Arbres de contexte, identification et adaptivité

Application à la modélisation

Aurélien Garivier

Université Paris Sud

Orsay

Plan de l'exposé

- Notions de théorie de l'information
- Le modèle des sources à arbres de contexte
- Identification par MDL
- Prédiction et mélanges

Notations

- Soit E un alphabet fini de lettres. Les **mots** sur E sont les éléments de $E^* = \bigcup_{n=0}^{\infty} E^n$.
- Pour $x, s \in E^*$, le **nombre d'occurences** de s dans x est $N_x(s) = \sum_{i=1}^{|x|-s+1} \mathbb{1}_{x_i^{i+|s|-1}=s}$
- Pour $P \in \mathbb{P}(E)$, l'**entropie** de P est $H(P) = \sum_{a \in E} P(a) \log \frac{1}{P(a)}$
- Pour $x \in E^n$, l'**entropie empirique** (coût de codage) de x est ($\log = \log_2$):

$$H(x) = nH(\hat{p}_x) = \sum_{a \in E} N_x(a) \log \frac{n}{N_x(a)}$$

Théorie de l'information - Shannon '48

- **Problème** : **code** $f_n : E^n \to \{0,1\}^*$ injectif
- **Hypothèse** : x est la réalisation d'une **source** P_n
- $\Longrightarrow$ **coût moyen** : $C(f_n) = \sum_{x \in E^n} P_n(x)|f_n(x)|$
- Théorème (optimisation)

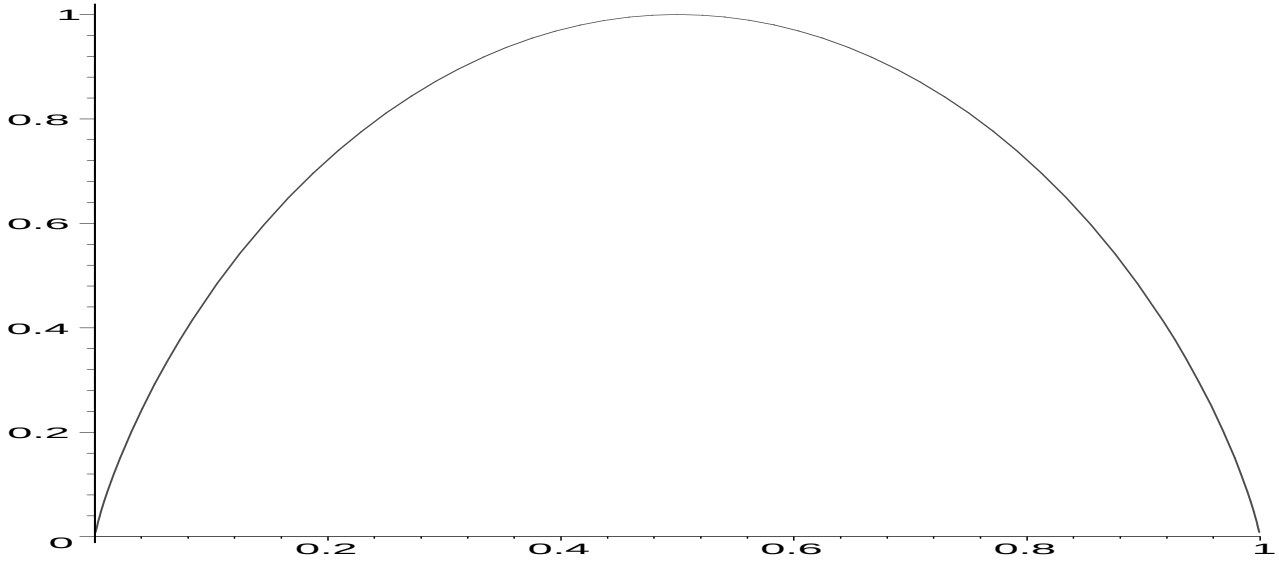
$$C(f_n) \geqslant \sum_{x \in E^n} P_n(x) \log \frac{1}{P_n(x)} \triangleq n H_n(P_n)$$

- Pour toutes les sources *stationnaires ergodiques*, $H_n(P_n) \to H(P)$ **entropie** de la source P.
- Il existe des codes s'approchant arbitrairement près du taux de codage entropique.

Exemple

$E = \{0, 1\}$, $P =$ source sans mémoire : $X_i \overset{iid}{\sim} B(p)$, $0 < p < 1$.

$$H_n(P) = H_1(p) = h(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$



Le codage arithmétique

- Idée : coder en calculant des probabilités selon une **Proba de codage** Q_n
- $f_n(x) =$ les $\lceil -\log Q_n(x) \rceil + 2$ premiers bits de $Q_n(X \leqslant x)$.

$$C(f_n) = \mathbb{E}_{P_n} -\log Q_n(X)_{(+2)} = nH_n(P) + nD_n\left(P_n | Q_n\right)_{(+2)}$$

- La meilleur proba de codage est le maximum de vraisemblance.
- La **redondance** (= perte par rapport au “coût idéal”) s’interprète mathématiquement comme une distance de Kullback-Leibler.

Codage universel

- **Problème :** si on ne connaît pas P ?
- Coder efficacement reste possible au prix d'un surcoût = redondance.
- Idée 1 (sources sans mémoire) :
  - on code la loi empirique $\rightarrow (|E|-1)\log n$ bits
  - puis on code x avec cette proba de codage $\rightarrow nH$ bits
- Idée 2 (sources sans mémoire) : Proba de codage Q=*mélange* des lois sans mémoire.

Le mélange de Krichevsky-Trofimov

- Notons $\Theta_E = \left\{\theta \in [0,1]^E : \sum_{a\in E} \theta_a = 1\right\}$ et $\mathbb{P}_\theta$ la loi du processus iid sur $E^{\mathbb{Z}}$ défini par $\mathbb{P}_\theta(X_0 = a) = \theta_a$.
- Mélange de Krichevski-Trofimov sur E^n :

$$\mathcal{KT}(x_1^n) = \int_{\theta\in\Theta_E} P_\theta(x_1^n)\, \frac{\Gamma\left(\frac{|E|}{2}\right)}{\sqrt{|E|}\Gamma(\frac{1}{2})^{|E|}} \prod_{a\in E} \theta_a^{-1/2} \mathrm{d}\theta_a$$

- Calcul itératif : comme un MV avec |E| “demi-expériences” avant de commencer : $\mathcal{KT}(011) = \frac{1/2}{1} \times \frac{1/2}{2} \times \frac{3/2}{3} = \frac{1}{16}$.

Optimalité

- Utilisation en codage itératif.
- Il vérifie la propriété fondamentale suivante : $\forall x \in E^n$,

$$-\log \mathcal{KT}(x) \leqslant \inf_{\theta \in \Theta} -\log P_\theta(x) + \frac{1}{2}\left(|E|-1\right)\log n + |E|/2$$

- $\Longrightarrow$ **redondance minimax** sur la classe des processus iid.

Plan de l'exposé

- Notions de théorie de l'information
- Le modèle des sources à arbres de contexte
- Identification par MDL
- Prédiction et mélanges

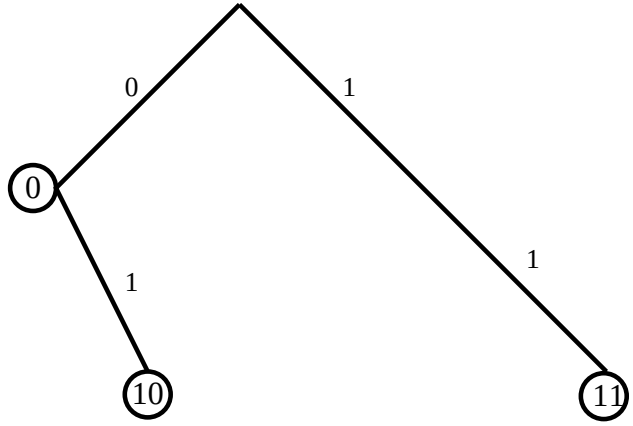
Dictionnaires complets de suffixes

- Un ensemble de mots $\mathcal{T}$ est un **dictionnaire complet de suffixes** (DCS) si $\forall x_{-\infty}^{0} \in E^{\mathbb{Z}-}, \exists! k \in \mathbb{N} : x_{-k}^{0} \in \mathcal{T}$.
- Exemples :
  - $\mathcal{T} = \{00, 10, 1\}$ en est un.
  - $\mathcal{T} = \{0, 10, 11\}$ n'en est pas un, car $\mathcal{T}(\ldots 010)$ n'est pas défini.
  - $\mathcal{T} = \{10^k : k \in \mathbb{N}\}$ en est un infini.
- On note $\mathcal{T}(y)$ l'unique suffixe d'un mot y dans $\mathcal{T}$.
- Pour $x = x_{-\infty}^{n}$ et $s \in \mathcal{T}$, le sous-mot de x qui apparaît dans le contexte s est $\mathcal{T}(x, s) = \bigodot_{i=1, \mathcal{T}(x_{-\infty}^{i-1})=s}^{n} x_i$

Tries et arbres de suffixes

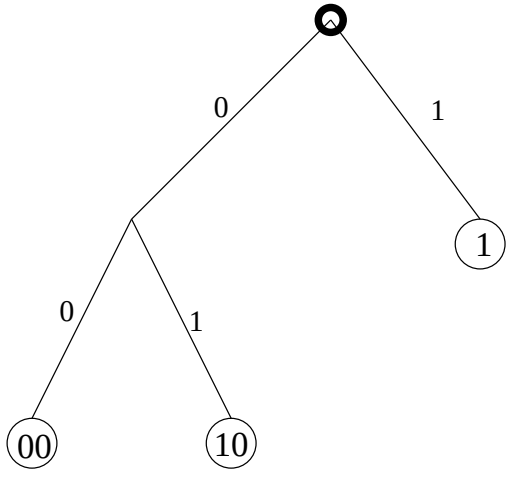
- Un **trie** est un arbre dont les arêtes sont étiquetées par des lettres.
- Aux noeuds d'un trie sont associés les mots obtenues en lisant les étiquettes *remontant* à la racine.
- Réciproquement, on peut associer un trie à tout ensemble de mots (comme ici $\{10, 0, 11\}$).

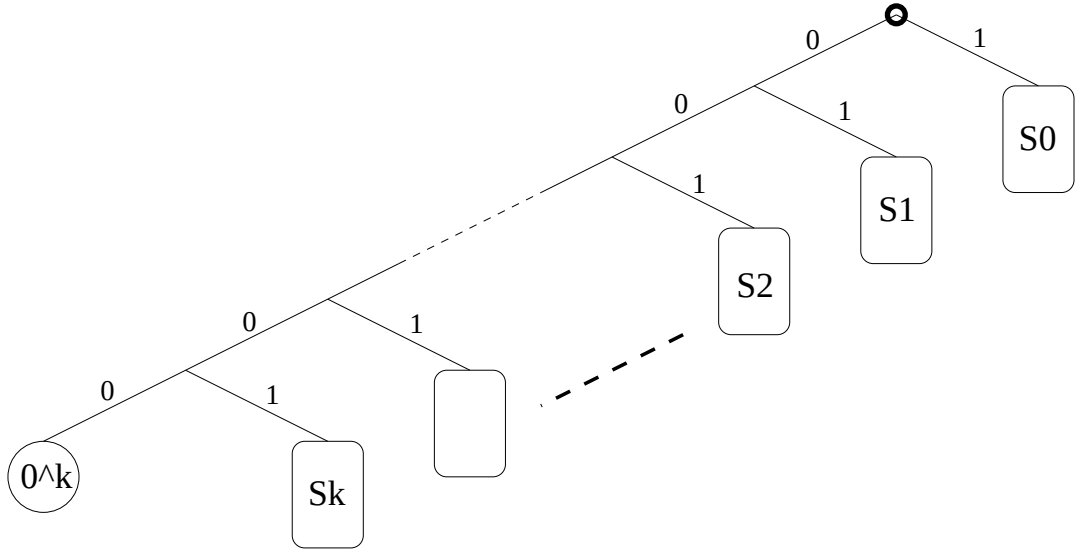
Mais il se peut que certains ne soient pas des feuilles !



Représentation des DCS comme trie

- A tout DCS $\mathcal{T}$ correspond un trie dont *les feuilles* sont associées les éléments de $\mathcal{T}$, et réciproquement.
- Exemples :





$$\mathcal{T} = \{00, 10, 1\} \qquad \mathcal{T} = \{0^k\} \cup \{10^j : 0 \leqslant j < k\}$$

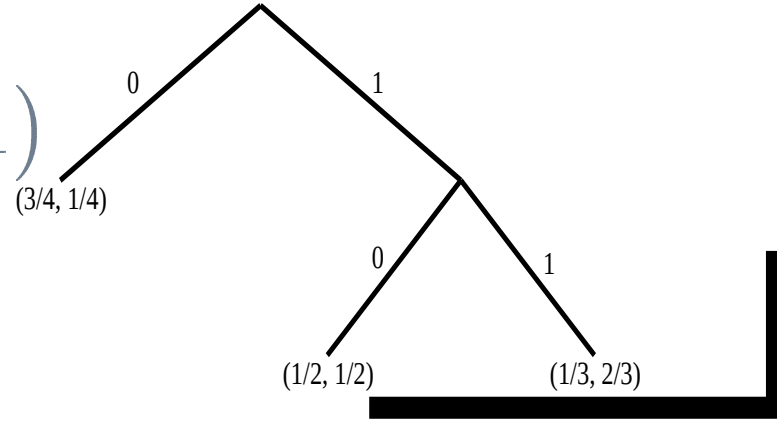
Sources à arbre de contexte

- Etant donné un DCS $\mathcal{T}$, appelé par la suite **arbre de contexte**, et $|\mathcal{T}|$ lois de probabilité sur E notées $p = (p(.|w))_{w\in\mathcal{T}}$, la **source à arbre de contexte** $\mathbb{P}_{\mathcal{T},p}$ est la loi stationnaire sur $E^{\mathbb{Z}}$ définie par

$$\mathbb{P}_{\mathcal{T},p}\left(X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0\right) = p\left(x_1 | \mathcal{T}\left(x_{-\infty}^0\right)\right).$$

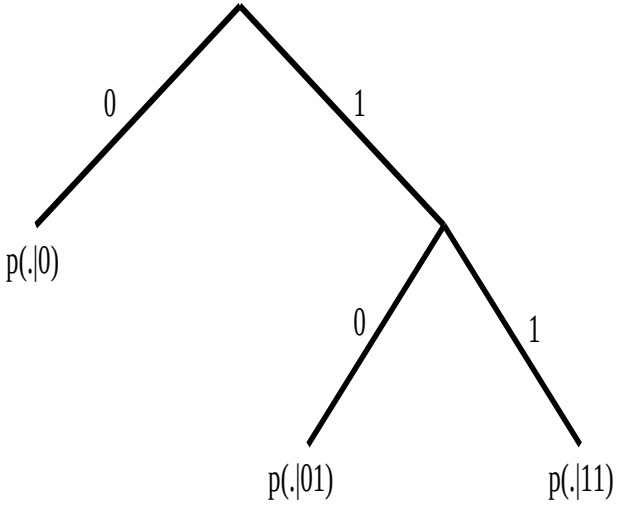
- Exemple :

$$\mathbb{P}\left(X_1^4 = 1001 | X_{-\infty}^0 = \ldots 01\right) = \frac{1}{2} \times \frac{1}{3} \times \frac{3}{4} \times \frac{1}{4}$$



Les CTS *finies* sont des chaînes de Markov

- La profondeur de l'arbre est alors l'ordre markovien.



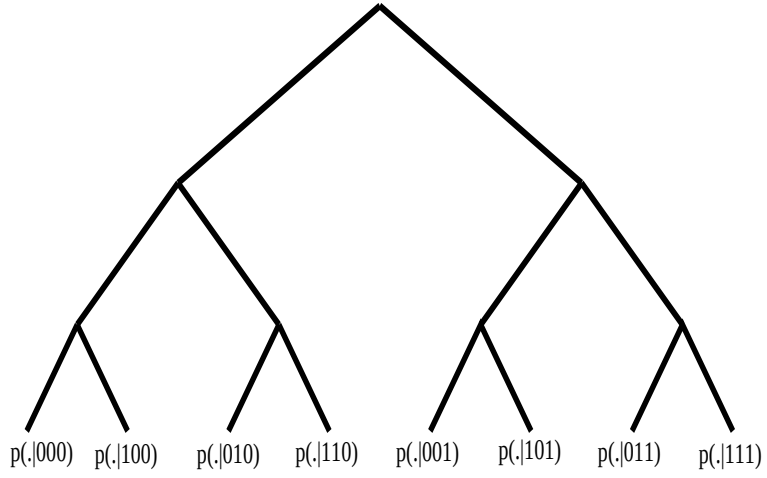
$$\rightarrow M = \begin{pmatrix} p(.|0) \\ p(.|0) \\ p(.|01) \\ p(.|11) \end{pmatrix}$$

- *Variable Length* Markov Chains : les CTS peuvent avoir moins de degrés de liberté pour une même taille de mémoire.

Les chaînes de Markov sont des CTS

- Le trie correspond est l'arbre plein de profondeur égal à l'ordre markovien :

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



- $\Longrightarrow$ les CTS ont le pouvoir d'universalité des CDM et peuvent approcher toutes les sources stationnaires ergodiques.
- pas plus difficile à utiliser.

Expression de la vraisemblance

- $x = \bigodot_{s \in \mathcal{T}} \mathcal{T}(x, s) \rightarrow$ vraisemblance s'écrit :

$$\begin{aligned} P_{\mathcal{T},p}(x_1^n | x_{-\infty}^0) &= \prod_{i=1}^{n} p(x_i | \mathcal{T}(x_{-\infty}^{i-1})) \\ &= \prod_{s \in \mathcal{T}} p_s\left(\mathcal{T}(x, s)\right) \end{aligned}$$

- D'où l'expression du maximum de vraisemblance :

$$-\log \hat{P}_{\mathcal{T}}(x) = \sum_{s \in \mathcal{T}} H(\mathcal{T}(x, s))$$

Mélange de $\mathcal{KT}$ pour un modèle

- On définit

$$\mathcal{KT}_{\mathcal{T}}(x_1^n|x_{-\infty}^0) = \prod_{s\in\mathcal{T}} \mathcal{KT}\left(\mathcal{T}(x,s)\right)$$

- **Théorème** : il existe une constante C telle que :

$$-\log_2 \mathcal{KT}_{\mathcal{T}}(x_1^n|x_{-\infty}^{-1}) \leqslant \inf_{\theta\in\Theta^{\mathcal{T}}} -\log_2 \mathbb{P}_{\mathcal{T},\theta}(x_1^n|x_{-\infty}^{-1}) + |\mathcal{T}|\frac{|E|-1}{2}\log\left(\frac{n}{|T|}\right) + C\,|\mathcal{T}|$$

- Cette redondance (ponctuelle) est minimax dans ce modèle.

Plan de l'exposé

- Notions de théorie de l'information
- Le modèle des sources à arbres de contexte
- Identification par MDL
- Prédiction et mélanges

Le principe MDL

- Idée du codage en 2 temps : le modèle $\mathcal{M}$ puis les données x sachant $\mathcal{M}$.
- Plus le modèle est riche, plus il peut facilement décrire les données.
- Mais il est lui-même plus difficile à décrire !
- $\Longrightarrow$ Minimum Description Length : choisir le modèle qui minimise le coût *total* de représentation des données.

MDL et estimateur BIC

- Coût “idéal” pour représenter x dans le modèle $\mathcal{T}$: $-\log \hat{P}_{\mathcal{T}}(x)$
- Nombre de bits nécessaires pour décrire les paramètres d’une source basée sur l’arbre $\mathcal{T}$: $|\mathcal{T}|\,(|E|-1)\log n$
- $\Longrightarrow$ On minimise en $\mathcal{T}$ le **critère BIC** :

$$\hat{\mathcal{T}}_{BIC}(x) = \min_{\mathcal{T}} -\log \hat{P}_{\mathcal{T}}(x) + \frac{1}{2}|\mathcal{T}|\,(|E|-1)\log n$$

MDL et estimateur $\mathcal{KT}$

- Le coût de x avec la loi de codage $\mathcal{KT}_{\mathcal{T}}$ est :

$$-\log \mathcal{KT}_{\mathcal{T}}(x)$$

- $\Longrightarrow$ On minimise en $\mathcal{T}$ le **critère** $\mathcal{KT}$:

$$\hat{\mathcal{T}}_{\mathcal{KT}}(x) = \min_{\mathcal{T}} -\log \mathcal{KT}_{\mathcal{T}}(x)$$

- Remarque : $\mathcal{KT}$ un peu moins que BIC car

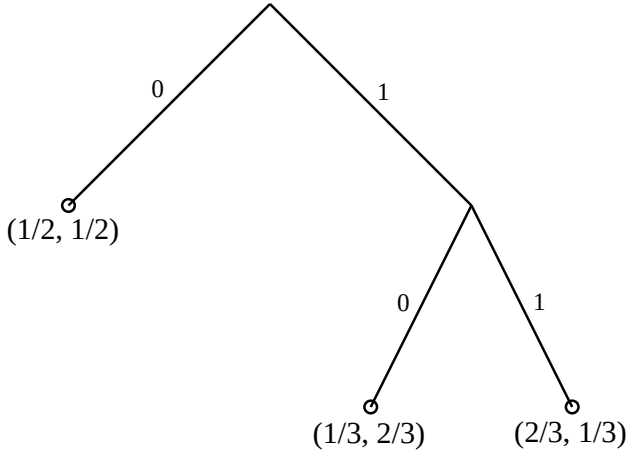
$$-\log_2 \mathcal{KT}_{\mathcal{T}}(x_1^n) \quad \leqslant \quad -\log_2 \hat{\mathbb{P}}_{\mathcal{T}}(x_1^n) + |\mathcal{T}| \frac{|E|-1}{2} \log\left(\frac{n}{|T|}\right) + C\,|\mathcal{T}|$$

Algorithme original `context` - Rissanen '81

- Pour tout $s \in \mathcal{T}$:

$$D(P_{s.}||P_s) = \sum_{a \in E} H\left(\mathcal{T}(x, sa)\right) - H\left(\mathcal{T}(x, s)\right) \geqslant 0$$

- Idée : casser les s tq $D(P_{s.}||P_s) \geqslant \frac{|E|-1}{2} \log n$
- Problème [Weinberger...]:



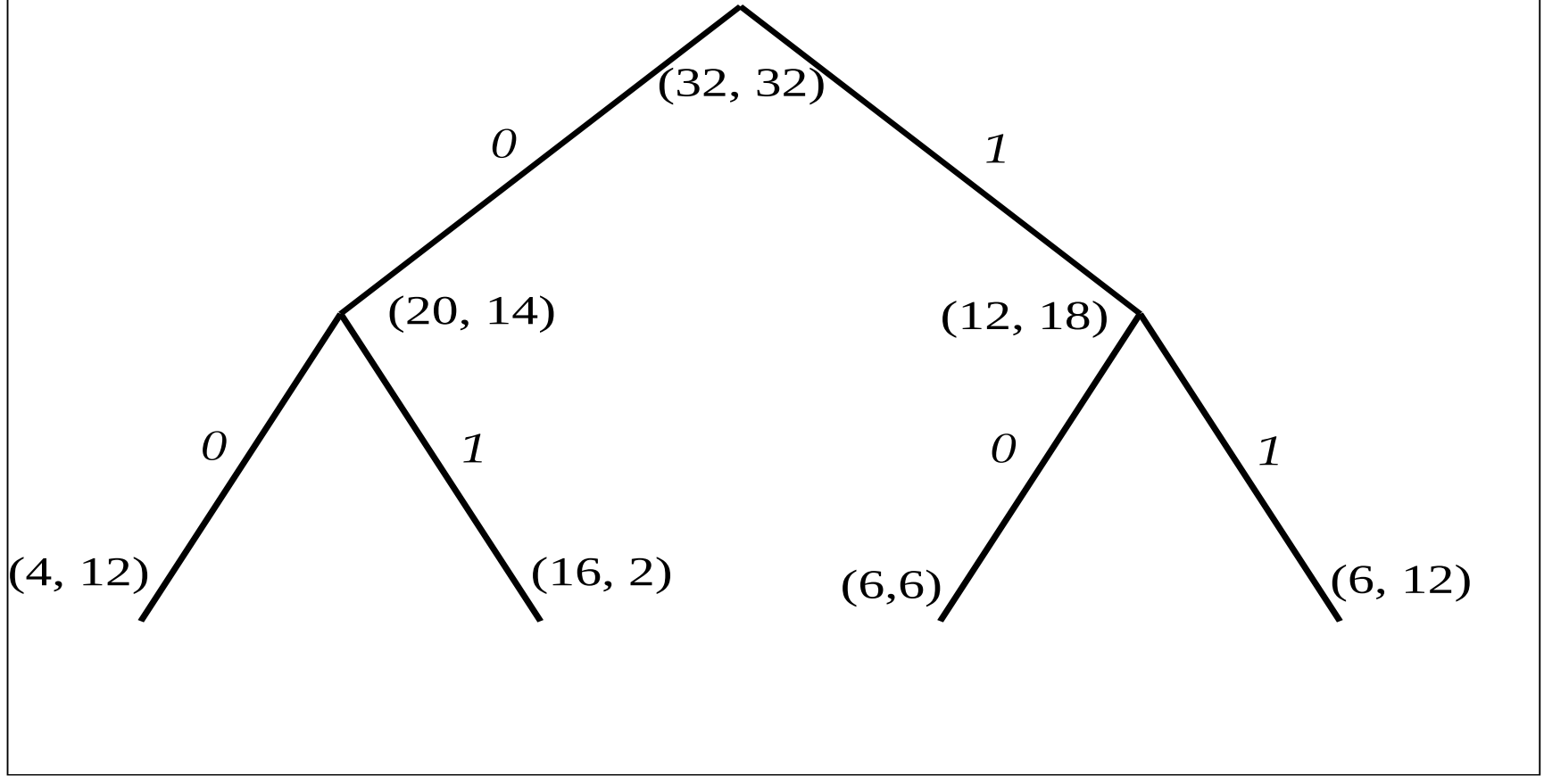
on ne casse
jamais la racine !

La méthode CTM - Willems, Shtarkov, Tjalkens '95

- Idée : partir du *bas* de l'arbre
- Pour chaque noeud s calculer :
  - son **cout propre** $CP(s) = H\left(\mathcal{T}(x,s)\right)$
  - son **sous-coût** $SC(s) = \sum_{a\in E} MC(sa) + \frac{1}{2}\log n$
  - et son **meilleur cout** $MC(s) = \min CP(s), SC(s)$
- Trouver les **noeuds actifs** s tq $MC(s) = SC(s)$
- Choix final : partant de la racine, casser tous les noeuds actifs.

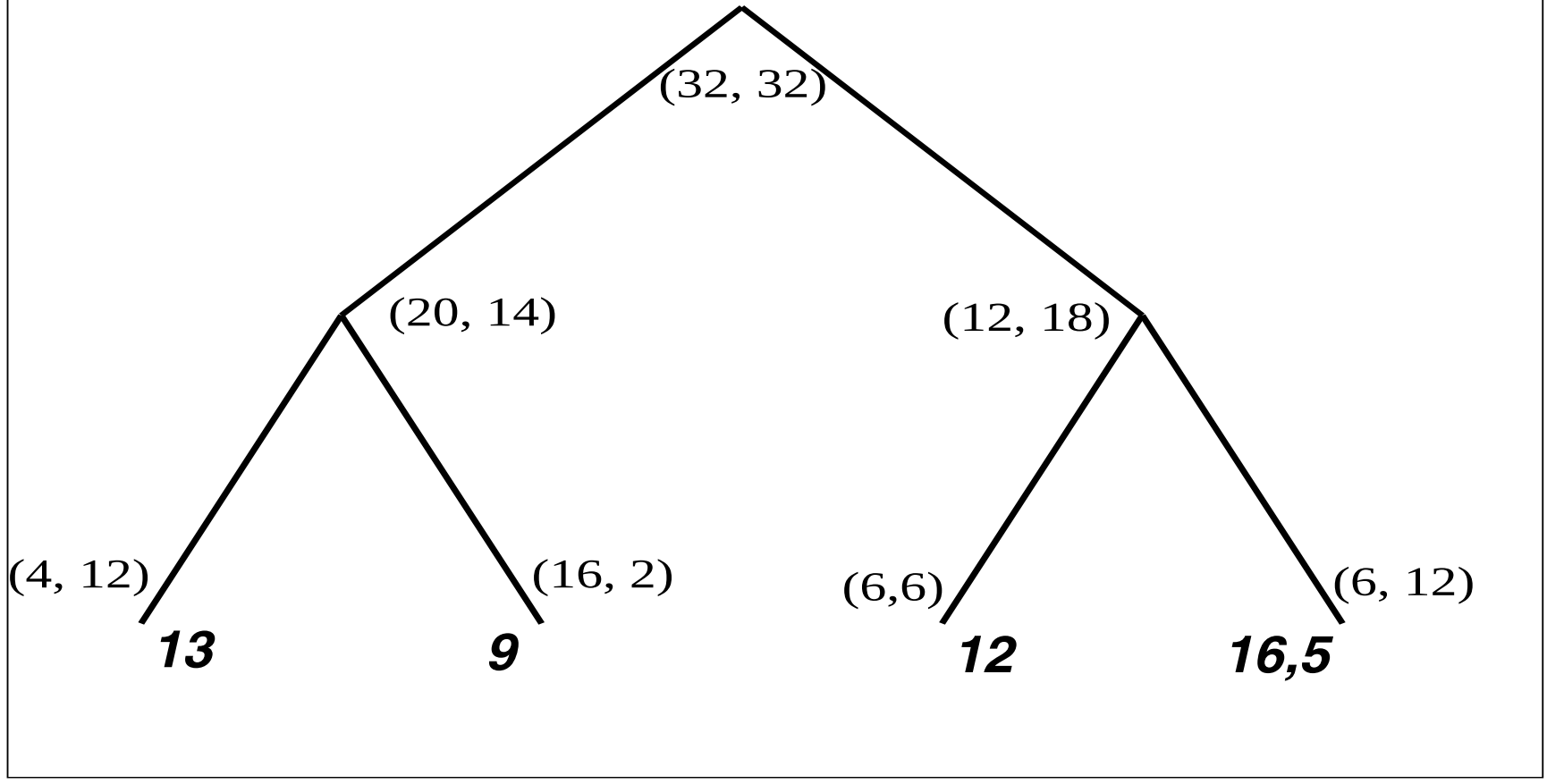
Exemple

$n = 64 \implies pen = \frac{1}{2} \log 64 = 3$



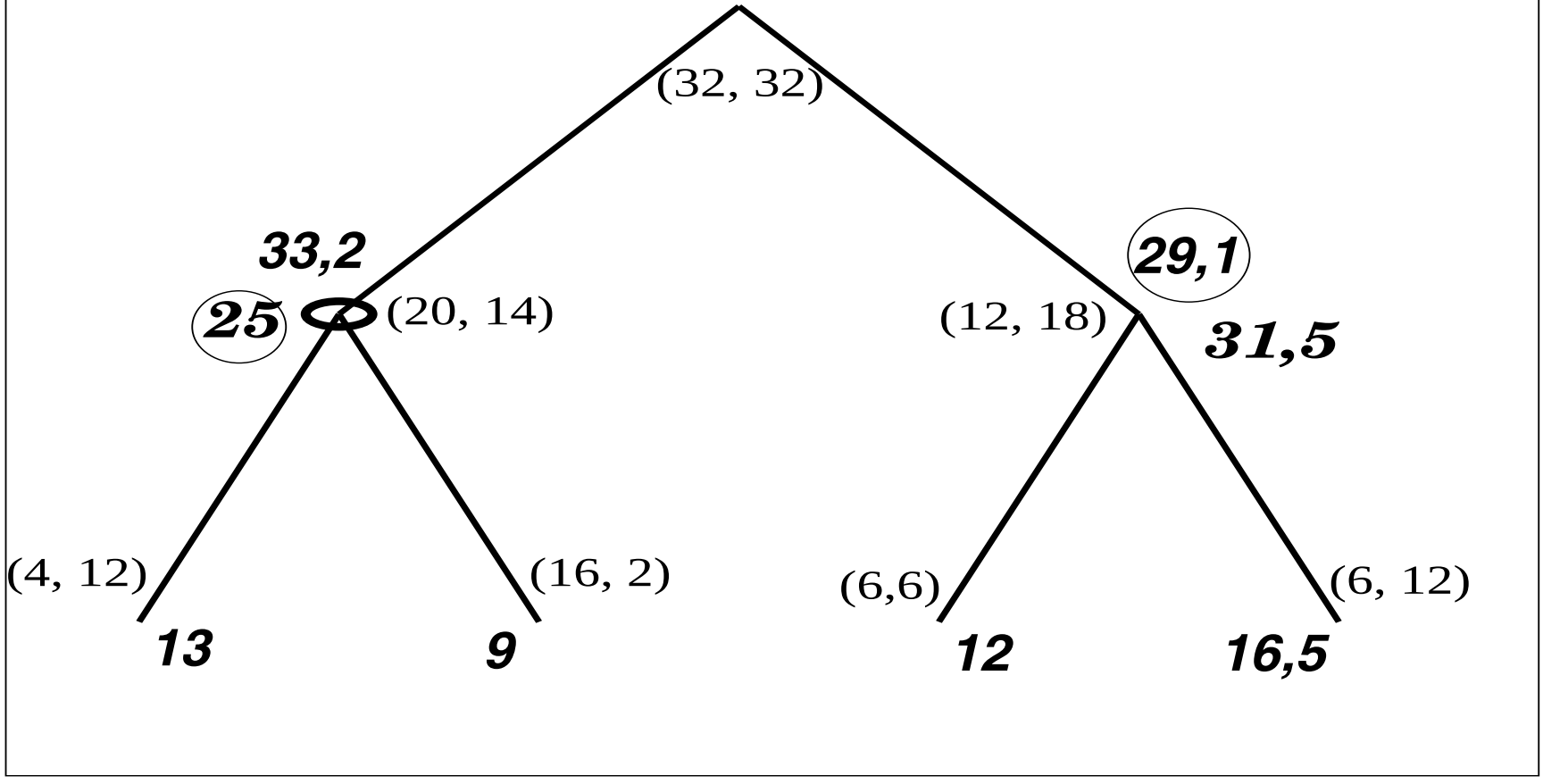
Exemple

$n = 64 \implies pen = \frac{1}{2} \log 64 = 3$



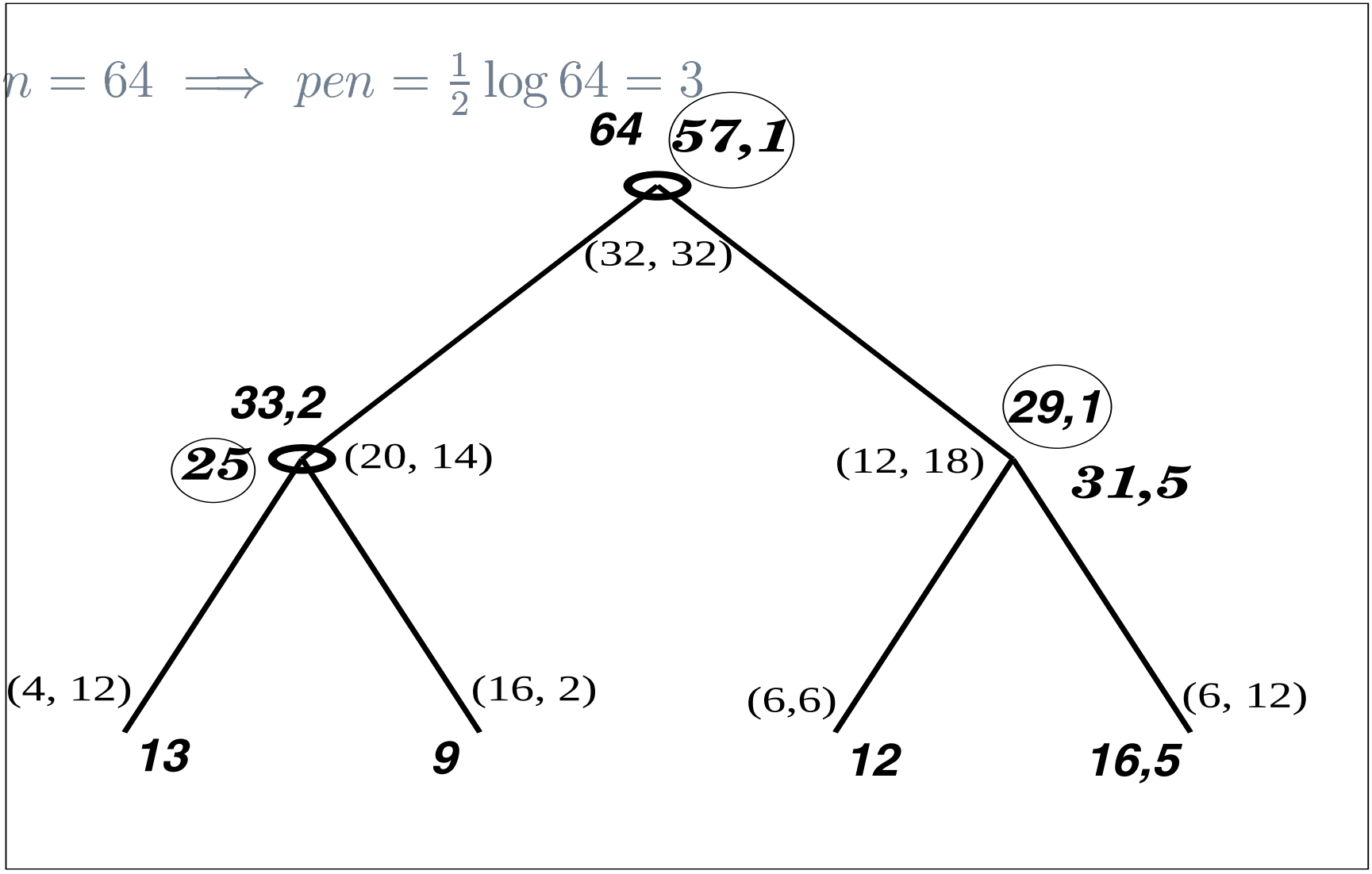
Exemple

$n = 64 \implies pen = \frac{1}{2} \log 64 = 3$



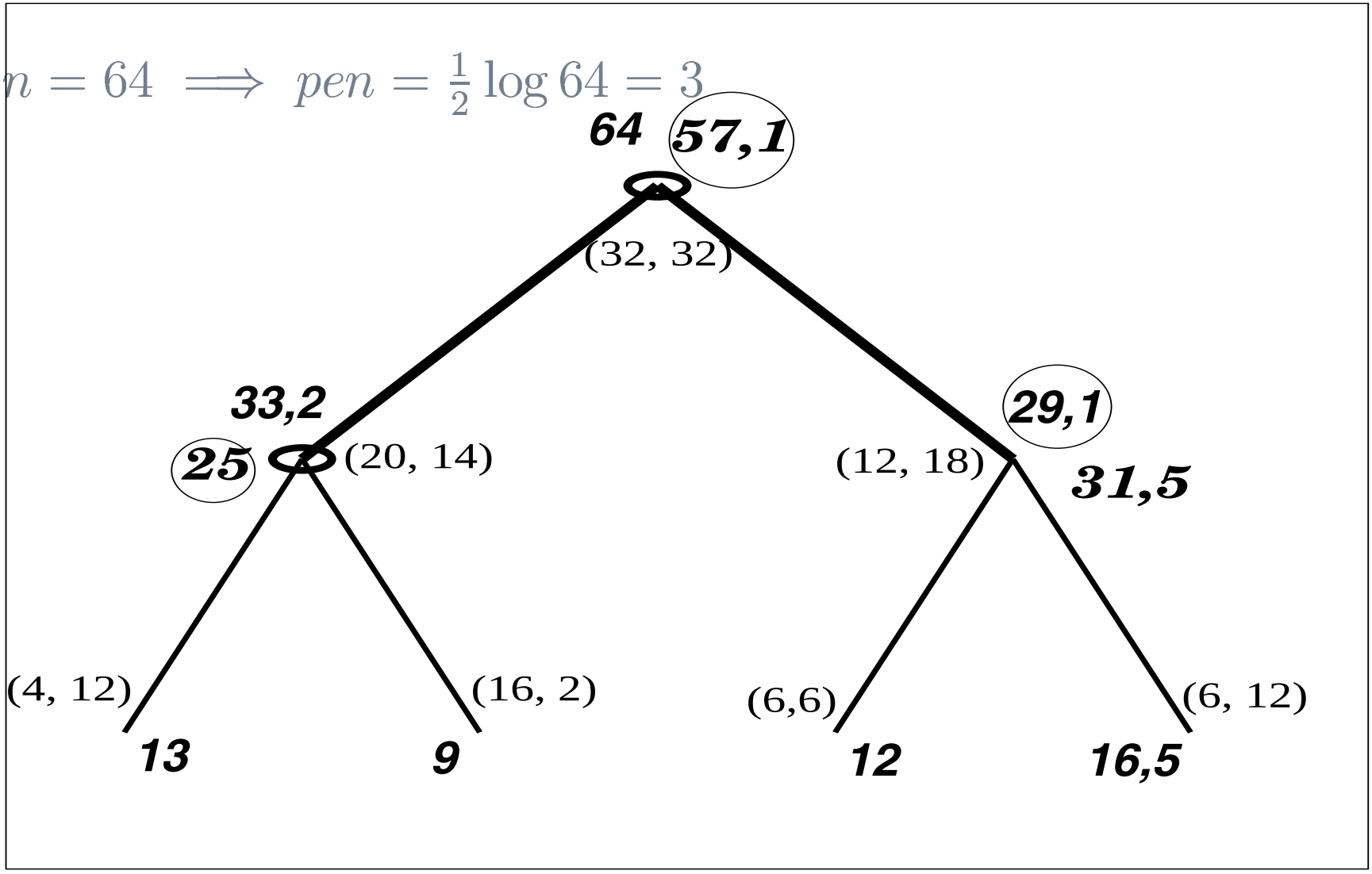
Exemple

$n = 64 \implies pen = \frac{1}{2} \log 64 = 3$



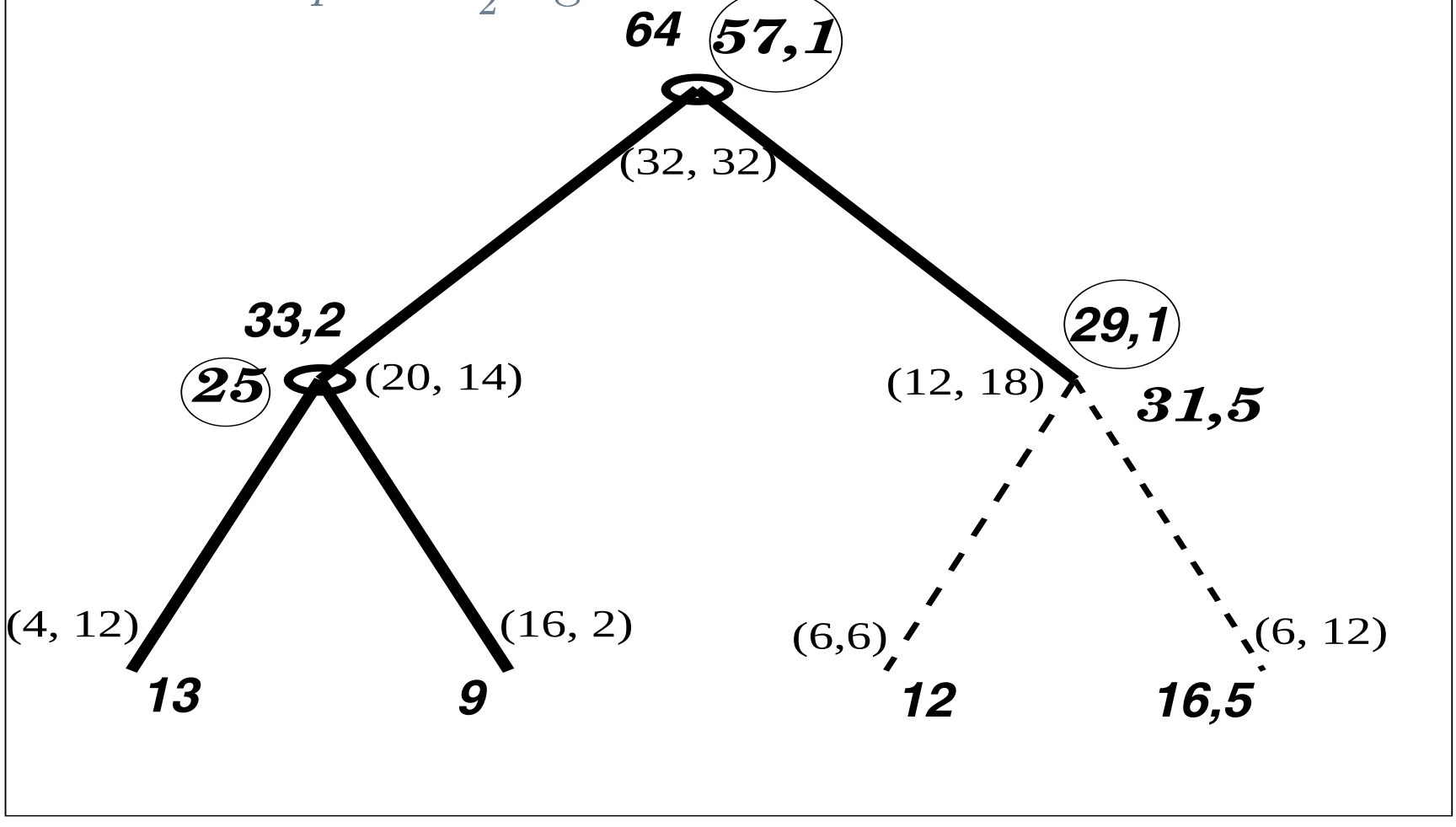
Exemple

$n = 64 \implies pen = \frac{1}{2} \log 64 = 3$



Exemple

$n = 64 \implies pen = \frac{1}{2} \log 64 = 3$



Théorèmes de consistance

- (Csiszár-Talata 2004, Garivier 2005) L'estimateur BIC choisit le bon CT ps à partir d'un certain rang.
- **Si on se limite** à chercher des arbres de hauteur $o(\log n)$, l'estimateur de Krichevsky-Trofimov est également consistant ps à partir d'un certain rang.
- Si la source est un CT infini, toute troncature est correctement identifiée à partir d'un certain rang, ps.

Idées pour la preuve pour BIC

- **Sous**estimation : les plus petits modèles sont strictement moins puissants au premier ordre.
- **Sur**estimation :
  - profondeur $< \log n$ typicalité

$$\forall \alpha > 0, \forall \delta > 0 : l(s) \ < \ \kappa \log n, N_n(s) \geqslant n^{\alpha}$$

$$\Longrightarrow \left| \frac{N_n(sa)}{N_n(s)} - Q(a|s) \right| \ < \ \sqrt{\frac{\delta \log N_n(s)}{N_n(s)}}$$

  - arbres avec plus de $\log n$ noeuds jamais choisis : argument de théorie de l'information.

Plan de l'exposé

- Notions de théorie de l'information
- Le modèle des sources à arbres de contexte
- Identification par MDL
- Prédiction et mélanges

Double mélange

- On présente le cas *binaire* pour simplifier
- Prior π sur les arbres: il y a $Catalan_s$ arbres à s feuilles d'où

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

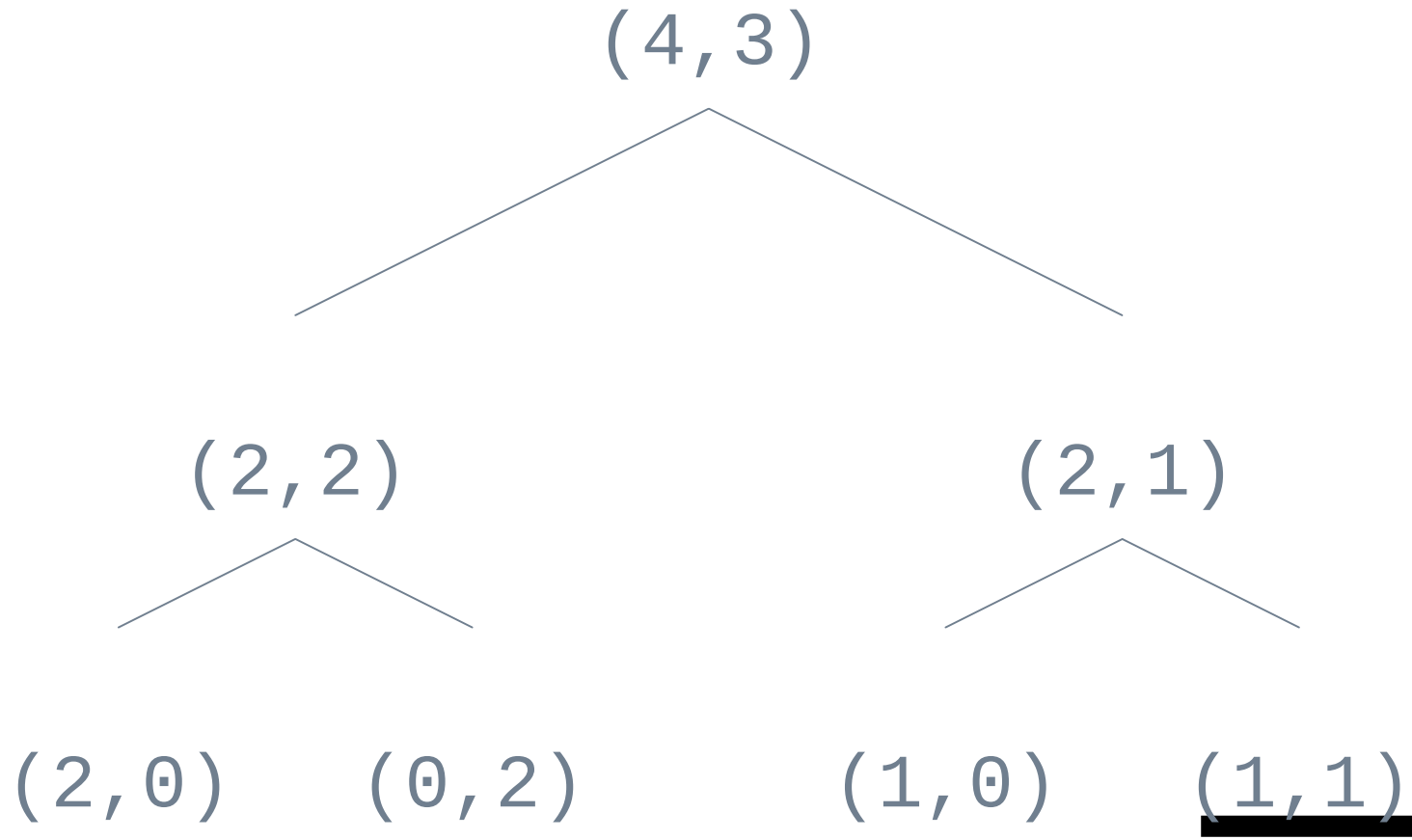
- On définit pour une collection $\mathbb{T}$ d'arbres

$$\mathcal{CTW}(x) = \sum_{\mathcal{T}\in\mathbb{T}} \mathcal{KT}_{\mathcal{T}}(x)\pi(\mathcal{T})$$

- C'est une loi de probabilité sur chaque E^n

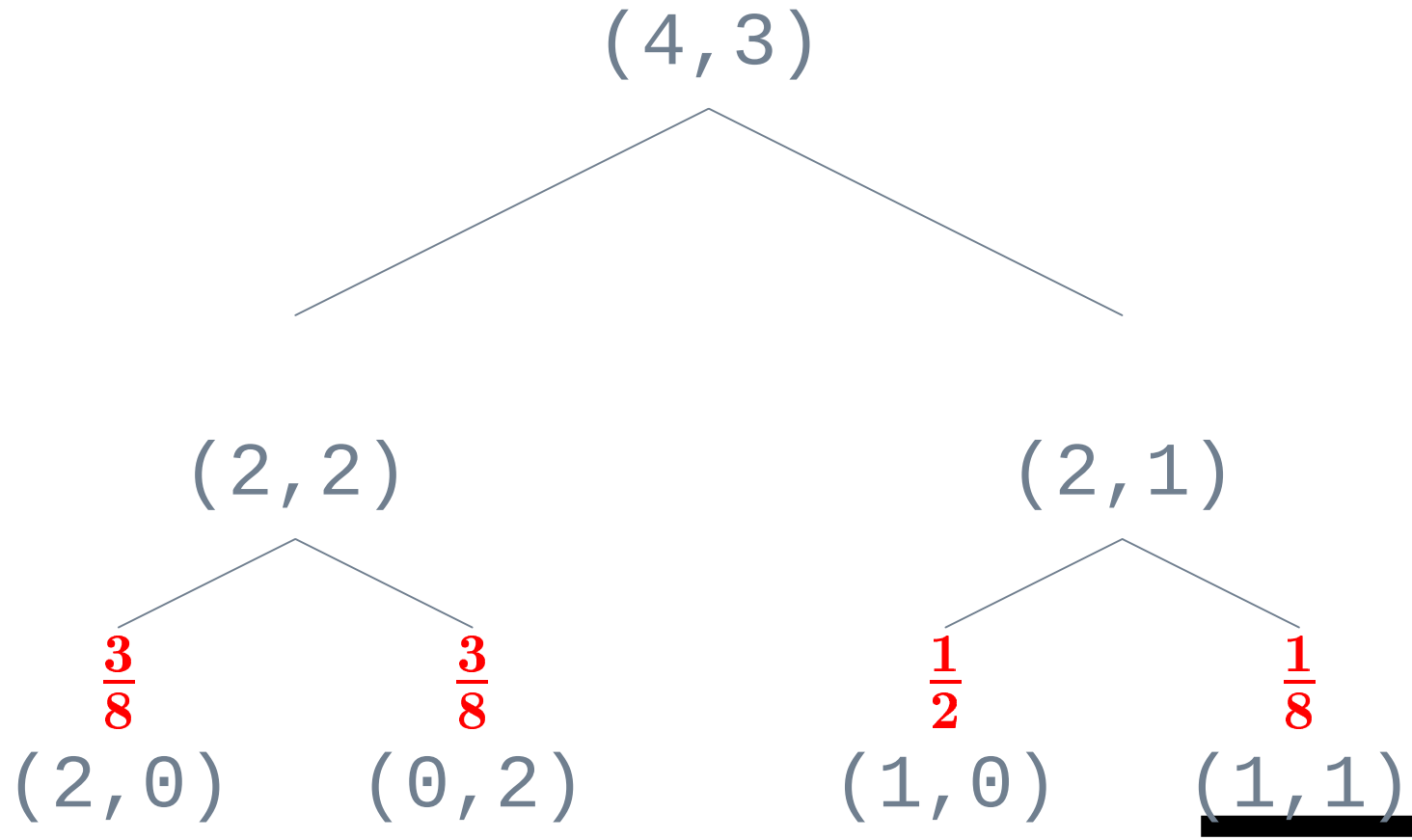
Cacul pratique de $\mathcal{CTW}$

En chaque noeud, on fait la moyenne du coût propre et du sous-coût.



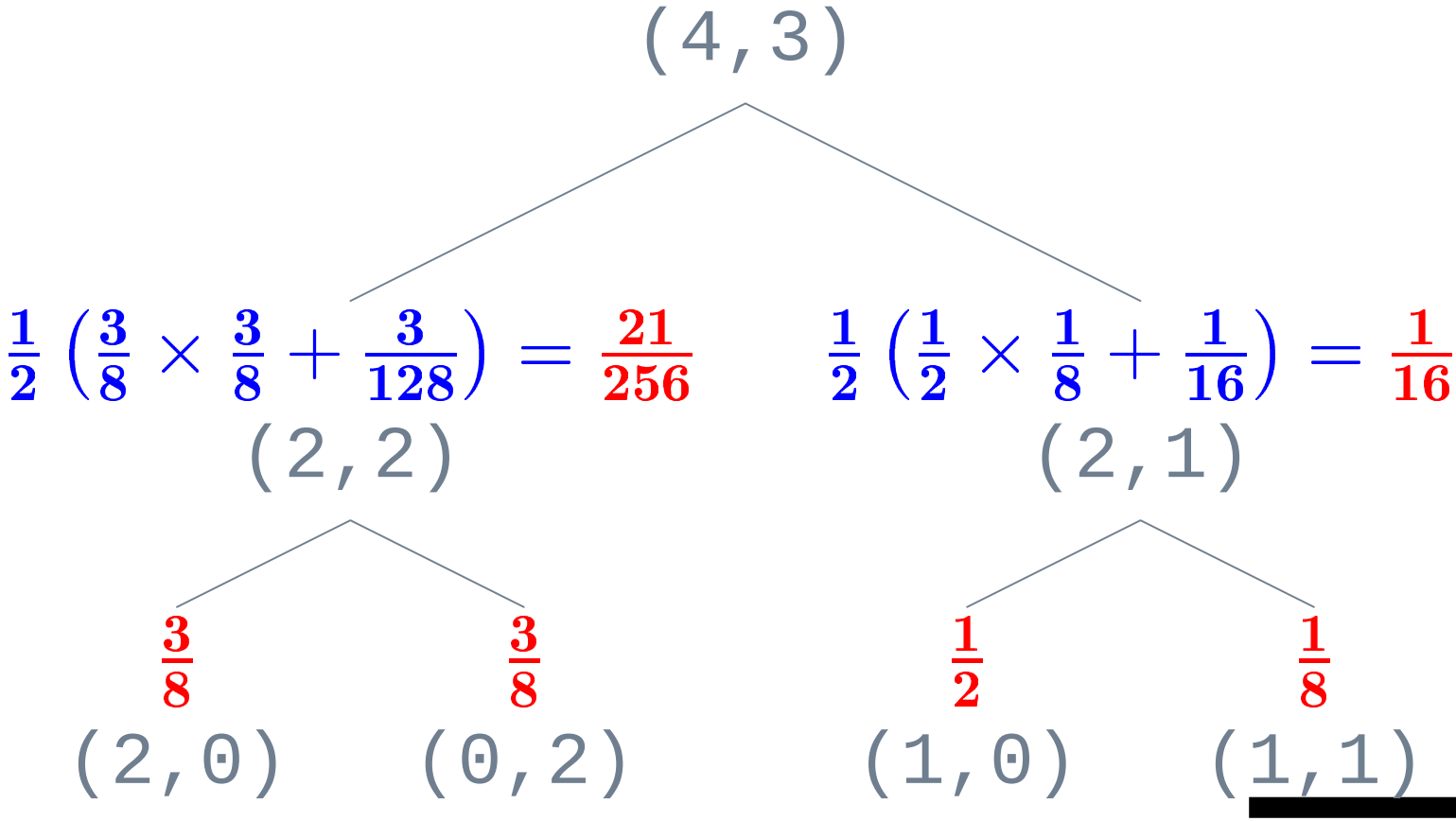
Cacul pratique de $\mathcal{CTW}$

En chaque noeud, on fait la moyenne du coût propre et du sous-coût.



Cacul pratique de $\mathcal{CTW}$

En chaque noeud, on fait la moyenne du coût propre et du sous-coût.



Cacul pratique de $\mathcal{CTW}$

En chaque noeud, on fait la moyenne du coût propre et du sous-coût.

$$\frac{1}{2}\left(\frac{21}{256}\times\frac{1}{16}+\frac{5}{2058}\right)=\frac{31}{8192}$$

(4,3)

$$\frac{1}{2}\left(\frac{3}{8}\times\frac{3}{8}+\frac{3}{128}\right)=\frac{21}{256}$$

(2,2)

$$\frac{3}{8}$$

(2,0)

$$\frac{3}{8}$$

(0,2)

$$\frac{1}{2}\left(\frac{1}{2}\times\frac{1}{8}+\frac{1}{16}\right)=\frac{1}{16}$$

(2,1)

$$\frac{1}{2}$$

(1,0)

$$\frac{1}{8}$$

(1,1)

Efficacité - optimalité

- **Théorème :** Pour tout $n \in \mathbb{N}$, pour tout $x \in E^n$:

$$-\log \mathcal{CTW}(x) \leqslant \inf_{\mathcal{T} \in \mathbb{T}} -\log \hat{P}_{\mathcal{T}}(x) + |\mathcal{T}| \log\left(\frac{n}{|\mathcal{T}|}\right) + 2|\mathcal{T}|$$

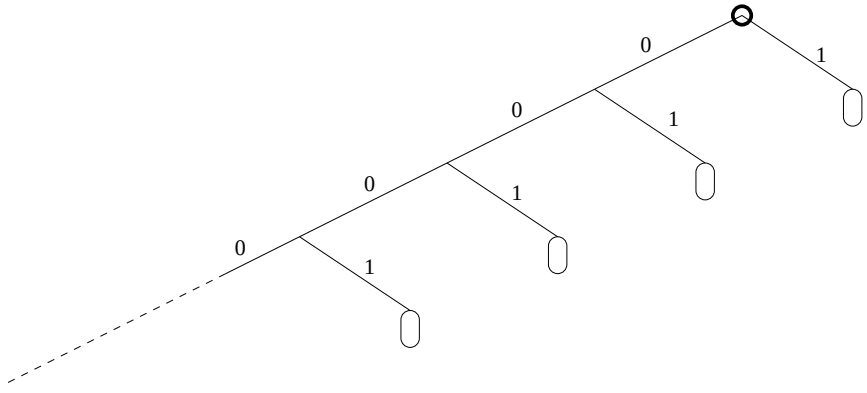
- Optimal au second ordre (minimax).
- Inégalité oracle non asymptotique.

Les processus de renouvellement

- Les cdm ont une redondance minimax en $O(\log n)$, la classe des processus stationnaire ergodique n'admet pas de redondance non triviale.
- Csiszár et Shields : classe des P.R. = la seule connue avec une complexité intermédiaire.
- **Définition :** sur $E = \{0, 1\}$, $(X_n)_{n\in\mathbb{Z}}$ est un P.R. si les distances entre les symboles '1' successifs de X sont i.i.d.
- **Théorème :** Csiszár et Shields - La redondance (moyenne et ponctuelle) minimax dans la classe des P.R. est de l'ordre de $\sqrt{n}$.
- classe non paramétrique !

Redondance de $\mathcal{CTW}$ sur les P.R.

- On peut visualiser les P.R. comme des sources à arbres de contexte infini



- **Théorème :** (Garivier 2004) Il existe des constantes positives C_1 et C_2 telles que pour tous les arbres de contexte $\mathcal{T}$ et tout $n \in \mathbb{N}$:

$$C_1\sqrt{n}\log n \leqslant R_n(\mathcal{CTW}, P.R.) \leqslant C_2\sqrt{n}\log n$$

Implications

- $\mathcal{CTW}$ est donc presque *adaptatif* dans cette classe de processus de complexité intermédiaire, à très longue mémoire.
- Note : On utilise vraiment des arbres profonds et très déséquilibrés : c'est toute la *souplesse* des modèles à arbres de contexte.
- Ce résultat suggère que $\mathcal{CTW}$ est adapté pour le traitement des processus plus irréguliers que des cdm, à longue mémoire.

Micro-bibliographie

- **Universal modeling and coding** - Rissanen, Langdon, IEEE-IT 1981
- **Universal coding, Information, Prediction, and Estimation** -Rissanen, IEEE-IT 1984
- **The Context-Tree Weighting Method : basic Properties** - Willems, Shtarkov, Tjalkens IEEE-IT 1995
- **Redundancy Rates for Renewal and Other Processes** - Csiszár, Shields - IEEE-IT 1996
- **Variable length Markov chains** - Bühlmann, Wyner, Abraham, Annals of Statistics 1999
- **Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL.** - Csiszár, Talata (Budapest), IEEE-IT 2004