (w_ν) , the ξ-prediction is

$$\xi(a|x) = \frac{\sum_\nu w_\nu \nu(xa)}{\sum_\nu w_\nu \nu(x)}$$

  for $a \in \{0, 1\}$.
- ξ is semimeasure
- "Committee of all models"

Minimum Description Length

- Minimum Description Length (MDL) estimator

$$\nu^x = \arg\max\{w_\nu \nu(x)\}$$
$$\varrho(x) = \max\{w_\nu \nu(x)\}$$

- ν^x is *maximizing element*
- $-\log\varrho(x) = \min\{-\log w_\nu - \log\nu(x)\}$
- $-\log w_\nu \ \leftrightarrow$ code of the model
- $-\log\nu(x) \ \leftrightarrow$ code of data given

Prediction using MDL

- Dynamic MDL predictor: $\varrho(a|x) = \frac{\varrho(xa)}{\varrho(x)}$
  not a semimeasure!
- Normalized dynamic MDL: $\varrho(a|x) = \frac{\varrho(xa)}{\varrho(x0)+\varrho(x1)}$
  measure
  search new model for each next symbol
- Static MDL predictor: $\varrho^x(a|x) = \frac{\nu^x(xa)}{\nu^x(x)}$
  (semi)measure
  find best model and use this for prediction
- $\Rightarrow$ Static MDL is computationally more efficient

Bayes Mixture Predictions

- **Theorem** (Solomonoff): Let $\mu \in \mathcal{C}$ be a measure, then

$$\sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \xi(a|x_{1:t})\big)^2 \leq \ln(w_\mu^{-1})$$

- $\Rightarrow$ The posteriors *almost surely* converge to the true probabilities *fast*
- Universal setup: μ must be a computable measure
- This requirement is (philosophically) very weak

Proof of Solomonoff's Theorem

$$\sum_{t=0}^{T} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \xi(a|x_{1:t})\big)^2$$

$$\leq \sum_{t=0}^{T} \mathbf{E} \sum_{a\in\{0,1\}} \mu(a|x_{1:t}) \ln \frac{\mu(a|x_{1:t})}{\xi(a|x_{1:t})} = \sum_{t=0}^{T} \mathbf{E} \ln \frac{\mu(x_t|x_{1:t})}{\xi(x_t|x_{1:t})}$$

$$= \mathbf{E} \ln \left(\prod_{t=0}^{T} \frac{\mu(x_t|x_{1:t})}{\xi(x_t|x_{1:t})} \right) = \mathbf{E} \ln \frac{\mu(x_{1:T+1})}{\xi(x_{1:T+1})} \leq \ln w_\mu^{-1}$$

Lemma:
The quadratic distance is bounded by the relative entropy.

Observation:
x dominates μ, i.e.
$x(x) \geq w_\mu\, \mu(x)$ for all x

MDL: Main Theorem

Theorem: $\mu \in \mathcal{C}$ measure, then

$$(i) \qquad \sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \varrho_{\text{norm}}(a|x_{1:t})\big)^2 \leq \ln w_\mu^{-1} + w_\mu^{-1},$$

normalized dynamic

$$(ii) \qquad \sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \varrho(a|x_{1:t})\big)^2 \leq 8 \cdot w_\mu^{-1},$$

dynamic

$$(iii) \qquad \sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \varrho^{x_{1:t}}(a|x_{1:t})\big)^2 \leq 21 \cdot w_\mu^{-1}$$

static

$\Rightarrow$ The posteriors *almost surely* converge to the true probabilities, but convergence is *slow* in general

Proof Idea

- For ϱ_{norm}:
  - use relative entropy bound
  - decompose ϱ_{norm} in ϱ and normalizer
  - ϱ-contribution bounded by $\ln w_\mu^{-1}$
  - normalizer contribution bounded by w_μ^{-1}
- Then bound the cumulative absolute difference $|\varrho - \varrho_{\mathrm{norm}}|$ by $2w_\mu^{-1}$
- Finally bound the cumulative absolute difference $|\varrho^x - \varrho|$ by $3w_\mu^{-1}$
- square distances may be chained

Loss Bounds

- **Theorem** (Hutter): $\mu \in \mathcal{C}$ measure $\Rightarrow$

$$L^{\xi}(T) \leq L^{\mu}(T) + 2\sqrt{L^{\mu}(T)\ln w_{\mu}^{-1}} + 2\ln w_{\mu}^{-1}$$

for $0/1$ los and arbitrary loss

- **Corollary**: For arbitrary loss,

$$L^{\varrho_{\text{norm}}}(T) \leq L^{\mu}(T) + O(\sqrt{L^{\mu}(T) w_{\mu}^{-1}}) + O(w_{\mu}^{-1})$$

Loss Bounds

- **Corollary**: For $0/1$ loss,

$$L^{\varrho}(T) \leq L^{\mu}(T) + O(\sqrt{L^{\mu}(T) w_{\mu}^{-1}}) + O(w_{\mu}^{-1})$$

$$L^{\varrho^x}(T) \leq L^{\mu}(T) + O(\sqrt{L^{\mu}(T) w_{\mu}^{-1}}) + O(w_{\mu}^{-1})$$

- Arbitrary loss open!
- Compare to prediction with expert advice: *worst-case* loss for *individual* sequences

$$L^{PEA}(T) \leq L^{\mu}(T) + 2\sqrt{2L^{\mu}(T) \ln w_{\mu}^{-1}} + O(\ln w_{\mu}^{-1})$$

Exponential Bounds are Sharp

- MDL bound exponentially worse than Bayes mixture
- This bound is sharp! Example $\nu_1,...,\nu_7,\nu_8 = \mu$ *deterministic*,

$$\begin{aligned}
\nu_1 &: 0\ 0\ 0\ 0\ 0\ 0\ 0 \ldots, & w_1 = \tfrac{1}{8} \\
\nu_2 &: 1\ 0\ 0\ 0\ 0\ 0\ 0 \ldots, & w_2 = \tfrac{1}{8} \\
\nu_3 &: 1\ 1\ 0\ 0\ 0\ 0\ 0 \ldots, & w_3 = \tfrac{1}{8} \\
\nu_4 &: 1\ 1\ 1\ 0\ 0\ 0\ 0 \ldots, & w_4 = \tfrac{1}{8} \\
\nu_5 &: 1\ 1\ 1\ 1\ 0\ 0\ 0 \ldots, & w_5 = \tfrac{1}{8} \\
\nu_6 &: 1\ 1\ 1\ 1\ 1\ 0\ 0 \ldots, & w_6 = \tfrac{1}{8} \\
\nu_7 &: 1\ 1\ 1\ 1\ 1\ 1\ 0 \ldots, & w_7 = \tfrac{1}{8} \\
\mu = \nu_8 &: 1\ 1\ 1\ 1\ 1\ 1\ 1 \ldots, & w_8 = \tfrac{1}{8}
\end{aligned}$$

Exponential Bounds are Sharp

- Then normalized dynamic MDL predicts probability of $\frac{1}{2}$ for $t = 1, \ldots, 7$
- $\rightarrow$ cumulative error $= O(w_\mu^{-1})$
- The bound is even sharp if $\mathcal{C}$ contains only Bernoulli distributions!
- But there under additional mild conditions, a good bound holds

Hybrid MDL predictions

- Hybrid MDL predictor: $\varrho^{hybrid}(a|x) = \frac{\nu^{xa}(xa)}{\nu^{x}(x)}$
- "Dynamic MDL but drop weights"
- Predictive properties? Poorer!
- Only converges if the maximizing element *stabilizes*
- This happens almost surely if
  - all (semi)measures in $\mathcal{C}$ are independent of the past (factorizable)
  - μ is uniformly stochastic, i.e. in each time step either deterministic or noisy with at least a certain amplitude

Complexity and Randomness

Universal case: $\mathcal{C} = \mathcal{M}$, and $\tilde{\mathcal{C}}$ is $\mathcal{C}$ restricted to computable measures

$$\Rightarrow 2^{Km(x)} \overset{\times}{=} \tilde{\varrho}(x) \overset{\times}{\leq} \tilde{\xi}(x) \overset{\times}{\leq} \varrho(x) \overset{\times}{=} \xi(x) \overset{\times}{=} M(x)$$

Gács: $\not\overset{\times}{=}$ $\Rightarrow$ which inequality is proper?

$\Rightarrow$ all quantities define Martin-Löf randomness by $f(x_{1:n}) \leq C\mu(x_{1:n})$ for all n and some C

Further Open Problems

- Between MDL and Bayes mixture?
- Active Learning?
- Other ideas?

- That's it, thank you!