Méthodes contextuelles et alphabets infinis en théorie de l'information

Aurélien Garivier, Université Paris Sud Orsay

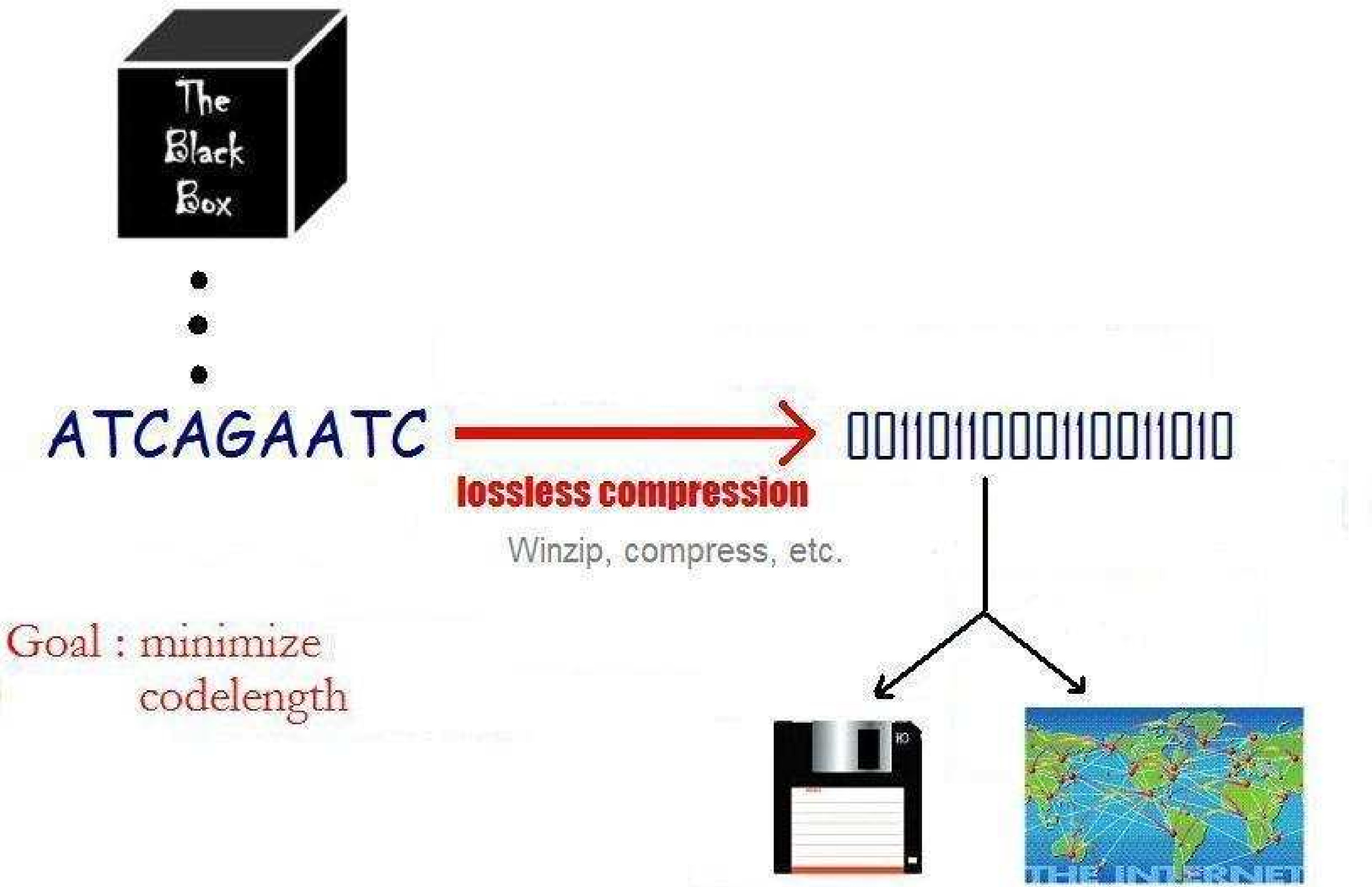
Sous la direction d'Elisabeth Gassiat (Paris XI) et Stéphane Boucheron (Paris VII)

Université Paris Sud Orsay.

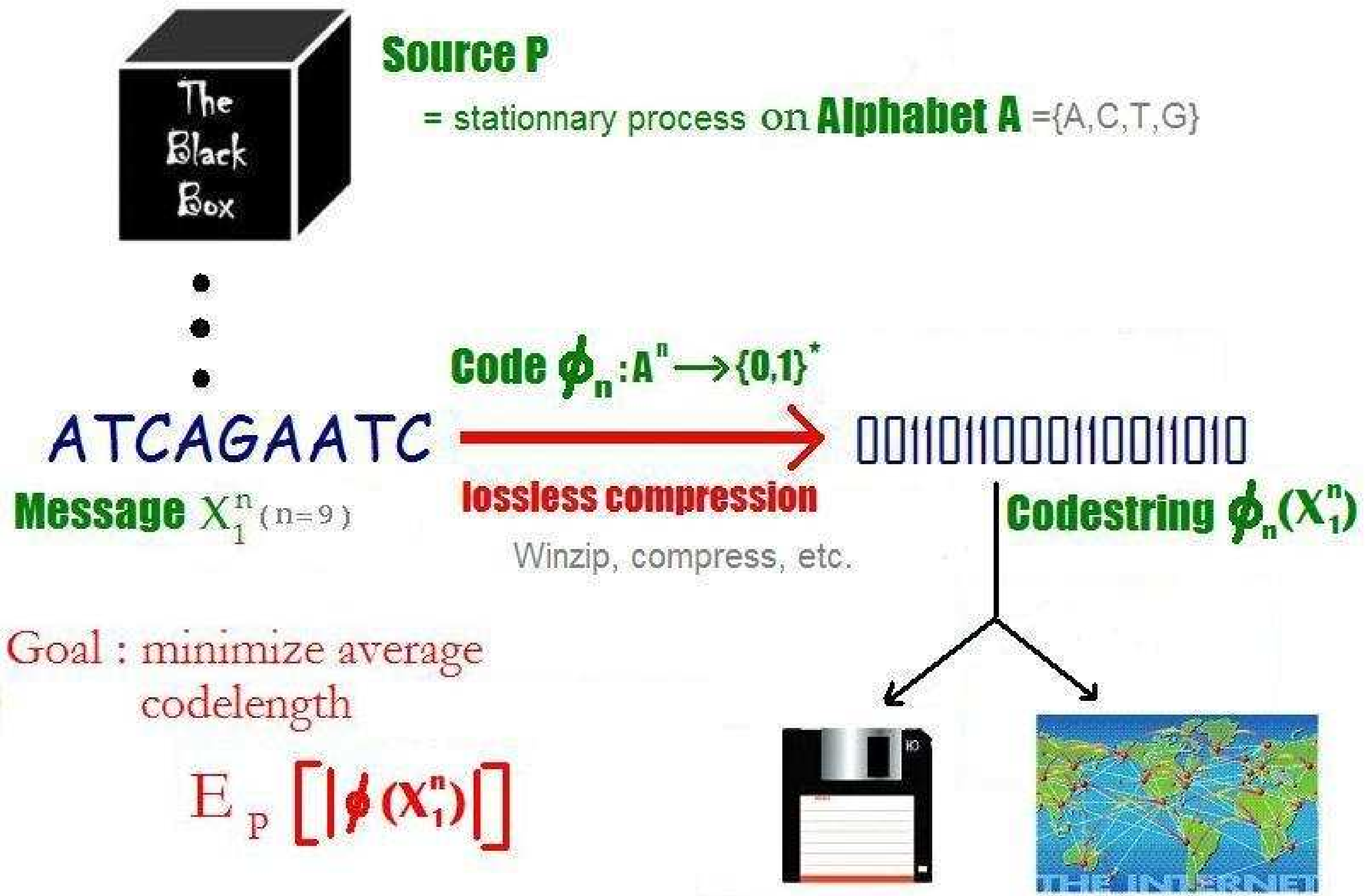
Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
- Application to model selection

Source coding



Source coding: Shannon model



Entropy

- **Theorem (Shannon '48)** :

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

  and there is a code reaching the bound (within 1 bit).

- Moreover,

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

  entropy rate of the source $\mathbb{P}$
  = minimal number of bits necessary per symbol.

Coding distribution

- Every code $\phi_n(x) \leftrightarrow$ can be associated the measure q_n on A^N such that

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

By the Kraft inequality, q_n is a (sub-)probability measure.

- Conversly, thru arithmetic coding, every (sub-)probability measure q_n on A^n can be associated a code ϕ_n such that $|\phi_n(\cdot)| = -\log q_n(\cdot) \quad (+Cte)$.

Conclusion: $\phi_n \leftrightarrow q_n$

in particular, $-\log q_n(x) =$ codelength.

- The Shannon '48 theorem expresses that the best coding distribution is the real probability !
- Coding distribution q_n suffers from **regret**
$-\log q_n\left(X_1^n\right) - \left(-\log P^n\left(X_1^n\right)\right) = \log \frac{P^n(X_1^n)}{q_n(X_1^n)}$.

Universal Coding

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

$\Longrightarrow$ We need a single coding distribution q_n for a whole class of sources

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

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

$\Longrightarrow$ unavoidable redundancy:

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

Küllback-Leibler information between P_θ and q_n

Two ideas for universal coding

1. **Two-step coding**
   - First transmit $\hat{\theta} = \arg\min_{\theta \in \Theta} -\log P_\theta\left(x_1^n\right) = \arg\max_{\theta \in \Theta} P_\theta^n\left(x_1^n\right)$.
   - Then code string x_1^n with coding distribution $\mathbb{P}_{\hat{\theta}}$.

   Ex: (memoryless model) $x_1^9 = AAATACAGT$: $\hat{\theta} = (5, 1, 2, 1)$

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

2. **Mixture coding** if ν is a probability measure on Θ, take

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

   Ex: Memoryless model. Choose $\nu =$ Dirichlet $\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$
   - Bayesian conjugate prior $\Longrightarrow$ easy computations.
   - Krichevsky-Trofimov mixture has also regret $\frac{|A|-1}{2} \log n$.

Measures of universality

1. **Maximal regret**:

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

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:

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

$$\Longrightarrow \; R^-\left(q_n, \Lambda\right) \leqslant R^+\left(q_n, \Lambda\right) \leqslant R^*\left(q_n, \Lambda\right)$$

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**:

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

$$\Longrightarrow R_n^-(\Lambda) \leqslant R_n^+(\Lambda) \leqslant R_n^*(\Lambda)$$

Theorem (Haussler ’97, Sion) $R_n^-(\Lambda) = R_n^+(\Lambda)$

Moreover, minimax redundancy is achived by a mixture.

Parametric Case

- **Theorem (Shtarkov & al.)** Let $\mathcal{I}_m$ be the class of memoryless processes over alphabet $\{1,\ldots,m\}$, then

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

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

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

Covers Markov Chains, HMM, VLMC, etc.

Non-parametric case ?

1. **Theorem (Kieffer '78)** Let $\mathcal{I}_\infty$ be the class of memoryless processes over the countably infinte alphabet $\mathbb{N}_+$, then $R_n^-(\mathcal{I}_\infty) = \infty$.
   $\Longrightarrow$ no universal coding possible in general.
2. **Theorem (Shields '93)** If $\mathcal{E}$ denotes the class of all stationnary ergodic processes over alphabet $\{0,1\}$, then $R_n^-(\mathcal{E}) = \infty$.
3. **Theorem (Csiszár and Shields '96)** Let $\mathcal{R}$ be the class of renewal processes on the binary alphabet $\{0,1\}$:

$$X = \cdots 1 \underbrace{00\cdots 1}_{N_i} \underbrace{00\cdots 1}_{N_{i+1}} \cdots, \qquad N_i \overset{iid}{\sim} \mu \in \mathfrak{M}_1(\mathbb{N}).$$

Then $R_n^-(\mathcal{R}) \sim R_n^*(\mathcal{R}) = \Theta\left(\sqrt{n}\right)$.
First example of an *intermediate complexity* class.

Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
  - Infinite alphabets and enveloppe classes
  - Pattern coding
  - Redundancy of CTW on Renewal Processes
- Application to model selection

Regret of memoryless classes

- **Proposition (Boucheron-G.-Gassiat '06):** 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 there exists $C > 0$ such that :

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

- **Proposition (Boucheron-G.-Gassiat '06):** Let Λ be a class of stationary memoryless sources over a countably infinite alphabet. Let $\hat{p}$ be defined by $\hat{p}(x_1^n) = \sup_{P\in\Lambda} P^n\{x_1^n\}$. Then

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

A counter-example

Proposition (Boucheron-G.-Gassiat '06) Let $f : \mathbb{N} \mapsto [0,1[$. For $k \in \mathbb{N}$, let $p_k \in \mathfrak{M}_1(\mathbb{N})$ be defined by:

$$p_k(l) = \begin{cases} 1 - f(k) & \text{if} l = 0; \\ f(k) & \text{if} l = k; \\ 0 & \text{otherwise.} \end{cases}$$

Let Λ be the class of memoryless sources with first marginal in $\{p_1, p_2, \ldots\}$. Then

$$R^+(\Lambda^n) < \infty \iff \sup_k f(k) \log k < \infty.$$

Donc, si $\sup_k f(k) \log k < \infty$ mais si $\sum_{k \in \mathbb{N}_+} f(k) = \infty$, on a

$$R^+(\Lambda^n) < \infty \text{ mais } R^*(\Lambda^n) = \infty.$$

Envelope classes

- **Definition** Let $f : \mathbb{N}_+ \mapsto [0,1]$. The **envelope class** Λ_f defined by function f is the collection of memoryless sources with first marginal dominated by f:

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

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

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

Power-law Envelope

Theorem (Boucheron-G.-Gassiat '06): Let $\alpha > 1$, $\zeta(\alpha) = \sum_{k \geqslant 1} \frac{1}{k^\alpha}$, and C be such that $C\zeta(\alpha) \geqslant 2^\alpha$. Let $\Lambda_{C \cdot^{-\alpha}}$ be the envelope class associated with the (slowly-) decreasing function

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

Then

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

where

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

Exponential Envelope

Theorem (Boucheron-G.-Gassiat '06): Let C and α denote positive real numbers satisfying $C > e^{2\alpha}$. Let $\Lambda_{Ce^{-\alpha \cdot}}$ be the envelope class associated with the (faster-) decreasing function

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

Then

$$\frac{1}{8\alpha}\log^2 n\,(1 - o(1)) \leqslant R^-(\Lambda^n_{Ce^{-\alpha\cdot}})$$

$$\leqslant R^*(\Lambda^n_{Ce^{-\alpha\cdot}}) \leqslant \frac{1}{2\alpha}\log^2 n + O(1).$$

Algorithm CensoringCode

Given a string $x \in \mathbb{N}_+^n$ and a cutoff strategy $(K_i)_{1\leqslant i \leqslant n}$, let $\tilde{x} \in \{0, K_n\}^n$ be defined by

$$\tilde{x}_i = \begin{cases} x_i & \text{if } x_i \leqslant K_i \\ 0 & \text{otherwise (the symbol is censored),} \end{cases}$$

Let also string $\check{x}$ be the subsequence of censored symbols, that is $(x_i)_{x_i > K_i, i \leqslant n}$.

Algorithm (Boucheron-G.-Gassiat '06): Code separately

- $\tilde{x}_i$ with an efficient universal coder on alphabet $\{0, K_n\}^n$,
- and $\check{x}$ with an Elias code.

Ex: if $\forall i, K_i = 6$ then

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 - Adaptivity

Theorem (Boucheron-G.-Gassiat '06) Let M and α be positive reals. Let the sequence of cutoffs $(K_i)_{i \leqslant n}$ be given by

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

The expected redundancy of procedure `CensoringCode` on class $\Lambda_{M\cdot^{-\alpha}}$ satisfies:

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

- almost optimal: within a factor $\log n$ from the lower bound for $n^{1/\alpha}\, A(\alpha)\, \log \lfloor C\zeta(\alpha) \rfloor \leqslant R^-\left(\Lambda_{M\cdot^{-\alpha}}\right)$.
- We propose an adaptive estimation of the cutoff by $\hat{K}_n = \mu C_n$, where $C_n = \text{Card}\{X_1, \ldots, X_n\}$ is the number of distinct symbols in the message.

Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
  - Infinite alphabets and enveloppe classes
  - Pattern coding
  - Redundancy of CTW on Renewal Processes
- Application to model selection

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: small

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

Pattern entropy & redundancy

- **Proposition (Orlitsky& al. '04, Gemelos& al '04)** Let the Pattern entropy be defined as $H\left(\Psi_1^n\right) = \mathbb{E}_{P^\Psi}\left[-\log P^\Psi\left(\Psi_1^n\right)\right]$. Then

$$\frac{1}{n} H\left(\Psi_1^n\right) \to H(\Psi) = H(X).$$

- For a pattern coding distribution q_n and a class $\Lambda = \{P_\theta, \theta \in \Theta\}$
  1. **Maximal pattern regret**: $R^*_\Psi\left(q_n, \Lambda\right) = \sup_{x_1^n \in A^n} \sup_{\theta \in \Theta} \log \frac{P_\theta^{\Psi,n}\left(x_1^n\right)}{q_n\left(x_1^n\right)}$.
  2. **Worst case pattern redundancy**: $R^+_\Psi\left(q_n, \Lambda\right) = \sup_{\theta \in \Theta} \mathbb{E}_{P_\theta^\Psi}\left[\log \frac{P_\theta^{\Psi,n}\left(\Psi_1^n\right)}{q_n\left(\Psi_1^n\right)}\right] = \sup_{\theta \in \Theta} KL\left(P_\theta^{\Psi,n}, q_n\right)$.
  3. **Expected pattern redundancy** with respect to prior π: $R^-_{\Psi,\pi}\left(q_n, \Lambda\right) = \mathbb{E}_\pi\left[\mathbb{E}_{P_\theta^\Psi}\left[\log \frac{P_\theta^{\Psi,n}\left(\Psi_1^n\right)}{q_n\left(\Psi_1^n\right)}\right]\right] = \mathbb{E}_\pi\left[KL\left(P_\theta^{\Psi,n}, q_n\right)\right]$.
- Minimax, Maximin redundancy.

A new lower-bound

- **Theorem (Orlitsky & al. '04)**

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

- **Theorem (G. '06)**

$$R^-_{\Psi,n}(\mathcal{I}_\infty) \geqslant 1.84\left(\frac{n}{\log n}\right)^{\frac{1}{3}}.$$

- The proof uses fine combinatorics on integer partitions with small summands.
- There is still a gap between lower- and upper-bounds.
  $\Longrightarrow$ are $R^-_{\Psi,n}(\mathcal{I}_\infty)$ and $R^*_{\Psi,n}(\mathcal{I}_\infty)$ of the same order of magnitude ?

Application to power-law classes

Algorithm: Code separately

- the dictionary $\Delta(x)$ with an Elias code,
- and the pattern $\psi(x)$ with an efficient pattern code.

- For $P \in \Lambda_{C \cdot^{-\alpha}}$, the code of the dictionary requires in average $O\left(n^{\frac{1}{\alpha}} \log n\right)$ bits.
- The second part has regret at most $O\left(\sqrt{n}\right)$, at least $\omega\left(n^{\frac{1}{3}}\right)$.

$\Longrightarrow$ Very simple procedure:

- efficient and adaptive for $1 < \alpha \leqslant 2$
- poor for $\alpha > 3$.

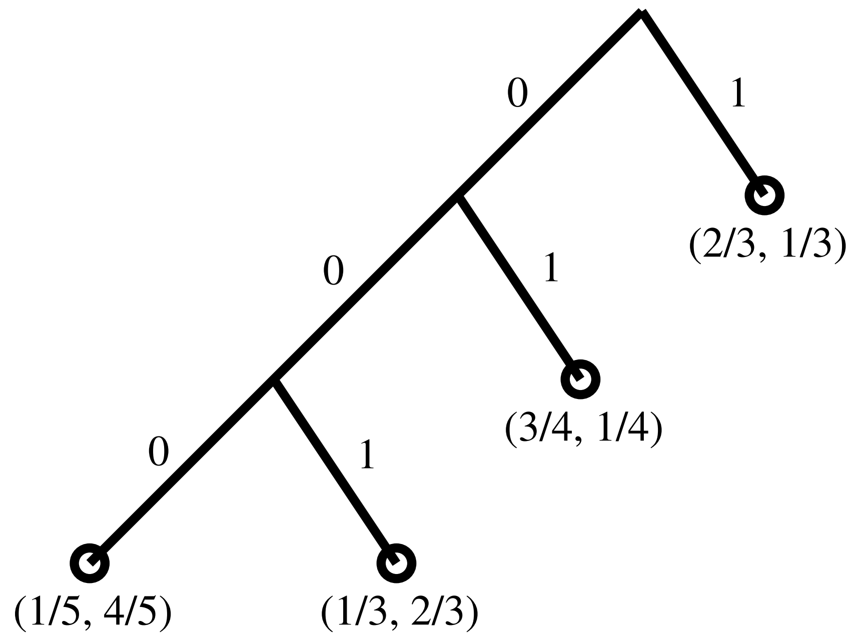
Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
  - Infinite alphabets and enveloppe classes
  - Pattern coding
  - Redundancy of CTW on Renewal Processes
- Application to model selection

Context Tree Sources

Informal Definition A Context tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$
\begin{aligned}
& P(X_1^4 = 00110 | X_{-1}^0 = 10) \\
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011)
\end{aligned}
$$



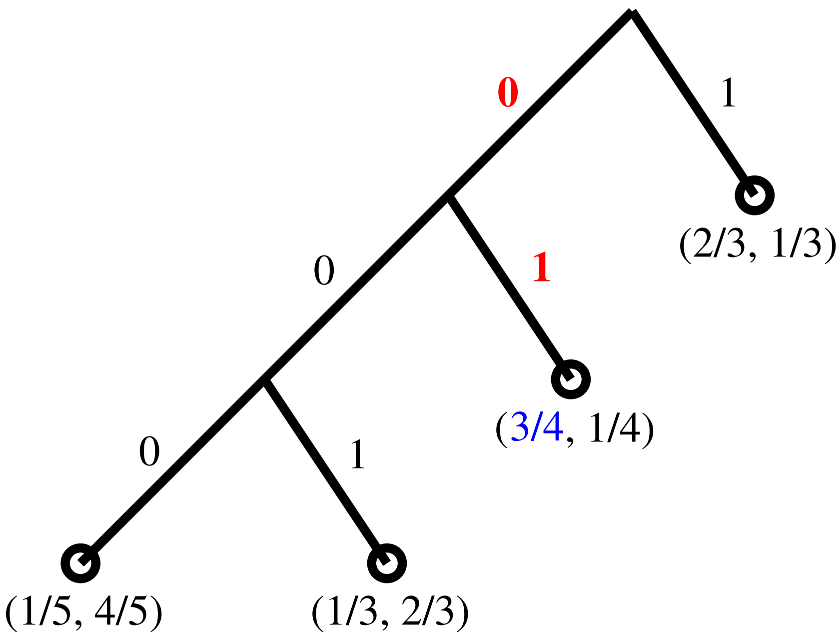
A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s \right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\}$$

Context Tree Sources

Informal Definition A Context tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$
\begin{aligned}
& P(X_1^4 = 00110 | X_{-1}^0 = 10) \\
= \quad & P(X_1 = 0 | X_{-1}^0 = \mathbf{10}) \qquad 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011)
\end{aligned}
$$



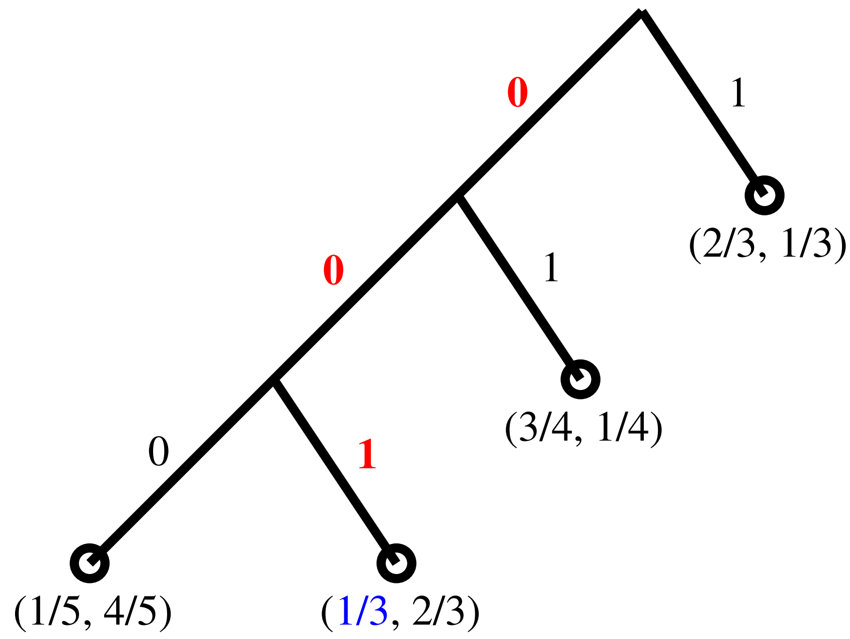
A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s \right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\}$$

Context Tree Sources

Informal Definition A Context tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$
\begin{aligned}
& P(X_1^4 = 00110 | X_{-1}^0 = 10) \\
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) \qquad 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = \mathbf{100}) \qquad 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011)
\end{aligned}
$$



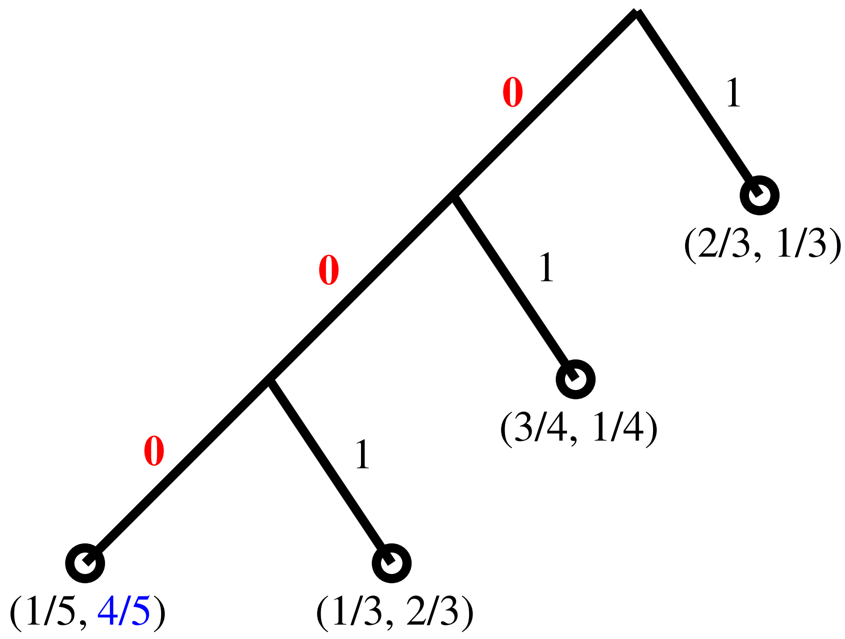
A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s \right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\}$$

Context Tree Sources

Informal Definition A Context tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$
\begin{array}{lll}
 & P(X_1^4 = 00110 | X_{-1}^0 = 10) & \\
= & P(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\
\times & P(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\
\times & P(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\
\times & P(X_4 = 1 | X_{-1}^3 = 10001) & \\
\times & P(X_5 = 0 | X_{-1}^4 = 100011) &
\end{array}
$$



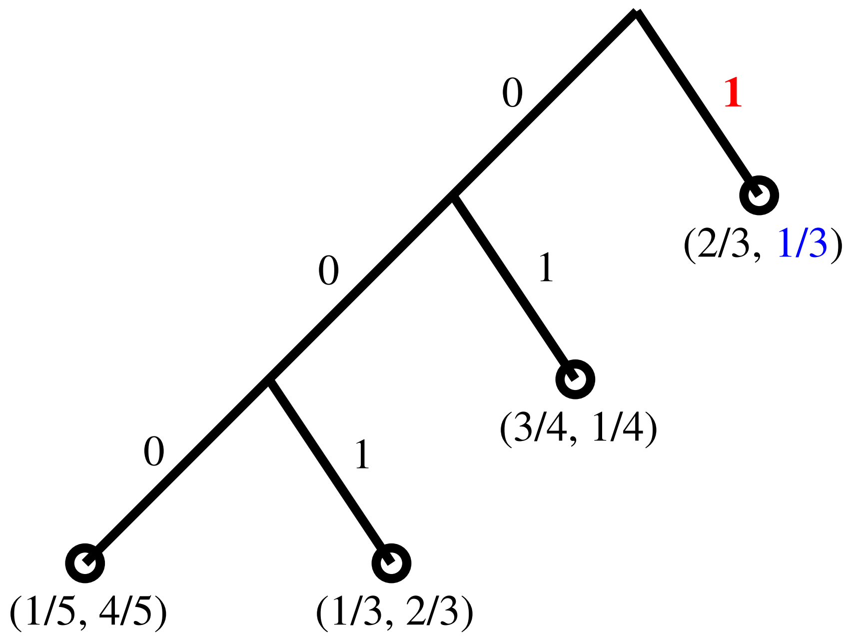
A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s\right) : s \in T, \sum_{i=1^{|A|}} \theta_i^s = 1 \right\}$$

Context Tree Sources

Informal Definition A Context tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$
\begin{array}{llr}
 & P(X_1^4 = 00110 | X_{-1}^0 = 10) & \\
= & P(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\
\times & P(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\
\times & P(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\
\times & P(X_4 = 1 | X_{-1}^3 = 10001) & 1/3 \\
\times & P(X_5 = 0 | X_{-1}^4 = 100011) &
\end{array}
$$



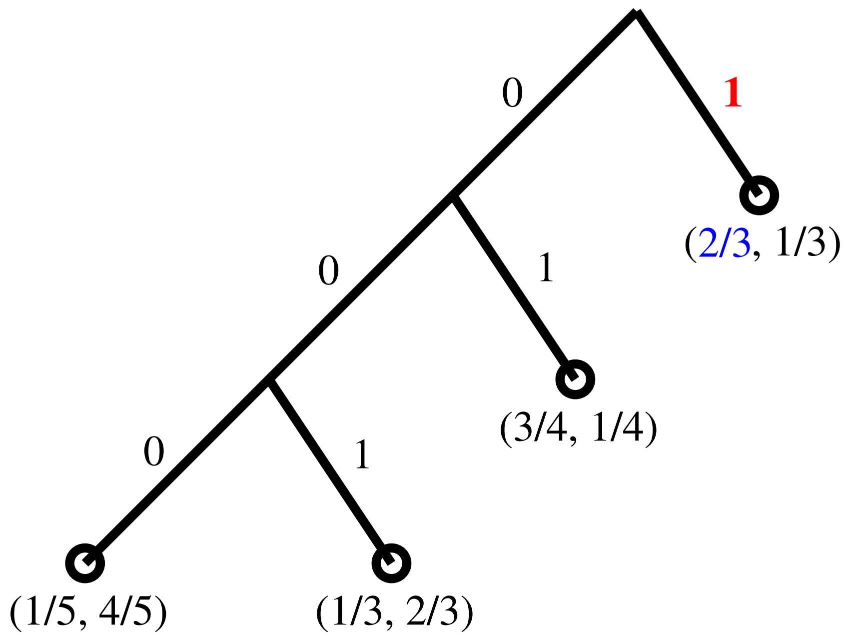
A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s \right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\}$$

Context Tree Sources

Informal Definition A Context tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$
\begin{aligned}
& P(X_1^4 = 00110 | X_{-1}^0 = 10) \\
= \quad & P(X_1 = 0 | X_{-1}^0 = 10) \quad && 3/4 \\
\times \quad & P(X_2 = 0 | X_{-1}^1 = 100) \quad && 1/3 \\
\times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) \quad && 4/5 \\
\times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) \quad && 1/3 \\
\times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) \quad && 2/3
\end{aligned}
$$



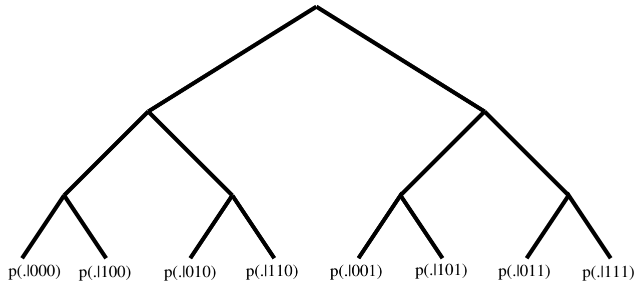
A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s\right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\}$$

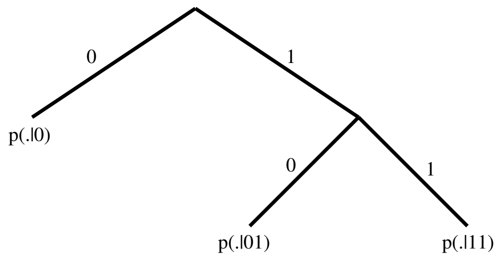
CTS versus Markov Chains

- Markov chains of order r are context tree sources corresponding to a complete tree of depth r.
  Markov chain of order 3 with transition matrix

$$M = \begin{pmatrix} p(.|000) \\ p(.|100) \\ \vdots \\ p(.|111) \end{pmatrix} \Rightarrow$$



- Finite context tree sources of depth d are Markov Chains of order d.



$$\Rightarrow \quad M = \begin{pmatrix} p(.|0) \\ p(.|0) \\ p(.|01) \\ p(.|11) \end{pmatrix}$$

$\Longrightarrow$ much more flexibility: large number of models per parameter space dimension.

Mixtures for VLMC

- Let T be a context tree with $|T|$ leaves, that is $|T|$ contexts.
- Using a product of T Dirichlet$\left(\frac{1}{2},\ldots,\frac{1}{2}\right)$ distributions as a prior for the parameter space Θ_T, we define the Krichevky-Trofimov mixture for contexts trees q_T^ν satisfying:

$$q_T^\nu(x_1^n|x_{-\infty}^0) = \prod_{s\in T} q^\nu\left(T(x,s)\right).$$

- **Proposition (Shtarkov&al '93)**

$$-\log q_T^\nu(x_1^n|x_{-\infty}^0) \leqslant \inf_{\theta\in\Theta_T} p_\theta(x_1^n|x_{-\infty}^0) + \frac{|A|-1}{2}|T|\log^+\frac{n}{T} + |T|\log m + m - 1.$$

CTW : a double mixture

- **Proposition (Sharkov & al '93)** Let T be a context tree and $|T|$ its number of leaves. Then

$$\pi(T) = 2^{-2|T|+1}$$

is a probability distribution on the set $\mathcal{T}$ of all context trees.

- **Context Tree Weighting** coding distribution :

$$q_n^{\mathrm{CTW}}(x_1^n) = \sum_T \pi(T) q_T^{\nu}(x_1^n)$$

can be computed efficiently and satisfies the oracle inequality:

$$-\log q_n^{\mathrm{CTW}}(x_1^n|x_{-\infty}^0) \leqslant \inf_{T\in\mathcal{T}} \inf_{\theta\in\Theta_T} p_\theta(x_1^n|x_{-\infty}^0)$$
$$+ \frac{|A|-1}{2}|T|\log^+\frac{n}{T} + |T|(2+\log m) + m - 2.$$

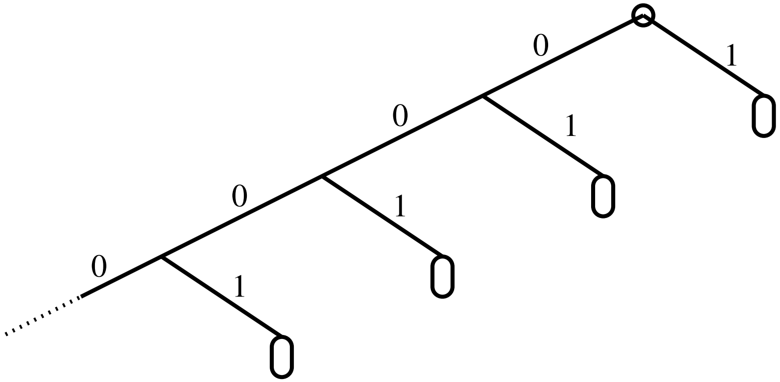
Questions on CTW

- Adaptivity on small parametric classes CTW is constructed on.
- What performance on more massive classes ?
- Csiszár and Shields result:

$$R_n^-(\mathcal{R}) \sim R_n^*(\mathcal{R}) = \Theta\left(\sqrt{n}\right)$$

was not constructive: is there a general-purpose algorithm performing well on renewal processes ?

- CTW is a good candidate since renewal processes are "infinite context tree sources".



Main redundancy result

- **Theorem (G. '04):** There exist constants C_1 and C_2 such that the regret of CTW over the class $\mathcal{R}$ of renewal processes satisfies:

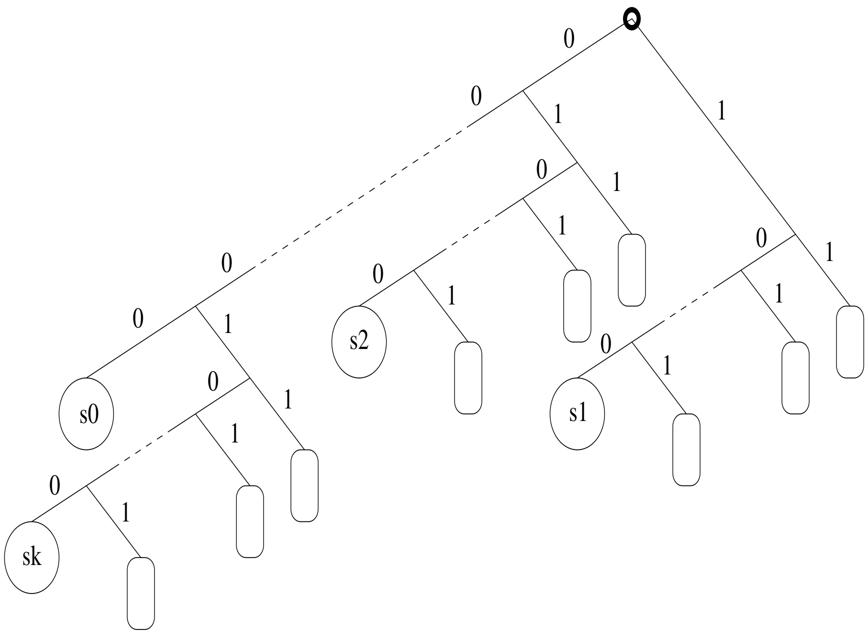
$$C_1\sqrt{n}\log n \leqslant R_n^*\left(\mathcal{R}\right) \leqslant c_2\sqrt{n}\log n.$$

- **Theorem (G. '04):** There exist constants C_3 and C_4 such that the regret of CTW over the class $\mathcal{MR}$ of Markovian renewal processes satisfies:

$$C_3 n^{\frac{2}{3}}\log n \leqslant R_n^*\left(\mathcal{MR}\right) \leqslant C_4 n^{\frac{2}{3}}\log n.$$

Comments

- Adaptivity result for CTW on a massive class.
- If the renewal distribution is bounded, CTW achieves regret $O(\log n)$ (contrary to ad-hoc coders).
- Requires deep contexts in the double mixtures ($\Longrightarrow$ the tree should not be cut off at depth $\log n$).
- Kind of non-parametric estimation: need for a balance between approximation and estimation.

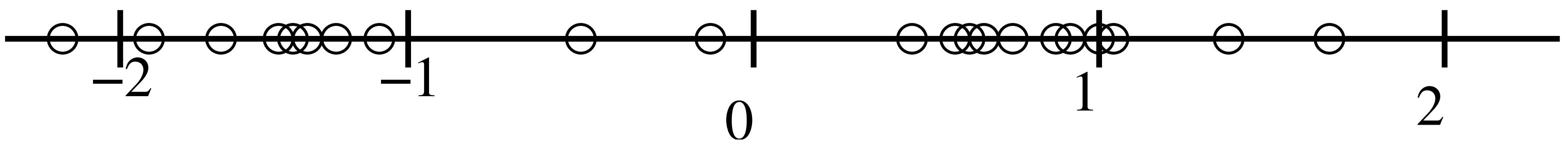


Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
- Application to model selection
  - The Minimum Description Length Principle
  - Consistency of the BIC Estimator for VLMC
  - HMM Order estimation for HMM with infinite emission

Model selection

- We are given data, say a string x_1^n,
- We suppose that it has been generated by a source that belongs to some model $\mathcal{M}_0, \mathcal{M}_1 \ldots$
- Goal: Identify that model $\mathcal{M}_j$ using data x_1^n.
  Examples :
  - DNA sequence : $x = ACCACTGACTACGACCT \ldots$
    Is it the realization of a Markov chain of order 0, 1, 2, . . . ? of which VLMC ?
  - Mixture of Gaussians with unknown number of componenents:



The MDL Principle

- Guillaume d'Ockham (XIV. century):

  Entia non sunt multiplicanda praeter necessitatem

- Jorma Rissanen ('78):

  Choose the model that gives the
  shortest description of data

- **Problem 1:** what is the description length of data in a model ?
  $\Longrightarrow$ need for an objective notion of description length.
- **Problem 1:** only a heuristic !
  $\Longrightarrow$ Provides estimators, the consistency remains to be proved.

Objective Description Length

- Information theory:

  objective codelength = codelength of a minimax coder.

- Estimator associated with optimal 2-step coder:

$$\arg\min_i \inf_{P \in M_i} -\log \hat{P}\left(x_1^n\right) + \frac{\dim M_i}{2} \log n.$$

  Coïncides with a penalized maximum likelihood estimator with a BIC penalty.

- Estimator associated with minimax mixtures $(\nu_i)_i$:

$$\arg\min_i -\log \int_{\theta \in \Theta_i} \hat{P}_\theta\left(x_1^n\right) \nu_i\left(d\theta\right).$$

Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
- Application to model selection
  - The Minimum Description Length Principle
  - Consistency of the BIC Estimator for VLMC
  - HMM Order estimation for HMM with infinite emission

Consistency

- **Theorem (Csiszár& Talata '04):** If the BIC and Mixture estimators are restricted to trees of depth smaller than $D(n)$, where $D(n) = o(\log n)$, then eventually almost-surely

$$\hat{T}_{\mathrm{BIC}\leqslant D} = \hat{T}_{\mathrm{Mix}\leqslant D} = T_0.$$

- relies on fine "typicality" results by Csiszár and Shields.
- **Theorem (G. '05):** Eventually almost-surely, the unlimited BIC estimator $\hat{T}_{\mathrm{BIC}}$ has size at most

$$\left|\hat{T}_{\mathrm{BIC}}\right| = o\left(\frac{\log n}{\log\log\log n}\right).$$

Comments

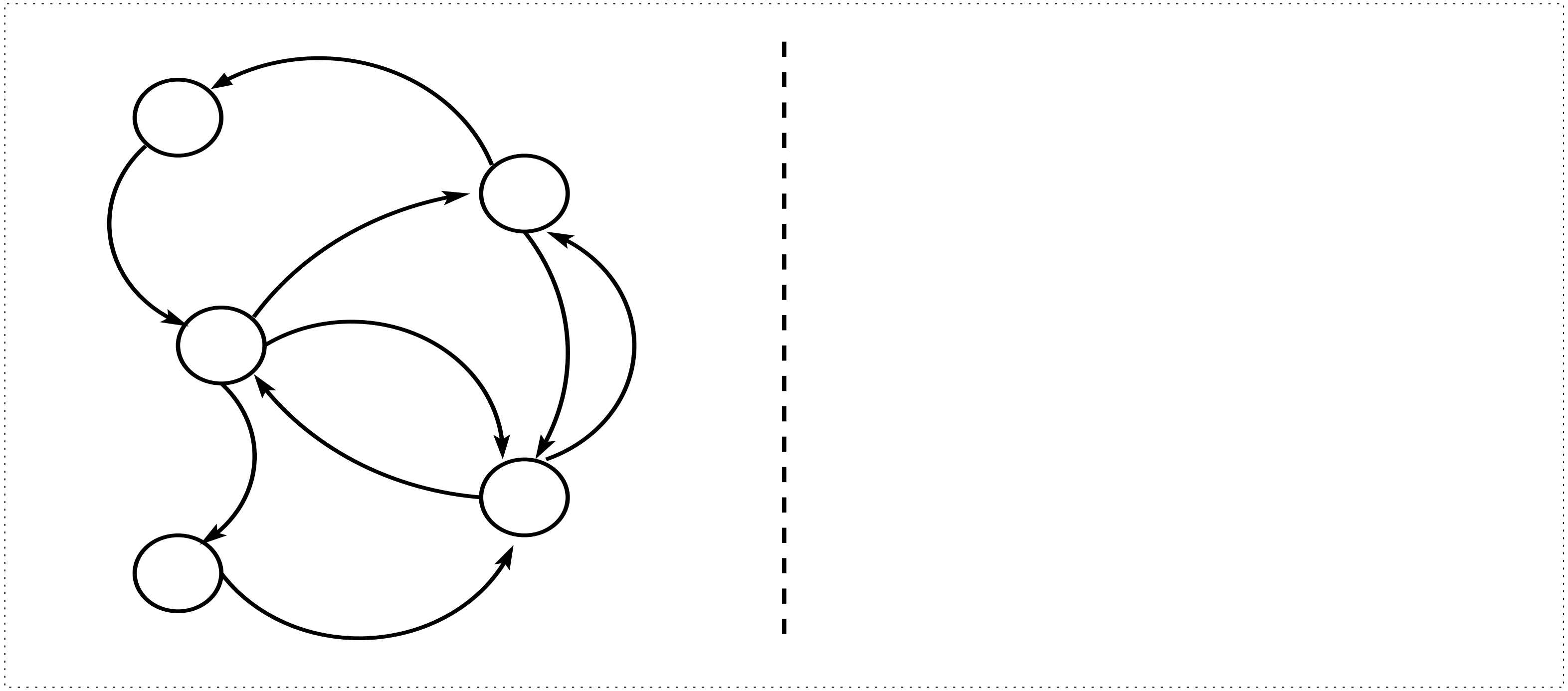
- The unlimited mixture estimator is not consistent: it fails to recognize $\mathcal{B}\left(\frac{1}{2}\right)$.
- There is an exponential number of models per dimension.
- However, there is sequential, time-linear algorithm for computing the unlimited estimators $\hat{T}_{\mathrm{BIC}}$ and $\hat{T}_{\mathrm{Mix}}$. It relies on the notion of **compact suffix tree**.

Presentation outline

- Lossless Source Coding
- Infinite alphabets, infinite memory
- Application to model selection
  - The Minimum Description Length Principle
  - Consistency of the BIC Estimator for VLMC
  - HMM Order estimation for HMM with infinite emission

Hidden Markov Models

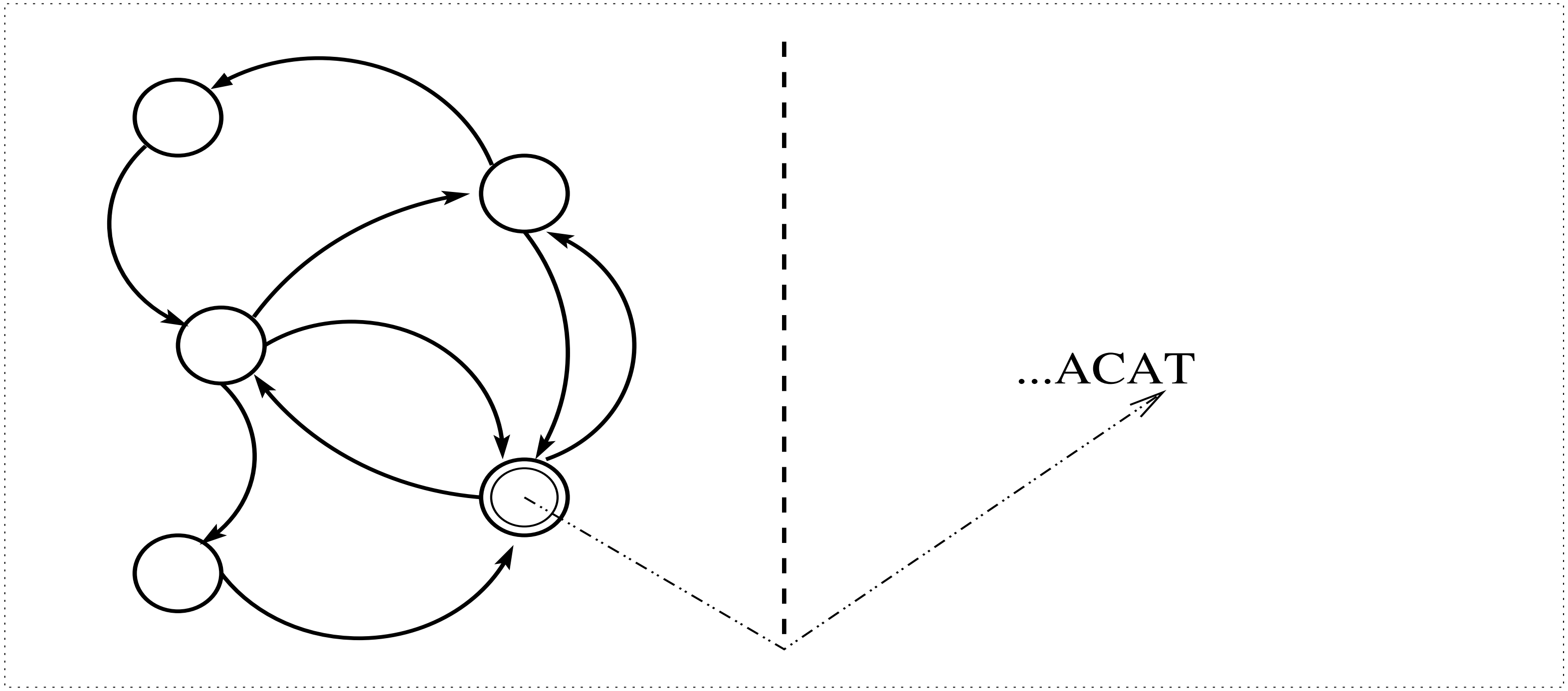
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

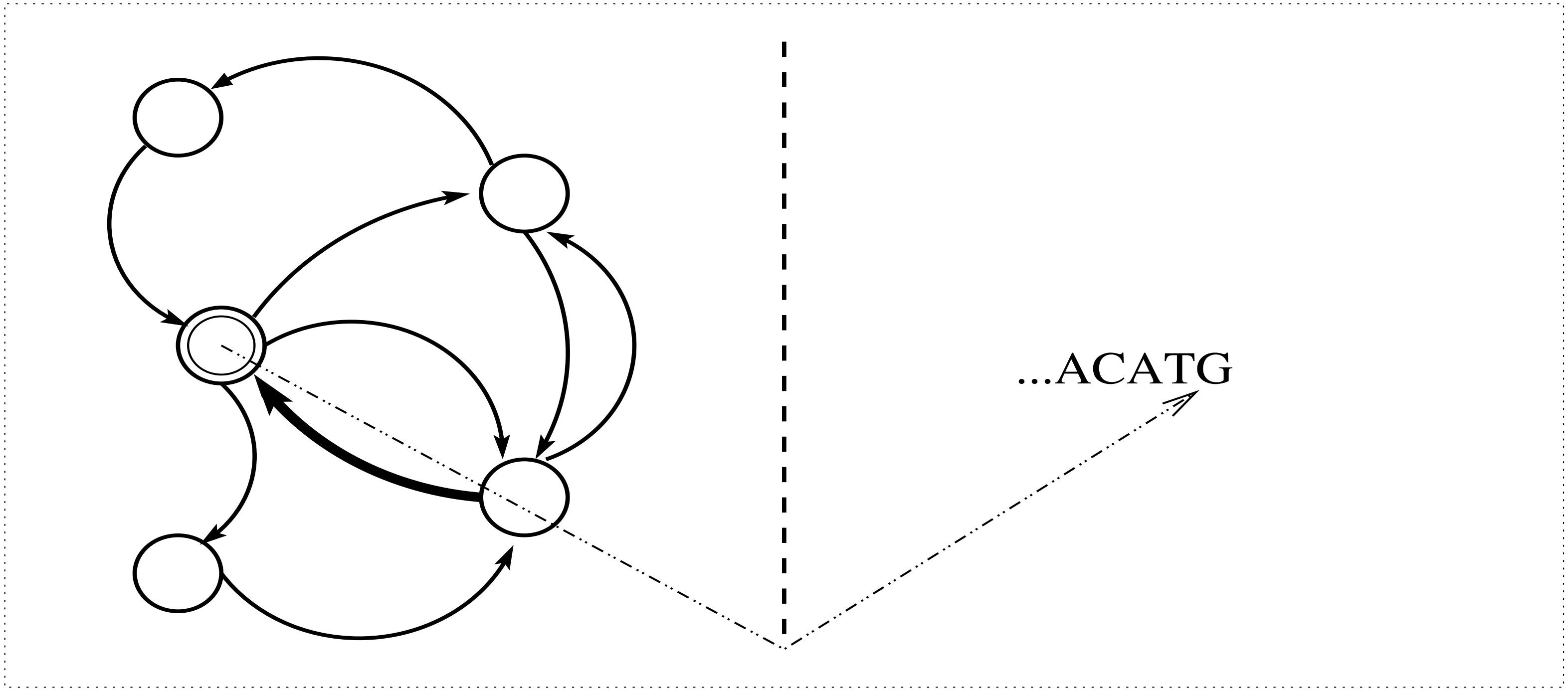
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

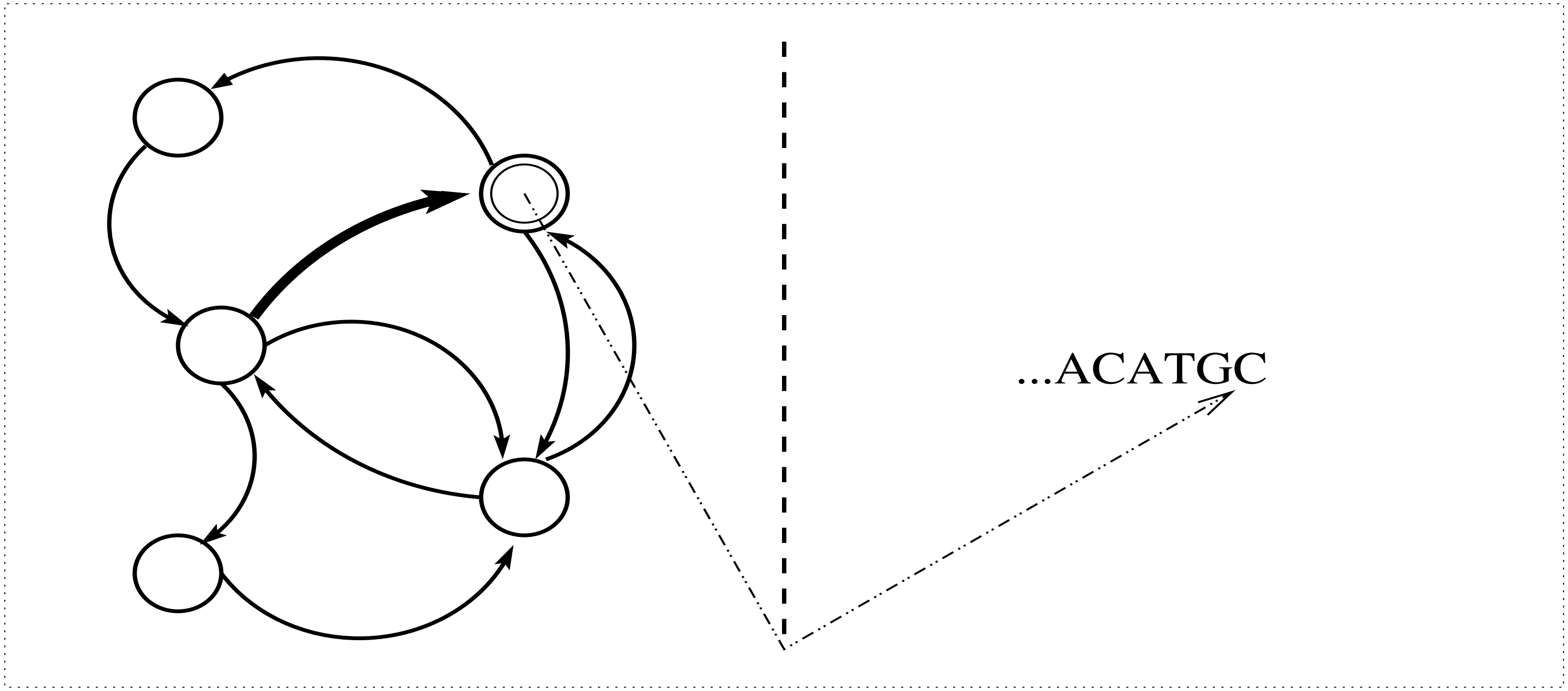
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

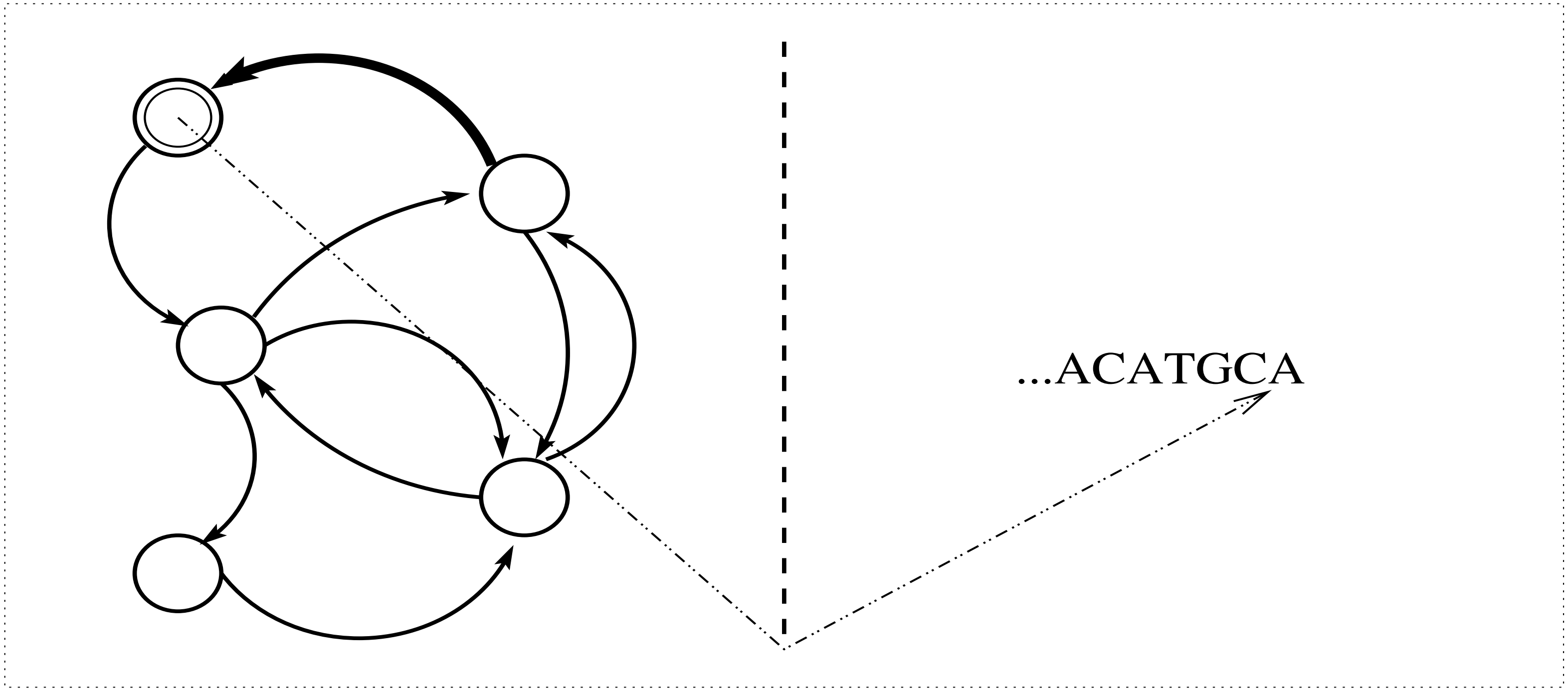
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

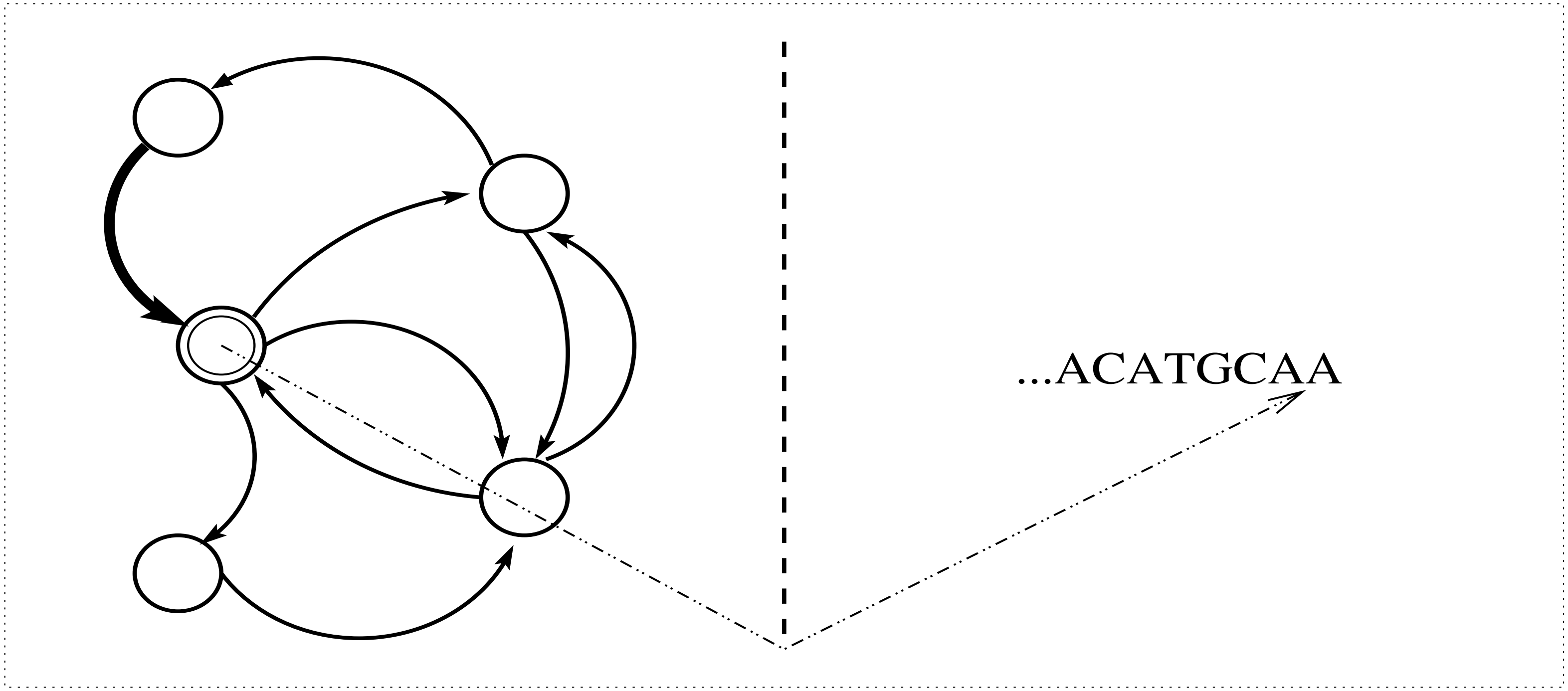
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

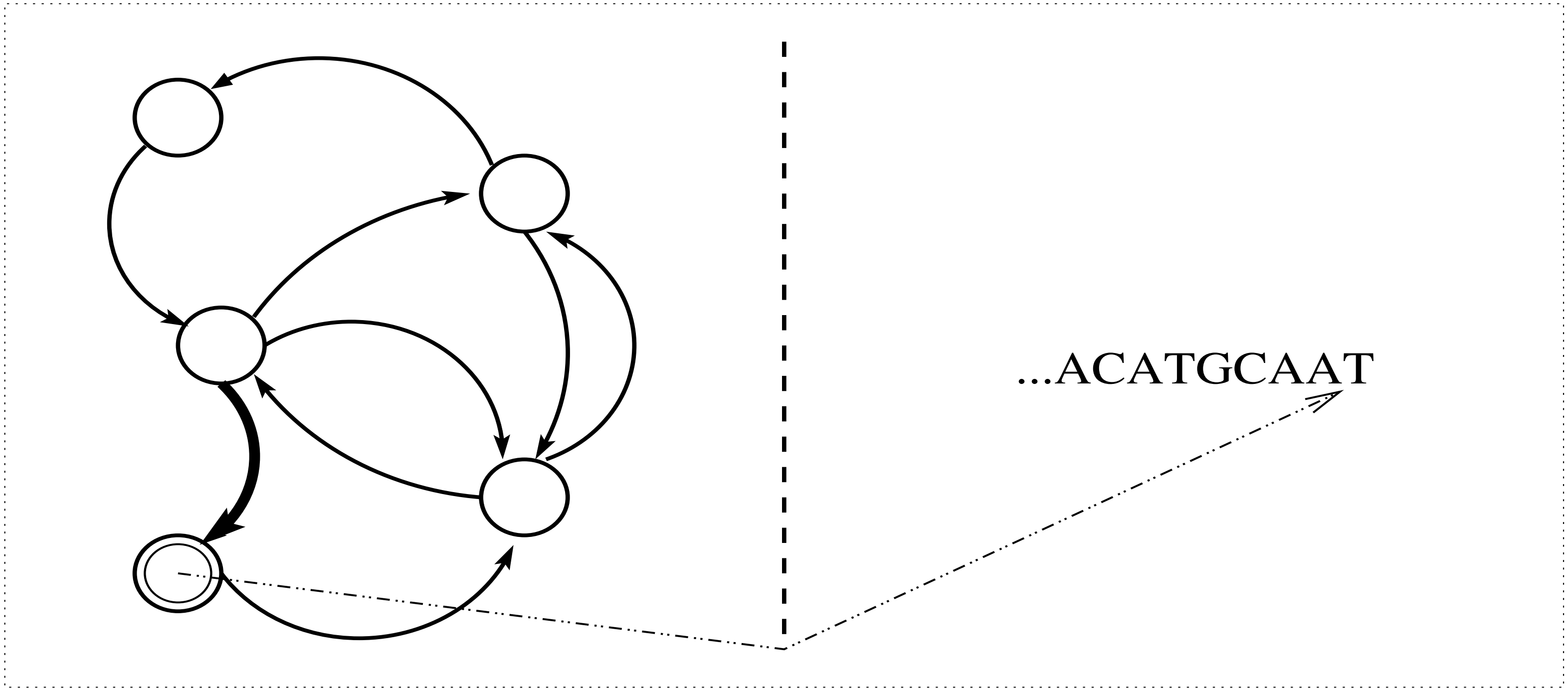
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

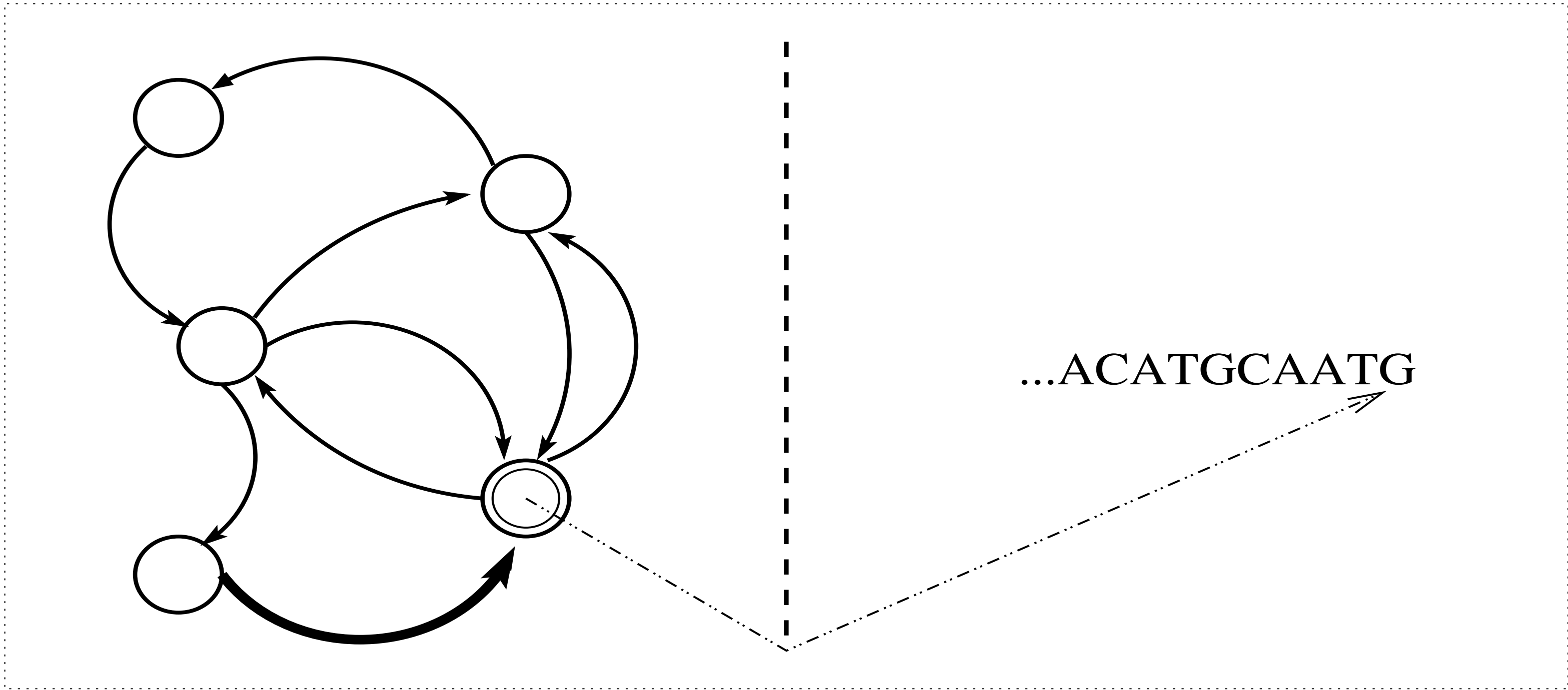
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

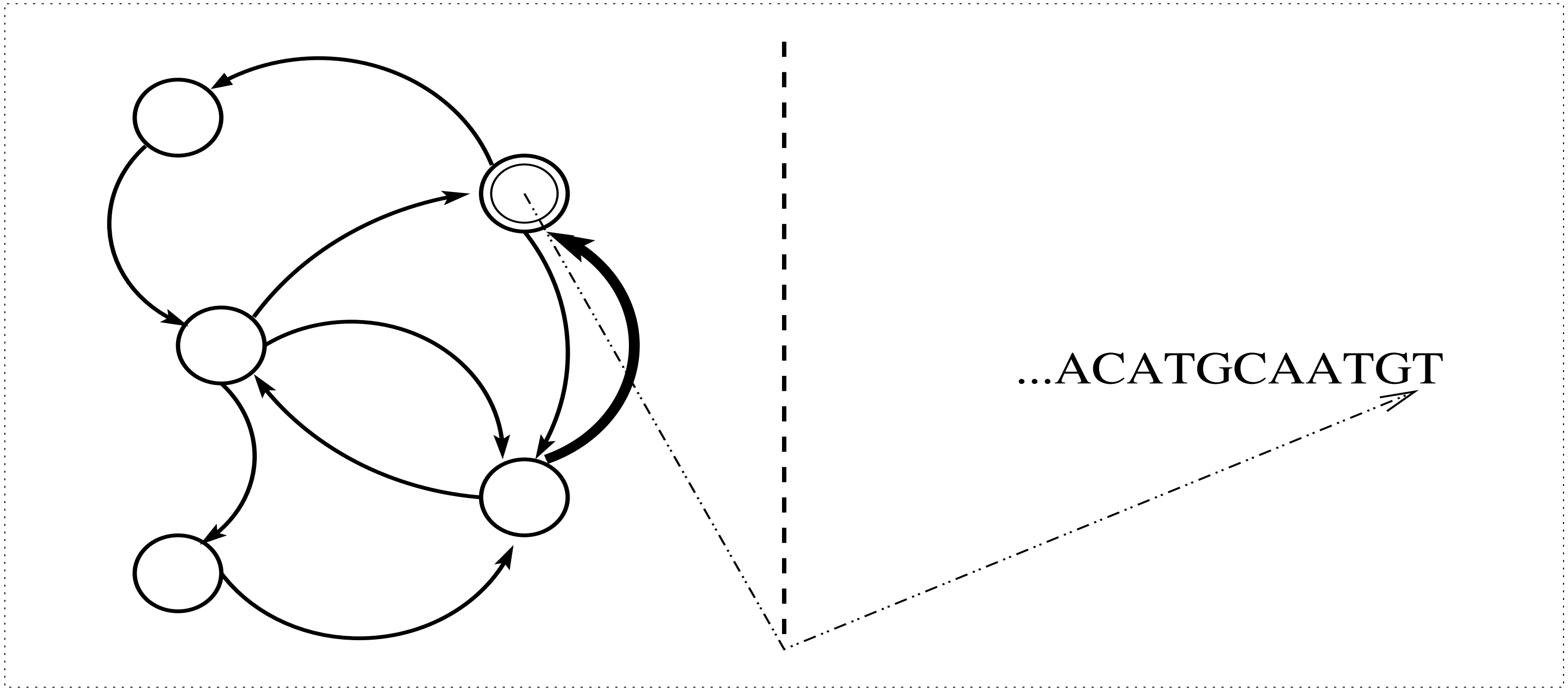
- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Hidden Markov Models

- Each hidden state z has its own emission distribution p_z.
- The process $(Z_n)_n$ of hidden states is Markovian.
- At every time n, one symbol is emitted independently with distribution p_{Z_n}.



$\Longrightarrow$ estimate the order = number of hidden states

Parameterization

Model $\mathcal{M}_k (k \in \mathbb{N})$ (of order k) : set of all HMM with k hidden states, parameterized by

$$\Theta_k = \left\{ (p_{jj'})_{1 \leqslant j,j' \leqslant k} : \sum_{j'=1}^{k} p_{jj'} = 1 \right\} \times \left\{ m = (m_1, \ldots, m_k) \in \mathbb{R}^k \right\}$$

- p is the transition kernel of the hidden Markov Chain,
- m_j is the expectation of the emission distribution in state j.

$$\dim \Theta_k = k(k-1) + k = k^2$$

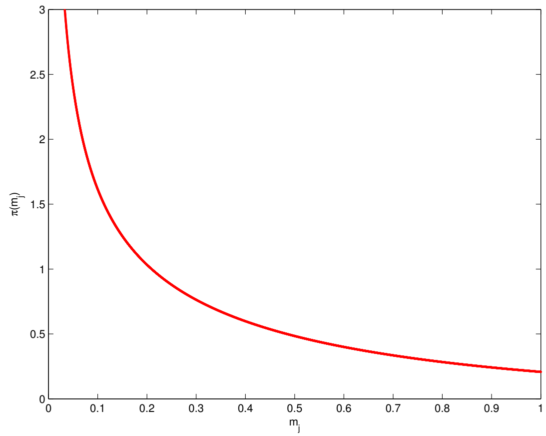
Poisson emission: conditionally on $Z_n = j$, $X_n \sim \mathcal{P}(m_j)$.

Gaussian emission: conditionally on $Z_n = j$, $X_n \sim \mathcal{N}(m_j, \sigma^2)$ where σ^2 is a fixed but unknown noise level.

Mixtures in model $\mathcal{M}_k$

Mixture q^k is obtained with prior ν_k over Θ_k such that, for a constant $\tau > 0$, we have under ν_k :

- p and m are independent;
- the iniation distribution $p^o_{j'} = 1/k$ for every $j' \leqslant k$ is deterministic,
- vectors $(p_{jj'} : j' \leqslant k)$ ($j \leqslant k$) are independent and Dirichlet$(1/2, \ldots, 1/2)$ distributed,
- parameters $m_1, \ldots, m_k$ are iid with $\mathcal{N}_{0,\tau}$ for Gaussian emission, and Gamma$(\tau, 1/2)$ for Poisson emission.



$\Longrightarrow$ we use conjugate priors with parameters inspired from Krichevsky-Trofimov mixtures.

Mixture inequalities

Proposition (Chambaz-G.-Gassiat '05): BIC-type mixture inequalities:

- Poisson emission:

$$0 \leqslant \sup_{\theta \in \Theta_k} \log \mathbb{P}_\theta(X_1^n) - \log q_n^k(X_1^n) \leqslant \frac{k^2}{2} \log n + k\tau X_{(n)} + c_{kn}.$$

- Gaussian emission:

$$0 \leqslant \sup_{\theta \in \Theta_k} \log f_\theta(X_1^n) - \log q_n^k(X_1^n) \leqslant \frac{k^2}{2} \log n + \frac{k}{2\tau^2} |X|_{(n)}^2 + d_{kn}.$$

Remark: can no longer be interpreted as codelength inequalities!

Two order estimators

- **Penalized Maximum Likelihood**:

$$\hat{k}_{ML} = \arg\min_{k \in \mathbb{N}} -\log \widehat{p_k}(x_1^n) + \mathrm{pen}(n,k).$$

- **Mixture** :

$$\hat{k}_{MIX} = \arg\min_{k \in \mathbb{N}} -\log q_k^n(x_1^n) + \mathrm{pen}(n,k).$$

- We need to penalize more than BIC (because of the maxima).
- We also need to penalize the mixture – it is often necessary.
  Ex: $B\,(1/2)$ for Markovian order

Consistency theorems

Theorem (Chambaz-G.-Gassiat '05) Let

$$S_{kn} = D_{kn} + k(k+1)\varphi_n \log n \quad \text{in the Gaussian case, and}$$
$$S_{kn} = E_{kn} + k(k+1)\frac{\log n}{\sqrt{\log\log n}} \quad \text{in the Poisson case.}$$

- If
$$\mathrm{pen}(n,k) = \sum_{\ell=1}^{k} \frac{\ell^2+\alpha}{2}\log n + C_{kn} + S_{kn},$$
then $\hat{k}_{ML} = k_0$ eventually almost surely.
- If
$$\mathrm{pen}(n,k) = \sum_{\ell=1}^{k-1} \frac{\ell^2+\alpha}{2}\log n + S_{kn}.$$
then $\hat{k}_{MIX} = k_0$ eventually almost surely.

Comments on the proofs

- Different behaviours of the maxima :
  - Poisson emission: $X_{(n)} = o(\log n)$ does not interfer with the BIC term.
  - Gaussian emission: $|X|^2_{(n)}$ is of order $\log n \implies$ we have to penalize significantly more than BIC.
- Underestimation is easy to avoid, not overestimation!
- The proofs are “imbricated”: we use the mixture inequalities even for $\hat{k}_{ML}$.
- Analog result for Gaussian and Poisson mixtures with $2i - 1$ degrees of freedom instead of i^2.
- Advantage : no need for a priori bounds on the order or on the emission parameters.
- Disadvantage: in practice, computationally difficult.