# Adaptive compression over countable alphabets

S. Boucheron,
joint work with A. Ben-Hammou, D. Bontemps, A. Garivier, E. Gassiat & M. Ohanessian

Microsoft-INRIA, Paul Sabatier, Paris-Diderot, ENS, Paris-Sud

September, 14th, 2015

# Lossless compression over a countable alphabet

## Lossless compression

Mapping messages (sequences of symbols from countable alphabet $\mathcal{X}$) to codewords (sequences of $\{0,1\}$), so as to minimize the expected length of codewords in a one-to-one and non-ambiguous way.

## Non-ambiguous codes satisfy Kraft-McMillan inequality

For $\lambda : A \to \mathbb{N}_+$,

$$\sum_{\omega \in A} 2^{-\lambda(\omega)} \leq 1, \text{ iff } \exists \text{ non-ambiguous code } f : A \to \{0,1\}^* \text{ with } \ell[f(\omega)] = \lambda(\omega)$$

## Kraft-Mac Millan inequality

provides a bridge between codes and probability distributions

- Any non-ambiguous code defines a (sub)-probability distribution over the set of messages
- Any probability distribution $Q$ over the set of messages defines a non-ambiguous encoding where codeword length is at most $-\log_2 Q(\omega) + 1$.

# Redundancy

**Definition (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 $P^n(\mathbf{x})$

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

$\Lambda^n$ is collection of probability distributions over messages of length $n$. Each probability distribution is called a source.

**Definition (Minimax redundancy)**

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

**Definition (Maximin redundancy)**

$\pi$ : prior distribution on sources

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

**MinMax Theorem** $R_+(\Lambda^n) = R^+(\Lambda^n)$

# Redundancies: alphabet size matters

$\Lambda$ : memoryless sources over finite alphabet with cardinality $k$

## Minimax redundancy

$$R^+(\Lambda^n) = \frac{k-1}{2} \log \frac{n}{2\pi e} + O(1)$$

Rissanen, Ryabko, Shtarkov, Krichevsky, Trofimov, Barron, Clarke, Xie et al..

## Krichevsky-Trofimov coding is asymptotically maximin and approximately minimax

$$\mathbb{KT}(X_{n+1} = a | X_{1:n} = x_{1:n}) = \frac{n_a(x_{1:n}) + \frac{1}{2}}{n + \frac{k}{2}},$$

## Countable alphabets

### Negative results

$$\exists (Q^n)_n, \quad \forall P \in \Lambda, \quad \lim_n \frac{1}{n} D(P^n, Q^n) = 0$$

iff

$$\exists P^*, \quad \forall P \in \Lambda, \quad \mathbb{E}_{P1}[-\log P^*(X)] < \infty$$

J. Kieffer (1993), Gyorfi, Pali van der Meulen (1993)

# Coding against countable alphabets

No analogue of Lempel-Ziv coding for stationary ergodic sources over a countable alphabet.

↪ There is no universal source code for an infinite source alphabet,
Györfi et al., IEEE IT, 1994

To obtain positive results... it is necessary to impose constraints on source classes

Envelope classes form one possible example of such constraints

# Envelop classes

## Envelop function

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

## Envelop class

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

## Envelope distribution

- $F(k) = 1 - \sum_{j>k} f(j)$ for $k \ge l_f := \max\{k : \sum_{j \ge k} f(j) \ge 1\}$ — envelope distribution
- $\overline{F} = 1 - F$ — tail envelope function
- $U(t) = \inf\{x : F(x) \ge 1 - 1/t\}$ — tail quantile (envelope) function
- $\vec{\nu}(x) = \left|\{i : f(i) \ge x\}\right|$ for $x \in ]0, \infty[$ — counting function

▸ Details on smoothed envelopes

# Envelopes

## Sub-exponential classes

- $F_c$ has non-decreasing hazard rate (ako log-concavity assumption)
- $U_c \circ \exp$ is concave.

## Example

- Exponential envelopes.
  $f(k) = \gamma e^{-\left(\frac{k}{\beta}\right)^{\alpha}}$. with $\alpha \geq 1, \beta > 0$ and $\gamma > 1$
- Poisson envelopes
  $f(k) = \gamma e^{-\beta}\beta^k/k!$ with $\beta > 0$ and $\gamma > 1$
- ...

## Regularly varying envelops

$F_c$ (resp. $U_c$) is regularly varying with index $-1/\gamma$ (resp. $\gamma > 0$)

$$\forall x > 0, \qquad \lim_t \frac{F_c(tx)}{F_c(t)} = x^{-1/\gamma}.$$

$$U_c(t) = t^{\gamma}\ell(t)$$

where $\ell$ is slowly varying

## Example

- Power-law envelopes:
  $U_c(t) = \kappa t^{\gamma}$
- Heavy-Tailed envelopes
  $U_c(t) = \kappa t^{\gamma}\ell(t)$

# Envelopes and profiles

The counting function $\vec{v}$ provides us with concise and precise characterizations of minimax redundancies.

## Regularly varying envelope

$\lim_{t\to 0} \frac{\vec{v}(tx)}{\vec{v}(t)}$ exists as a function of $x$.

## Then, for somme $\alpha \in [0, 1]$

$\vec{v}(x) = x^{-\alpha} L(1/x)$ with $L$ slowly varying at infinity.

There is a correspondence between the regular variation properties of $\vec{v}$ and the regular variation properties of $U$

▸ More about $\vec{v}$

# Bounds on minimax redundancy

**Theorem (BGG, 2009)**

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

$$R^+(\Lambda^n) \leq \inf_{u:u\leq n} \left[ n\bar{F}_{\Lambda^1}(u) \log_2 e + \frac{u-1}{2} \log_2 n \right] + 2 \,.$$

**Suggestion**

If the envelop is known, choose threshold $\tau$ as the solution of $\bar{F}_{\Lambda^1}(u) = \frac{u}{n}$.

i) Encode symbols over threshold using Elias penultimate code

ii) Encode other symbols using Krichevsky-Trofimov mixture over alphabet $\{1, \ldots, \tau\}$.

If the envelop is not known, look for a data-driven threshold

▸ Lower bounds

# Minimax regret

## Another worst-case oriented risk measure for coding

$$R^*(\Lambda^n) = \inf_{Q \in \mathfrak{X}^n} \sup_{\mathbf{x} \in \mathcal{X}^n} \sup_{P^n \in \Lambda^n} \log \frac{P^n(\mathbf{x})}{Q(\mathbf{x})}$$

## Obvious

$$R^+(\Lambda^n) \leq R^*(\Lambda^n)$$

Minimax regret fits into the individual sequences framework (see Cesa-Bianchi & Lugosi)

# Risk bounds for regular envelope classes

## For any envelope class

$$R^*(\Lambda_f^n) \leq \underbrace{1 + \vec{\nu}(1/n)\log(3) + \int_{1/n}^{1} \frac{\vec{\nu}(x)}{2x}\mathrm{d}x}_{(I)} + \underbrace{n\nu_1([0,1/n[)}_{(II)}$$

starting from Acharya, Jafarpour, Orlitsky, and Suresh ArXiV.1405.7460

## Interpretation

(I) $\rightarrow$ Cost of encoding symbols with envelope probability larger than $1/n$

(II) $\rightarrow$ Cost of encoding symbols with envelope probability smaller than $1/n$

## For regular envelope classes

The shares of I) and II) depends on the index of regularity

$\int_{1/n}^{1} \frac{\vec{\nu}(x)}{2x}\mathrm{d}x$ is the leading term

# Regularly varying envelopes: $\alpha \in ]0,1[$

- $\vec{v}(x) = x^{-\alpha} L(1/x)$ with $L$ slowly varying at infinity
- $\nu_1[0,x] \sim \frac{\alpha}{1-\alpha} x^{1-\alpha} L(1/x) = \frac{\alpha}{1-\alpha} x \vec{v}(x)$
- $\lim_{n\to\infty} \frac{\vec{v}(1/n)}{\int_1^n \vec{v}(1/x)/x \mathrm{d}x} = \alpha$

Karamata integration theorem

Recall

$$R^+(\Lambda_f^n) \leq \underbrace{1 + \vec{v}(1/n)\log(3) + \int_{1/n}^1 \frac{\vec{v}(x)}{2x}\mathrm{d}x}_{(I)} + \underbrace{n\nu_1([0,1/n[)}_{(II)}$$

For regular envelope classes

$$\hookrightarrow \qquad R^+(\Lambda_f^n) \leq \left(1 + \log(3) + \frac{1}{2\alpha} + \frac{\alpha}{1-\alpha}\right)\vec{v}(1/n)$$

as (I) and (II) scale like $\vec{v}(1/n)$.

$n\nu_1([0,1/n[)$ is (almost) the reciprocal of the hazard rate at $\vec{v}(1/n)$

# Regularly varying envelopes: $\alpha = 0$

- In general $\nu_1[0,x] \lll x\vec{\nu}(x)$, but $\nu_1([0,x])$ may not be regularly varying near 0.
- $\lim_{n\to\infty} \frac{\vec{\nu}(1/n)}{\int_1^n \vec{\nu}(1/x)/x\mathrm{d}x} = 0$

## In all cases

(I) is asymptotically large with respect to (II)

## Two cases deserve special investigations

- The envelope distribution has non-decreasing hazard rate ($U \circ \exp$ concave)

  Bontemps, B. & Gassiat, IEEE IT 2014
- The envelope distribution has decreasing hazard rate ($U \circ \exp$ convex)

In both cases the limiting behaviour of $n\nu_1(]0,1/n[)$ is well understood.

# If the envelope distribution has non-decreasing hazard rate

Alternatively, if $\nu_1[0,x] \sim x$ as $x \to 0$,

$\hookrightarrow$ (II) is upper bouded by a constant that depends on the envelope distribution but not on $n$.

The minimax redundancy is upper bounded by

$$\int_1^n \frac{\vec{v}(1/x)}{2x} \mathrm{d}x$$

while $\vec{v}(x) \sim U(1/x)$.

$\hookrightarrow$ We recover (up to a constant) the equivalent determined in (BBG14))

# If the envelope distribution has decreasing hazard rate

Alternatively, if $\nu_1([0,x]) \sim xL_0(1/x)$ where $L_0$ is slowly varying and tending to infinity, then

$$\vec{v}(x) \sim \int_1^{1/x} L_0(u)/u \mathrm{d}u$$

$R^+(\Lambda_f^n)$ scales like

$$\int_1^n \frac{\int_1^x \frac{L_0(u)}{u}\mathrm{d}u}{x}\mathrm{d}x \leq \vec{v}(1/n)\log(n)$$

The larger $L_0$, the poorer the upper bound

# Lower bounds

## For minimax regret

If $R^*(\Lambda_f^n) \leq n/16$ and $n_1 := \lfloor n - 3\sqrt{nR^*(\Lambda_f^n)} \rfloor$ then

$$R^*(\Lambda_f^n) \geq \sum_{i=l_f}^{\vec{\nu}(1/n_1)} \frac{1}{2} \log \frac{2n_1 f_i + 2}{\pi} + \sum_{i > \vec{\nu}(1/n_1)} \log\left(2 - \mathrm{e}^{-n_1 f_i}\right)$$

Acharya et al. 2014

## Corollary

$$R^*(\Lambda_f^n) \geq \log \sqrt{\frac{4}{\pi}} \vec{\nu}(1/n_1) + \frac{1}{4} \int_1^{n_1} \frac{\vec{\nu}(1/x)}{x} \mathrm{d}x + \frac{n_1 \nu_1([0, 1/n_1])}{\mathrm{e}}$$

# Flavors of adaptivity

## For collections of small classes

### Definition (Asymptotic adaptivity)

$(Q^n)_n$ is asymptotically adaptive with respect to $(\Lambda_m)_{m\in\mathcal{M}}$ if

$$\forall m \in \mathcal{M}, \quad R^+(Q^n, \Lambda_m^n) = \sup_{\mathbb{P}\in\Lambda_m} D(\mathbb{P}^n, Q^n) \leq (1 + o(1))R^+(\Lambda_m^n)$$

## For collections of massive envelop classes

### Definition (Weak asymptotic adaptivity)

$(Q^n)_n$ is asymptotically weakly adaptive with respect to $(\Lambda_m)_{m\in\mathcal{M}}$

$$\forall m \in \mathcal{M}, \quad R^+(Q^n, \Lambda_m^n) \leq o(\log n)R^+(\Lambda_m^n)\,.$$

# Censuring codes: sketch

## AC-code : Thresholding above last record

$m_i = \max_{1 \le j \le i} x_j$.
The $j^{\text{th}}$ record is denoted by $\widetilde{m}_j$ ($\widetilde{m}_0 = 0$)
Let $\widetilde{\mathbf{m}} = (\widetilde{m}_i - \widetilde{m}_{i-1} + 1)1$.
Symbols from $\widetilde{\mathbf{m}}$ encoded using Elias penultimate code.

## Progressive KT coding below the last record

$\widetilde{x}_i = x_i \mathbb{I}_{x_i \le m_{i-1}}$.
$C_M$ : progressive $\mathbb{KT}$- encoding of $\widetilde{x}_{1:n}0$

$$Q_{i+1}(\widetilde{X}_{i+1} = j | X_{1:i} = x_{1:i}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{m_i+1}{2}} \quad \text{if} \quad 1 \le j \le m_i,$$

$$Q_{i+1}(\widetilde{X}_{i+1} = 0 | X_{1:i} = x_{1:i}) = \frac{1/2}{i + \frac{m_i+1}{2}},$$

where $n_i^j$ is the number of occurrences of symbol $j$ in $x_{1:i}$, $n_i^0 = 0$.

Example

# Light-tailed envelopes

The AC-code is adaptive with respect to source classes defined by envelopes with finite and non-decreasing hazard rate.

**Theorem (B., Bontemps, Gassiat, 2014)**

$Q^n$ : the coding probability associated with the AC-code,
If $f$ is an envelope with non-decreasing hazard rate,

$$R^+(Q^n; \Lambda_f^n) \leq (1 + o(1)) R^+(\Lambda_f^n)$$

while

$$R^+(\Lambda_f^n) = (1 + o(1))(\log e) \int_1^n \frac{U_c(x)}{2x} \mathrm{d}x$$

Details

# Envelopes with heavier tails

If the tail envelope distribution is heavier than exponential, thresholding at maximum does not lead to (weakly) adaptive coding

## Proxy threshold: $m_c$ solution of

$$t\overline{F}_c(u) = u \text{ or } u = U_c\left(\frac{t}{u}\right)$$

## Properties

- $m_c$ is non-decreasing.
- $m_c(t) \nearrow \infty$
- $m_c(t)/t \searrow 0$
- If $U_c$ is $\gamma$-regularly varying, $m_c$ is $\gamma/(\gamma+1)$-regularly varying.
- $m_c(t)$ scales like $\vec{v}(1/t)$.

## Empirical theshold

$$M_n = \min\left(n, \{k \ : \ X_{k,n} \leq k\}\right)$$

# Weak adaptivity of ETAC encoding

If $\overline{F}_c \in MDA(-1/\gamma)$ with $\gamma > 0$,

$\forall \epsilon > 0$, for sufficiently large $n$,

$$\mathbb{E}X_{M_n,n} \leq m_n(1+\epsilon) \qquad R^+(\Lambda_f^n) \geq \frac{m_n}{2}\,.$$

If $Q^n$ is the coding probability associated with the ETAC code

$$R^+(Q^n, \Lambda_n) \leq (5 + o_\Lambda(1))\,\frac{m_n}{2}\log n + 2$$

B., Gassiat, Ohannessian, 2014

For power law envelopes $U_c(t) = \kappa t^\gamma$ (Acharya et al. 2014)

$$R^+(\Lambda_f^n) \sim \left(\frac{\kappa^{1/\gamma}}{\gamma} n\right)^{\frac{\gamma}{\gamma+1}} \left(\frac{1}{\gamma} + \gamma \log \mathrm{e} + c\right)$$

Details

# ETAC thresholding when $\vec{v}$ is regularly varying with index $\alpha \in ]0,1[$

$i_n := \vec{v}(1/n)$

$$\overline{F}(i_n) \sim \frac{\alpha}{1-\alpha} \frac{i_n}{n}$$

Both $m_n$ (the ETAC threshold) and $i_n$ (threshold motivated by Poissonized analysis) scale like $\vec{v}(1/n)$

Thanks

# References


- A. Ben-Hamou, S. Boucheron, M.I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications Bernoulli, to appear.
- S. Boucheron and E. Gassiat : A Bernstein-von Mises theorem for discrete probability distributions Electronic Journal of Statistics. **3** (2009) 114-148.
- S. Boucheron and A. Garivier and E. Gassiat : Coding over Infinite Alphabet s IEEE Trans. on Inform. Theory **55** (2009 ) 358 - 373.
- D. Bontemps : Universal coding on infinite alphabets: exponentially decreasing envelopes. IEEE Trans. Inform. Theory 57 (2011), no. 3, 1466–1478.
- D. Bontemps, S. Boucheron and E. Gassiat : Adaptive compression against a countable alphabet. IEEE Trans. Inform. Theory 60 (2014), 808-821.
- S. Boucheron, E. Gassiat & M. Ohanessian : About Adaptive Coding on Countable Alphabets: Max-Stable Envelope Classes IEEE Trans. on Information Theory, 61, 9, pp. 4948-4967 (2015)
- S. Boucheron, M. Thomas : Concentration inequalities for order statistics. Electronic Communications in Probability. 17 (2012). 1-12

# Envelop classes

## Smoothed distribution function

- $F_c$ has piecewise constant hazard rate,
- $\overline{F}_c(n) = \overline{F}(n)$
- $U_c(t) = \inf\{x : 1/\overline{F}_c(x) \geq t\}$.

If $X \sim F_c$ then $\lfloor X \rfloor + 1 \sim F$ and $U(t) = \lfloor U_c(t) \rfloor + 1$ for $t > 1$.

## Lemma (Stochastic comparison by quantile coupling)

There exists a probability space where $X \sim G \in \Lambda_f$, $Y \sim F_c$ such that

$$\mathbb{P}\{X \leq Y\} = 1$$

Return

# Envelopes and profiles

The counting function $\vec{v}$ provides us with concise and precise characterizations of minimax redundancies.
Mass of symbols with probability in $A$

$$\nu_1(A) := \sum_{i:f(i)\in A} f(i)$$

Counting measure:

$$\nu(A) := \sum_i \mathbb{I}_{\{f(i)\in A\}}$$

## Useful identities

$$\begin{aligned} \nu_1([0,x[) &= \sum_{j>\vec{v}(x)} f_j \\ \sum_{i\leq\vec{v}(x)} \phi(f_i) &= \phi(x)\vec{v}(x) + \int_x^1 \phi'(v)\vec{v}(v)\mathrm{d}v \end{aligned}$$

Return

# Bounds on minimax redundancy

## Redundancy-Capacity theorem

For any prior $\mu$ on $\Lambda^1(f)$

$$R^+(\Lambda^n) \quad = \quad I(\theta; X_{1:n})$$

## For an ad hoc prior

$$I(\theta; X_{1:n}) \geq \mathbb{E} Z_n$$

where $Z_n$ is the number of distinct symbols in $X_{1:n}$

$$\mathbb{E} Z_n \geq m_n$$

where $m_n$ satisfies $\overline{F}_c(m_n) \approx \frac{m_n}{n}$

## For light-tailed envelopes

$$R^+(\Lambda_f^n) \sim \log(e) \int_1^n \frac{U_c(x)}{2x} \mathrm{d}x \, (1+o(1))$$

Bontemps, B. & Gassiat, 2014 using Haussler & Opper, AoS, 1997

Return

# Censuring codes: sketch

$x_{1:n}$

5 15 8 1 30 7 1 2 1 8 4 7 15 1 5 17 13 4 12 12

$m_{1:n}$

5 15 15 15 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30

$\tilde{x}_{1:n} \curvearrowright$ progressive KT encoding

0 0 8 1 0 7 1 2 1 8 4 7 15 1 5 17 13 4 12 12

$\tilde{m} \curvearrowright$ Elias encoding

6 11 16

Return

# Light-tailed envelopes

## Decomposing redundancy of AC-code

## Decomposing pointwise redundancy

$$-\log Q^n(X_{1:n}) + \log \mathbb{P}^n(X_{1:n}) \quad = \quad \underbrace{\ell(C_E)}_{\text{I}} + \underbrace{\ell(C_M) + \log \mathbb{P}^n(X_{1:n})}_{\text{II}} .$$

## Establishing main theorem in (BBG, 2014)

$\hookrightarrow$

- (I) (Elias encoding of increments between records) is negligible with respect to $R^+(\Lambda_f^n)$, uniformly for $\mathbb{P} \in \Lambda_f$,
- The expected value of (II) is upper bounded, uniformly for $\mathbb{P} \in \Lambda_f$, by a term which is equivalent to $R^+(\Lambda_f^n)$.

Return

# Light-tailed envelopes

## Stochastic behavior of $M_n$

Let $X_1, \ldots, X_n \sim_{i.i.d.} P \in \Lambda_f^1$, let $M_n = \max(X_1, \ldots, X_n)$, then,

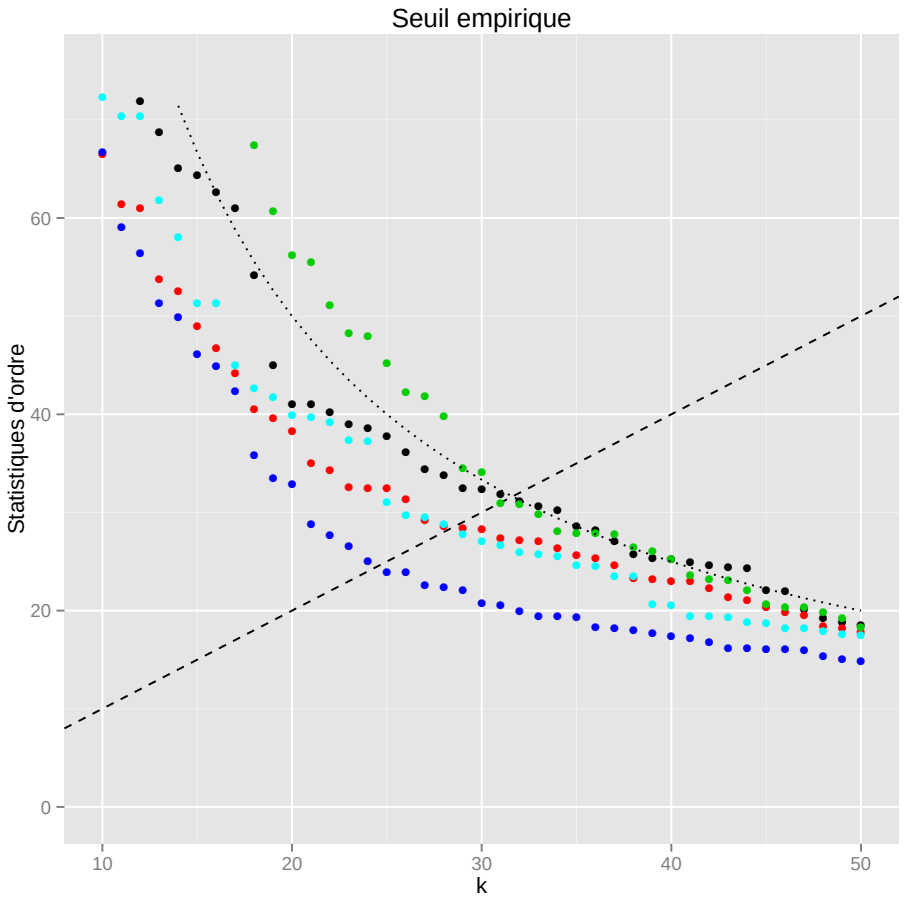
$$\begin{aligned} \mathbb{E}M_n &\leq U_c(en) + 1 \\ \mathbb{E}[M_n \log M_n] &\leq [U_c(en) + 1] \log[U_c(en) + 1] + 2/b^2. \end{aligned}$$

## Ingredients of proof

- Rényi's representation of order statistics & concavity of $U \circ \exp$
- Sub-additivity of relative entropy (see Ledoux, 2001, Massart, 2006)
- The entropy method $\rightarrow$ sharp tail and moment bounds for order statistics (B. & Thomas, 2012)

Return

# Weak adaptivity of ETAC encoding




$$M_n = \min\left(n, \{k \ : \ X_{k,n} \leq k\}\right)$$

$F_c \in \mathrm{MDA}(\gamma), \gamma > 0$

- $\frac{M_n}{m_n} \xrightarrow{P} 1\,.$
- $\frac{X_{M_n,n}}{m_c(n)} \xrightarrow{P} 1\,.$

$M_n$ is self-bounded

$$\mathbb{P}\left\{|M_n - \mathbb{E}M_n| \geq t\right\} \leq 2e^{\left(-\frac{t^2}{2(\mathbb{E}M_n + t)}\right)}.$$

Return