

Computable Bayesian Compression for Uniformly Discretizable Statistical Models

Łukasz Dębowski
debowski@cwi.nl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

ALT 2009, Porto

Introduction

- Statistical model **kernel** $\mathbf{P} : \Theta \times \mathbb{X}^* \ni (\theta, x) \mapsto \mathbf{P}_\theta(x)$.
- **Prior** \mathbf{Q} with the **support set** $\Theta \subset \mathbb{X}^{\mathbb{N}}$.
- Suppose that \mathbf{P} and \mathbf{Q} are **recursive**.

Consider the **Bayesian measure** $\mathbf{Y} = \int \mathbf{P}_\theta d\mathbf{Q}(\theta)$.

- 1 Vovk and V'yugin (1993, 1994), Takahashi (2008):

\mathbf{Y} gives the best enumerable
compression of **algorithmically**
 \mathbf{P}_θ -**random** data



parameter θ is
algorithmically
Q-random

— *Weaker results: van Lambalgen (1987).*

- 2 We wanted to show the **converse** (\iff),
which holds if the parameter is '**effectively learnable**'.

— *A similar result can be inferred from V'yugin (2007).*

The paradigm of computation for mathematical statistics

Uniform computation with an oracle:

- 1 \mathbb{X} is a **countable** alphabet, $\mathbb{Y} \subset \mathbb{X}$ is a **finite** alphabet;
- 2 **data** $\mathbf{x} \in \mathbb{X}^{\mathbb{N}}$ and their finite prefix $x^n \in \mathbb{X}^n$;
 - a real vector is coded as a sequence;
- 3 **parameter** $\theta \in \mathbb{X}^{\mathbb{N}}$ and its finite prefix $\theta^m \in \mathbb{X}^m$;
 - a real vector is coded as a sequence;
- 4 **programs** are strings from a **prefix-free** subset of \mathbb{Y}^* ;
 - \log — logarithm to the **base** $|\mathbb{Y}|$;
- 5 **algorithmic complexity** $K(x^n|\delta)$ is the minimal length of a program that computes x^n **given** $\delta \in \mathbb{X}^* \cup \mathbb{X}^{\mathbb{N}}$;
- 6 function $f : \mathbb{X}^* \cup \mathbb{X}^{\mathbb{N}} \rightarrow \mathbb{R}$ is called
 - **recursive** if some program, **given any** α and d , computes $f(\alpha)$ up to a finite precision d ;
 - **enumerable** if some program, **given any** α and k , computes k -th lower approximation of $f(\alpha)$.

Martin-Löf random sequences for a measure \mathbf{Y}

$$\mathcal{L}_{\mathbf{Y}} = \left\{ \mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \inf_{n \in \mathbb{N}} [K(x^n) + \log \mathbf{Y}(x^n)] > -\infty \right\}$$

The set of M-L random sequences $\mathcal{L}_{\mathbf{Y}}$ is the set of sequences which cannot be better compressed by another enumerable semimeasure.

We have $\mathcal{L}_{\mathbf{Y}} := \{ \mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \mathcal{I}(\mathbf{x}; \mathbf{Y}) < \infty \}$, where

impossibility level $\mathcal{I}(\mathbf{x}; \mathbf{Y}) := \inf_n |\mathbb{Y}|^{-K(x^n)} / \mathbf{Y}(x^n)$ satisfies

$$\mathcal{I}(\mathbf{x}; \mathbf{Y}) \stackrel{*}{<} \liminf_{n \rightarrow \infty} \frac{\mathbf{M}(x^n)}{\mathbf{Y}(x^n)} \stackrel{*}{<} \sup_{n \in \mathbb{N}} \frac{\mathbf{M}(x^n)}{\mathbf{Y}(x^n)} \stackrel{*}{<} [\mathcal{I}(\mathbf{x}; \mathbf{Y})]^{1+\epsilon}$$

for an $\epsilon > 0$ and the **universal enumerable semimeasure** \mathbf{M} .

We have $\mathbf{U}(x^n) \stackrel{*}{<} \mathbf{M}(x^n)$ for any enumerable (semi)measure \mathbf{U} .

Algorithmic randomness for $\mathbf{Y} = \int \mathbf{P}_\theta d\mathbf{Q}(\theta)$

The set of Martin-Löf random sequences for a measure \mathbf{Y} :

$$\mathcal{L}_{\mathbf{Y}} = \left\{ \mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \inf_{n \in \mathbb{N}} [K(x^n) + \log \mathbf{Y}(x^n)] > -\infty \right\}.$$

We have $\mathbf{Y}(\mathcal{L}_{\mathbf{Y}}) = 1$ by Barron's inequality and hence

$$\mathbf{P}_\theta(\mathcal{L}_{\mathbf{Y}}) = 1 \text{ for almost all } \theta.$$

What does $\mathbf{P}_\theta(\mathcal{L}_{\mathbf{Y}})$ equal for a particular θ ?

The sets of conditionally Martin-Löf random data

The set of **conditionally** random sequences for a measure \mathbf{P}_θ :

$$\mathcal{L}_{\mathbf{P}_\theta} = \left\{ \mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \inf_{n \in \mathbb{N}} [K(x^n | \theta) + \log \mathbf{P}_\theta(x^n)] > -\infty \right\}.$$

Since $\mathbf{P}_\theta(\mathcal{L}_{\mathbf{P}_\theta}) = 1$ by Barron's inequality, we will obtain

$$\mathbf{P}_\theta(\mathcal{L}_Y) = \begin{cases} 1 & \text{if } \mathcal{L}_{\mathbf{P}_\theta} \subset \mathcal{L}_Y, \\ 0 & \text{if } \mathcal{L}_{\mathbf{P}_\theta} \cap \mathcal{L}_Y = \emptyset. \end{cases}$$

The chain rule for impossibility levels

Vovk & V'yugin, 1993:

For recursive \mathbf{P} and \mathbf{Q} , we have

$$\inf_{\theta \in \Theta} [\mathcal{I}(\mathbf{x}; \mathbf{P}|\theta) \mathcal{I}(\theta; \mathbf{Q})] <^* \mathcal{I}(\mathbf{x}; \mathbf{Y}) <^* \inf_{\theta \in \Theta} [\mathcal{I}(\mathbf{x}; \mathbf{P}|\theta) [\mathcal{I}(\theta; \mathbf{Q})]^{1+\epsilon}]$$

with **cond. impossibility level** $\mathcal{I}(\mathbf{x}; \mathbf{P}|\theta) := \inf_n |\mathbb{Y}|^{-K(x^n|\theta)} / \mathbf{P}_\theta(x^n)$.

Hence

$$\mathcal{L}_Y \supset \bigcup_{\theta \in \mathcal{L}_Q} \mathcal{L}_{\mathbf{P}|\theta}$$

and

$$\mathbf{P}_\theta(\mathcal{L}_Y) = 1 \text{ if } \theta \in \mathcal{L}_Q.$$

Further results for M-L random sequences

The **decomposition** of \mathcal{L}_Y (Takahashi, 2008):

$$\mathbf{P} \text{ and } \mathbf{Q} \text{ are recursive} \implies \mathcal{L}_Y = \bigcup_{\theta \in \mathcal{L}_Q} \mathcal{L}_{\mathbf{P}|\theta}$$

An **effectively strictly consistent estimator** (V'yugin, 2007):

$$\forall \theta \in \Theta, \epsilon > 0, \delta > 0 \quad \mathbf{P}_\theta \left(\mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \sup_{n \geq N(\epsilon, \delta)} |\mathbf{E}(x^n) - \theta| > \epsilon \right) < \delta$$

for model \mathbf{P} , estimator \mathbf{E} , and sample size N being **recursive**

$$\implies \mathcal{L}_{\mathbf{P}|\theta} \subset \{ \mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \lim_{n \rightarrow \infty} \mathbf{E}(x^n) = \theta \}$$

\implies sets $\mathcal{L}_{\mathbf{P}|\theta}$ are disjoint for distinct θ

$$\text{Then, } \mathbf{P}_\theta(\mathcal{L}_Y) = \begin{cases} 1 & \text{if } \theta \in \mathcal{L}_Q, \\ 0 & \text{if } \theta \in \Theta \setminus \mathcal{L}_Q. \end{cases}$$

Uniformly discretizable statistical models

For $A(y^m) := \{\theta \in \Theta : y^m \text{ is the prefix of } \theta\}$, use measures

$$\mathbf{T}(x^n, y^m) := \int_{A(y^m)} \mathbf{P}_\theta(x^n) d\mathbf{Q}(\theta),$$

$$\mathbf{Y}(x^n) := \mathbf{T}(x^n, \lambda) = \int \mathbf{P}_\theta(x^n) d\mathbf{Q}(\theta).$$

A **Bayesian model** (\mathbf{P}, \mathbf{Q}) is called (μ, ν) -**uniformly discretizable** if functions $\mu, \nu : \mathbb{N} \rightarrow \mathbb{R}$ are nondecreasing and

$$\lim_{n \rightarrow \infty} \frac{\log [\mathbf{Q}(\theta^m) \mathbf{P}_\theta(x^n) / \mathbf{T}(x^n, \theta^m)]}{\log m} = 0, \quad m \geq \mu(n), \quad (\text{likelihood d.})$$
$$\lim_{m \rightarrow \infty} \mathbf{T}(x^n, \theta^m) / \mathbf{Y}(x^n) = 1, \quad n \geq \nu(m), \quad (\text{posterior d.})$$

for all $\theta \in \Theta$ and \mathbf{P}_θ -almost all \mathbf{x} .

$\mu(n)$ -uniformly discretizable = (μ, ν) -u.d. with $\mu(\nu(m)) \leq m^\alpha$.
 $\mu(n)$ appears to be close to **redundancy** $-\log \mathbf{Y}(x^n) + \log \mathbf{P}_\theta(x^n)$.

Examples of uniformly discretizable models

- 1 Exponential families

$$\mathbf{P}_\theta(x^n) := \prod_{i=1}^n \exp[\beta T(x_i) - \psi(\beta)] p(x_i), \quad \theta := \psi'(\beta),$$

are $(\frac{1+\epsilon}{2}) \log n$ -uniformly discretizable.

- 2 Measures of processes $(X_i)_{i \in \mathbb{Z}} = (K_i, \theta_{K_i})_{i \in \mathbb{Z}}$, i.e.,

$$\mathbf{P}_\theta(x^n) := \prod_{i=1}^n \mathbf{1}_{\{z_i = \theta_{k_i}\}} k_i^{-\alpha} / \zeta(\alpha), \quad x_i = (k_i, z_i), \quad \alpha > 1,$$

are $n^{2\beta/(1-\beta)+\epsilon}$ -uniformly discretizable, $\beta = \alpha^{-1}$, $\beta \in (0, 1)$.

(Mutual information $\mathbf{E} [-\log \mathbf{Y}(x^n) + \log \mathbf{P}_\theta(x^n)] = \Theta(n^\beta)$.)

- 3 The model $\mathbf{P}_\theta(x^n) := \mathbf{1}_{\{x^n = \theta^n\}}$ is n -uniformly discretizable.
- 4 The singleton model is 0-uniformly discretizable.

Results for approximately random sequences

$$\mathcal{L}_{\mathbf{Y},g(n)} = \left\{ \mathbf{x} \in \mathbb{X}^\infty : \inf_{n \in \mathbb{N}} \frac{K(x^n) + \log \mathbf{Y}(x^n)}{g(n)} > -\infty \right\}$$

Theorem

For a (μ, ν) -*uniformly discretizable* Bayesian model (\mathbf{P}, \mathbf{Q}) with $\mu(\nu(m)) \leq m^\alpha$ and *recursive* $\mathbf{P}, \mathbf{Q}, \nu$, we have

- 1 $\mathbf{P}_\theta(\mathcal{L}_{\mathbf{P}, \log \mu(n)}) = \begin{cases} 1 & \text{if } \theta \in \mathcal{L}_{\mathbf{Q}, \log n}, \\ 0 & \text{if } \theta \in \Theta \setminus \mathcal{L}_{\mathbf{Q}, \log n}. \end{cases}$
- 2 $\mathbf{P}_\theta(\mathcal{L}_{\mathbf{P}}) = 0$ if $\theta \in \Theta \setminus \mathcal{L}_{\mathbf{Q}}$.

The proof applies *chain rules* for *algorithmic complexity* and *Shannon-Fano coding*. Coding the *length* of discretized parameter and *complexity of its complexity* implies losing $(3 + \epsilon) \log \mu(n)$ bits.

A comparison of definitions

Assume $\theta \in \mathbb{Y}^{\mathbb{N}}$ and put $|\theta - \theta'| := \left| \sum_{k=1}^{\infty} (\theta_k - \theta'_k) |\mathbb{Y}|^{-k} \right|$.

- **Uniformly discretizable models:** Suppose that for \mathbf{P}_{θ} -almost all \mathbf{x}

$$\lim_{m \rightarrow \infty} \mathbf{T}(x^n, \theta^m) / \mathbf{Y}(x^n) = 1, \quad n \geq \nu(m).$$

For $\sigma(x^n) := \sup \{m : n \geq \nu(m)\}$ set the **discrete MAP estimator**

$$\mathbf{E}(x^n) \in \operatorname{argmax}_{\theta \in \Theta} \mathbf{T}(x^n, \theta^{\sigma(x^n)}).$$

- **Effectively strictly consistent estimators:**

$$\forall \theta \in \Theta, \epsilon > 0, \delta > 0 \quad \mathbf{P}_{\theta} \left(\mathbf{x} \in \mathbb{X}^{\mathbb{N}} : \sup_{n \geq N(\epsilon, \delta)} |\mathbf{E}(x^n) - \theta| > \epsilon \right) < \delta.$$

Put $\sigma(x^n) := \sup \{m : n \geq N(|\mathbb{Y}|^{-m}, f(m))\}$ for $\sum_m f(m) < \infty$.

In both cases, $\limsup_n |\mathbf{E}(x^n) - \theta| / |\mathbb{Y}|^{-\sigma(x^n)} \leq 1$ for \mathbf{P}_{θ} -almost all \mathbf{x} .

But this need not hold for **all** $\mathbf{x} \in \mathcal{L}_{\mathbf{P}|\theta}$ in the first case.

Effective Borel-Cantelli lemmas work for concrete models (Davie, 2001).

Countable unions of models

Let models $(\mathbf{P}^i, \mathbf{Q}^i)$ be (μ_i, ν_i) -uniformly discretizable on Θ^i , $\mathbf{P}^i \perp \mathbf{P}^j$.

- $\Theta := \bigcup_{i \in A} c(i)\Theta^i$ for a prefix code $c : A \rightarrow \mathbb{X}^+$,
- $\text{idx}(\theta) := i$ and $\text{trn}(\theta) := \vartheta$ for $\theta = c(i)\vartheta \in \Theta$,
- $\mathbf{P}_\theta(x) := \mathbf{P}_{\text{trn}(\theta)}^{\text{idx}(\theta)}(x)$ for $\theta \in \Theta$,
- $\mathbf{Q} := \sum_{i \in A} w(i)(\mathbf{Q}^i \circ \text{trn})$ for a metaprior $w(i) > 0$, $\sum_{i \in A} w(i) = 1$.

The model (\mathbf{P}, \mathbf{Q}) is (μ, ν) -uniformly discretizable provided

$$\infty > \mu(n) := \sup_{i \in A} (|c(i)| + \mu_i(n)), \quad (1)$$

$$\infty > \nu(m) := \sup_{i \in A} \nu_i(m - |c(i)|), \quad (2)$$

$$\lim_{n \rightarrow \infty} \frac{w(\text{idx}(\theta)) \mathbf{Y}^{\text{idx}(\theta)}(x^n)}{\mathbf{Y}(x^n)} = 1 \text{ for } \mathbf{P}_\theta\text{-almost all } \mathbf{x} \text{ and all } \theta \in \Theta. \quad (3)$$

Takahashi (2008) \implies (3) holds true for θ that are \mathbf{Q} -random.

Csiszar & Shields (2000) \implies (3) may fail for θ that are not \mathbf{Q} -random.

Does (1) and (2) imply (3) for all $\theta \in \Theta$?

Conclusion

When a Bayesian statistician **believes** that θ is **distributed** according to the prior \mathbf{Q} , it simply means that they **assume** θ to be **algorithmically random** with respect to \mathbf{Q} .

A prior on **learnable** parameters should resume **falsifiable beliefs** about the **algorithmic complexity** of parameters **to be encountered**.

Open (a bit informal) questions:

- 1 Are models with **effectively identifiable** parameters more general than models with **effectively consistent** estimators?
- 2 Is condition $\mu(n), \nu(m) < \infty$ related to consistency of Bayesian model selection?
- 3 Are models used in density estimation uniformly discretizable?
- 4 When do **normalized minimax** codes with (what?) **luckiness** converge in length to Bayesian codes?

Doesn't it resemble pulling oneself up from the mud?



bootstrap - To load and initialise the operating system on a computer. Normally abbreviated to **boot**. From the curious expression **to pull oneself up by one's bootstraps**, one of the legendary feats of Baron von Münchhausen.

Free Online Dictionary of Computing

But we have got one new language to paraphrase a few old stories.