Codage et entropie métrique : l'exemple des classes enveloppes.

Aurélien Garivier, CNRS et ENST

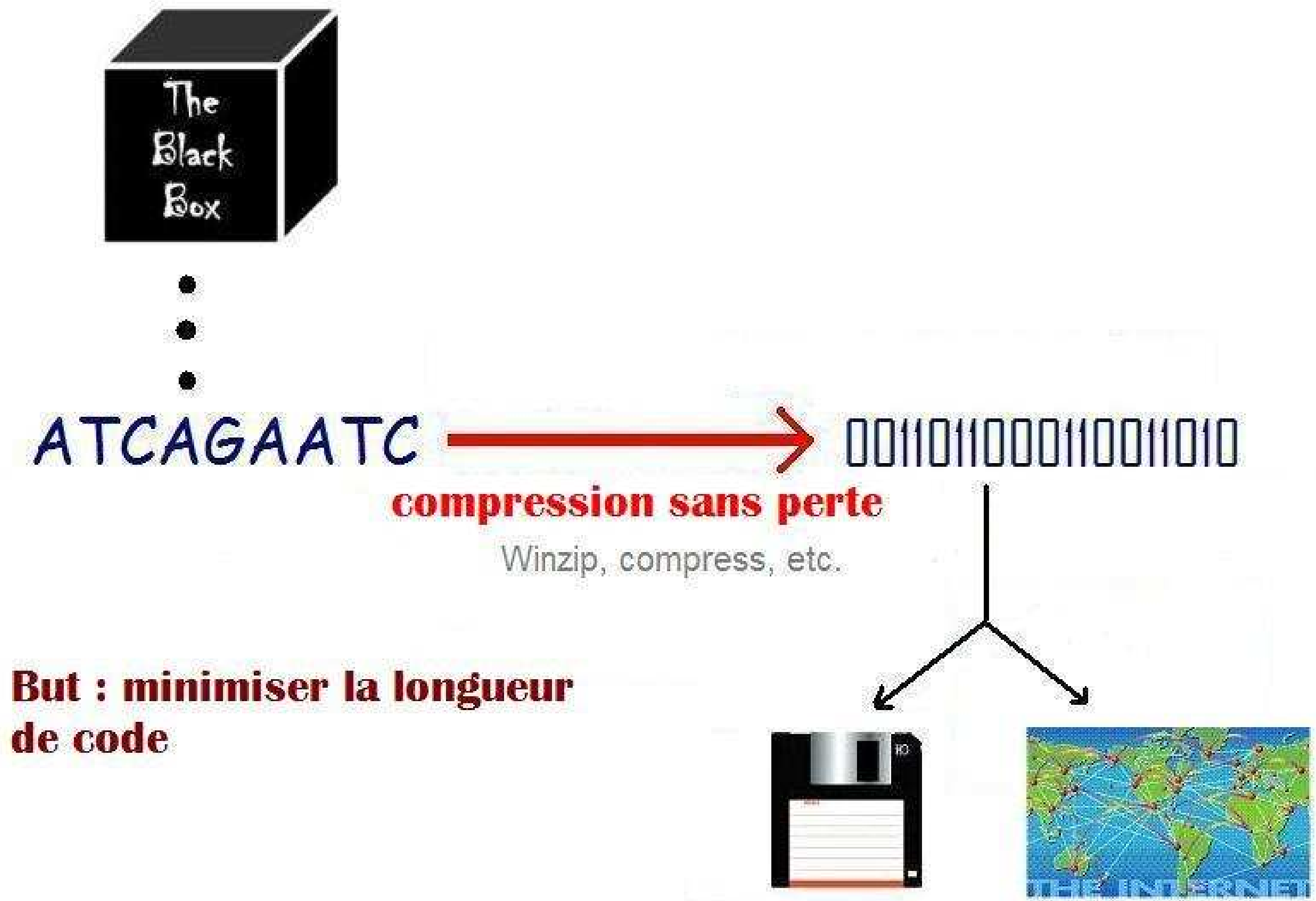
Réunion de rentrée de l'équipe MAFIA

Université Paul Sabatier - Toulouse.

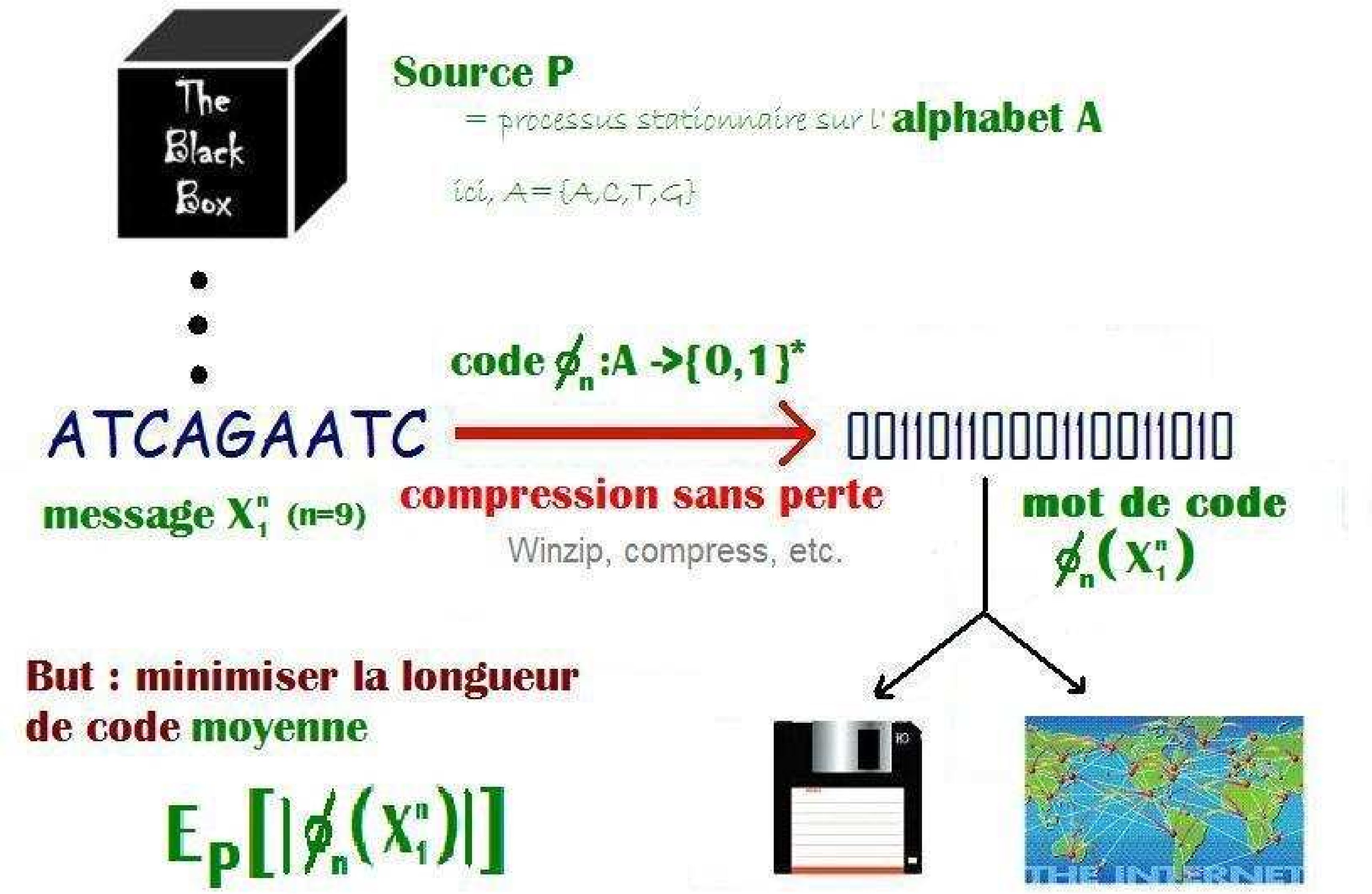
Plan de l'exposé

- .Codage source sans perte
- .Entropie métrique
- .Le cas des classes enveloppe

Codage source



Codage source : Shannon



Entropie

- **Théorème (Shannon '48)** :

$$\mathbb{E}_{\mathbb{P}}\left[|\phi_n(x)|\right] \geqslant H_n(\mathbb{P}) \triangleq \mathbb{E}_{\mathbb{P}}\left[-\log \mathbb{P}^n(X_1^n)\right]$$

  ...et il existe un code qui atteint ce taux (à 1 bit près).

- En outre,

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

  taux entropique de la source $\mathbb{P}$
  =nombre minimal de bit par symbole.

Loi de codage

- A chaque code $\phi_n(x)$ on peut associer une (sous-)probabilité q_n sur A^n telle que

$$q_n(\cdot) = 2^{-|\phi_n(\cdot)|}$$

- Inversement, grâce au codage arithmétique, on peut associer à toute (sous-)probabilité q_n sur A^n un code ϕ_n tel que $|\phi_n(\cdot)| = -\log q_n(\cdot) \quad (+Cte)$.

Conclusion: code $\phi_n \iff$ loi de codage q_n

En particulier, $-\log q_n(x) =$ longueur de code.

- le théorème de Shannon '48 dit que la meilleure loi de codage est la loi du processus !
- La perte obtenue en utilisant une autre loi q_n est appelée regret

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

Codage universel

- Que faire si la loi de la source est inconnue ??
- ...et si on veut un seul code pour plusieurs sources à la fois ?

$\Longrightarrow$ on a besoin d'une seule loi de codage q_n pour tout une classe de sources

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

c'est donc un problème d'estimation de densité...

$\Longrightarrow$ Redondance inévitable :

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

...avec perte logarithmique = information de Küllback-Leibler entre P_θ et q_n.

Exemples de codeurs universels

1. **Codage deux-temps**
   - Code d'abord $\hat{\theta}\left(X_1^n\right) = \arg\min_{\theta\in\Theta} -\log P_\theta\left(X_1^n\right)\ldots$
   - $\ldots$puis X_1^n avec la loi de codage $\mathbb{P}_{\hat{\theta}}^n$.

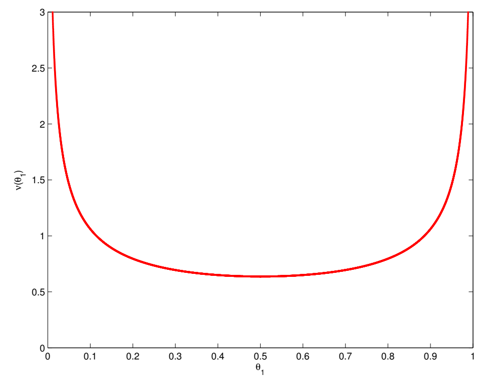
   Ex: (i.i.d.)

   $$X_1^9 = AAATACAGT : \hat{\theta} = (5,1,2,1)/9$$

   $\Longrightarrow$ regret $\frac{|A|-1}{2}\log n.$

2. **Codage par mélange** si ν est un prior sur Θ, on prend

$$q_n^\nu(x_1^n) = \int_\Theta P_\theta\left(x_1^n\right)\ d\nu\left(\theta\right)$$



Ex: (i.i.d.) $\rightarrow \nu =$ Dirichlet $\left(\frac{1}{2},\ldots,\frac{1}{2}\right)$

- Krichevsky-Trofimov inégalité de mélange

$$-\log q_n^\nu(x_1^n) \leqslant \inf_\theta -\log P_\theta^n(x_1^n) + \frac{|A|-1}{2}\log n.$$

Mesures d'universalité

1. **Redondance dans le pire des cas :**

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

2. **Redondance bayésienne** avec le prior π:

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

$$\Longrightarrow \; R^{-}(q_n, \Theta) \leqslant R^{+}(q_n, \Theta)$$

Mesures de complexité

1. **Redondance minimax**:

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

2. **Redondance bayésienne**:

$$R_{\pi,n}^-(\Theta) = \min_{q_n} \mathbb{E}_\pi [KL(P_\theta^n, q_n)]$$

$$R_n^-(\Theta) = \max_{\pi} \min_{q_n} \mathbb{E}_\pi [KL(P_\theta^n, q_n)]$$

$$\Longrightarrow \quad R_n^-(\Theta) \leqslant R_n^+(\Theta).$$

Theorème maximin (Haussler ’97, Sion)

$$R_n^-(\Theta) = R_n^+(\Theta).$$

Plan de l'exposé

- .Codage source sans perte
- .Entropie métrique
- .Le cas des classes enveloppe

Affinités et "distances"

- Pour $\alpha \neq 1$, la *D-divergence d'ordre* α entre P et Q est définie par

$$D_\alpha(P||Q) = \frac{1}{1-\alpha}\left(1 - \int (dP)^\alpha \, (dQ)^{(1-\alpha)}\right).$$

- Pour $\alpha = 1$, la *D-divergence d'ordre* 1 est l'information de Küllback-Leibler :

$$D_1(P||Q) = KL(P||Q) = \int dP \log \frac{dP}{dQ}.$$

- D_α est positive et croissante en α.
- $D_{\frac{1}{2}} = D^2(P,Q)$ définit une vraie métrique : le carré de la distance de Hellinger.

La borne de Haussler-Opper

On se concentre maintenant sur le cas i.i.d.

Pour tout prior π sur Θ, pour tout $\alpha \in]0,1[$ et pour tout $n \geqslant 1$:

$$-\int_\Theta d\pi(\theta^*) \log \int_\Theta d\pi(\tilde{\theta}) e^{-n(1-\alpha)D_\alpha(P_{\theta^*}||P_{\tilde{\theta}})} \leqslant R^-_{n,\pi} \leqslant -\int_\Theta d\pi(\theta^*) \log \int_\Theta d\pi(\tilde{\theta}) e^{-nD_1(P^*_\theta||Q_{\tilde{\theta}})}.$$

Entropie métrique

- $\mathcal{D}_\epsilon(\Theta, D_{\frac{1}{2}})$ = cardinal minimal d'une partition de Θ de D-diamètre ϵ.
- $\mathcal{M}_\epsilon(\Theta, D_{\frac{1}{2}})$ = cardinal maximal d'une partie de Θ ϵ-séparée.
- **Propriété :**

$$\mathcal{M}_\epsilon(\Theta, D) \leqslant \mathcal{D}_\epsilon(\Theta, D) \leqslant \mathcal{M}_{\frac{\epsilon}{2}}(\Theta, D).$$

- Si la limite existe quand $\epsilon \longrightarrow 0$,

$$\frac{\log \mathcal{D}_\epsilon\left(\Theta, D_{\frac{1}{2}}\right)}{\log \frac{1}{\epsilon}} \longrightarrow \dim(\Theta) = \text{dimension métrique de } \Theta.$$

Bornes sur la redondance

- Soit

$$b(\epsilon) = \sup\left\{ \frac{D_1(P_{\tilde{\theta}}||P_{\theta^*})}{D_{\frac{1}{2}}(P_{\tilde{\theta}}||P_{\theta^*})} : \tilde{\theta}, \theta^* \in \Theta \text{ et } D_{\frac{1}{2}}(P_{\tilde{\theta}}||P_{\theta^*}) \leqslant \epsilon \right\}.$$

- Alors

$$\sup_{\epsilon>0} \min\left\{ \log \mathcal{D}_\epsilon(\Theta, D_{\frac{1}{2}}), n\frac{\epsilon^2}{8} \right\} - \log 2 \leqslant R^-_{n,\pi}$$
$$\leqslant \inf_{\epsilon>0} \left\{ \log \mathcal{D}_\epsilon(\Theta, D_{\frac{1}{2}}) + b(\epsilon) n \epsilon^2 \right\}.$$

Applications

- Si $\Theta < \infty$ alors
$$R_n^+ \longrightarrow \log|\Theta| \text{ quand } n \longrightarrow \infty.$$
- Si $\dim(\Theta, D_{\frac{1}{2}}) = 0$ alors
$$R_n^+ = o\left(logn\right).$$
- Si $\dim(\Theta, D_{\frac{1}{2}}) = k < \infty$ alors
$$R_n^+ \sim \frac{k}{2}\log n.$$
- Si $\log \mathcal{D}_\epsilon(\Theta, D_{\frac{1}{2}}) \sim \left(\frac{1}{\epsilon}\right)^\alpha$ alors
$$R_n^+ \approx n^{\frac{\alpha}{\alpha+2}}.$$

Plan de l'exposé

- .Codage source sans perte
- .Entropie métrique
- .Le cas des classes enveloppe

Les classes enveloppes

- **Définition** Soit $f : \mathbb{N}_+ \mapsto [0,1]$. La **classe enveloppe** Λ_f associée à la fonction f est la collection de sources sans mémoire dont chaque marginale est dominée par f :

$$\Lambda_f = \left\{ P \; : \; \forall x \in \mathbb{N}_+, \; P^1(x) \leqslant f(x) \;, \text{ et } P \text{ sans mémoire} \right\}.$$

- **Théorème (Boucheron-G.-Gassiat '06)**

$$R_n^+(\Lambda_f) < \infty \iff \sum_{k \in \mathbb{N}_+} f(k) < \infty.$$

Enveloppe algébrique

Theorem (Boucheron-G.-Gassiat '06): Soit $\alpha > 1$, $\zeta(\alpha) = \sum_{k\geqslant 1} \frac{1}{k^\alpha}$, et C tel que $C\zeta(\alpha) \geqslant 2^\alpha$. Soit $\Lambda_{C\,\cdot^{-\alpha}}$ la classe enveloppe associée à la fonction lentement décroissante :

$$f_{\alpha,C} : x \mapsto \frac{C}{x^\alpha}.$$

Alors

$$n^{1/\alpha} A(\alpha) \log \lfloor C\zeta(\alpha) \rfloor \leqslant R_n^-(\Lambda_{C\,\cdot^{-\alpha}}) \leqslant \left(\frac{2Cn}{\alpha - 1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1)\,.$$

avec

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

Enveloppe exponentielle

Theorème (Boucheron-G.-Gassiat '06): Soit C et α une constante positive telle que $C > e^{2\alpha}$. Soit $\Lambda_{Ce^{-\alpha\cdot}}$ la classe enveloppe associée à la fonction plus rapidement décroissante

$$f_{\alpha,C} : x \mapsto C\mathrm{e}^{-\alpha x}.$$

Alors

$$R_n^+(\Lambda_{Ce^{-\alpha\cdot}}) \sim \frac{1}{4\alpha}\log^2 n.$$

Algorithme CensoringCode

Etant donnée une chaîne $x \in \mathbb{N}_+^n$ et une stratégie de censure $(K_i)_{1 \leqslant i \leqslant n}$, soit $\tilde{x} \in \{0, K_n\}^n$ défini par

$$\tilde{x}_i = \begin{cases} x_i & \text{if } x_i \leqslant K_i \\ 0 & \text{otherwise (le symbole est censuré)} \end{cases}$$

Soit également $\check{\mathrm{x}}$ la séquence des symboles censurés, c'est-à-dire $(x_i)_{x_i > K_i, i \leqslant n}$.

Algorithme (Boucheron-G.-Gassiat '06): Code séparément

- $\tilde{x}_i$ avec un codeur efficace sur l'alphabet $\{0, K_n\}^n$,
- et $\check{x}$ avec un code d'Elias pour les entiers

Ex: si $\forall i, K_i = 6$ alors

x	2	1	7	3	5	9	2	1	1	2	3	6	2	2	9	8	1	2	...
$\tilde{x}$	2	1	0	3	5	0	2	1	1	2	3	6	2	2	0	0	1	2	...
$\check{x}$			7			9									9	8			...

Performance - Adaptativité

Theorème (Boucheron-G.-Gassiat '06) Soit C et α des nombres positifs. On considère la stratégie de censure $(K_i)_{i \leqslant n}$ donnée par

$$K_i = \left\lfloor \left(\frac{2C\,i}{\alpha - 1} \right)^{1/\alpha} \right\rfloor .$$

La redondance moyenne de l'algorithme `CensoringCode` sur la classe $\Lambda_{C\cdot^{-\alpha}}$ vérifie :

$$R_n^+ \left(\texttt{CensoringCode}, \Lambda_{C\cdot^{-\alpha}}\right) \leqslant \left(\frac{2Cn}{\alpha - 1} \right)^{\frac{1}{\alpha}} \log n \,(1 + o(1)) \,.$$

- presque optimal: à un facteur $\log n$ de la borne inf $n^{1/\alpha} A(\alpha) \log \lfloor C\zeta(\alpha) \rfloor \leqslant R_n^- \left(\Lambda_{C\cdot^{-\alpha}}\right)$.
- nous proposons aussi une estimation adaptative de la censure avec $\hat{K}_n = \mu C_n$, où $C_n = \text{Card}\{X_1, \ldots, X_n\}$ est le nombre de symboles différents dans le message.