Infinite urn models and applications

Collegio Carlo Alberto Seminar-Torino

Stéphane Boucheron

January 27th 2017

Infinite urn models, missing mass and applications

Infinite urn models

Setting

Probability distribution over $\mathbb{N}$

1. Sequence $U_1, \ldots, U_n, \ldots$, i.i.d. over $\mathbb{N}_+ := \mathbb{N} \setminus \{0\}$ (the *symbols*)
2. (Unknown) distribution defined by p.m.f. : $(p_j)_{j=1}^{\infty}$
3. Counts: $X_{n,j} = \sum_{i=1}^{n} \mathbb{I}_{U_i=j}$, number of occurrences of j among the first n symbols

Inference problems (species sampling)

1. Estimation of $(p_j)_{j=1}^{\infty}$
2. Estimation of missing mass $M_{n,0} = \sum_{j=1}^{\infty} p_j \mathbb{I}_{X_{n,j}=0}$
3. Estimation of mass of rare symbols $M_{n,r} = \sum_{j=1}^{\infty} p_j \mathbb{I}_{X_{n,j}=r}$
4. Estimation of discovery probability $M_{n+m,r}$ from $U_1, \ldots, U_n$

Occupancy counts, profile

$K_{n,r}$

For $r \geq 1$, $K_{n,r}$: number of symbols with r occurrences in sample

$K_{n,\bar{r}}$

For $r \geq 1$, $K_{n,\bar{r}}$: number of symbols with at least r occurrences in sample

K_n

K_n : number of distinct symbols in the sample

Profile, histogram

$(K_{n,r})_{r\geq 1}$
$(K_{n,r}/K_n)_r$ defines an (empirical) distribution over integers

$$K_n = \sum_{r\geq 1} K_{n,r} \qquad n = \sum_{r\geq 1} K_{n,r} \times r \qquad \text{obvious!}$$

Motivations for mass estimators

Connections between masses and occupancy counts

$$\mathbb{E}M_{n,0} = \mathbb{E}\frac{K_{n+1,1}}{n+1} \qquad \Rightarrow \qquad \mathbb{E}\frac{K_{n,1}}{n} = \mathbb{E}M_{n-1,0} \geq \mathbb{E}M_{n,0}$$

$$\mathbb{E}M_{n,0} \leq \mathbb{E}\frac{K_{n,1}}{n} = \mathbb{E}\frac{K_{n+1}}{n+1} + \frac{2}{n}\frac{\mathbb{E}K_{n+1,2}}{n+1} = \mathbb{E}M_{n,0} + \frac{1}{n}\mathbb{E}M_{n,1}$$

$$\mathbb{E}M_{n,0} \leq \mathbb{E}\frac{K_{n,1}}{n} \leq \mathbb{E}M_{n,0} + \frac{1}{n}$$

Mass estimation

Good-Turing estimator of the missing mass

$$\widehat{M}_{n,0} := \frac{K_{n,1}}{n}$$

Good-Turing mass-estimators

$$\widehat{M}_{n,r} := (r+1)\frac{K_{n,r+1}}{n}$$

Harder problems

- Discovery probabilities $\mathbb{E}(M_{n+m,r} \mid \mathcal{F}_n)$
- Number of new discoveries $\mathbb{E}(K_{n+m} \mid \mathcal{F}_n) - K_n$

Assessing Good-Turing estimators

Providing non-asymptotic/distribution-free guarantees for GT estimators

$$\text{How large is} \quad \left| \frac{\widehat{M}_{0,n}}{M_{0,n}} - 1 \right| \quad ?$$

- Needs to get a handle on the tail behavior of $\frac{K_{n,1}}{n}$ and of $M_{n,0}$

Concentration inequalities

Classical exponential inequalities

$X_1, \ldots, X_n$ independent centered random variables
$Z := \sum_{i=1}^n X_i$

Hoeffding $a_i \leq X_i \leq b_i$

$$\mathbb{P}\{Z \geq t\} \leq \mathrm{e}^{-\frac{4t^2}{2\sum_{i=1}^n (b_i - a_i)^2}} \qquad \text{for } t > 0$$

Bennett $X_i \leq b$ $v := \mathrm{var}(Z)$

$$\mathbb{P}\{Z \geq t\} \leq \mathrm{e}^{-\frac{v}{b^2} h\left(\frac{bt}{v}\right)} \qquad \text{for } t > 0 \text{ with } h(x) = (1+x)\log(1+x) - x$$

Bernstein $\sum_{i=1}^n \mathbb{E}(X_i)_+^q \leq \frac{q!}{2} v c^{q-2}$ $v \geq \mathrm{var}(Z)$

$$\mathbb{P}\{Z \geq t\} \leq \mathrm{e}^{-\frac{t^2}{2(v+ct)}} \qquad \text{for } t > 0$$

Concentration of measure phenomenon

Any function of many independent random variables that does not depend too much on any of them is strongly concentrated around its expectation (or its median)
M. Talagrand

Gaussian concentration (1975)

$X \sim \mathcal{N}(0, \mathrm{Id}_n)$
$f : \mathbb{R}^n \to \mathbb{R}$, L-Lipschitz, $Z := f(X_1, \ldots, X_n)$

$$\mathrm{var}(Z) \leq \mathbb{E}\left[\|\nabla f\|^2\right] \leq L^2$$

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \mathrm{e}^{-\frac{t^2}{2L^2}}$$

Poisson/binomial scenarii

Poissonizing sample/message length

- $(X_j)_{j\geq 1}$: a collection of independent Poisson Point Processes on $(0,\infty)$ with intensity p_j ($X_j(t) \sim \mathrm{Poi}_{tp_j}$)
- $K_r(t)$: sum of independent Bernoulli random variables $\mathbb{I}_{X_j(t)=r} \sim$ Bernoulli with success probability $\mathrm{Poi}_{tp_j}(r)$
- $M_r(t)$: sum of weighted independent Bernoulli random variables $M_r(t) = \sum_{j\geq 1} p_j \mathbb{I}_{X_j(t)=r}$

Concentration inequalities come (almost) for free

$K_r(n)$ are sums of independent Bernoulli random variables
$M_r(n)$ are weighted sums of independent Bernoulli random variables

Poissonized identities for the missing mass

$$\mathbb{E}M_0(t) = \frac{\mathbb{E}K_1(t)}{t} \qquad \text{and} \qquad \mathrm{var}(M_0(t)) = 2\frac{\mathbb{E}K_2(t)}{t^2} - \frac{\mathbb{E}K_2(2t)}{2t^2}$$

Concentration inequalities

Variance inequalities

For the number of distinct symbols

$$\frac{\mathbb{E}K_1(2t)}{2t} \leq \mathrm{var}(K(t)) \leq \mathbb{E}K_1(t) \qquad \text{(Poisson setting)}$$

$$\mathrm{var}(K_n) \leq \mathbb{E}\left[(1 - M_{n,0})K_{n,1}\right] \leq \mathbb{E}K_{n,1} \leq \mathbb{E}K_n \qquad \text{(Binomial setting)}$$

Tight in the birthday paradox setting

$p_j = 1/n^2$ for $1 \leq j \leq n^2$, $1 - M_{n,0} = K_n/n^2$

$$\mathrm{var}(K_n) \leq \mathbb{E}\left[K_{n,1}\frac{K_n}{n^2}\right] \leq 1$$

Efron-Stein-Steele inequalities (1981-86)

Thm. : the expected Jackknife estimate of variance is an upper bound on variance

- $Z = f(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are indenpendent.
- $Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ (f_i any function)

$$\mathrm{var}(Z) \leq \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z_i)^2\right]$$

Efron-Stein-Steele inequalities : sketch of proof

$X_1, \ldots, X_n$ independent random variables
$X'_1, \ldots, X'_n$ independent "copies" ($X'_i \sim X_i$)
$f, g : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$

$X^{(i)} := X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n$
$X^{(i)} := X'_1, \ldots, X'_i, X_{i+1}, \ldots, X_n$ $\qquad X := X^{(0)}, X' := X^{(n)}$

$$\begin{aligned}
\operatorname{cov}(f(X), g(X)) &= = \mathbb{E}\left[f(X)(g(X) - g(X'))\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[f(X)\left(g(X^{(i-1)}) - g(X^{(i)})\right)\right] \\
&= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\left(f(X) - f(X^{(i)})\right)\left(g(X^{(i-1)}) - g(X^{(i)})\right)\right]
\end{aligned}$$

$$\operatorname{cov}(f(X), f(X)) \leq \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\left(f(X) - f(X^{(i)})\right)^2\right] \qquad \text{(Cauchy-Schwarz)}$$

Variance inequalities for occupancy counts and missing mass

$$\mathrm{var}(K_{n,\bar{r}}) \leq r\mathbb{E}K_{n,r}$$

$$\mathrm{var}(K_{n,r}) \leq r\mathbb{E}K_{n,r} + (r+1)\mathbb{E}K_{n,r+1}$$

$$\mathrm{var}(M_{n,0}) \leq \sum_{j=1}^{\infty} p_j^2 \,\mathrm{var}\left(\mathbb{I}_{X_{n,j}=0}\right) \leq 2\frac{\mathbb{E}K_2(n)}{n^2}$$

Negative association

Definition

Real-valued random variables $Z_1, \ldots, Z_K$ are said to be *negatively associated* if, for any two disjoint subsets A and B of $\{1, \ldots, K\}$, and any two real-valued functions $f : \mathbb{R}^{|A|} \mapsto \mathbb{R}$ and $g : \mathbb{R}^{|B|} \mapsto \mathbb{R}$ that are both either coordinate-wise non-increasing or coordinate-wise non-decreasing :

$$\mathbb{E}\left[f(Z_A).g(Z_B)\right] \leq \mathbb{E}\left[f(Z_A)\right].\mathbb{E}\left[g(Z_B)\right] .$$

Examples

In the Bernoulli scenario $(X_{n,j})_{j\geq 1}$ are negatively associated

Monotone functions of negatively associated random variables are negatively associated

Concentration inequalities

K_n is sub-Poisson

$$\log \mathbb{E} e^{\lambda(K_n - \mathbb{E}K_n)} \leq \mathbb{E}K_{n,1}\phi(\lambda) \qquad \text{with} \qquad \phi(\lambda) = e^{\lambda} - \lambda - 1$$

K_n satisfies a Bennett inequality with variance factor $\mathbb{E}K_{n,1}$

$K_{n,\bar{r}}$

Is a sum of negatively associated Bernouilli random variables
$K_{n,\bar{r}}$ satisfies a Bennett inequality with variance factor $r\mathbb{E}K_{n,r}$

$K_{n,r}$

$K_{n,r} = K_{n,\bar{r}} - K_{n,\overline{r+1}}$

Missing mass

Generic concentration inequalities for missing mass

Lower tail (sub-Gaussian with sharp bound on variance, McAllester-Ortiz)

$$v_n^- := \frac{\mathbb{E}K_2(n)}{n^2}$$

$$\log \mathbb{E}e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq v_n^- \frac{\lambda^2}{2} \qquad \lambda \leq 0$$

Upper tail (sub-Gamma with looser bound on variance)

$$v_n^+ := \frac{\mathbb{E}K_2(n)}{n^2}$$

$$\log \mathbb{E}e^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq v_n^+ \frac{\lambda^2}{2(1-\lambda/n)} \qquad \lambda \geq 0$$

Proof

$\phi(x) := \exp(x) - x - 1$

$$\log \mathbb{E}\left[e^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] \leq \sum_{j=1}^{\infty} e^{-np_j} \phi(\lambda p_j) \qquad \text{neg. assoc and Bennett ineq.}$$

$$\begin{aligned}
\sum_{j=1}^{\infty} e^{-np_j} \phi(\lambda p_j) &= \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \sum_{j=1}^{\infty} e^{-np_j} \frac{(np_j)^r}{r!} \\
&= \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n) \\
&\leq \mathbb{E}K_{\overline{2}}(n) \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \\
&= \lambda^2 \frac{\mathbb{E}K_{\overline{2}}(n)/n^2}{1 - \lambda/n},
\end{aligned}$$

Sub-Gaussian versus Sub-Gamma ?

Uniformity

A collection $\mathcal{C}$ of random variables is *uniformly* sub-Gaussian if there exists a constant c, such that satisfies

$$\forall X \in \mathcal{C}, \qquad \log \mathbb{E} e^{\lambda(X - \mathbb{E}X)} \leq c \operatorname{var}(X) \frac{\lambda^2}{2}$$

Bernoulli random variables are *not* uniformly sub-Gaussian

If $X_p \sim \mathrm{Ber}(p)$, the best possible upper-bound for all $p \in (0,1)$ is

$$\log e^{\lambda(X_p - \mathbb{E}X_p)} \leq \operatorname{var}(X_p) \frac{\lambda^2}{2(2p(1-p)c_{\mathrm{LS}}(p))} \qquad \text{where} \qquad c_{\mathrm{LS}}(p) = \frac{\log((1-p)/p)}{1-2p}$$

Sub-Gaussian bounds for the right-tail of $M_{n,0}$

Best variance proxy

$$w_n := \sum_{j=1}^{\infty} \frac{p_j^2}{2c_{\mathrm{LS}}((1-p_j)^n)} \leq \frac{1}{2n}$$

If $\max_j p_j \leq 1/4$

$$w_n \geq \frac{3}{32n} \left(1 - \sum_{j:p_j<1/n} p_j \right)$$

w_n may be order of magnitudes larger than $\mathbb{E}K_2(n)/n^2$

Karlin's setting

Motivations

Predicting the number of discoveries during next campaign

Ideally

$K_{\lceil nx \rceil}/K_n$ should converge in probability toward a function of x
At least : $\lim_n \mathbb{E}K_{\lceil nx \rceil}/\mathbb{E}K_n \in (0, \infty)$

Estimating the missing mass

Counting functions and regular variation

Counting function $\vec{\nu}$ and counting measure ν

$$\vec{\nu}(x) := \#\{j : p_j \geq x\} \qquad \nu(A) = \sum_{j \geq 1} \mathbb{I}_{p_j \in A}$$

$$\nu_1(0, x) := \sum_{j=1}^{\infty} p_j \mathbb{I}_{p_j \leq x}$$

Karlin's assumption

Regular variation assumption on $\vec{\nu}$ in the neighborhood of 0

$$\lim_{x \searrow 0} \frac{\vec{\nu}(sx)}{\vec{\nu}(x)} = s^{-\alpha} \qquad \text{for some} \qquad \alpha \in (0, 1)$$

Why?

For $\alpha \in (0, 1)$ Karlin's assumption implies (among other things)

- $\mathbb{E}K_n \sim \Gamma(1-\alpha)\vec{\nu}(1/n)$
- $\mathbb{E}K_{n,r} \sim \frac{\alpha\Gamma(r-\alpha)}{r!}\vec{\nu}(1/n)$
- $\mathrm{var}(M_{n,0}) \sim \alpha\Gamma(2-\alpha)(1-2^{\alpha-2}\frac{\vec{\nu}(1/n)}{n^2}$

Chinese restaurant processes (Poisson-Dirichlet processes)

A Poisson-Dirichlet process with parameters $\alpha \in (0, 1)$ and $\theta > -\alpha$ generates a distribution over $\mathbb{N} \setminus \{0\}$ that is a.s. α-regularly varying

Challenging concentration inequalities for $M_{n,0}$: $\alpha \in (0,1)$

If $\vec{\nu}$ satisfies the regular variation condition with index $\alpha \in (0,1)$, then the missing mass $M_{n,0}$ (or $M_0(n)$) is sub-Gaussian on the left tail with variance factor $v_n^- = 2\mathbb{E}K_2(n)/n^2$ and sub-gamma on the right tail with variance factor $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$.

The variance factors satisfy

$$\lim_n \frac{v_n^-}{\mathrm{var}(M_{n,0})} = \frac{1}{1-2^{\alpha-2}},$$

$$\lim_n \frac{v_n^+}{\mathrm{var}(M_{n,0})} = \frac{2}{\alpha(1-2^{\alpha-2})},$$

and thus

$$\lim_n \frac{v_n^-}{v_n^+} = \frac{\alpha}{2}.$$

Challenging concentration inequalities for $M_{n,0}$: $\alpha = 0$

Two settings : $\mathbb{E}K_{n,1} \to \infty$ or $\mathbb{E}K_{n,1} = O(1)$

- $\mathbb{E}K_{n,1} = O(1)$ for Geometric, negative binomial, Poisson, ... distributions
- $\nu_1(0,x) \sim_{x\to 0} x\ell_1(1/x)$ with $\ell_1 \in \mathrm{RV}_0$ and $\ell_1(t) \nearrow \infty$

de Haan's regular variation (Extended Regular Variation)

$\vec{\nu}(1/\cdot) \in \Pi_a$ where a is slowly varying if

$$\lim_{t\to\infty} \frac{\vec{\nu}(1/(tx) - \vec{\nu}(1/t))}{a(t)} = \log(x) \qquad \text{for } x > 0$$

If $\vec{\nu}(1/\cdot) \in \Pi_a$ with $a \in \mathrm{RV}_0$ and $a(t) \nearrow \infty$

For each $r \geq n$

- $K_n \overset{\mathbb{P}}{\sim} \mathbb{E}K_n \underset{+\infty}{\sim} \ell(n)$
- $K_{n,r} \overset{\mathbb{P}}{\sim} \mathbb{E}K_{n,r} \underset{+\infty}{\sim} \frac{a(n)}{r}$
- $K_{n,\bar{r}} \overset{\mathbb{P}}{\sim} \mathbb{E}K_{n,\bar{r}} \underset{+\infty}{\sim} \ell(n)$
- $M_{n,r} \overset{\mathbb{P}}{\sim} \mathbb{E}M_{n,r} \underset{+\infty}{\sim} \frac{a(n)}{n}$.

Missing masses are uniformly sub-gamma

- $v(n) := \frac{12a(n)}{n^2}$
- $\mathrm{var}(M_{n,0}) \sim \frac{3a(n)}{4n^2}$
- $\exists n_0, \forall n > n_0,$

$$\log \mathbb{E}\left[e^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] \leq \frac{v_n \lambda^2}{2(1 - \lambda/n)} \qquad \lambda > 0$$

Excessively slow variation

Geometric distribution ($p_j = q(1-q)^{j-1}$)

No hope to obtain uniform sub-gamma tail behavior

What is the missing mass made of?

- Essentially symbols with probability of order $1/n$
- The *concentration on two points* phenomenon: with probability tending to 1, $M_{0,n}$ tends to be distributed over a pair of values (that depend on n)

Estimation

Estimating the missing mass

One needs to make assumptions on the sampling distribution to guarantee the consistency of the Good-Turing estimator.

There is no hope to find a universally consistent estimator of the missing mass without any such restrictions, as shown recently by Mossel & Ohannessian (2015)

If $\mathbb{E}K_{n,2}/\mathbb{E}K_{n,1}$ remains bounded and $\mathbb{E}K_{n,1} \to +\infty$

$$\frac{M_{n,0}}{\mathbb{E}M_{n,0}} \xrightarrow{\mathbb{P}} 1\,,$$

and

$$\frac{\widehat{M}_{n,0}}{M_{n,0}} \xrightarrow{\mathbb{P}} 1\,.$$

Poissonized setting

$$\mathrm{var}(\widehat{M}_0(t) - M_0(t)) = \frac{1}{t^2}\left(\mathbb{E}K_1(t) + 2\mathbb{E}K_2(t)\right) . \tag{1}$$

$\widehat{M}_0(t) - M_0(t)$ is sub-gamma

1. on the right tail with variance factor $\mathrm{var}(\widehat{M}_0(t) - M_0(t))$ and scale factor $1/t$,
2. on the left tail with variance factor $3\mathbb{E}K(t)/t^2$ and scale factor $1/t$.

For all $\lambda \geq 0$,

1. $\log \mathbb{E}e^{\lambda(\widehat{M}_0(t) - M_0(t))} \leq \mathrm{var}(\widehat{M}_0 t - M_0(t))t^2\phi\left(\frac{\lambda}{t}\right)$
2. $\log \mathbb{E}e^{\lambda(M_0(t) - \widehat{M}_0 t)} \leq \frac{3\mathbb{E}K(t)}{2t^2}\frac{\lambda^2}{1-\lambda/t}$

Lossless Compression over countable alphabets

Statistical coding

Redundancy of coding probability Q^n with respect to source P^n

Expected difference between codelengths obtained by feeding an arithmetic coder with $Q^n(\mathbf{x})$ rather than with the correct source statistics $P^n(\mathbf{x})$

$$D(P^n, Q^n) = \mathbb{E}_{P^n} \log \frac{P^n(X_{1:n})}{Q^n(X_{1:n})}$$

Λ^n : collection of probability distributions over messages of length n. Each probability distribution is called a source.

Minimax redundancy

$$\overline{R}(\Lambda^n) = \inf_Q \sup_{P \in \in \Lambda} D(P^n, Q^n)$$

Maximin redundancy

π : prior distribution on sources

$$\underline{R}(\Lambda^n) = \sup_\pi \inf_Q \mathbb{E}_\pi D(P^n, Q^n)$$

MinMax Theorem $\underline{R}(\Lambda^n) = \overline{R}(\Lambda^n)$

Envelope classes

Envelop function

$f : \mathbb{N} \to \mathbb{R}_+$ with $1 < \sum_{j>0} f(j) < \infty$.

Envelop class

$$\Lambda_f = \left\{ \mathbb{P} \ : \ \forall x \in \mathbb{N}, \ \mathbb{P}^1\{x\} \leq f(x) \text{ and } \mathbb{P} \text{ is stationary and memoryless.} \right\}$$

Envelope distribution

1. $F(k) := 1 - \sum_{j>k} f(j)$ for $k \geq l_f := \max\{k : \ \sum_{j \geq k} f(j) \geq 1\}$ envelope distribution
2. $\overline{F} = 1 - F$ tail envelope function
3. $\vec{\nu}(x) = |\{i : f(i) \geq x\}|$ for $x \in]0, \infty[$ counting function

PCC Code (I)

{Mixture coding of the censored sequence}

$$\widetilde{Q}^{n+1}(\widetilde{x}_{1:n}) = \prod_{i=0}^{n-1} \widetilde{Q}_{i+1}(\widetilde{x}_{i+1} \mid \widetilde{x}_{1:i}) \,,$$

Predictive probabilities $\widetilde{Q}_{i+1}(\cdot \mid \widetilde{x}_{1:i})$

given by Krichevsky-Trofimov mixtures on the alphabet $\{0, 1, \ldots, K_i\}$,

$$\widetilde{Q}_{i+1}\left(\widetilde{X}_{i+1} = k \mid \widetilde{X}_{1:i} = \widetilde{x}_{1:i}\right) = \frac{\widetilde{n}_i^k + \frac{1}{2}}{i + \frac{K_i+1}{2}} \,,$$

where, for $0 \leq k \leq K_i$, $\widetilde{n}_i^k$ is the number of occurrences of symbol k in $\widetilde{x}_{1:i}$.

For $0 \leq k \leq K_i$,

$$\widetilde{n}_i^k = \begin{cases} K_i & \text{if } k = 0 \,, \\ n_i^j - 1 & \text{if } 1 \leq k \leq K_i \text{ and } \langle j, k \rangle \in \mathcal{D}_i \,, \end{cases}$$

where n_i^j is the number of occurrences of symbol j in $x_{1:i}$.

PCC Code (II)

1. Dictionary initialization: $\mathcal{D}_0 \leftarrow \{\langle 0, 0\rangle\}$
   For $i > 0$, $\mathcal{D}_i$ contains 0 and symbols from $x_{1:i}$ with rank of insertion
2. Create a censored sequence $\widetilde{x}_{1:n}$

$$\widetilde{x}_i = \begin{cases} k & \text{if } \exists k \leq K_{i-1}, \quad \langle x_i, k\rangle \in \mathcal{D}_{i-1} \\ 0 & \text{otherwise, (newly discovered symbol)} \end{cases}$$

   where K_i := number of distinct symbols in $x_{1:i}$.
3. Dictionary update: when $\widetilde{x}_i = 0$, $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \langle x_i, K_i\rangle$, with $K_i = K_{i-1} + 1$.
4. Let K := number of *censored* input symbols and let $i_{1:K}$ be their indices. Extract subsequence $x_{i_{1:K}}$
5. C_M := Mixture/arithmetic coding of $\widetilde{x}_{1:n}$
6. C_E := Elias / integer coding of each redacted symbol in $x_{i_{1:K}}$ individually
7. Interleave the coded censored symbols of C_E just after each coded 0 symbol of C_M, to form the overall PCC-code.

Weak adaptivity of PCC codes

$$\mathbf{R}_f(n) := \log(\mathrm{e}) \int_1^n \frac{\vec{\nu}_f(1/t)}{2t} \mathrm{d}t \,.$$

Let Λ_f be an envelope source class, with $f \in \mathrm{RV}_\alpha$ and $\alpha \in (0, 1($. Then

$$\overline{R}(\Lambda_f^n) \asymp \mathbf{R}_f(n) \,.$$

Theorem

Let (Q_n) : coding distribution associated to the Pattern Censoring Code. \
$\forall \alpha \in (0, 1(, \forall f$ with $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha, \exists a_f, b_f > 0$:

$$(a_f + o_f(1))\mathbf{R}_f(n) \leq \overline{R}(\Lambda_f^n) \leq \overline{R}(Q_n, \Lambda_f^n) \leq (b_f + o_f(1))\mathbf{R}_f(n) \log\log n \,.$$

{The Pattern Censoring Code is adaptive, within $\log\log n$, w.r.t. $\left((\Lambda_f)_{f \in \mathrm{RV}_\alpha}\right)_{\alpha \in (0,1)}$ }

Redundancy of PCC code

Redundancy of the PCC code with respect to a source P over $\mathbb{N} \setminus \{0\}$

$$D(P^n \mid Q^n) = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_P\left[\log \frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})} \middle| X_{1:i}\right]\right]$$

Decomposing instantaneous redundancy

$$\begin{aligned}
&\mathbb{E}_P\left[\log \frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})} \middle| X_{1:i}\right] \\
&= \underbrace{\sum_{j\geq 1} p_j \mathbb{I}_{N_i^j>0} \log\left(\frac{p_j\left(i+\frac{K_i+1}{2}\right)}{N_i^j - \frac{1}{2}}\right)}_{\text{(I)}} \\
&\quad + \underbrace{\sum_{j\geq 1} p_j \mathbb{I}_{N_i^j=0} \log\left(\frac{p_j\left(i+\frac{K_i+1}{2}\right)}{K_i + \frac{1}{2}}\right)}_{\text{(II)}} \\
&\quad + \underbrace{\sum_{j\geq 1} p_j \mathbb{I}_{N_i^j=0} \left(\log(j+1) + 2\log\log(j+1)\right)}_{} .
\end{aligned}$$

Redundancy of PCC code}

Distribution-free bounds

$$\begin{aligned}\mathbb{E}_P\left[\log\frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})}\right] &\leq \kappa\left(\nu_1(0,1/i)+\frac{\vec{\nu}(1/i)}{i}\right)\\ &\quad+\sum_{j\geq 1}p_j\mathbb{P}\{N_i^j=0\}\left(\log\left(\frac{j+1}{\mathbb{E}K_i}\right)+2\log\log(j+1)\right),\end{aligned}$$

where κ is (another) universal constant

Using regular variation of the envelope distribution

For each Regularly varying envelope, $\exists i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$

$$\begin{aligned}\sum_{j\geq 1}p_j\mathbb{I}_{N_i^j=0}\left(\log\log(j+1)\right) &\leq \frac{5}{2}\left(\Gamma(1-\alpha)+\frac{1}{1-\alpha}\right)\frac{\log\log(i)\vec{\nu}_f(1/i)}{i}\end{aligned}$$

Techniques

Everything boils down to controlling inverse binomials, inverse of number of distinct symbols K_i, missing mass $M_{i,0}$, and some random integrals

The distribution and moments of these stochastic quantities ultimately depend on details of the sampling distribution, but ...

Some non-asymptotic inequalities are distribution-free

- $\mathbb{E}M_{i,0} \leq \mathbb{E}\frac{K_{i,1}}{i}$
- $\mathrm{var}(K_i) \leq \mathbb{E}K_{i,1}$
- $\mathbb{E}\frac{1}{K_i} \leq \frac{3}{\mathbb{E}K_i}$
- ...

The analysis is made modular by delaying invocation of regular variation arguments

References

- A. Ben-Hamou, B., M. Ohannessian: *Concentration inequalities in the infinite urn scheme with occupancy counts and the missing mass, with applications*. Bernoulli. **23**, 249-287 (2017)
- A. Ben-Hamou, B., E. Gassiat: *Censoring meets pattern-coding*. ArXiv.