On the Convergence Speed of MDL Predictions for Bernoulli Sequences

or

Is MDL Really So Bad?

Jan Poland and Marcus Hutter

IDSIA • Lugano • Switzerland

Bernoulli Classes

- Set of *parameters* $\Theta = \{\vartheta_1, \vartheta_2, \ldots\} \subset [0,1]$
- *Weights* w_ϑ for each $\vartheta \in \Theta$
- Weights correspond to *codes*: $w_\vartheta = 2^{-\ell(\texttt{Code}_\vartheta)}$

ϑ	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$	1
w	$\frac{1}{4}$	$\frac{1}{64}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{4}$	$\frac{1}{64}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{4}$
`Code`	`00`	`111000`	`1100`	`111001`	`10`	`111010`	`1101`	`111011`	`01`

$$\texttt{Code} = \underbrace{\texttt{111}}_{\texttt{1+\#bits}}\ \underbrace{\texttt{0}}_{\texttt{stop}}\ \underbrace{\texttt{10}}_{\texttt{data}}$$

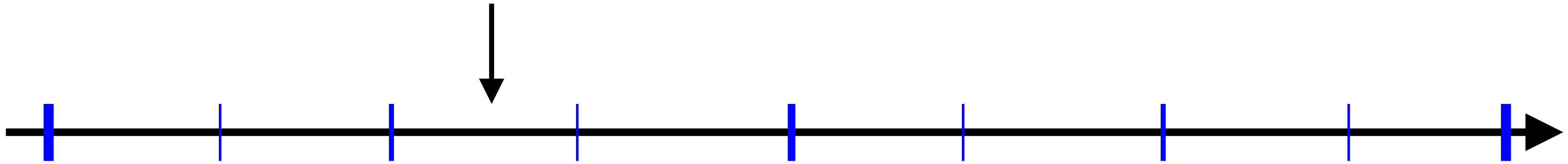
Estimators

- Given observed sequence $x = x_1 x_2 \ldots x_n$
- Probability of x given ϑ:
  $$p_\vartheta(x) = \vartheta^{\#\text{ones}(x)}(1-\vartheta)^{n-\#\text{ones}(x)}$$
- *Posterior weights* $w_\vartheta(x) = \dfrac{w_\vartheta p_\vartheta(x)}{\sum_\vartheta w_\vartheta p_\vartheta(x)}$
- *Bayes mixture* $\xi(x) = \sum_\vartheta w_\vartheta(x)\vartheta$
- *MDL/MAP* $\vartheta^*(x) = \arg\max_\vartheta w_\vartheta(x)\vartheta$
- *Maximum Likelihood* (ML): Same as MAP, but with prior weights set to 1

An Example Process

True parameter

$\vartheta_0 = \frac{5}{16} = 0.3125$



Sequence x		0	1	0	0000011	...(32)....(640)...	
Bayes mixture ξ	0.5	0.21	0.5	0.45	0.4	0.27	0.3
ML estimate	0	0	0.5	0.34	5/16	0.25	5/16
MAP (MDL) θ^*	0	0	0.5	0.5	0.5	0.25	5/16

What We Know

- Let $\vartheta_0 \in \Theta$ be the true parameter with weight w_0
- ξ *converges* to ϑ_0 *almost surely* and *fast*,

  precisely $\sum_{t=0}^{\infty} \mathbf{E}(\xi - \vartheta_0)^2 \leq \ln(w_0^{-1})$
- ϑ^* converges to ϑ_0 almost surely and *in general slow*,

  precisely $\sum_{t=0}^{\infty} \mathbf{E}(\vartheta^* - \vartheta_0)^2 \leq O(w_0^{-1})$
- Even true for arbitrary non-i.i.d. (semi-) measures!
- The ML estimates converge to ϑ_0 almost surely,
  no such assertion about convergence speed possible

Is MDL Really So Bad?

- Bayes mixture bound is $\texttt{description length}(\vartheta_0)$
- MDL bound is $\exp(\texttt{description length}(\vartheta_0))$
- $\Rightarrow$ MDL is exponentially worse *in general*
- This is also a loss bound!
- How about *simple* classes?
- Deterministic classes: can show bound

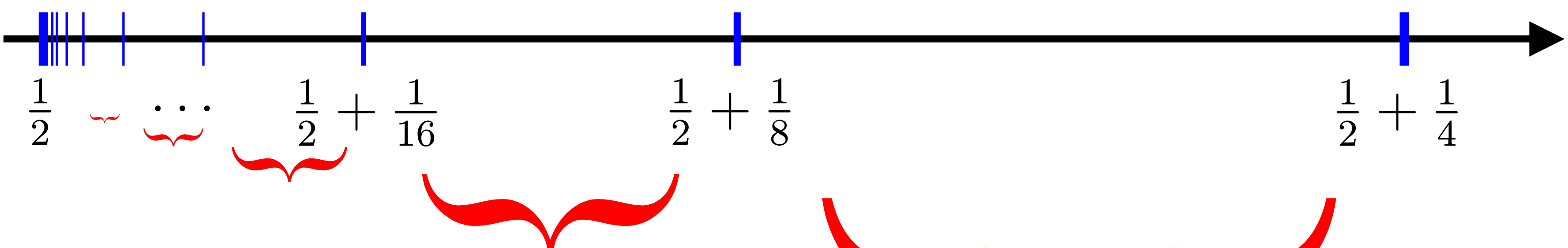
$$\texttt{huge constant} \times (\texttt{description length}(\vartheta_0))^3$$

- Simple stochastic classes, e.g. *Bernoulli*?

MDL Is Really So Bad!

$\sum_t \mathbf{E}(\vartheta^* - \vartheta_0)^2 = O(w_0^{-1})$ in the following *example*:

N parameters, $w_\vartheta = \frac{1}{N}$ for all ϑ, $\vartheta_0 = \frac{1}{2}$



$$\sum_t \mathbf{E}(\vartheta^* - \vartheta_0)^2 \mathbf{1}_{\vartheta^* \in [\frac{1}{2}+\frac{1}{16}, \frac{1}{2}+\frac{1}{8}]} = O(1)$$

$$\sum_t \mathbf{E}(\vartheta^* - \vartheta_0)^2 \mathbf{1}_{\vartheta^* \in [\frac{1}{2}+\frac{1}{8}, \frac{1}{2}+\frac{1}{4}]} = O(1)$$

MDL Is Not That Bad!

- The *instantaneous* loss bound is good,

  precisely $\mathbf{E}\,(\vartheta^* - \vartheta_0)^2 \le \frac{1}{n} O\big(\ln(w_0^{-1})\big)$

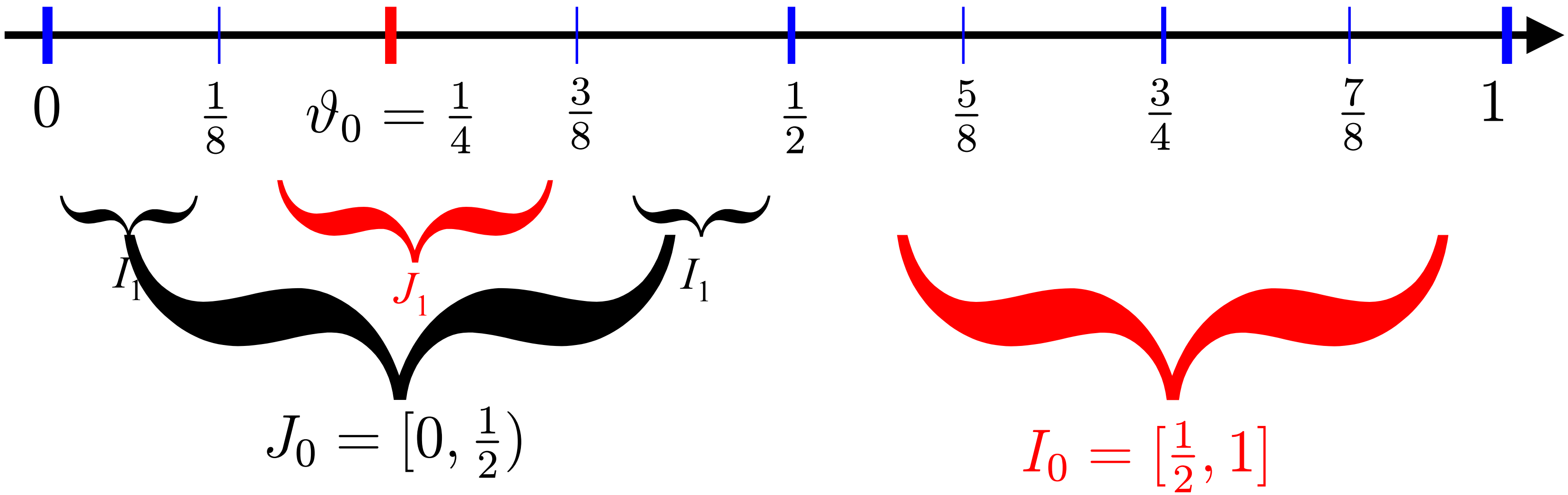
- This does not imply a finitely bounded cumulative loss!

- The cumulative loss bound is good for certain *nice classes* (parameters+weights)

- Intuitively: Bound is good if parameters of equal weights are uniformly distributed

Prepare Sharper Upper Bound

- Define interval construction (I_k, J_k) which exponentially contracts to ϑ_0
- Let $K(I_k)$ be the shortest description length of some $\vartheta \in I_k$



Sharper Upper Bound

- Let $K(J_k)$ be the shortest description length of some $\vartheta \in J_k$
- Let $\Delta(k) = \max\{K(I_k) - K(J_k), 0\}$
- Theorem:

$$\sum_t \mathbf{E}(\vartheta^* - \vartheta_0)^2 \leq O\Big(\ln w_0^{-1} + \sum_{k=1}^{\infty} 2^{-\Delta(k)} \sqrt{\Delta(k)}\Big)$$

- Corollaries: “Uniformly distributed weights $\Rightarrow$ good bounds

The Universal Case

- $\Theta = \{$all computable $\vartheta \in [0,1]\}$
- $w_\vartheta = 2^{-K(\vartheta)}$, where K denotes the prefix Kolmogorov complexity
- $\sum_k 2^{-\Delta(k)} \sqrt{\Delta(k)} = \infty \;\Rightarrow$ Theorem not applicable
- Conjecture: $\sum\limits_t \mathbf{E}(\vartheta^* - \vartheta_0)^2 \leq O\big(\ln w_0^{-1} + \sum\limits_{k=1}^{\infty} 2^{-\Delta(k)}\big)$
- $\Rightarrow$ bound `huge constant`$\times$`polynomial` holds for incompressible ϑ_0
- Compare to determistic case

Conclusions

- Cumulative and instantaneous bounds are *incompatible*
- Main positive generalizes to arbitrary i.i.d. classes
- Open problem: good bounds for more general classes?
- Thank you!