Universal densities for stationary processes

Łukasz Dębowski

Institute of Computer Science
Polish Academy of Sciences

European Meeting of Statisticians,
Warsaw, 3rd–7th July 2023

Introduction

- Consider ergodic measures over a countable alphabet.
- It is known that universal measures, i.e., those consistently estimating the entropy rate, exist for any finite alphabet.
- A simple example is the PPM (prediction by partial matching) measure, also called the $\boldsymbol{R}$-measure, constructed gradually by Cleary and Witten (1984) and by Ryabko (1988, 2008).
- Alas, universal measures or codes do not exist for a countably infinite alphabet (Kieffer, 1978; Györfi et al., 1994).
- It may seem that a finite alphabet is necessary in general.

In this talk, we will disprove this hypothesis by constructing universal densities with respect to a given reference measure.

Ł. Dębowski. Universal densities exist for every finite reference measure. IEEE Transactions on Information Theory, 2023.

General setting

- $(\mathbb{X}, \mathcal{X}, \mu)$ — a countably generated space with a σ-finite μ:
  - counting measure $\mu(A) = \gamma(A) := \operatorname{card} A$ for a countable $\mathbb{X}$,
  - Lebesgue measure $\mu([a, b]) = \lambda([a, b]) := b - a$ for $\mathbb{X} = \mathbb{R}$.
- Product space $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$.
- Random variables $X_k : \mathbb{X}^{\mathbb{Z}} \ni (x_i)_{i \in \mathbb{Z}} \mapsto x_k \in \mathbb{X}$.
- The tuples of points are $x_{j:k} := (x_j, x_{j+1}, \ldots, x_k)$.
- For a probability measure R on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$, we denote its finite-dimensional restrictions $R_n(A) := R(X_{1:n} \in A)$.
- If $R_n \ll \mu^n$ then we write the densities

$$R_\mu(x_{1:n}) := \frac{dR_n}{d\mu^n}(x_{1:n}). \tag{1}$$

- The space of stationary ergodic measures on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ with respect to the shift operation will be denoted as $\mathbb{E}$.

Differential entropy

- We define the block entropy

$$\begin{aligned} h_\mu(n) &:= \mathrm{E}\left[-\log P_\mu(X_{1:n})\right] \\ &= -\int P_\mu(x_{1:n}) \log P_\mu(x_{1:n}) d\mu^n(x_{1:n}). \end{aligned} \tag{2}$$

- $h_\mu(n) \geq 0$ if μ is the counting measure.
- $h_\mu(n) \leq 0$ if μ is a probability measure.
- By stationarity and by the Jensen inequality, the block entropy is subadditive. Hence by the Fekete lemma, sequence $h_\mu(n)/n$ is decreasing and there exists the entropy rate

$$h_\mu := \lim_{n\to\infty} \frac{h_\mu(n)}{n} = \inf_{n\geq 1} \frac{h_\mu(n)}{n}. \tag{3}$$

Asymptotic equipartition

- Class of stationary ergodic measures with a finite entropy rate,

$$\mathbb{E}(\mu) := \{P \in \mathbb{E} : P_n \ll \mu^n \text{ and } |h_\mu| < \infty\}. \quad (4)$$

- As shown by Barron (1985a), for $P \in \mathbb{E}(\mu)$ we have

$$\lim_{n\to\infty} [-\log P_\mu(X_{1:n})] / n = h_\mu \text{ a.s.}, \quad (5)$$

- This follows by the Breiman ergodic theorem since

$$P_\mu(X_0|X_{-\infty:-1}) := \lim_{n\to\infty} P_\mu(X_0|X_{-n:-1}) \text{ a.s.} \quad (6)$$

and $\mathrm{E}\sup_{n\in\mathbb{N}} |\log P_\mu(X_0|X_{-n:-1})| < \infty$, whereas

$$h_\mu = \mathrm{E}[-\log P_\mu(X_0|X_{-\infty:-1})]. \quad (7)$$

Universal measures

Definition

A probability measure R where $R_n \ll \mu^n$ is called universal with respect to μ if for any $P \in \mathbb{E}(\mu)$,

$$\lim_{n\to\infty} \left[-\log R_\mu(X_{1:n})\right]/n = h_\mu \text{ a.s.,} \tag{8}$$

$$\lim_{n\to\infty} \mathrm{E}\left[-\log R_\mu(X_{1:n})\right]/n = h_\mu. \tag{9}$$

- For the counting measure $\mu(A) = \gamma(A) := \operatorname{card} A$ for $A \subset \mathbb{X}$, we speak of measures that are universal with respect to alphabet $\mathbb{X}$, respectively.
- In this case, we drop the subscript γ: $P_\gamma(x) \to P(x)$, $R_\gamma(x) \to R(x)$, and $h_\gamma \to h$.

Finite alphabet

Definition (PPM density)

Let card $\mathbb{X} = D$. The PPM density of order $k \geq 0$ is

$$\mathrm{PPM}_k^D(x_{1:n}) := \begin{cases} D^{-k-1} \prod_{i=k+2}^{n} \frac{N(x_{i-k:i}|x_{1:i-1}) + 1}{N(x_{i-k:i-1}|x_{1:i-2}) + D}, & k \leq n-2, \\ D^{-n}, & k \geq n-1, \end{cases} \quad (10)$$

where the frequency of a substring $w_{1:k}$ in a string $x_{1:n}$ is

$$N(w_{1:k}|x_{1:n}) := \sum_{i=1}^{n-k+1} 1\{x_{i:i+k-1} = w_{1:k}\}. \quad (11)$$

Subsequently, we define the (total) PPM density as

$$\mathrm{PPM}^D(x_{1:n}) := \sum_{k=0}^{\infty} w_k \, \mathrm{PPM}_k^D(x_{1:n}), \qquad w_k := \frac{1}{k+1} - \frac{1}{k+2}. \quad (12)$$

The total PPM measure is universal.

Inspiration for the NPD density

- Feutrill and Roughan (2021) considered a problem of estimating the differential entropy rate h_λ (with respect to the Lebesgue measure λ) of Gaussian processes with long memory.
- They observed that the differential entropy rate can be roughly estimated via their NPD entropy rate estimator, which reads

$$\hat{h}_{\text{NPD}}(x_1, \ldots, x_n) = \hat{H}\left(\lceil kx_1 \rceil, \ldots, \lceil kx_n \rceil\right) - \log k, \tag{13}$$

  where $\hat{H}$ is a consistent estimator of the entropy rate for a countable alphabet by Kontoyiannis et al. (1998).
- Feutrill and Roughan (2021) tried to argue that the NPD estimator tends to h_λ for $k \to \infty$ and $n \to \infty$ but their treatment of the joint limit was not rigorous.

Countably generated finite measure space

Definition (NPD density)

Let $(\mathbb{X}, \mathcal{X}, \mu)$ be a countably generated finite measure space. Let $\mathcal{X}_l \uparrow \mathcal{X}$ where $l = 0, 1, 2, \ldots$ be a filtration where the σ-fields $\mathcal{X}_l$ are finite with $\mathcal{X}_0 = \{\mathbb{X}, \emptyset\}$. Let χ_l be the finite partitions that generate σ-fields $\mathcal{X}_l$ respectively. We introduce quantizations of points $x \in \mathbb{X}$ as symbols $x^l := A$ for $x \in A \in \chi_l$. Moreover, for $l = 0, 1, 2, \ldots$, let R^l be universal measures for alphabets χ_l. We define the NPD density of order $l \geq 0$ as

$$\mathrm{NPD}^l_\mu(x_{1:n}) := \frac{R^l(x^l_{1:n})}{\prod_{i=1}^n \mu(x^l_i)}. \tag{14}$$

Subsequently, we define the (total) NPD density as

$$\mathrm{NPD}_\mu(x_{1:n}) := \sum_{l=0}^{\infty} w_l \, \mathrm{NPD}^l_\mu(x_{1:n}), \qquad w_l := \frac{1}{l+1} - \frac{1}{l+2}. \tag{15}$$

The total NPD measure is universal.

Why is NPD universal? (I)

- Since NPD_μ is a probability density then by Barron (1985a,b),

$$\liminf_{n\to\infty} \frac{[-\log \mathrm{NPD}_\mu(X_{1:n})]}{n} \ge h_\mu \text{ a.s.} \tag{16}$$

- Denote quantized block entropies

$$h_\mu^l(n) := \mathrm{E}\left[-\log P_n(X_{1:n}^l)\right] - n\,\mathrm{E}\left[-\log \mu(X_i^l)\right]. \tag{17}$$

- By the universality of measures R^l, we have

$$\lim_{n\to\infty} \frac{\left[-\log \mathrm{NPD}_\mu^l(X_{1:n})\right]}{n} = h_\mu^l := \inf_{n\ge 1} \frac{h_\mu^l(n)}{n} \text{ a.s.} \tag{18}$$

- Since $\mathrm{NPD}_\mu(x_{1:n}) \ge w_l\, \mathrm{NPD}_\mu^l(x_{1:n})$ then

$$\limsup_{n\to\infty} \frac{[-\log \mathrm{NPD}_\mu(X_{1:n})]}{n} \le \inf_{l\ge 0} h_\mu^l \text{ a.s.} \tag{19}$$

- It remains to show $\inf_{l\ge 0} h_\mu^l = h_\mu$.

Why is NPD universal? (II)

Lemma (Dębowski, 2021, Chapter 3, Problem 4)

For an interval A, let $f : A \to [0, \infty]$ be a nonnegative, continuous, and convex measurable function, let $\nu \ll \rho$ be two finite measures on a measurable space, and let $\mathcal{G}_n \uparrow \mathcal{G}$ be a filtration. We have

$$\lim_{n\to\infty} \int f\left(\frac{d\nu|_{\mathcal{G}_n}}{d\rho|_{\mathcal{G}_n}}\right) d\rho = \int f\left(\frac{d\nu|_{\mathcal{G}}}{d\rho|_{\mathcal{G}}}\right) d\rho, \tag{20}$$

where the sequence on the left hand side is increasing.

Why is NPD universal? (III)

So as to show that $\inf_{l\geq 0} h_\mu^l = h_\mu$, we observe that

$$\frac{P_n(x_{1:n}^l)}{\prod_{i=1}^n \mu(x_i^l)} = \frac{dP_n|\mathcal{X}_l^n}{d\mu^n|\mathcal{X}_l^n}(x_{1:n}). \tag{21}$$

Hence we have

$$h_\mu^l = \inf_{n\geq 1} \frac{h_\mu^l(n)}{n} = \inf_{n\geq 1} \frac{1}{n} \int \eta\left(\frac{dP_n|\mathcal{X}_l^n}{d\mu^n|\mathcal{X}_l^n}\right) d\mu^n, \tag{22}$$

where $\eta(x) := -x \log x$. We switch the order of infimums,

$$\inf_{l\geq 0} h_\mu^l = \inf_{n\geq 1} \frac{1}{n} \inf_{l\geq 0} \int \eta\left(\frac{dP_n|\mathcal{X}_l^n}{d\mu^n|\mathcal{X}_l^n}\right) d\mu^n \tag{23}$$

and we apply the lemma to function $f(x) = \log 2 - \eta(x)$. Hence

$$\inf_{l\geq 0} h_\mu^l = \inf_{n\geq 1} \frac{1}{n} \int \eta\left(\frac{dP_n}{d\mu^n}\right) d\mu^n = \inf_{n\geq 1} \frac{h_\mu(n)}{n} = h_\mu. \tag{24}$$

Conditional density estimation

- For $P \in \mathbb{E}(\mu)$, denote the conditional measure

$$P_\mu^{(\infty)}(x) := \lim_{n\to\infty} P_\mu(x|X_{-n:-1}). \quad (25)$$

- Obviously, $P^{(\infty)} = P_1$ if P is a memoryless source.
- For R where $R_n \ll \mu^n$, the Cesàro mean measure is

$$\bar{R}_\mu^{(n)}(x) := \frac{1}{n}\sum_{i=0}^{n-1} R_\mu(x|X_{n-i:n-1}). \quad (26)$$

- Total variation: $\delta(P, R) := \frac{1}{2}\int |P_\mu(x) - R_\mu(x)|\, d\mu(x)$.

Theorem (idea of Györfi et al., 1994 + Pinsker inequality)

If R is universal with respect to μ then for any $P \in \mathbb{E}(\mu)$,

$$\lim_{n\to\infty} \delta(P^{(\infty)}, \bar{R}^{(n)}) = \lim_{n\to\infty} \delta(P^{(n)}, \bar{R}^{(n)}) = 0 \text{ a.s.} \quad (27)$$

Prediction with the $0-1$ loss

- The predictor $\boldsymbol{f_P}$ induced by a measure $\boldsymbol{P}$ is the maximizer

$$\boldsymbol{f_P}(\boldsymbol{x}_{1:\boldsymbol{n}-1}) = \arg\max_{\boldsymbol{x_n}\in\mathbb{X}} \boldsymbol{P}(\boldsymbol{x_n}|\boldsymbol{x}_{1:\boldsymbol{n}-1}). \tag{28}$$

- Predictor $\boldsymbol{f} : \mathbb{X}^* \to \mathbb{X}$ is called universal with respect to $\boldsymbol{\mu}$ if for any $\boldsymbol{P} \in \mathbb{E}(\boldsymbol{\mu})$,

$$\lim_{\boldsymbol{n}\to\infty} \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} 1\{\boldsymbol{X_i} \neq \boldsymbol{f}(\boldsymbol{X}_{1:\boldsymbol{i}-1})\} = \boldsymbol{u} \text{ a.s.}, \tag{29}$$

where $\boldsymbol{u} := \mathrm{E}\left[1 - \max_{\boldsymbol{x}\in\mathbb{X}} \boldsymbol{P}(\boldsymbol{x}|\boldsymbol{X}_{-\infty:-1})\right]$.

Theorem (strengthens Dębowski and Steifer, 2022)

Consider a countable alphabet $\mathbb{X}$. Suppose that measure $\boldsymbol{R}$ is universal with respect to $\boldsymbol{\mu}$. The Cesàro mean predictor $\boldsymbol{f_{\bar{R}}}$ is universal with respect to $\boldsymbol{\mu}$.

Natural numbers and real line

Theorem

Let $\mathbb{X} = \mathbb{N}$ *and a probability measure* μ *with* $\mu(x) > 0$ *for all* $x \in \mathbb{N}$. *Let* $P \in \mathbb{E}$ *with* $H(P_1) + D(P_1||\mu) < \infty$. *Then*

$$\lim_{n\to\infty} \frac{1}{n} \left[-\log \mathrm{NPD}_\mu(X_{1:n}) - \sum_{i=1}^{n} \log \mu(X_i) \right] = h \text{ a.s.} \tag{30}$$

Theorem

Let $\mathbb{X} = \mathbb{R}$, $\mu \sim N(m, \sigma^2)$, *and the Lebesgue measure* λ. *Let* $P \in \mathbb{E}$ *with* $P_n \ll \lambda^n$ *and* $|\mathrm{E}\, X_i|, \mathrm{Var}\, X_i, |h_\lambda| < \infty$. *Then*

$$\lim_{n\to\infty} \frac{1}{n} \left[-\log \mathrm{NPD}_\mu(X_{1:n}) + \left[\sum_{i=1}^{n} \frac{(X_i - m)^2}{2\sigma^2} \right] \log e \right] + \log \sigma\sqrt{2\pi} = h_\lambda \text{ a.s.} \tag{31}$$

References I

A. R. Barron. The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *The Annals of Probability*, 13: 1292–1303, 1985a.

A. R. Barron. *Logically Smooth Density Estimation*. PhD thesis, Stanford University, 1985b.

J. G. Cleary and I. H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32:396–402, 1984.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. New York: Wiley & Sons, 2021.

Ł. Dębowski and T. Steifer. Universal coding and prediction on ergodic random points. *The Bulletin of Symbolic Logic*, 28(2):387–412, 2022.

A. Feutrill and M. Roughan. NPD entropy: A non-parametric differential entropy rate estimator. `https://arxiv.org/abs/2105.11580`, 2021.

L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for infinite alphabet. *IEEE Transactions on Information Theory*, 40:267–271, 1994.

References II

J. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24:674–682, 1978.

I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44: 1319–1327, 1998.

B. Ryabko. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series. In *2008 IEEE Information Theory Workshop, Porto*, pages 271–275. Institute of Electrical and Electronics Engineers, 2008.

B. Y. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24(2):87–96, 1988.