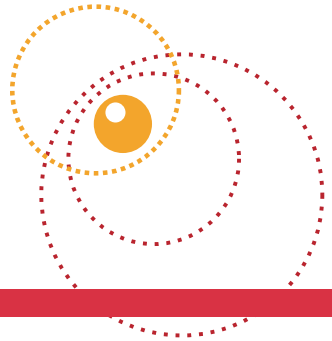
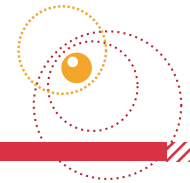


Coding on Countably Infinite Alphabets

Non-parametric Information Theory

20 Juillet 2009
Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat

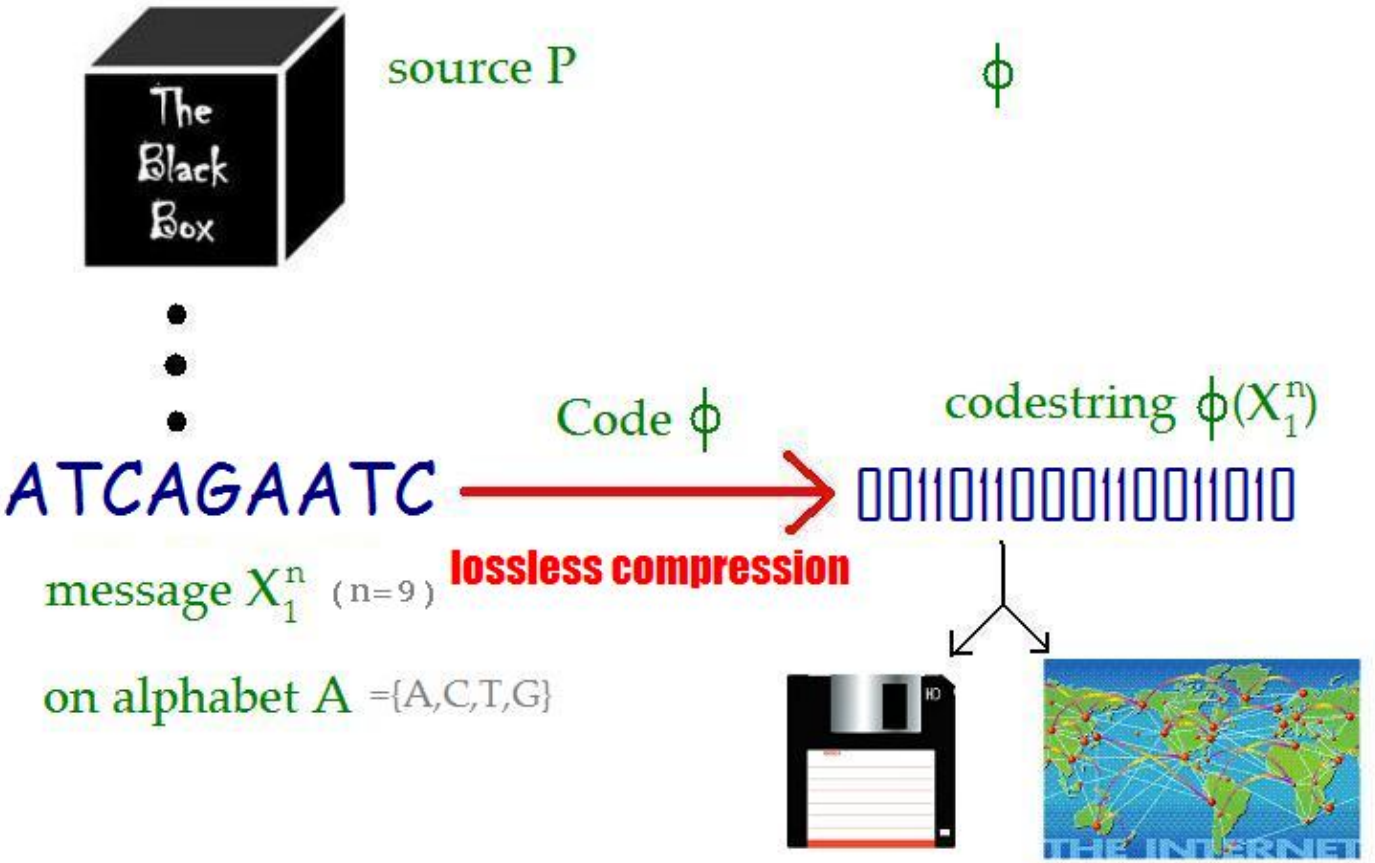
ParisTech
GRANDES ECOLES D'INGENIEURS DE PARIS
PARIS INSTITUTE OF TECHNOLOGY



Outline

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Data Compression : Shannon Modelization



Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Data Compression : Shannon Modelization

Source P

= stationnary process on **Alphabet A** ={A,C,T,G}

Code $\phi_n : A^n \longrightarrow \{0,1\}^*$

ATCAGAATC → 001101100011001101O

Message $X_1^n (n=9)$

lossless compression

Winzip, compress, etc.

Codestring $\phi_n(X_1^n)$

Goal : minimize average codelength

$$E_P\left[|\phi(X_1^n)|\right]$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Outline

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Source Entropy

- Shannon ('48) :

$$\mathbb{E}_P\left[|\phi_n(x)|\right] \geq H_n(P) = \mathbb{E}_P\left[-\log P^n(X_1^n)\right]$$

  n-block entropy and there is a code reaching the bound (within 1 bit).

- Moreover, if P is stationary and ergodic :

$$\frac{1}{n} H_n(P) \to H(P)$$

  entropy rate of the source P
  = minimal number of bits necessary per symbol.

Coding Distributions

- Kraft Inequality :

$$\sum_{x_1^n} 2^{-|\phi_n(x)|} \leq 1$$

  and reciprocal.
- Code $\phi \iff$ coding distribution $Q_n^{\phi}(\mathbf{x}) = 2^{-|\phi_n(x)|}$.
- The Shannon '48 theorem expresses that in average the best coding distribution is P !
- Otherwise, coding distribution Q_n suffers from the regret

$$-\log Q_n\left(X_1^n\right) - \left(-\log P^n\left(X_1^n\right)\right) = \log \frac{P^n\left(X_1^n\right)}{Q_n\left(X_1^n\right)}.$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Outline

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Universal Coders

- What if the source statistics are unknown ?
- What if we need versatile code ?

$\Longrightarrow$ Need a single coding distribution Q_n for a class of sources

$$\Lambda = \{P_\theta, \theta \in \Theta\}$$

Ex : memoryless processes, Markov chains, HMM...

$\Longrightarrow$ unavoidable redundancy :

$$\begin{aligned} \mathbb{E}_{P_\theta}\left[|\phi\left(X_1^n\right)|\right] - H\left(X_1^n\right) &= \mathbb{E}_{P_\theta}\left[\log Q_n\left(X_1^n\right) + \log P_\theta\left(X_1^n\right)\right] \\ &= KL\left(P_\theta, Q_n\right) \end{aligned}$$

= Küllback-Leibler divergence between P_θ and Q_n.

Coders : Measures of Universality

1. **Maximal regret** :

$$R^*\left(Q_n,\Lambda\right)=\sup_{x_1^n\in A^n}\sup_{\theta\in\Theta}\log\frac{P_\theta\left(x_1^n\right)}{Q_n\left(x_1^n\right)}$$

2. **Worst case redundancy** :

$$R^+\left(Q_n,\Lambda\right)=\sup_{\theta\in\Theta}\mathbb{E}_{P_\theta}\left[\log\frac{P_\theta^n\left(X_1^n\right)}{Q_n\left(X_1^n\right)}\right]=\sup_{\theta\in\Theta}KL\left(P_\theta,Q_n\right)$$

3. **Expected redundancy** with respect to prior π :

$$R_\pi^-\left(Q_n,\Lambda\right)=\mathbb{E}_\pi\left[\mathbb{E}_{P_\theta}\left[\log\frac{P_\theta^n\left(X_1^n\right)}{Q_n\left(X_1^n\right)}\right]\right]=\mathbb{E}_\pi\left[KL\left(P_\theta,Q_n\right)\right]$$

$$\Longrightarrow\ R^-\left(Q_n,\Lambda\right)\leq R^+\left(Q_n,\Lambda\right)\leq R^*\left(Q_n,\Lambda\right)$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Classes : Measures of Complexity

1. **Minimax regret** :

$$R_n^*(\Lambda) = \inf_{Q_n} R^*(Q_n, \Lambda) = \min_{Q_n} \max_{x_1^n, \theta} \log \frac{P_\theta^n(x_1^n)}{Q_n(x_1^n)}$$

2. **Minimax redundancy** :

$$R_n^+(\Lambda) = \inf_{Q_n} R^+(Q_n, \Lambda) = \min_{Q_n} \max_{\theta} KL(P_\theta^n, Q_n)$$

3. **Maximin redundancy** :

$$R_n^-(\Lambda) = \sup_{\pi} \inf_{Q_n} R_\pi^-(Q_n, \Lambda) = \max_{\pi} \min_{Q_n} \mathbb{E}_\pi [KL(P_\theta^n, Q_n)]$$

$$\Longrightarrow \quad R_n^-(\Lambda) \leq R_n^+(\Lambda) \leq R_n^*(\Lambda)$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

General Standard Results

- Minimax result (Haussler '97, Sion) : $R_n^- (\Lambda) = R_n^+ (\Lambda)$.
- Shtarkov's NML : $R_n^* (\Lambda) = R^* (Q_{\mathrm{NML}}^n, \Lambda)$ with $Q_{\mathrm{NML}}^n(\mathbf{x}) = \frac{\hat{p}(\mathbf{x})}{\sum_{\mathbf{y} \in \mathcal{X}^n} \hat{p}(\mathbf{y})}$ and $\hat{p}(\mathbf{x}) = \sup_{P \in \Lambda} P^n(\mathbf{x})$, so that and hence

$$R_n^* (\Lambda) = \log \left(\sum_{\mathbf{y} \in \mathcal{X}^n} \hat{p}(\mathbf{y}) \right) .$$

- Channel capacity method : if $d\mu(\theta, x) = d\pi(\theta) dP_\theta(x)$ then

$$\inf_{Q_n} R_\pi^- (Q_n, \Lambda) = H_\mu(X_1^n) - H_\mu(X_1^n | \theta) = H_\mu(\theta) - H_\mu(\theta | X_1^n) .$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Finite Alphabets : Standard Results

- **Theorem (Shtarkov, Barron, Spankowski,...)** If $\mathcal{I}_m$ denotes the class of memoryless processes over alphabet $\{1, \ldots, m\}$, then

$$R_n^+(\mathcal{I}_m) = \frac{m-1}{2} \log \frac{n}{2e} + \log \frac{\Gamma(1/2)^m}{\Gamma\left(\frac{m}{2}\right)} + o_m(1)$$

$$R_n^*(\mathcal{I}_m) = \frac{m-1}{2} \log \frac{n}{2} + \log \frac{\Gamma(1/2)^m}{\Gamma\left(\frac{m}{2}\right)} + o_m(1) \leq \frac{m-1}{2} \log n + 2$$

- **Theorem (Rissanen '84)** If $\dim \Theta = k$, and if there exists a $\sqrt{n}$-consistent estimator of θ given X_1^n, then

$$\liminf_{n \to \infty} R_n^-(\Lambda) \geq \frac{k}{2} \log n.$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Outline

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Infinite Alphabets : Issues

Motivations : integer coding, lossless image coding, text coding on words, mathematical curiosity.

1. Understand general structural properties of minimax redundancy and minimax regret.
2. Characterize those source classes that have finite minimax regret.
3. Quantitative relations between minimax redundancy or regret and integrability of the envelope function.
4. Develope effective coding techniques for source classes with known non-trivial minimax redundancy rate.
5. Develope adaptive coding schemes for collections of source classes that are too large to have redundancy rates.

Sanity-check Properties

1. $R^*(\Lambda^n) < +\infty \iff Q^n_{\text{NML}}$ is well-defined and given by

$$Q^n_{\text{NML}}(\mathbf{x}) = \frac{\hat{p}(\mathbf{x})}{\sum_{\mathbf{y}\in\mathcal{X}^n} \hat{p}(\mathbf{y})} \text{ for } \mathbf{x} \in \mathcal{X}^n$$

   where $\hat{p}(\mathbf{x}) = \sup_{P\in\Lambda} P^n(\mathbf{x})$.

2. $R^+(\Lambda^n)$ and $R^*(\Lambda^n)$ are non-decreasing, sub-additive (or infinite) functions of n. Thus,

$$\lim_{n\to\infty} \frac{R^+(\Lambda^n)}{n} = \inf_{n\in\mathbb{N}_+} \frac{R^+(\Lambda^n)}{n} \leq R^+\left(\Lambda^1\right),$$

   and

$$\lim_{n\to\infty} \frac{R^*(\Lambda^n)}{n} = \inf_{n\in\mathbb{N}_+} \frac{R^*(\Lambda^n)}{n} \leq R^*\left(\Lambda^1\right).$$

Memoryless Sources

1. Let Λ be a class of stationary memoryless sources over a countably infinite alphabet. Let $\hat{p}$ be defined by $\hat{p}(x) = \sup_{P \in \Lambda} P\{x\}$. The minimax regret with respect to Λ^n is finite if and only if the normalized maximum likelihood (Shtarkov) coding probability is well-defined and :

$$R^*(\Lambda^n) < \infty \Leftrightarrow \sum_{x \in \mathbb{N}_+} \hat{p}(x) < \infty.$$

2. We give an example of a memoryless class Λ such that $R^+(\Lambda^n) < \infty$, but $R^*(\Lambda^n) = \infty$.

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Outline

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Redundancy and Regret Equivalence Property

Definition

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

Theorem

Let f be a non-negative function from $\mathbb{N}_+$ to $[0,1]$, let Λ_f be the class of stationary memoryless sources defined by envelope f. Then

$$R^+\left(\Lambda_f^n\right) < \infty \iff R^*\left(\Lambda_f^n\right) < \infty \iff \sum_{k \in \mathbb{N}_+} f(k) < \infty .$$

Idea of the Proof

The only thing to prove is that

$$\sum_{k\in\mathbb{N}_+} f(k) = \infty \Rightarrow R^+\left(\Lambda_f^n\right) = \infty\,.$$

Let the sequence of integers $(h_i)_{i\in\mathbb{N}}$ be defined recursively by $h_0 = 0$ and

$$h_{i+1} = \min\left\{ h \;:\; \sum_{k=h_i+1}^{h} f(k) > 1 \right\}\,.$$

We consider the class $\Lambda = \{P_\theta, \theta \in \Theta\}$ where $\Theta = \mathbb{N}$. The memoryless source P_i is defined by its first marginal P_i^1 which is given by

$$P_i^1(m) = \frac{f(m)}{\sum_{k=h_i+1}^{h_{i+1}} f(k)} \text{ for } m \in \{p_i + 1, ..., p_{i+1}\}\,.$$

Taking any infinite entropy prior π over Θ the $\Lambda^1 = \{P_i^1 \;;\; i \in \mathbb{N}_+\}$ shows that

$$R^+\left(\Lambda^1\right) \geq R_\pi^-\left(\Lambda^1\right) = H(\theta) - H(\theta|X_1) = \infty\,.$$

Upper-bounding the Regret

Theorem

If Λ is a class of memoryless sources, let the tail function $\bar{F}_{\Lambda^1}$ be defined by $\bar{F}_{\Lambda^1}(u) = \sum_{k>u} \hat{p}(k)$, then :

$$R^*(\Lambda^n) \leq \inf_{u:u\leq n} \left[n\bar{F}_{\Lambda^1}(u) \log e + \frac{u-1}{2} \log n \right] + 2 \,.$$

Corollary

Let Λ denote a class of memoryless sources, then the following holds :

$$R^*(\Lambda^n) < \infty \Leftrightarrow R^*(\Lambda^n) = o(n) \text{ and } R^+(\Lambda^n) = o(n) \,.$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Lower-bounding the Redundancy

Theorem

Let f denote a non-increasing, summable envelope function. For any integer p, let $c(p) = \sum_{k=1}^{p} f(2k)$. Let $c(\infty) = \sum_{k\geq 1} f(2k)$. Assume furthermore that $c(\infty) > 1$. Let $p \in \mathbb{N}_+$ be such that $c(p) > 1$. Let $n \in \mathbb{N}_+$, $\epsilon > 0$ and $\lambda \in]0,1[$ be such that $n > \frac{c(p)}{f(2p)} \frac{10}{\epsilon(1-\lambda)}$. Then

$$R^+(\Lambda_f^n) \geq C(p,n,\lambda,\epsilon) \sum_{i=1}^{p} \left(\frac{1}{2} \log \frac{nf(2i)\,(1-\lambda)\,\pi}{2c(p)e} - \epsilon \right),$$

where $C(p,n,\lambda,\epsilon) = \frac{1}{1+\frac{c(p)}{\lambda^2 nf(2p)}} \left(1 - \frac{4}{\pi}\sqrt{\frac{5c(p)}{(1-\lambda)\epsilon nf(2p)}}\right).$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Power-law Classes

Theorem

Let α denote a real number larger than 1, and C be such that $C\zeta(\alpha) \geq 2^\alpha$. The source class $\Lambda_{C,-\alpha}$ is the envelope class associated with the decreasing function $f_{\alpha,C} : x \mapsto 1 \wedge \frac{C}{x^\alpha}$ for $C > 1$ and $\alpha > 1$. Then :

1.

$$n^{1/\alpha} A(\alpha) \log \left\lfloor (C\zeta(\alpha))^{1/\alpha} \right\rfloor \leq R^+(\Lambda^n_{C,-\alpha})$$

where

$$A(\alpha) = \frac{1}{\alpha} \int_1^\infty \frac{1}{u^{1-1/\alpha}} \left(1 - e^{-1/(\zeta(\alpha)u)}\right) \mathrm{d}u\,.$$

2.

$$R^*(\Lambda^n_{C,-\alpha}) \leq \left(\frac{2Cn}{\alpha - 1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1)\,.$$

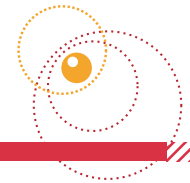
Exponential Classes

Theorem

Let C and α denote positive real numbers satisfying $C > e^{2\alpha}$. The class $\Lambda_{Ce^{-\alpha\cdot}}$ is the envelope class associated with function $f_\alpha : x \mapsto 1 \wedge Ce^{-\alpha x}$. Then

$$\frac{1}{8\alpha}\log^2 n\,(1 - o(1)) \leq R^+(\Lambda^n_{Ce^{-\alpha\cdot}})$$
$$\leq \quad R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \leq \frac{1}{2\alpha}\log^2 n + O(1)$$

cf Dominique Bontemps : $R^+(\Lambda^n_{Ce^{-\alpha\cdot}}) \sim \frac{1}{4\alpha}\log^2 n$



Outline

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

The CensoringCode Algorithm

Algorithm CensoringCode

$K \leftarrow 0$
$counts \leftarrow [1/2, 1/2, \ldots]$
for i from 1 to n **do**
  $cutoff \leftarrow \left\lfloor \left(4\frac{Ci}{\alpha-1}\right)^{1/\alpha} \right\rfloor$
  if $cutoff > K$ **then**
    for $j \leftarrow K+1$ to $cutoff$ **do**
      $counts[0] \leftarrow counts[0] - counts[j] + 1/2$
    end for
    $K \leftarrow cutoff$
  end if
  if $x[i] \leq cutoff$ **then**
    ArithCode($x[i]$, $counts[0 : cutoff]$)
  else
    ArithCode(0, $counts[0 : cutoff]$)
    $C1 \leftarrow C1 \cdot$EliasCode($x[i]$)
    $counts[0] \leftarrow counts[0] + 1$
  end if
  $counts[x[i]] \leftarrow counts[x[i]] + 1$
end for
$C2 \leftarrow$ ArithCode()
$C_1 \cdot C_2$

Theorem

Let C and α be positive reals. Let the sequence of cutoffs $(K_i)_{i \leq n}$ be given by

$$K_i = \left\lfloor \left(\frac{4Ci}{\alpha - 1} \right)^{1/\alpha} \right\rfloor .$$

Then

$$R^+(CensoringCode, \Lambda^n_{C,-\alpha}) \leq \left(\frac{4Cn}{\alpha - 1} \right)^{\frac{1}{\alpha}} \log n \, (1 + o(1)) .$$

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Approximately Adaptive Algorithms

A sequence $(Q^n)_n$ of coding probabilities is said to be *approximately asymptotically adaptive* with respect to a collection $(\Lambda_m)_{m\in\mathcal{M}}$ of source classes if for each $P \in \cup_{m\in\mathcal{M}}\Lambda_m$, for each Λ_m such that $P \in \Lambda_m$:

$$D(P^n, Q^n)/R^+(\Lambda_m^n) \in O(\log n)\,.$$

- A modification of `CensoringCode` choosing the $n+1$th cutoff K_{n+1} according to the number of *distinct* symbols in $\mathbf{x}$ is approximately adaptive on

$$\mathcal{W}_\alpha = \left\{P \,:\, P \in \Lambda_{\cdot-\alpha}, \right.$$
$$\left. 0 < \liminf_k k^\alpha P^1(k) \leq \limsup_k k^\alpha P^1(k) < \infty \right\}\,.$$

- Pattern coding is approximately adaptive if $1 < \alpha \leq 5/2$.

Stéphane Boucheron, Aurélien Garivier, Elisabeth Gassiat
<garivier@enst.fr>

Patterns

The information conveyed in a message x can be separated into

1. a dictionary $\Delta = \Delta(x)$: the sequence of distinct symbols occurring in x in order of appearance ;
2. a pattern $\psi = \psi(x)$ where ψ_i is the rank of $\mathbf{x}_i$ in dictionary Δ.

Example :

Message	x	$=$	a	b	r	a	c	a	d	a	b	r	a
Pattern	$\psi(x)$	$=$	1	2	3	1	4	1	5	1	2	3	1
Dictionary	$\Delta(x)$	$=$	a	b	r		c		d				

$\Longrightarrow$ A random process $(X_n)_n$ with distribution P induces a random pattern process $(\Psi_n)_n$ on $\mathbb{N}_+$ with distribution :

$$P^{\Psi}\left(\Psi_1^n = \psi_1^n\right) = \sum_{x_1^n : \psi(x_1^n) = \psi_1^n} P\left(X_1^n = x_1^n\right).$$

Pattern Coding

- The pattern entropy rate exists and coincides with the process entropy rate :

$$\frac{1}{n} H\left(\Psi_1^n\right) = \frac{1}{n} \mathbb{E}_{P^\Psi}\left[-\log P^\Psi\left(\Psi_1^n\right)\right] \to H(\Psi) = H(X).$$

- Pattern redundancy satisfies :

$$1.84 \left(\frac{n}{\log n}\right)^{\frac{1}{3}} \leq R^-_{\Psi,n}\left(\mathcal{I}_\infty\right) \leq R^*_{\Psi,n}\left(\mathcal{I}_\infty\right) \leq \left(\pi\sqrt{\frac{2}{3}\log e}\right)\sqrt{n}.$$

  The proof of the lower-bound uses fine combinatorics on integer partitions with small summands (see Garivier '09). Upper-bounds in $O(n^{2/5})$ have been given (see Shamir '07), but there is still a gap between lower- and upper-bounds.

- For memoryless coding, the pattern redundancy is neglectible wrt. the codelength of the dictionary whenever $n < 5/2$.