Construction de peignes, mélange de peignes et marches aléatoires

Peggy Cénac-Guesdon - Institut de Mathématiques de Bourgogne

Travaux en collaboration avec B.Chauvin, F.Paccaut, N.Pouyanne, S.Herrmann et P.Vallois

Journées ALEA - 8 Mars 2012

Plan

Chaîne de Markov à mémoire variable

Introduction

- Motivation : modéliser une chaîne de caractères par de l'aléatoire
  - $\ldots AACGTGACCATTGAGA\ldots$
  - $\ldots 011010100011010001110\ldots$

  ADN, analyse textuelle, compression de données, physique statistique et mesure de Gibbs...
- Ici, l'alphabet est $\mathcal{A} = \{0, 1\}$. On va définir des chaînes de Markov à valeur dans $\mathcal{L} = \mathcal{A}^{-\mathbb{N}}$

$$\begin{aligned} U_0 &= \ldots X_{-2}X_{-1}X_0 \\ U_1 &= \ldots X_{-2}X_{-1}X_0X_1 \\ &\textit{etc} \end{aligned}$$

(U_n) est une chaîne de Markov sur les mots infinis à gauche $\mathcal{L}$.

Introduction

- Motivation : modéliser une chaîne de caractères par de l'aléatoire
  - $\ldots AACGTGACCATTGAGA\ldots$
  - $\ldots 011010100011010001110\ldots$

  ADN, analyse textuelle, compression de données, physique statistique et mesure de Gibbs...
- Ici, l'alphabet est $\mathcal{A} = \{0, 1\}$. On va définir des chaînes de Markov à valeur dans $\mathcal{L} = \mathcal{A}^{-\mathbb{N}}$

$$\begin{aligned} U_0 &= \ldots X_{-2}X_{-1}X_0 \\ U_1 &= \ldots X_{-2}X_{-1}X_0X_1 \\ &\textit{etc} \end{aligned}$$

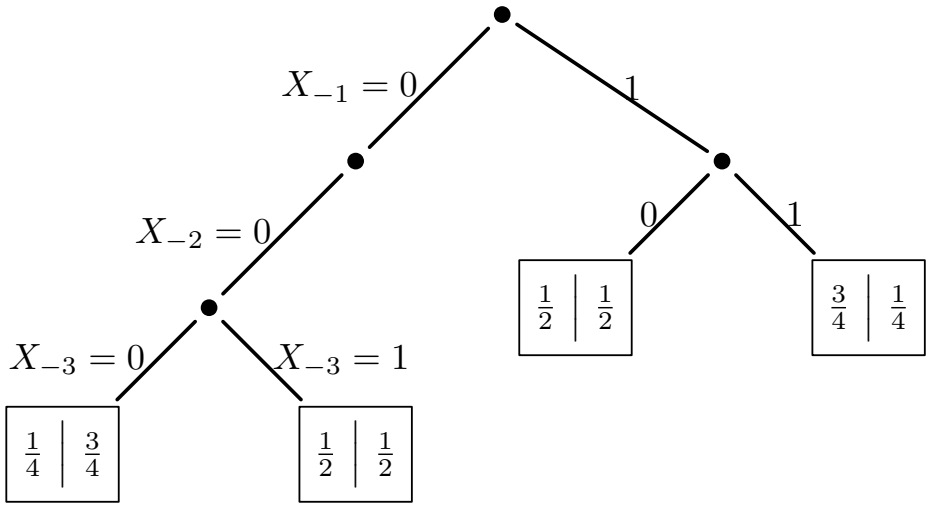
(U_n) est une chaîne de Markov sur les mots infinis à gauche $\mathcal{L}$.

Chaîne de Markov à mémoire variable

- Deux points de vue de modélisation :
  - VLMC pour "*Variable Length Markov Chain*" ;
  - Source dynamique au sens de la théorie de l'information.
- Les chaînes de Markov de mémoire variable ont été introduites par Rissanen (1983) ; la longueur de leur mémoire dépend du passé.

- La partie du passé nécessaire à la prédiction du prochain symbole s'appelle contexte. L'ensemble des contextes est regroupé dans l'arbre des contextes
- La transition de U_n vers U_{n+1} dépend d'un suffixe de U_n.

Exemple



- Si on observe le mot $U_0 = ...000$, la probabilité que U_1 finisse par 0 est $1/4$. Pour le mot ...100, cette probabilité vaut $1/2$.
- Ces chaînes permettent d'éviter l'estimation d'un nombre exponentiellement croissant de paramètres nécessaires à la description d'une chaîne de Markov d'ordre inconnu.

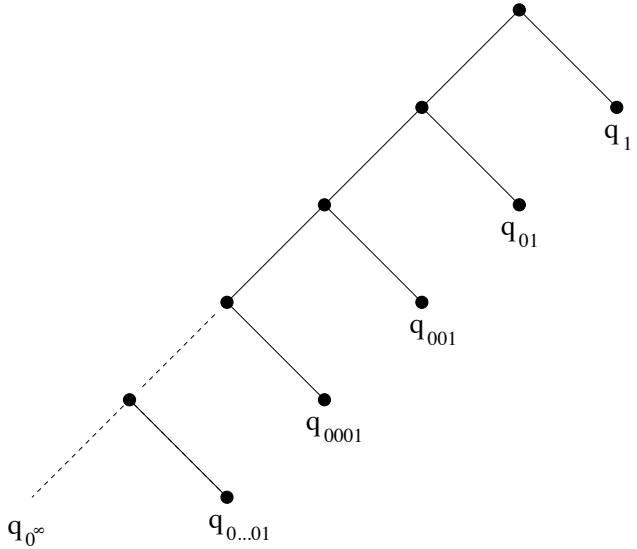
Remarque sur le caractère Markovien

- On rappelle que c'est la suite (U_n) à valeur dans $\mathcal{L}$ qui est une chaîne de Markov. X_n est la dernière lettre de U_n.
- Si l'arbre est fini, (X_n) est aussi une chaîne de Markov (d'ordre la hauteur de l'arbre) ;
- Si l'arbre est infini, (X_n) n'est pas une chaîne de Markov.
- Questions :
  - Quelle est la loi stationnaire de (U_n) ?
  - Quelles sont les propriétés de mélange de (X_n) ?
  - A quel point le comportement de X_n diffère-t'il d'une chaîne de Markov ?

Bibliographie

- Chaînes d'ordre infini :
  - Harris, On chains of infinite order, *Pacific J. Math.*, 1955 ;
  - Comets, Fernandez,Ferrari, Processes with long memory : Regenerative construction and perfect simulation, *Ann. of Appl. Prob.*, 2002 ;
  - Gallo, Garcia, N., Perfect simulation for stochastic chains of infinite memory : relaxing the continuity assumptions, 2010 ;
- Statistique et estimation de l'arbre des contextes :
  - Bühlmann, Wyner, Variable length Markov chains, *Ann. Statist.*, 1999 ;
  - survey de Galves, Löcherbach, Stochastic chains with memory of variable length, *TICSP Series*, 2008.

Le peigne infini



- (X_n) n'est pas une chaîne de Markov.
- $u_n = \ldots 11001000$ contexte : $0001 =: \overleftarrow{\text{pref}}(u_n)$ et $\mathbb{P}(U_{n+1} = U_n\alpha | U_n) = q_{0001}(\alpha)$
- Cas général :

$$\mathbb{P}(U_{n+1} = U_n\alpha | U_n) = q_{\overleftarrow{\text{pref}}(U_n)}(\alpha)$$

Le peigne infini (2)

- On note $c_0 = 1$ et pour $n \geqslant 1$,

$$c_n := \prod_{k=0}^{n-1} q_{0^k 1}(0).$$

Proposition (Cénac, Chauvin, Paccaut, Pouyanne (2011))

Dans le cas irréductible i.e. $q_{0^\infty}(0) \neq 1$, il existe une unique mesure invariante π sur $\mathcal{L}$ pour $(U_n)_n$ si et seulement si la série $\sum c_n$ converge.

- On suppose maintenant cette condition satisfaite. On note

$$S(x) := \sum_{n \geqslant 0} c_n x^n.$$

Des exemples de peignes infinis

- Pour tout mot fini w, on note $\pi(w) := \pi(\mathcal{L}w)$ et on peut vérifier que pour $n \geqslant 0$,

$$\pi(10^n) = \frac{c_n}{S(1)} \quad \text{and} \quad \pi(0^n) = \frac{\sum_{k \geqslant n} c_k}{S(1)}.$$

- Example 1 : le peigne logarithmique

$$c_0 = 1 \quad \text{et pour } n \geq 1, \quad c_n = \frac{1}{n(n+1)(n+2)(n+3)}.$$

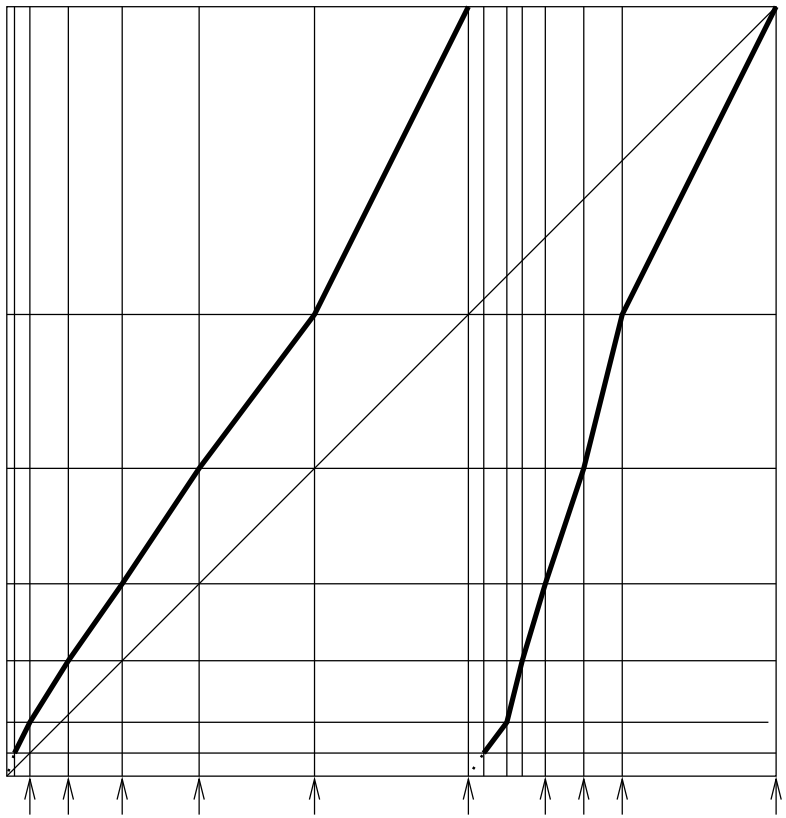
Les probabilités conditionnelles correspondantes sont

$$q_1(0) = \frac{1}{24} \quad \text{et pour } n \geqslant 1, \quad q_{0^n1}(0) = 1 - \frac{4}{n+4}.$$

- Example 2 : le peigne factoriel

$$\text{Pour } n \geq 0, \quad q_{0^n1}(0) = \frac{1}{n+2} \quad \text{et } c_n = \frac{1}{(n+1)!}.$$

Système dynamique pour le peigne Infini



$$T'(0) = \lim_{n\to+\infty} \frac{1}{q_{0^n1}(0)}, \quad T'(a_1) = \lim_{n\to+\infty} \frac{1}{q_{0^n1}(1)}$$

Une VLMC stationnaire définit une source dynamique

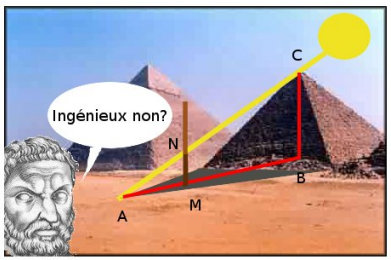
Théorème

Soit (U_n) une VLMC stationnaire, de mesure stationnaire π (sur $\mathcal{L}$). ρ et T sont définies ci-dessus. On note ξ une va uniforme sur $[0,1]$ et $Y_n := \rho(T^n\xi)$, $n \geq 0$. Alors :

(i) *La mesure de Lebesgue sur $[0,1]$ est invariante par T.*

(ii) *Pour un mot w, les probabilités des intervalles fondamentaux de la source sont liés à la VLMC stationnaire par*

$$\mathbb{P}(Y_1 \dots Y_N = w) = \pi(\overline{w}) = \mathbb{P}(\overline{U_n} = w \dots)$$



Preuve. Thalès

Mélange et trie des suffixes

Melange

- Une suite $(U_n)_{n\geqslant 0}$ de mesure stationnaire π est dite mélangeante si pour tous les mots A et B, on a

$$\lim_{n\to\infty} \sum_{w,|w|=n} \pi(AwB) = \pi(A)\pi(B).$$

- On s'intéresse ici à un type de mélange, le *ψ-mélange*. On définit les coefficients de mélange par :

$$\psi(n,A,B) := \frac{\sum_{|w|=n} \pi(AwB) - \pi(A)\pi(B)}{\pi(A)\pi(B)},$$

et on dit la séquence est *ψ-mélangeante* si :

$$\lim_{n\to\infty} \sup_{A,B} |\psi(n,A,B)| = 0.$$

Bibliographie

- Pour la définition du ψ-mélange :
  Doukhan,
  Mixing : properties and examples, Lecture Notes in Stat. **85**, 1994 ;
- Pour des propriétés de mélange de processus de renouvellement :
  Isola,
  Renewal sequences and intermittency, *J. Statist. Phys.*, 1999.

Mélange pour nos exemples de peignes

Proposition (peigne logarithmique)

La VLMC définie par le peigne logarithmique a un coefficient de mélange polynômial non uniforme : il existe une constante positive $C_{A,B}$ telle que

$$|\psi(n, A, B)| \leqslant \frac{C_{A,B}}{n^3}.$$

$C_{A,B}$ ne peut pas être majorée uniformément : $\psi(n, 0, 0^n) \to \frac{13}{6}$.

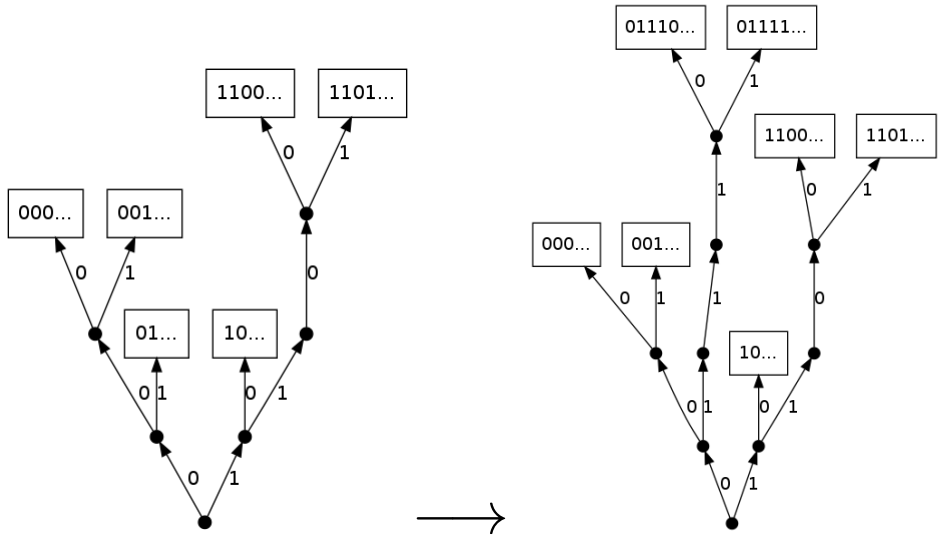
Proposition (peigne factoriel)

La VLMC définie par le peigne factoriel a un mélange exponentiel uniforme : il existe une constante positive C telle que

$$|\psi(n, A, B)| \leqslant \frac{C}{(2\pi)^n}.$$

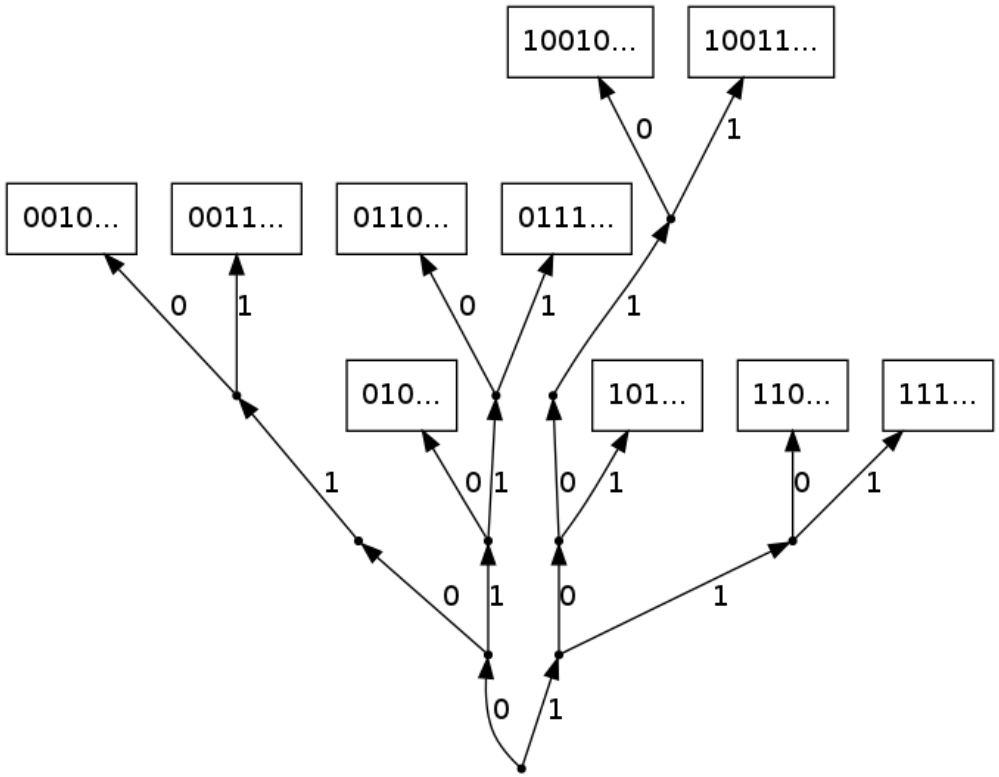
Trie et Trie des suffixes

- On considère le processus de tries des suffixes $(\mathcal{T}_n)_n$ associé à une séquence infinie générée par un peigne infini.
- Un trie est un arbre planaire contenant les mots que l'on souhaite classer dans ses feuilles. Il est obtenu par construction récursive.



Trie construit à partir de
$(000\ldots, 10\ldots, 1101\ldots, 001, \ldots, 01110\ldots, 1100\ldots, 01111\ldots)$.

Trie des suffixes



Trie des Suffixes $\mathcal{T}_{10}$ associé au mot $1001011001110\ldots$.

Bibliographie

- Trie des suffixes d'une source symétrique sans mémoire :
  - sur la taille moyenne de l'arbre : Blumer *et al.* Average sizes of suffix trees and DAWGs, 1989 ;
  - sur la hauteur : Devroye *et al.* A note on the height of suffix trees, 1992 ;
- pour une source asymétrique sans mémoire :
  Fayolle, Compression de données sans perte et combinatoire analytique, 2006 ;
- pour une source avec mémoire :
  - sur les arbres des suffixes : Szpankowski, Asymptotic properties of data compression and suffix trees, 1993 ;
  - sur les tries : Clément, Flajolet, Vallee, Dynamical sources in Information Theory : Analysis of general tries, 2001.

Définitions

- On définit la hauteur et le niveau de saturation

$$\begin{aligned} H_n &= \max_{u\in\mathcal{T}_n\setminus\partial\mathcal{T}_n}\{|u|\} \\ \ell_n &= \max\left\{j\in\mathbb{N}|\ \#\{u\in\mathcal{T}_n\setminus\partial\mathcal{T}_n,|u|=j\}=2^j\right\}. \end{aligned}$$

- On définit également les constantes h_+ et h_- dans $[0,+\infty]$ par

$$\begin{aligned} h_+ &= \lim_{n\to+\infty}\frac{1}{n}\max_{w,|w|=n,\pi(w)>0}\left\{\ln\left(\frac{1}{\pi(w)}\right)\right\}, \\ h_- &= \lim_{n\to+\infty}\frac{1}{n}\min_{w,|w|=n,\pi(w)>0}\left\{\ln\left(\frac{1}{\pi(w)}\right)\right\}. \end{aligned}$$

- Hypothèses classiques :
  - La séquence à insérer est générée par une source i.i.d. voire Markovienne d'ordre 1 ;
  - $h_+ < +\infty$;
  - $h_- > 0$.

Peignes logarithmique et factoriel

- Les 3 hypothèses ne sont pas vérifiées pour nos deux exemples : le peigne logarithmique et le peigne factoriel :
  - (X_n) n'est pas markovienne ;
  - Pour le peigne logarithmique : $\pi(10^n)$ est d'ordre n^{-4} et donc

$$h_- \leqslant \lim_{n\to+\infty} \frac{1}{n} \ln\left(\frac{1}{\pi\left(10^{n-1}\right)}\right) = 4 \lim_{n\to+\infty} \frac{\ln n}{n} = 0;$$

  - Pour le peigne factoriel : $\pi(10^n)$ est d'ordre $\frac{1}{(n+1)!}$ et donc

$$h_+ \geqslant \lim_{n\to+\infty} \frac{1}{n} \ln\left(\frac{1}{\pi\left(10^{n-1}\right)}\right) = \lim_{n\to+\infty} \frac{n!}{n} = +\infty.$$

Des arbres originaux

Résultat (CCPP (2012))

Soit $\mathcal{T}_n$ le trie des suffixes construit à partir des n premiers suffixes d'une suite générée par un peigne logarithmique. Alors la hauteur H_n de $\mathcal{T}_n$ satisfait

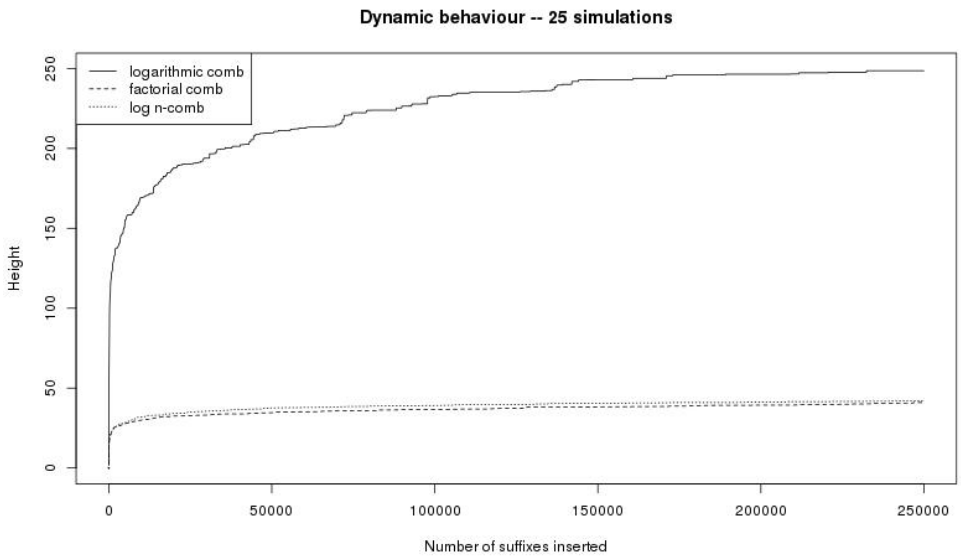
$$\forall \delta > 0, \qquad \frac{H_n}{n^{\frac{1}{4}-\delta}} \underset{n\to\infty}{\longrightarrow} +\infty \quad \text{en probabilité.}$$

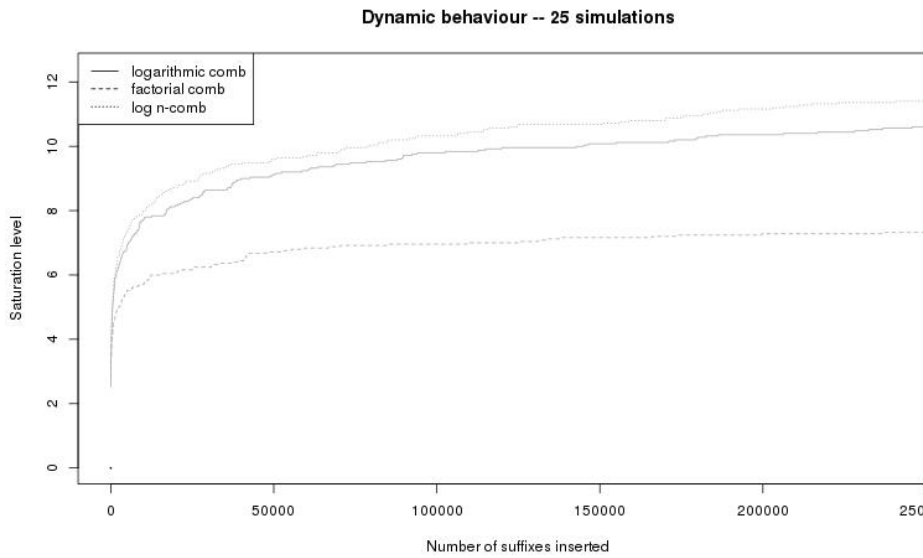
Résultat (CCPP(2012))

Soit $\mathcal{T}_n$ le trie des suffixes construit à partir des n premiers suffixes de la suite générée par un peigne factoriel. Alors le niveau de saturation ℓ_n de $\mathcal{T}_n$ satisfaît : pour tout $\delta > 1$, p.s., quand n tend vers l'infini,

$$\ell_n \in o\left(\frac{\log n}{(\log\log n)^{\delta}}\right).$$

Simulations





Hauteur et niveau de saturation pour le peigne logarithmique (trait plein), le peigne factoriel (longs pointillés) et pour un log n-peigne (courts pointillés).

Quelques clés pour la preuve

- On utilise un argument de dualité à la Pittel entre les instants où les branches poussent et les longueurs des branches ;
- Une branche pousse lorsque l'on a deux suffixes qui ont le même préfixe, autrement dit, lorsqu'une deuxième occurrence de motif apparaît dans la séquence ;
- Grâce aux calculs permettant d'établir le mélange, on peut écrire la fonction génératrice de la deuxième occurence d'un motif 10^k :

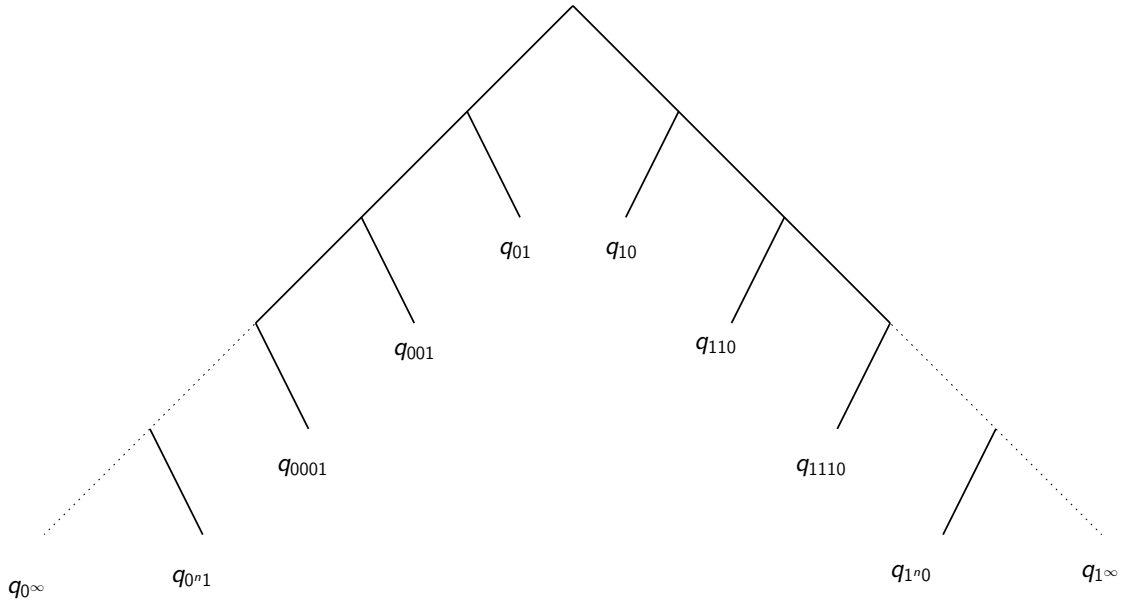
$$\Phi(x) = \frac{c_k^2 x^{2k+1}\big(U(x)-1\big)}{S(1)(1-x)\big[1+c_k x^k(U(x)-1)\big]^2},$$

avec $U(x) = [(1-x)S(x)]^{-1}$;

- Théorème de transfert et analyse de singularités ;
- Inégalités de concentration et Borel-Cantelli permettent de conclure.

Marches aléatoires persistantes

Le double peigne



- $u_n = \ldots 11001000$ contexte : 0001 et $\mathbb{P}(U_{n+1} = U_n\alpha | U_n) = q_{0001}(\alpha)$
- $u_n = \ldots 1100011$ contexte : 110 et $\mathbb{P}(U_{n+1} = U_n\alpha | U_n) = q_{110}(\alpha)$

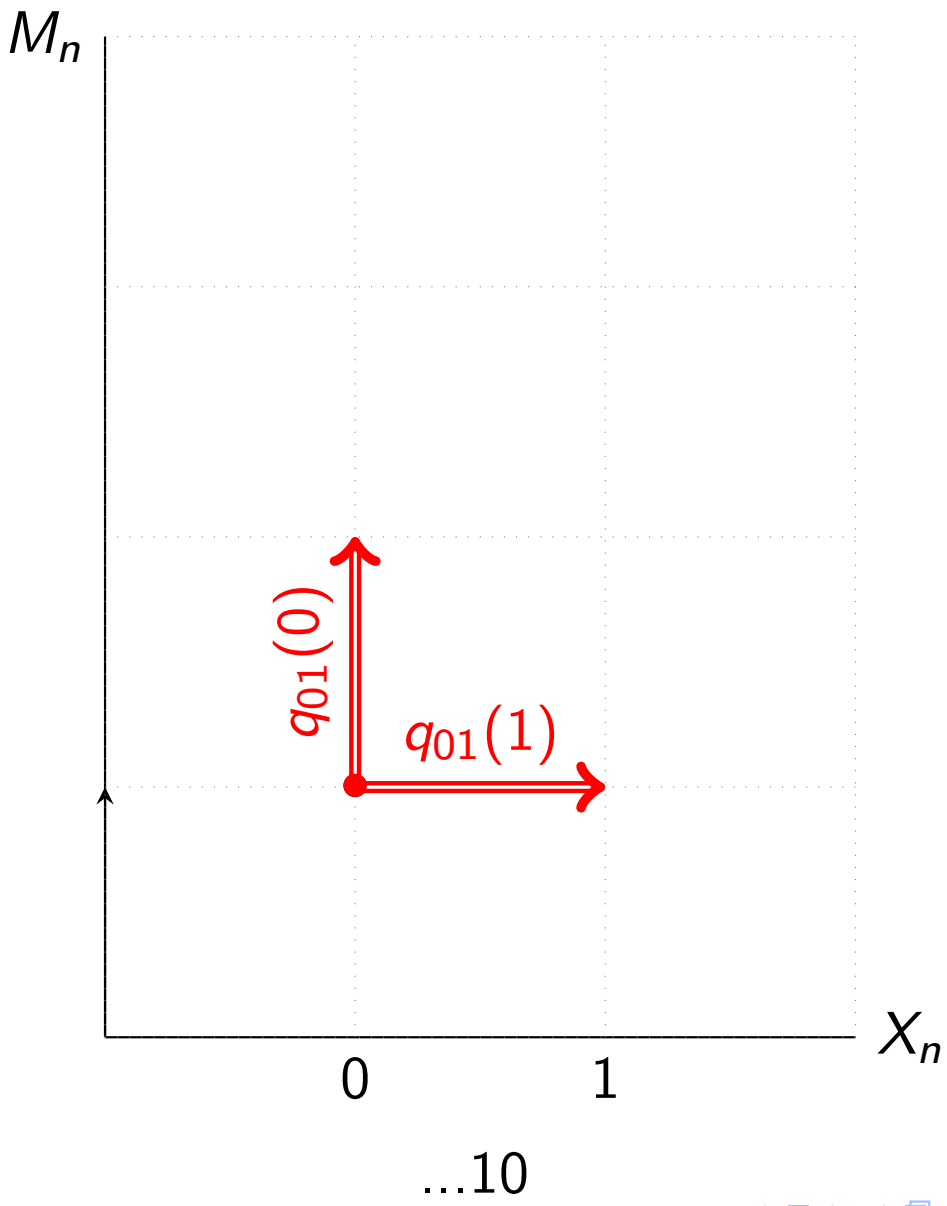
Prise en compte de la mémoire

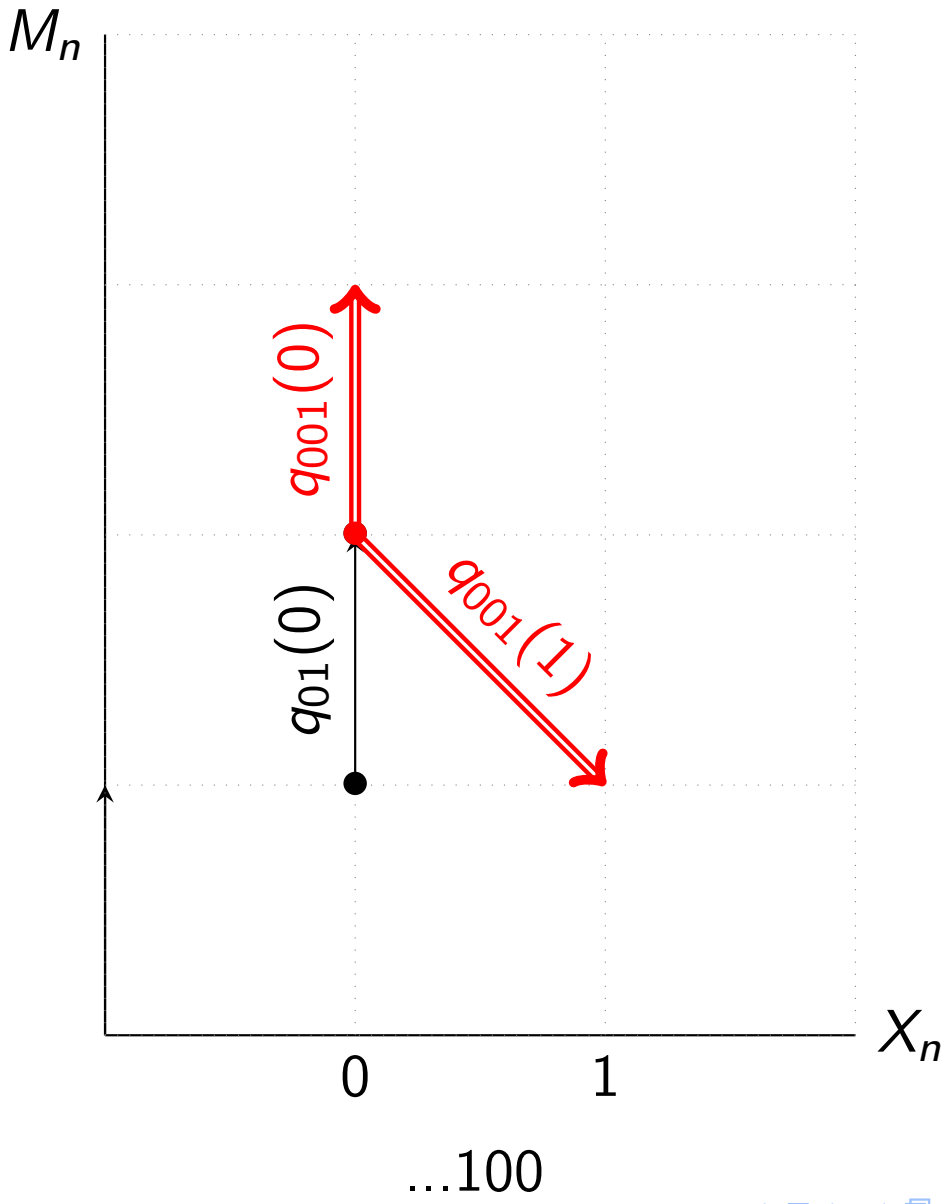
- On a vu que la suite des lettres (X_n) n'est pas une chaîne de Markov.
- On introduit une variable aléatoire mémoire notée M_n définie comme étant la longueur du run de '0' ou de '1' contenant la n-ième lettre (et jusqu'à la n-ième lettre) :

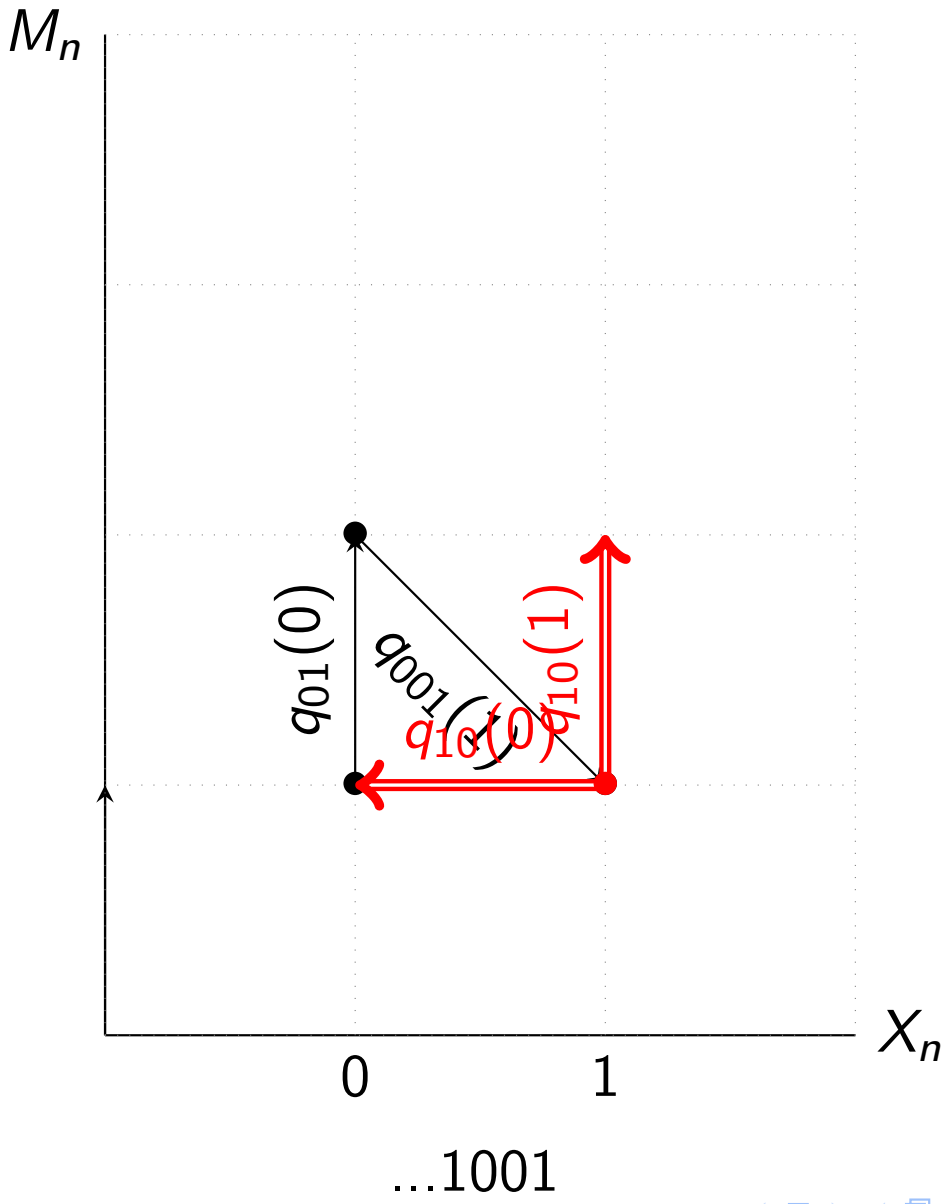
$$M_n = 1 + \sup\{0 \leq i \leq n : \ \forall j \in \{0, \dots, i\} \ X_{n-j} = X_n\}.$$

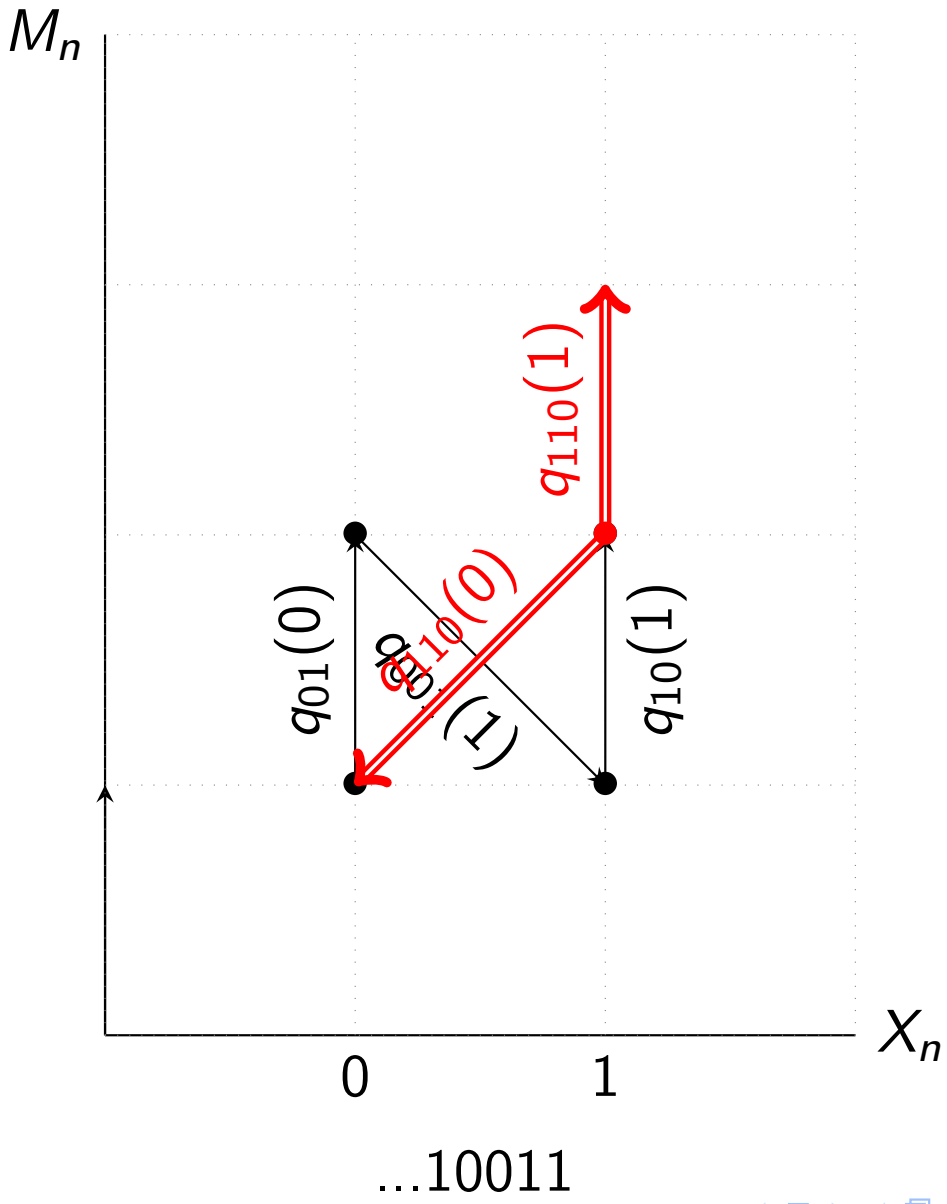
- Le couple (X_n, M_n) est une chaîne de Markov, d'ordre 1, à espace d'états $\{0,1\} \times \mathbb{N}^*$ et de noyau de transition :

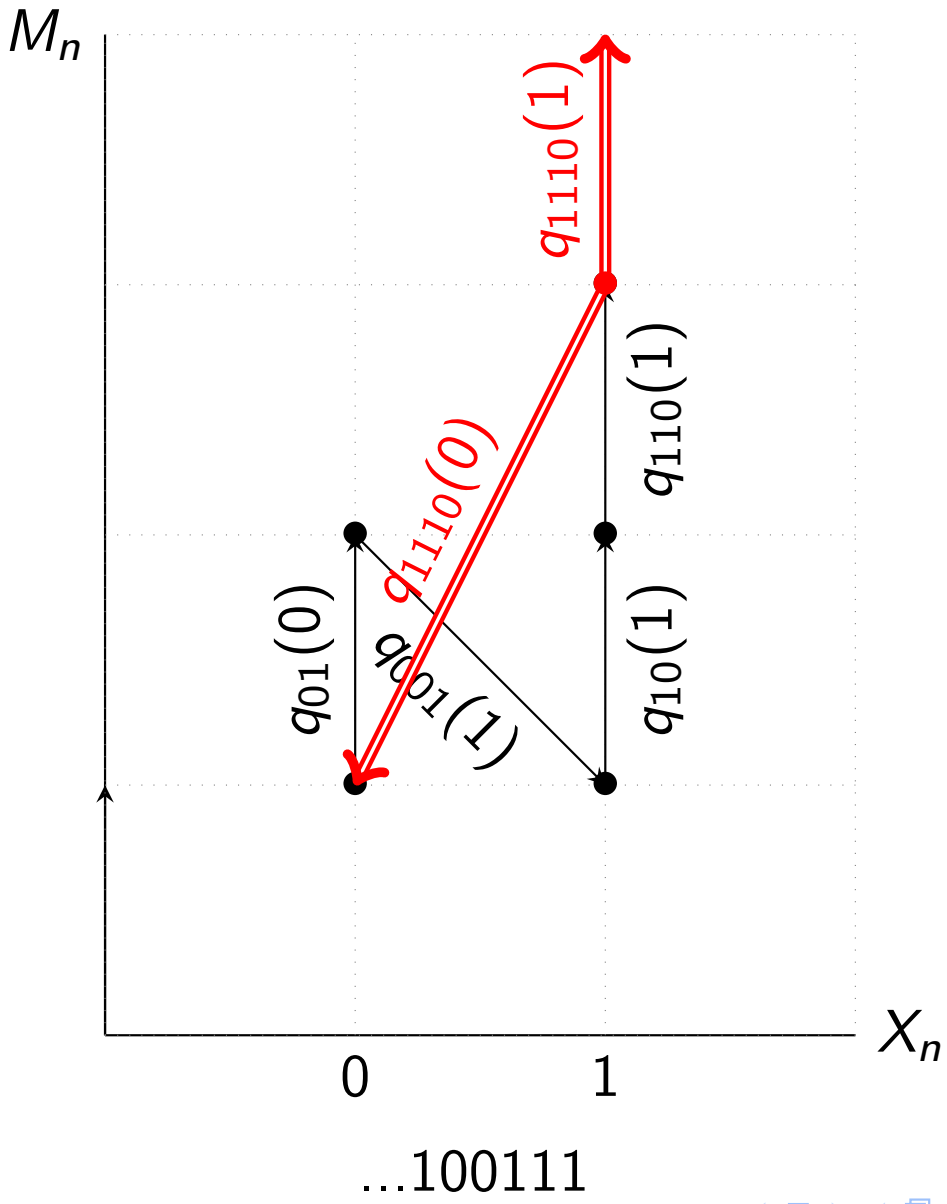
$$\begin{cases} Q\Big((a_i, k+1)|(a_i, k)\Big) = q_{a_i^k a_j}(a_i), \quad 1 \leq i \neq j \leq 2, \quad k \geq 1, \\ Q\Big((a_j, 1)|(a_i, k)\Big) = q_{a_i^k a_j}(a_j). \end{cases}$$

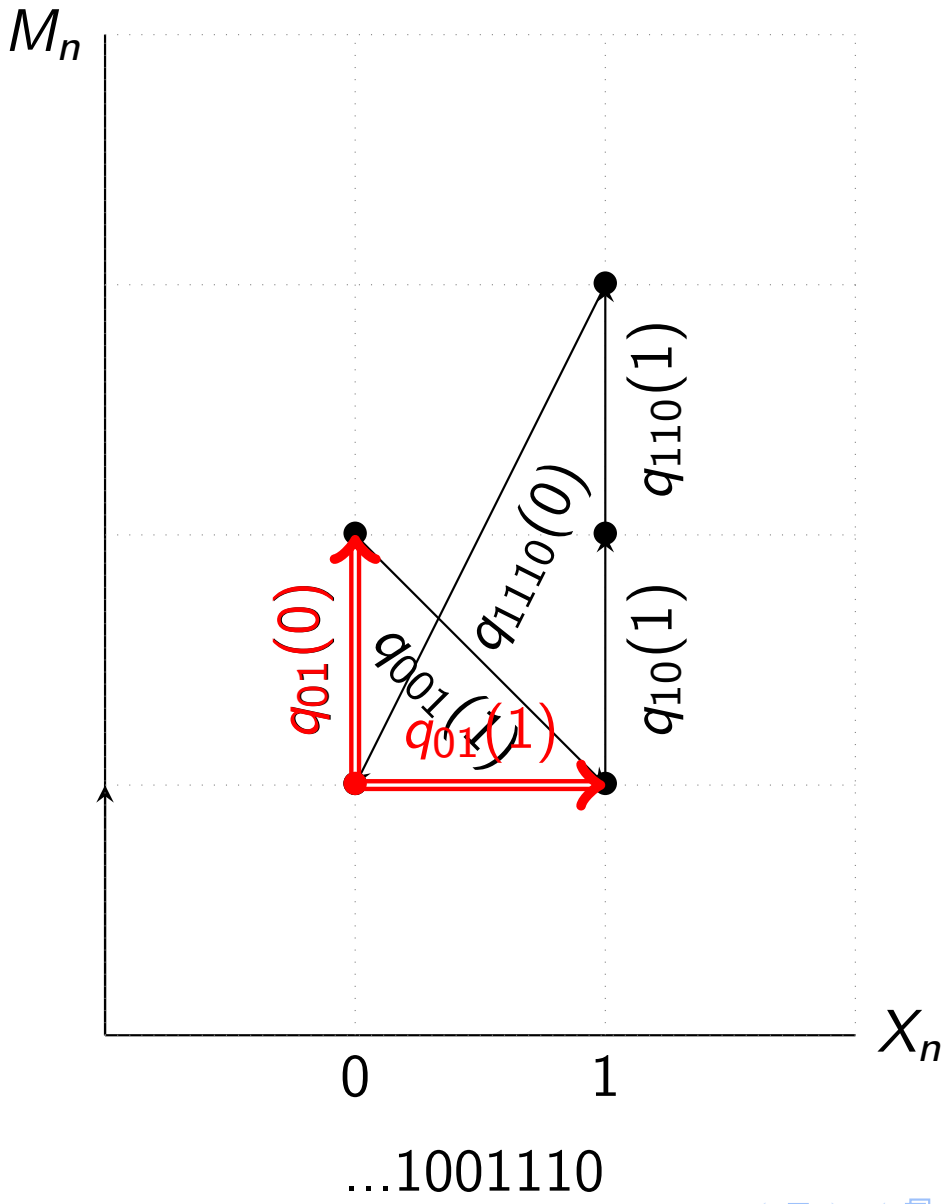

M_n
$q_{01}(0)$
$q_{01}(1)$
X_n
0
1
...10


M_n
$q_{001}(0)$
$q_{01}(0)$
$q_{001}(1)$
X_n
0
1
...100


M_n
$q_{01}(0)$
$q_{001}(1)$
$q_{10}(0)$
$q_{10}(1)$
X_n
0
1
...1001


M_n
$q_{110}(1)$
$q_{10}(0)$
$q_{01}(0)$
$q_{01}(1)$
$q_{10}(1)$
X_n
0
1
...10011


M_n
$q_{1110}(1)$
$q_{1110}(0)$
$q_{110}(1)$
$q_{01}(0)$
$q_{01}(1)$
$q_{10}(1)$
X_n
0
1
...100111


M_n
$q_{110}(1)$
$q_{1110}(0)$
$q_{01}(0)$
$q_{001}(1)$
$q_{01}(1)$
$q_{10}(1)$
X_n
0
1
...1001110

A propos de la mesure invariante

- Souvenez-vous... pour le peigne, on a dit qu'il existait une mesure invariante si et seulement si

$$\sum_{n\in\mathbb{N}} c_n = \sum_{n\in\mathbb{N}} \prod_{k=0}^{n-1} q_{0^k1}(0) < \infty.$$

- On note maintenant

$$\Theta_1 = \sum_{n\geq 0} \prod_{k=0}^{n-1} q_{0^k1}(0),$$

et

$$\Theta_2 = \sum_{n\geq 0} \prod_{k=0}^{n-1} q_{1^k0}(1).$$

A propos de la mesure invariante

- Souvenez-vous... pour le peigne, on a dit qu'il existait une mesure invariante si et seulement si

$$\sum_{n\in\mathbb{N}} c_n = \sum_{n\in\mathbb{N}} \prod_{k=0}^{n-1} q_{0^k1}(0) < \infty.$$

- On note maintenant

$$\Theta_1 = \sum_{n\geq 0} \prod_{k=0}^{n-1} q_{0^k1}(0),$$

et

$$\Theta_2 = \sum_{n\geq 0} \prod_{k=0}^{n-1} q_{1^k0}(1).$$

CNS d'existence de la mesure invariante

Proposition (Mesure stationnaire pour le double peigne)

On suppose que $q_{0^\infty}(0) \neq 1$ *et* $q_{1^\infty}(1) \neq 1$. *Le processus de Markov* $(U_n)_{n\geq 0}$ *admet une mesure invariante sur les mots infinis à gauche* $\mathcal{L}$ *si et seulement si* Θ_1 *et* Θ_2 *convergent. Dans ce cas, la mesure stationnaire est unique et satisfait :*

$$\pi(0) = \frac{\Theta_1}{\Theta_1 + \Theta_2} \quad , \quad \pi(10) = \frac{1}{\Theta_1 + \Theta_2}.$$

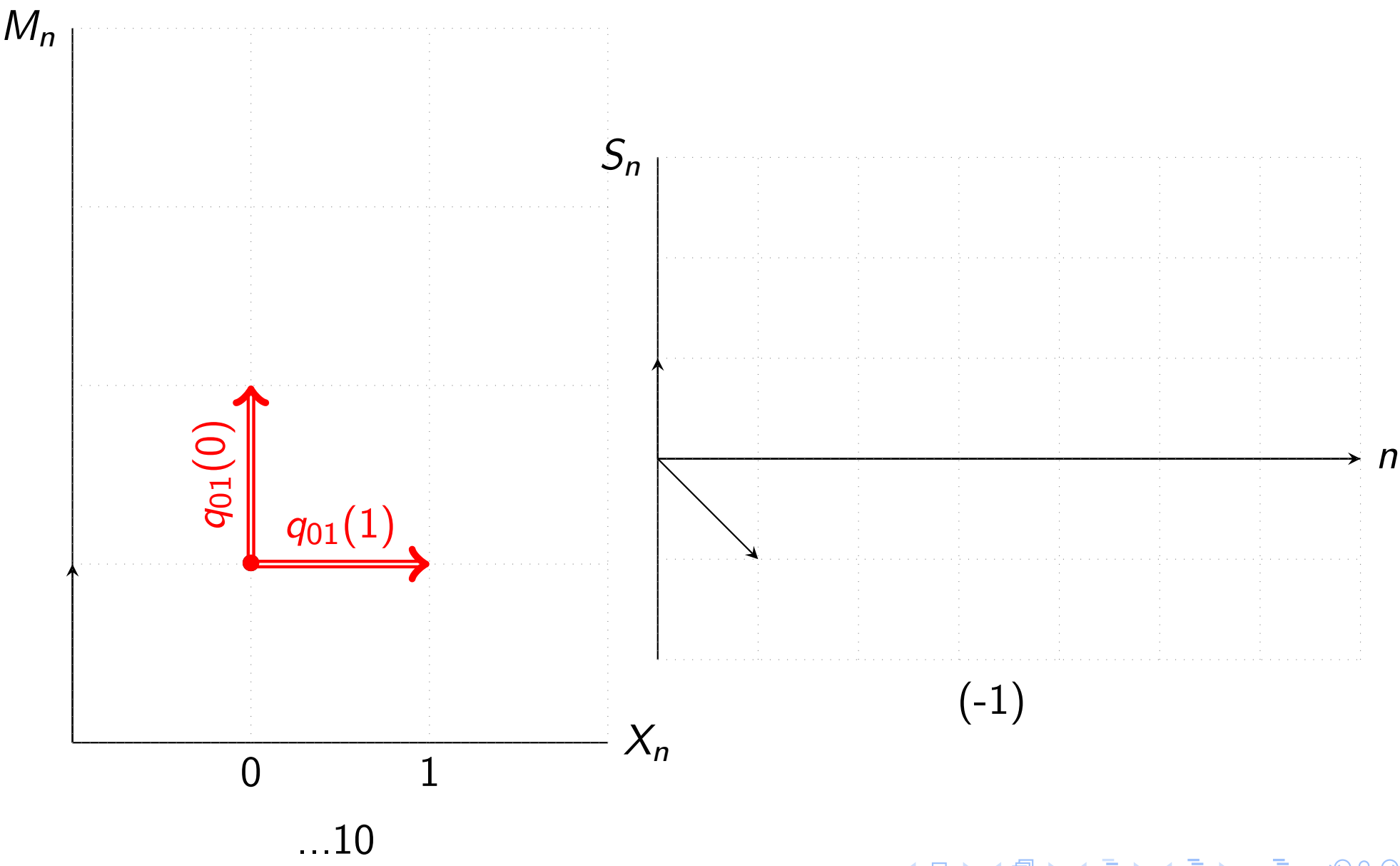
Remarque : On peut retrouver ce résultat en utilisant la chaîne de Markov (X_n, M_n), en calculant sa mesure invariante et en cherchant la loi de la première marginale.

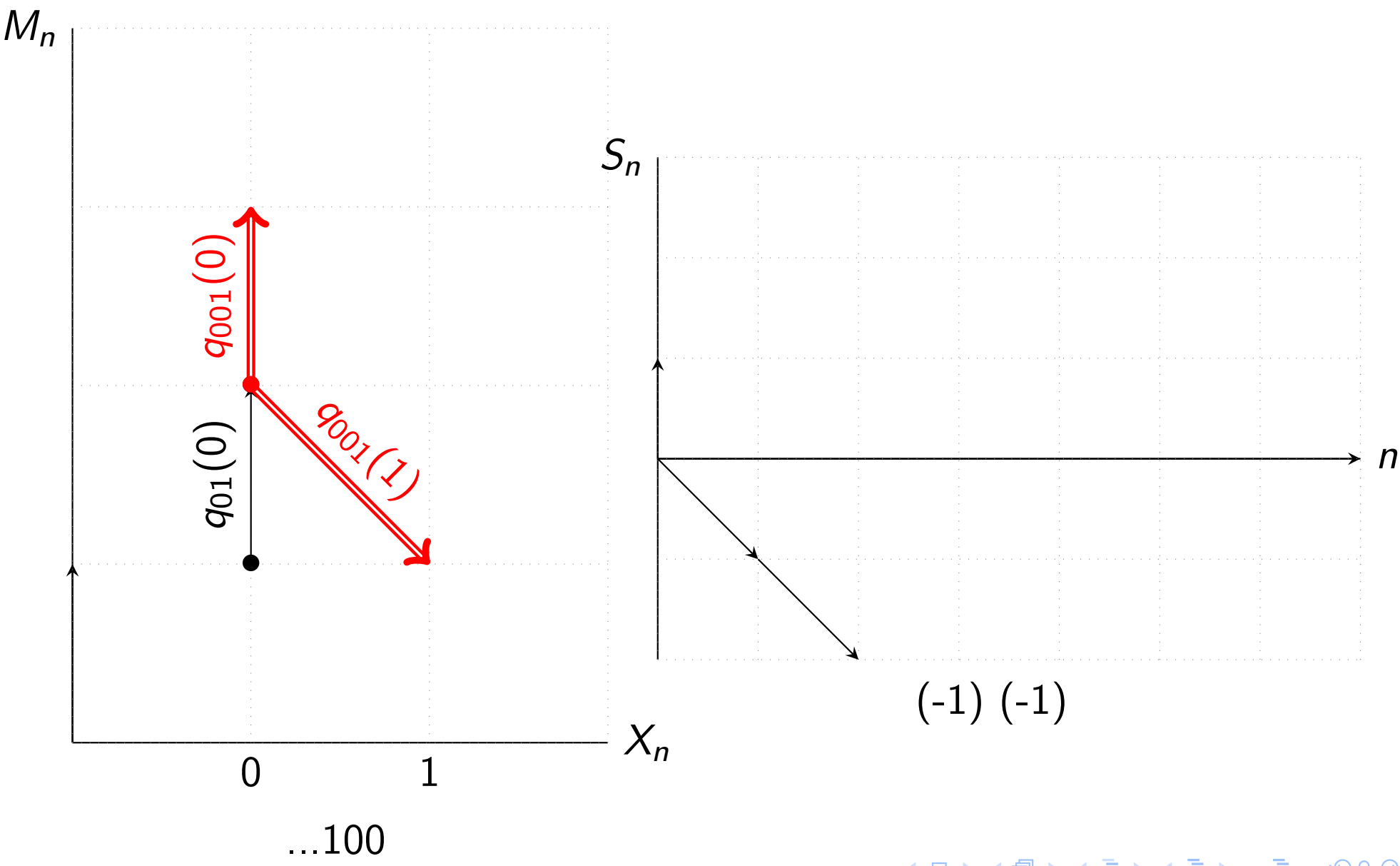
La marche aléatoire associée

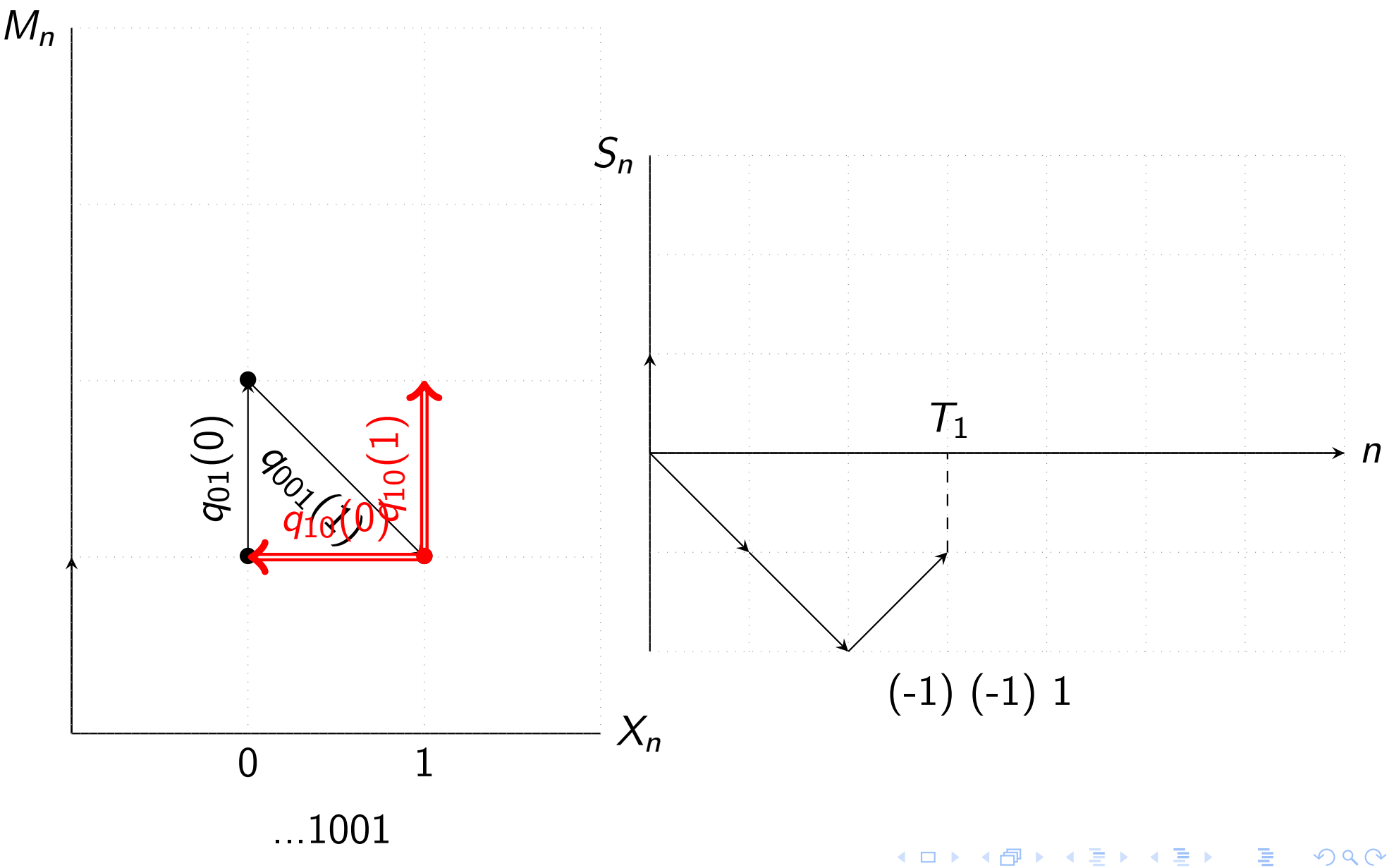
- On change le codage dans la suite (X_n) : $0 \longrightarrow -1$.
- On s'intéresse à la marche aléatoire dont les incréments sont la première marginale de la chaîne de Markov (X_n, M_n) :

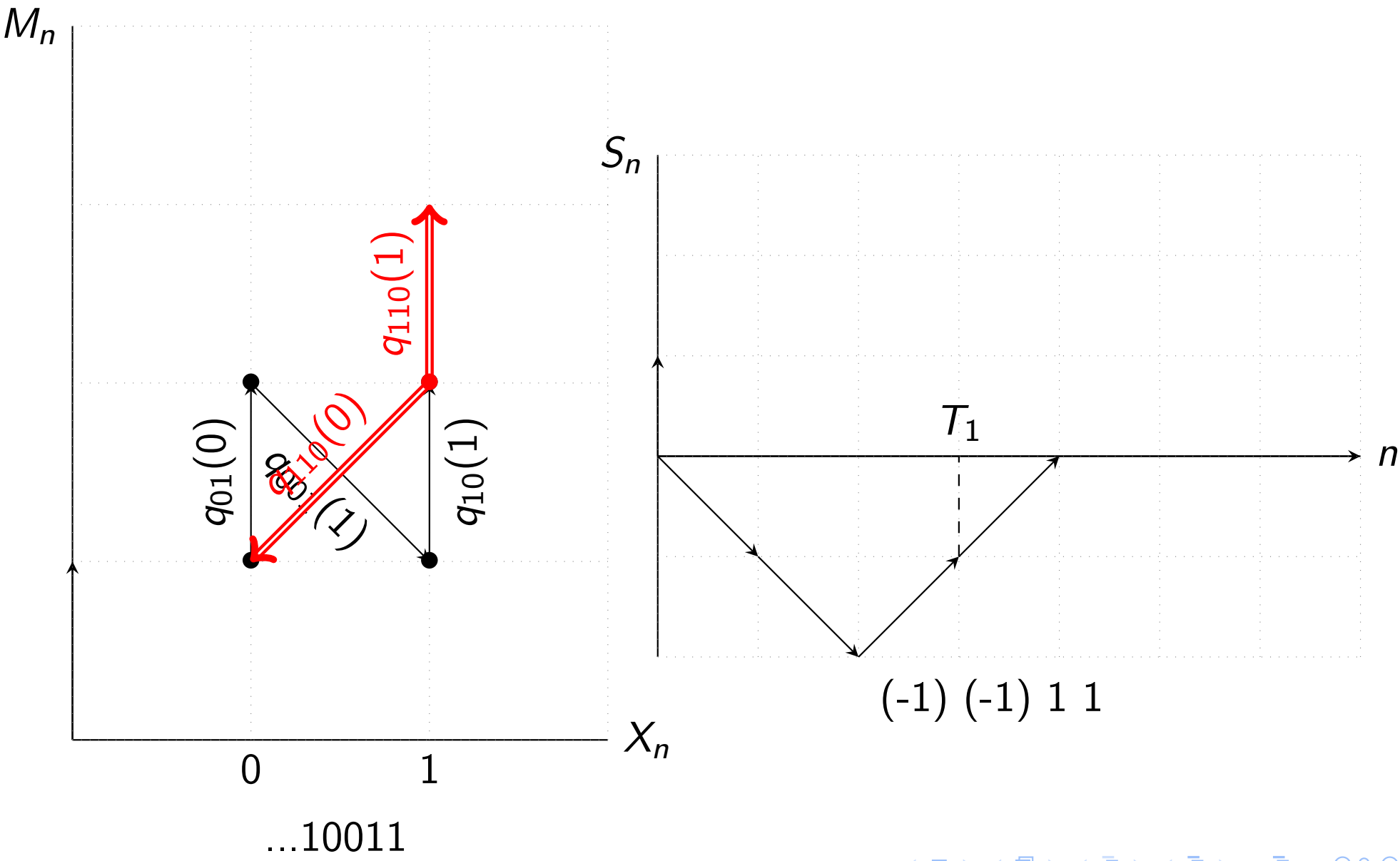
$$S_n = \sum_{k=0}^{n} X_k.$$

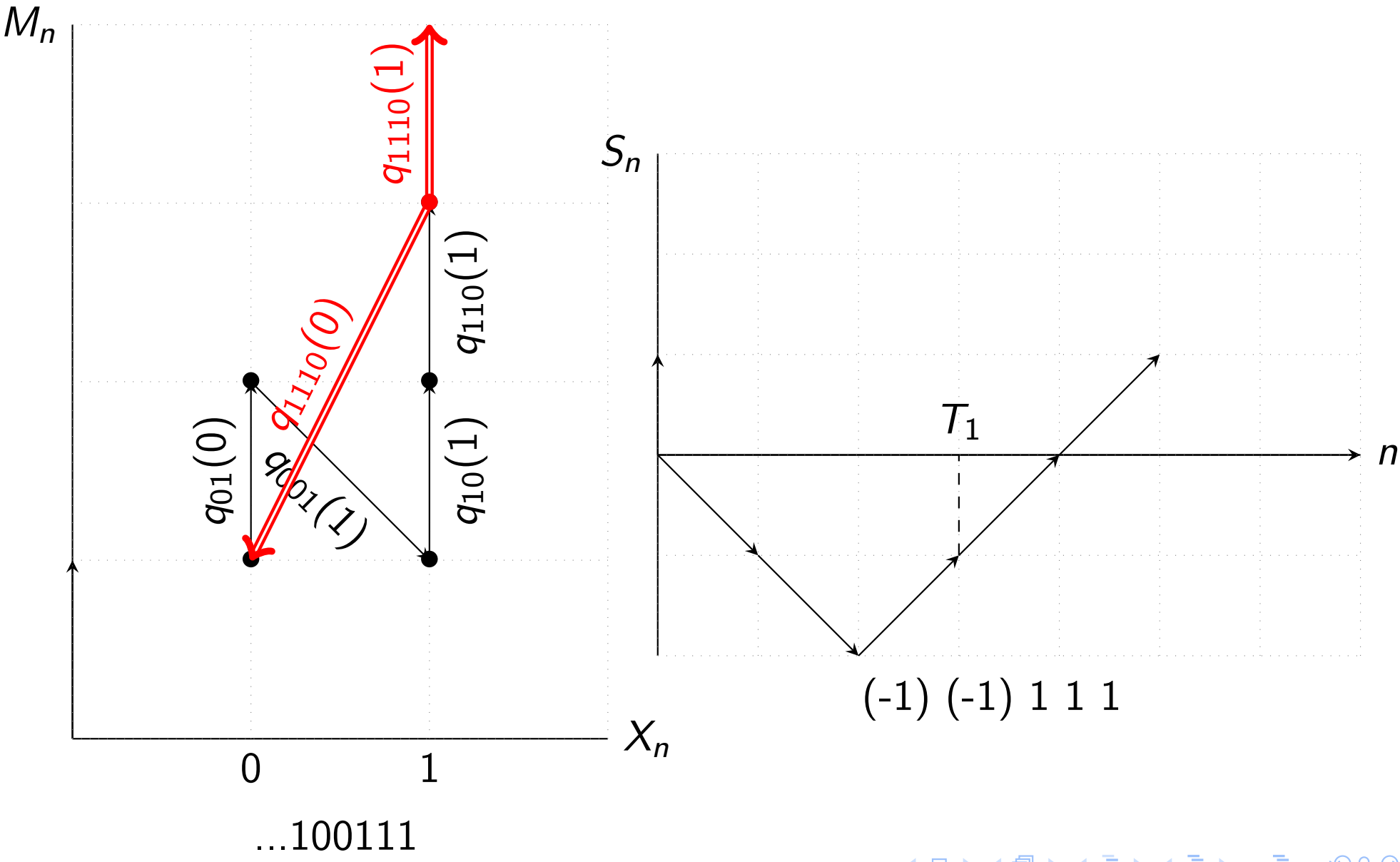
- Les incréments ont de la mémoire : on parle de marche aléatoire persistante.

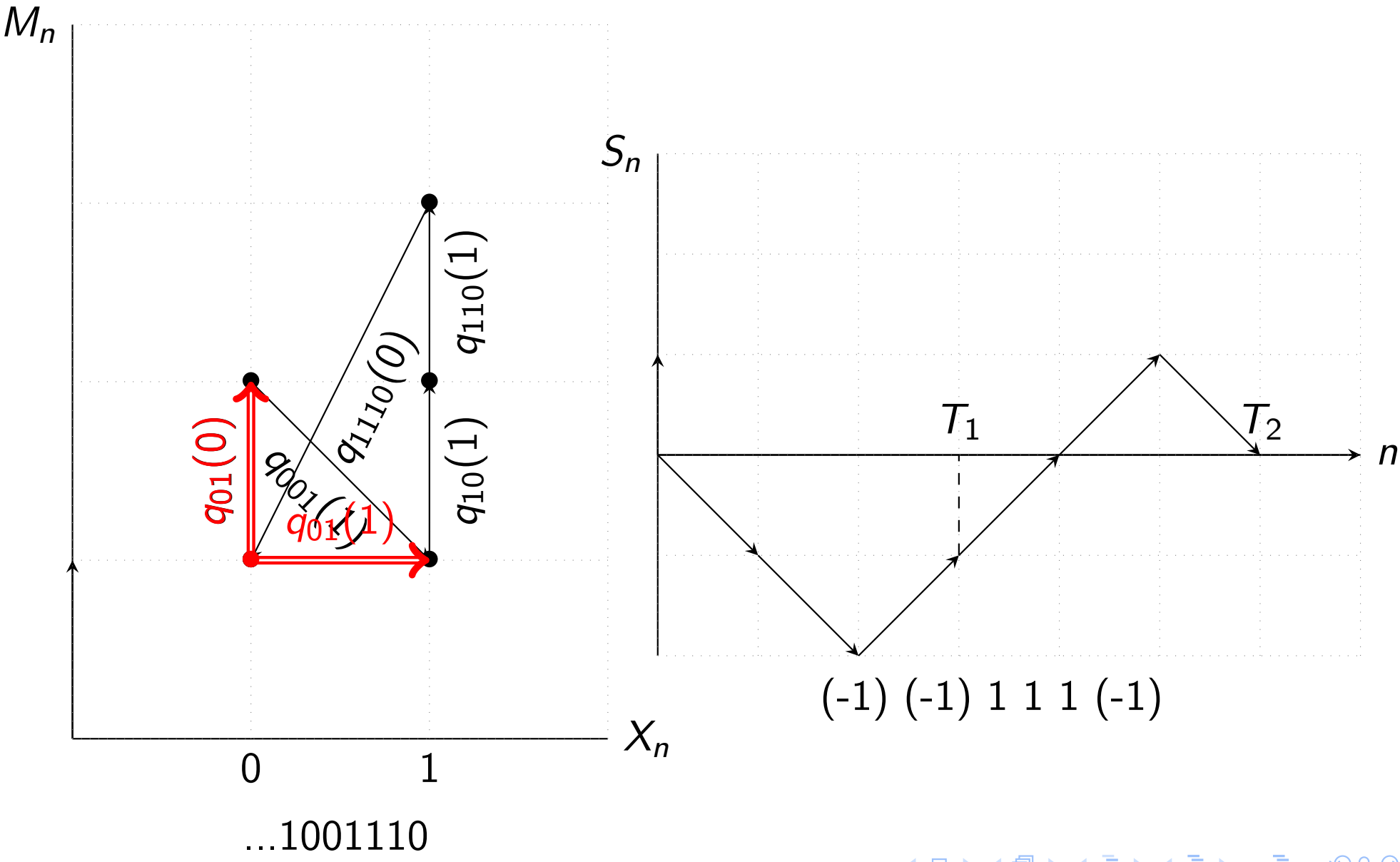

M_n
S_n
n
$q_{01}(0)$
$q_{01}(1)$
X_n
0
1
...10
(-1)


M_n
$q_{001}(0)$
$q_{01}(0)$
$q_{001}(1)$
0
1
...100
X_n
S_n
n
(-1) (-1)


M_n
$q_{01}(0)$
$q_{001}(0)$
$q_{10}(0)$
$q_{10}(1)$
0
1
X_n
...1001
S_n
T_1
n
(-1) (-1) 1


M_n
$q_{110}(1)$
$q_{10}(0)$
$q_{01}(0)$
$q_{01}(1)$
$q_{10}(1)$
X_n
0
1
...10011
S_n
T_1
n
(-1) (-1) 1 1


M_n
$q_{1110}(1)$
$q_{1110}(0)$
$q_{110}(1)$
$q_{01}(0)$
$q_{01}(1)$
$q_{10}(1)$
X_n
0
1
...100111
S_n
T_1
n
(-1) (-1) 1 1 1


M_n
$q_{01}(0)$
$q_{001}(1)$
$q_{01}(1)$
$q_{1110}(0)$
$q_{110}(1)$
$q_{10}(1)$
X_n
0
1
...1001110
S_n
T_1
T_2
n
(-1) (-1) 1 1 1 (-1)

Pour des incréments Markoviens...

Herrmann, Vallois, From persistent random walk to the telegraph noise, *Stochastics and Dynamics*, 2010.

- Construction d'un processus à temps continu à partir de la marche.
- Le processus limite est l'ITN (Integrated Telegraph Noise), lié à à une EDP, l'équation du télégraphe :

$$\frac{\partial^2 u}{\partial t^2}(x,t) = a^2 \frac{\partial^2}{\partial x^2} u(x,t) - 2c \frac{\partial u}{\partial t} u(x,t),$$

où $u(x,t)$ représente la tension dans un câble au point x et à l'instant t.

- But : étendre ce travail avec des incréments possédant une plus grande mémoire et étudier le processus limite.

Vers le passage au continu

- On définit les instants de saut T_n par

$$T_n = \inf\left\{n \geq T_{n-1} : X_n \neq X_{T_{n-1}}\right\}.$$

- On suppose $S_0 = 0$ et $X_1 = -1$. La trajectoire diminue de 1 en 1 jusqu'à $T_1 - 1$ où la marche atteint un minimum local. Puis elle augmente de 1 en 1 jusqu'à $T_2 - 1$ ...
- On note N_t le nombre de sauts jusqu'à t : $N_t = \sum_n \mathbb{1}_{\{T_n \leq t\}}$ et on a

$$S_t = -\sum_{n=1}^{t} (-1)^{N_n}.$$

- La trajectoire de $(S_t)_{t\geq 0}$ est l'interpolation linéaire entre les points

$$\begin{aligned}(W_n, Z_n) &= (T_n - 1, S_{T_n} - 1 + (-1)^n) \\ &= \left(T_n - 1, \sum_{k=1}^{n} (-1)^k (T_k - T_{k-1})\right).\end{aligned}$$

Vers le passage au continu (2)

- On introduit un paramètre d'échelle $\varepsilon > 0$ et on note

$$\begin{aligned} q_{0^k1}(1) &= f_1(k\varepsilon)\varepsilon + \varepsilon\alpha_{1,k,\varepsilon}, \\ q_{1^k0}(0) &= f_2(k\varepsilon)\varepsilon + \varepsilon\alpha_{2,k,\varepsilon}, \end{aligned}$$

avec $\lim_{\varepsilon\to 0} \sup_{i,k} |\alpha_{i,k,\varepsilon}| = 0$ et où f_1 et f_2 sont deux fonctions positives telles que

$$\int_0^\infty f_i(u)du = \infty.$$

- On définit le processus

$$S^\varepsilon(t) = \varepsilon S_k, \quad t = k\varepsilon,$$

et on prolonge par interpolation linéaire pour tout $t \geq 0$ réel.

Vers le passage au continu (3)

- Soit une suite (e_n) de variables aléatoires indépendantes telles que

$$\mathbb{P}(e_{2n-1} > t) = \exp - \int_0^t f_2(u)du, \quad \mathbb{P}(e_{2n} > t) = \exp - \int_0^t f_1(u)du.$$

- On définit le processus de comptage

$$N_t^0 = \sum_{n \geq 1} \mathbb{1}_{\{e_1 + \ldots + e_n \leq t\}}, \quad \text{pour tout } t \geq 0.$$

Théorème (CCHV)

On suppose $S^\varepsilon(0) = \varepsilon$ et $M_0 = 1$. Alors $(S^\varepsilon(t), t \geq 0)$ converge en loi lorsque ε tend vers 0 vers le processus défini par

$$S^0(t) = \int_0^t (-1)^{N_s^0} ds, \quad t \geq 0.$$

Merci !