Bestiaire de chaînes de Markov à mémoire variable et marches aléatoires persistantes

Peggy Cénac-Guesdon - IMB

INSTITUT DE MATHÉMATIQUES DE BOURGOGNE

UNIVERSITÉ DE BOURGOGNE

Travaux en collaboration avec B. Chauvin, B. De Loynes, A. Le Ny, Y. Offret, F. Paccaut, N. Pouyanne.

Journées MAS 2016 - 29 août 2016

Plan

1. Chaîne de Markov à mémoire variable
   - Introduction
   - Définition d'une VLMC
   - Premiers exemples
   - Bestiaire

2. Marches aléatoires persistantes
   - Le double peigne
   - La marche aléatoire associée
   - Récurrence/transience

Chaîne de Markov à mémoire variable

Introduction

- Motivation : modéliser une chaîne de caractères par de l'aléatoire
  - $\ldots AACGTGACCATTGAGA \ldots$
  - $\ldots 0110101000110110001110 \ldots$
- Exemples d'applications :
  - modélisation de séquences biologiques (ADN, protéines)
  - modélisation du langage,
  - analyse des mouvements humains (Wang et Liu, thèse de Thierry Dumont),
  - modélisation des regroupements de segments chromosomiques prenant en compte les déséquilibres de liaison (thèse de Laval Jacquin),...
- Illustration du défaut de la modélisation markovienne quand l'ordre de la chaîne est grand ; pour un alphabet de taille 4 :

ordre	0	1	2	3	4	5	10
paramètres	3	12	48	192	768	3072	$3,1.10^6$

Introduction

- Motivation : modéliser une chaîne de caractères par de l'aléatoire
  - $\dots AACGTGACCATTGAGA\dots$
  - $\dots 0110101000110110001110\dots$
- Exemples d'applications :
  - modélisation de séquences biologiques (ADN, protéines)
  - modélisation du langage,
  - analyse des mouvements humains (Wang et Liu, thèse de Thierry Dumont),
  - modélisation des regroupements de segments chromosomiques prenant en compte les déséquilibres de liaison (thèse de Laval Jacquin),...
- Illustration du défaut de la modélisation markovienne quand l'ordre de la chaîne est grand ; pour un alphabet de taille 4 :

ordre	0	1	2	3	4	5	10
paramètres	3	12	48	192	768	3072	$3,1.10^6$

Introduction

- Motivation : modéliser une chaîne de caractères par de l'aléatoire
  - $\ldots AACGTGACCATTGAGA \ldots$
  - $\ldots 0110101000110110001110 \ldots$
- Exemples d'applications :
  - modélisation de séquences biologiques (ADN, protéines)
  - modélisation du langage,
  - analyse des mouvements humains (Wang et Liu, thèse de Thierry Dumont),
  - modélisation des regroupements de segments chromosomiques prenant en compte les déséquilibres de liaison (thèse de Laval Jacquin),...
- Illustration du défaut de la modélisation markovienne quand l'ordre de la chaîne est grand ; pour un alphabet de taille 4 :

ordre	0	1	2	3	4	5	10
paramètres	3	12	48	192	768	3072	$3,1.10^6$

Chaîne de Markov à mémoire variable

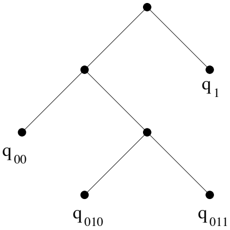
- Ici, pour simplifier, l'alphabet est $\mathcal{A} = \{0, 1\}$ ou $\mathcal{A} = \{u, d\}$. On va définir des chaînes de Markov à valeur dans $\mathcal{L} = \mathcal{A}^{-\mathbb{N}}$. On part de U_0 tiré selon une loi initiale sur $\mathcal{L}$: $U_0 = \ldots X_{-2}X_{-1}X_0$. A chaque pas de temps, on ajoute aléatoirement une lettre à droite :

$$\forall n \geq 0, \quad U_{n+1} = U_n X_{n+1} = \ldots X_0 X_1 \ldots X_n X_{n+1}.$$

  (U_n) est une chaîne de Markov sur les mots infinis à gauche $\mathcal{L}$.
- 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, introduit par Rissanen (1983). L'algorithme CONTEXT permet d'estimer les contextes et les transitions.
- On suppose que l'ensemble des contextes est dénombrable.

Exemple : le petit bambou

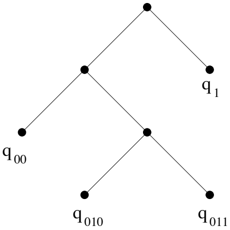
On se donne un arbre des contextes probabilisés :



$$U_0 = \cdots 00 \qquad \mathbb{P}(X_1 = 1 | U_0) = q_{00}(1)$$

Exemple : le petit bambou

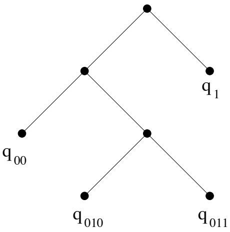
On se donne un arbre des contextes probabilisés :



$$
\begin{aligned}
U_0 &= \cdots 00 \qquad & \mathbb{P}(X_1 = 1 | U_0) = q_{00}(1) \\
U_1 &= \cdots 001 \qquad & \mathbb{P}(X_2 = 0 | U_1) = q_1(0)
\end{aligned}
$$

Exemple : le petit bambou

On se donne un arbre des contextes probabilisés :



$$
\begin{aligned}
U_0 &= \cdots 00 & \qquad \mathbb{P}(X_1 = 1 | U_0) &= q_{00}(1) \\
U_1 &= \cdots 001 & \qquad \mathbb{P}(X_2 = 0 | U_1) &= q_{1}(0) \\
U_2 &= \cdots 0010 & \qquad \mathbb{P}(X_3 = 0 | U_2) &= q_{010}(0)
\end{aligned}
$$

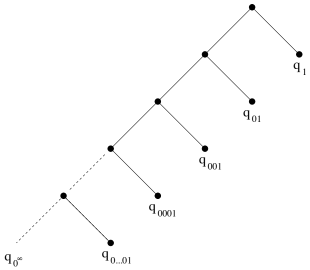
La transition de U_n vers U_{n+1} dépend d'un suffixe de U_n :

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

Remarques sur le caractère Markovien

- La suite (U_n) à valeur dans $\mathcal{L}$ est une chaîne de Markov.
- 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 mesure de probabilités invariante de (U_n) ?
  - Y'a t-il convergence vers cette mesure de probabilités ?
  - Quelles sont les propriétés de mélange de (X_n) ?
  - ...

Le peigne infini

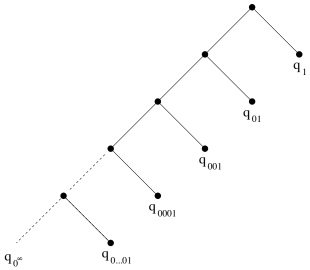


Si $U_n = \cdots 110\color{red}{1000}$, le contexte est $\color{red}{0001}$ et

$$\mathbb{P}(X_{n+1} = \alpha | U_n) = q_{\color{red}{0001}}(\alpha).$$

C'est la longueur du dernier "run" de '0' qui définit la loi de transition. (X_n) n'est pas markovienne a priori.

Le peigne infini



On note pour $n \geqslant 1$, $c_n := \displaystyle\prod_{k=0}^{n-1} q_{0^k1}(0)$ et $c_0 = 1$.

Proposition (C., 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.

Vers le bestiaire

- Résultat de Gallo Paccaut (2013) : S'il y a un nombre fini de contextes infinis et si il existe ε tel que

$$\forall c \in \mathcal{C}, \forall \alpha \in \mathcal{A} \quad \varepsilon < q_c(\alpha) < 1 - \varepsilon,$$

  alors il existe une mesure stationnaire pour la VLMC.
- On suppose que

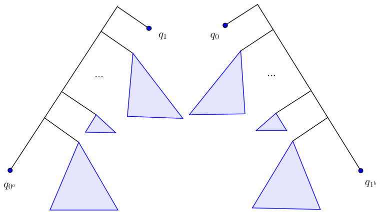
$$\forall c \in \mathcal{C}, \forall \alpha \in \mathcal{A}, q_c(\alpha) \neq 0, 1.$$

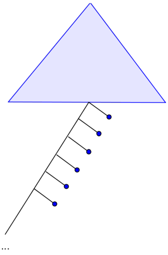
  Si 1 et 0^a sont contextes (ou si 0 et 1^a) sont contextes, alors il existe une unique mesure de probabilité invariante. (se montre en appliquant les résultats généraux de Meyn and Tweedies sur les CM à espaces d'états non dénombrables)

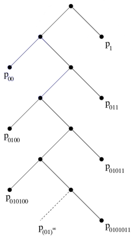
Au zoo (1)

On suppose que $\forall c \in \mathcal{C}, \forall \alpha \in \mathcal{A}, q_c(\alpha) \neq 0, 1$.

- Mesure de probabilités invariante sans condition ;



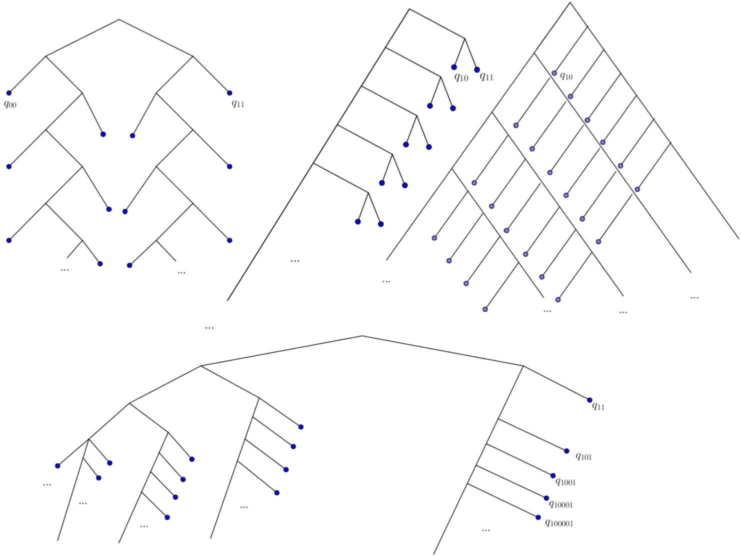




Cénac*, Chauvin, De Loynes, Le Ny, Offret, Paccaut, Pouyanne.

Au zoo (2)

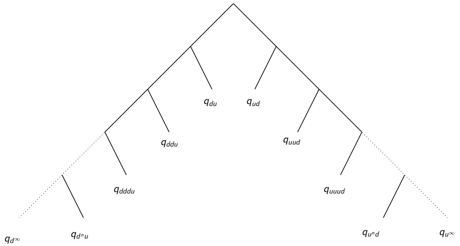
- CNS d'existence et unicité de probabilité invariante :



Marches aléatoires persistantes

Cénac*, Chauvin, De Loynes, Le Ny, Offret, Paccaut, Pouyanne.

Le double peigne



- $u_n = \dots uudd\,uddd$ contexte : $dddu$ et $\mathbb{P}(U_{n+1} = U_n\alpha | U_n) = q_{dddu}(\alpha)$
- $u_n = \dots uudd\,duu$ contexte : uud et $\mathbb{P}(U_{n+1} = U_n\alpha | U_n) = q_{uud}(\alpha)$

Le modèle : prise en compte de la mémoire

- On considère l'alphabet $\mathcal{A} = \{u, d\}$, où u désigne une montée et est associée à un saut de longueur $+1$ et d désigne une descente associée à -1.
- On construit une VLMC $(U_n)_{n\in\mathbb{N}}$ à partir du double peigne. On note $X_n := +1$ si la dernière lettre de U_n est u, et $X_n = -1$, si la dernière lettre de U_n est d.
- On définit la marche aléatoire persistante

$$S_n := \sum_{\ell=1}^{n} X_\ell.$$

- Pour tout $n \geq 1$, $m \geq 0$,

$$\begin{aligned} \mathbb{P}\left(S_{m+1} = S_m + 1 | U_m = \ldots ud^n\right) &= q_{d^n u}(u) \\ \mathbb{P}\left(S_{m+1} = S_m - 1 | U_m = \ldots du^n\right) &= q_{u^n d}(d). \end{aligned}$$

Le modèle : prise en compte de la mémoire

- On considère l'alphabet $\mathcal{A} = \{u, d\}$, où u désigne une montée et est associée à un saut de longueur $+1$ et d désigne une descente associée à -1.
- On construit une VLMC $(U_n)_{n\in\mathbb{N}}$ à partir du double peigne. On note $X_n := +1$ si la dernière lettre de U_n est u, et $X_n = -1$, si la dernière lettre de U_n est d.
- On définit la marche aléatoire persistante

$$S_n := \sum_{\ell=1}^{n} X_\ell.$$

- Pour tout $n \geq 1$, $m \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(S_{m+1} = S_m + 1 | U_m = \dots ud^n\right) &= q_{d^n u}(u) \\
\mathbb{P}\left(S_{m+1} = S_m - 1 | U_m = \dots du^n\right) &= q_{u^n d}(d).
\end{aligned}$$

Le modèle : prise en compte de la mémoire

- On considère l'alphabet $\mathcal{A} = \{u, d\}$, où u désigne une montée et est associée à un saut de longueur $+1$ et d désigne une descente associée à -1.
- On construit une VLMC $(U_n)_{n\in\mathbb{N}}$ à partir du double peigne. On note $X_n := +1$ si la dernière lettre de U_n est u, et $X_n = -1$, si la dernière lettre de U_n est d.
- On définit la marche aléatoire persistante

$$S_n := \sum_{\ell=1}^{n} X_\ell.$$

- Pour tout $n \geq 1$, $m \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(S_{m+1} = S_m + 1 | U_m = \dots ud^n\right) &= q_{d^n u}(u) \\
\mathbb{P}\left(S_{m+1} = S_m - 1 | U_m = \dots du^n\right) &= q_{u^n d}(d).
\end{aligned}$$

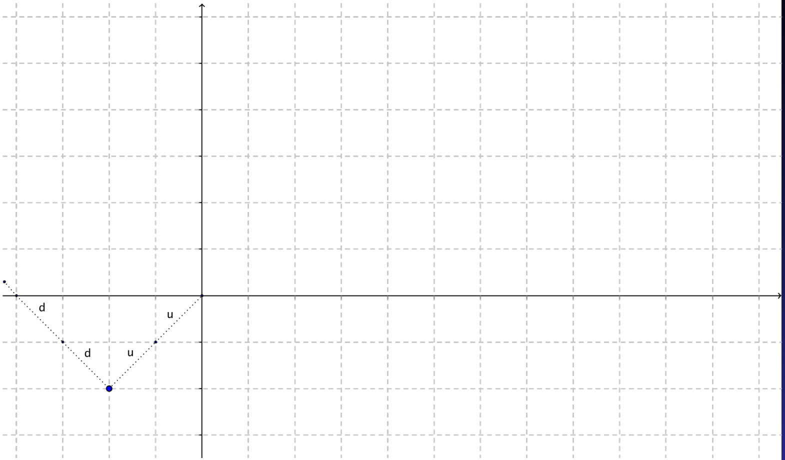
Le modèle : prise en compte de la mémoire

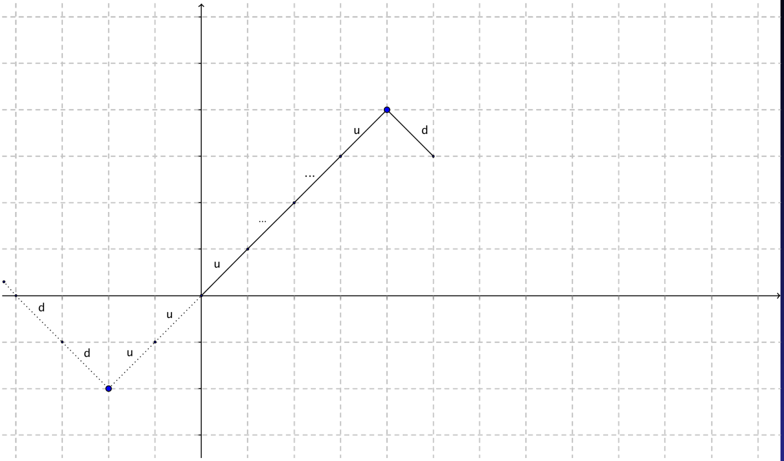
- On considère l'alphabet $\mathcal{A} = \{u, d\}$, où u désigne une montée et est associée à un saut de longueur $+1$ et d désigne une descente associée à -1.
- On construit une VLMC $(U_n)_{n\in\mathbb{N}}$ à partir du double peigne. On note $X_n := +1$ si la dernière lettre de U_n est u, et $X_n = -1$, si la dernière lettre de U_n est d.
- On définit la marche aléatoire persistante

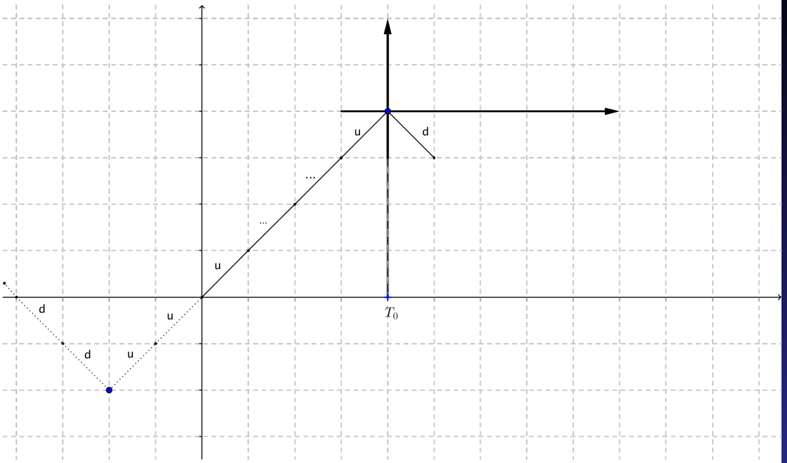
$$S_n := \sum_{\ell=1}^{n} X_\ell.$$

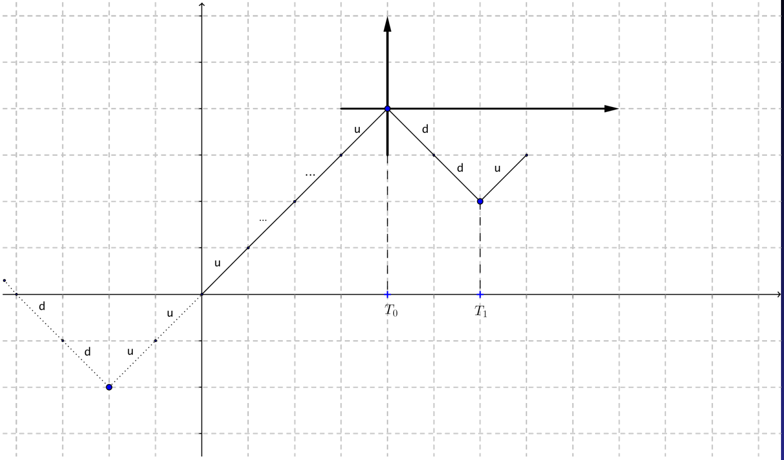
- Pour tout $n \geq 1$, $m \geq 0$,

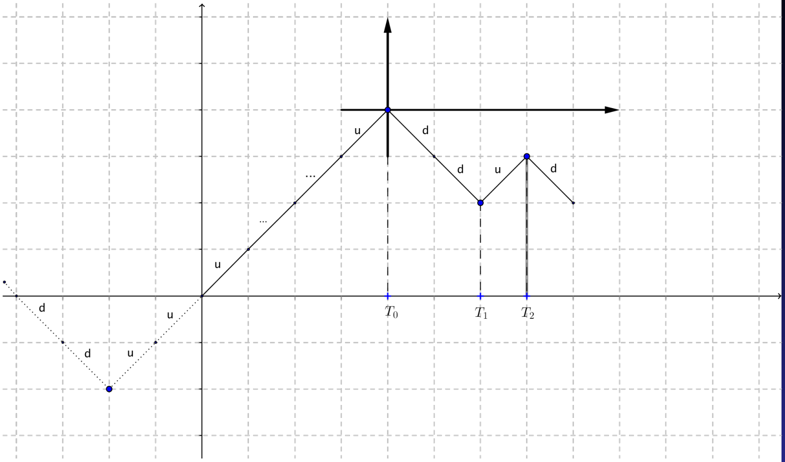
$$\begin{aligned}\mathbb{P}\left(S_{m+1} = S_m + 1 | U_m = \ldots ud^n\right) &= q_{d^n u}(u)\\ \mathbb{P}\left(S_{m+1} = S_m - 1 | U_m = \ldots du^n\right) &= q_{u^n d}(d).\end{aligned}$$


d
d
u
u

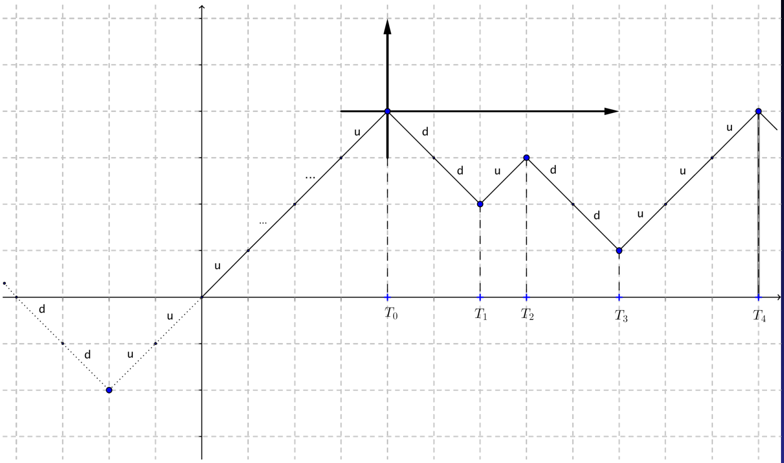

d
d
u
u
u
...
...
u
d

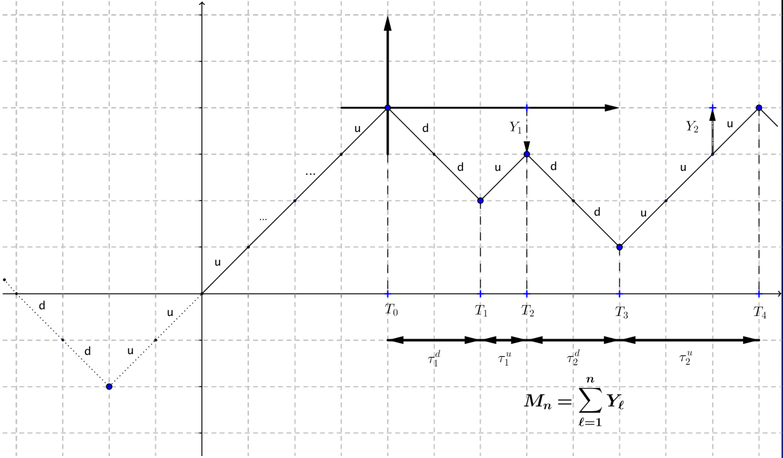

u
d
...
...
u
d
u
d
u
T_0


u
d
...
d
u
...
u
T_0
T_1
d
u
d
u



u
d
d
u
d
...
...
u
T_0
T_1
T_2
d
u
d
u


u
u
...
...
u
d
d
u
d
d
u
u
u
d
d
u
d
u
T_0
T_1
T_2
T_3
T_4



$$\boldsymbol{M_n = \sum_{\ell=1}^{n} Y_\ell}$$

A propos de la mesure invariante

- 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 $\Theta_d = \sum_n \prod_{k=0}^{n-1} q_{d^ku}(d)$ et $\Theta_u = \sum_n \prod_{k=0}^{n-1} q_{u^kd}(u)$.
En réalité ici $\Theta_d = \mathbb{E}[\tau_1^d]$ et $\Theta_u = \mathbb{E}[\tau_1^u]$.

Proposition (Mesure stationnaire pour le double peigne)

On suppose $q_{d^\infty}(d) \neq 1$ et $q_{u^\infty}(u) \neq 1$. Le processus de Markov $(U_n)_{n\geq 0}$ admet une mesure invariante sur $\mathcal{L}$ si et seulement si Θ_u et Θ_d convergent. Dans ce cas, la mesure stationnaire est unique.

Drift et probabilités de changements de cap

- Cas où le drift est bien défini : $(\Theta^d < \infty \text{ ou } \Theta^u < \infty)$

$$\mathbf{d}_M := \underbrace{\Theta^d - \Theta^u}_{\in[-\infty,\infty]} \quad \text{et} \quad \mathbf{d}_S := \frac{\Theta^u - \Theta^d}{\Theta^u + \Theta^d} \in [-1, 1].$$

Par exemple, Θ^ℓ est finie si

$$q_{\ell^n \bar{\ell}}(\bar{\ell}) \geq \frac{1}{n} + \frac{1}{n\log(n)} + \cdots + \frac{1+\varepsilon}{n\log(n)\cdots\log_{[p]}(n)},$$

- Le drift n'est pas défini : par exemple pour $\ell \in \{u, d\}$,

$$q_{\ell^n \bar{\ell}}(\bar{\ell}) \leq \frac{1}{n} + \frac{1}{n\log(n)} + \cdots + \frac{1}{n\log(n)\cdots\log_{[p]}(n)}.$$

Probabilités de changements de cap et queue de distribution des sauts

Exemple : Si les probabilités de changements de cap s'écrivent avec des paramètres $\lambda_0, \lambda_1, \cdots, \lambda_p$, pour n assez grand sous la forme

$$q_{\ell^n \bar{\ell}}(\bar{\ell}) = \frac{\lambda_0}{n} + \frac{\lambda_1}{n\log(n)} + \cdots + \frac{\lambda_p}{n\log(n)\cdots\log_{[p]}(n)},$$

alors la queue de distribution des longueurs de saut $\mathbb{P}(\tau_1^\ell \geq n)$ est de l'ordre

$$\frac{1}{n^{\lambda_0}(\log(n))^{\lambda_1}\cdots(\log_{[p]}(n))^{\lambda_p}}.$$

Récurrence/transience

- Avec le double peigne, on construit des exemples de marches récurrentes et non symétriques sur $\mathbb{Z}$.
- Le comportement asymptotique de la marche est déterminé par les queues de distributions $\mathcal{T}^\ell(n) := \mathbb{P}(\tau^\ell \geq n)$
- Si le drift $\mathbf{d}_M$ est bien défini, la marche est récurrente ssi $\mathbf{d}_M = 0$. Dans ce cas, $\lim_{n\to\infty} \frac{S_n}{n} = \mathbf{d}_S = 0$.
  Une légère perturbation des paramètres q_c conduit en général à de la transience. Le critère est global.
- Que se passe-t'il dans le cas où le drift n'est pas défini ?

Récurrence/transience

- Avec le double peigne, on construit des exemples de marches récurrentes et non symétriques sur $\mathbb{Z}$.
- Le comportement asymptotique de la marche est déterminé par les queues de distributions $\mathcal{T}^{\ell}(n) := \mathbb{P}(\tau^{\ell} \geq n)$
- Si le drift $\mathbf{d}_M$ est bien défini, la marche est récurrente ssi $\mathbf{d}_M = 0$. Dans ce cas, $\lim_{n\to\infty} \frac{S_n}{n} = \mathbf{d}_S = 0$.
  Une légère perturbation des paramètres q_c conduit en général à de la transience. Le critère est global.
- Que se passe-t'il dans le cas où le drift n'est pas défini ?

Récurrence/transience

- Avec le double peigne, on construit des exemples de marches récurrentes et non symétriques sur $\mathbb{Z}$.
- Le comportement asymptotique de la marche est déterminé par les queues de distributions $\mathcal{T}^{\ell}(n) := \mathbb{P}(\tau^{\ell} \geq n)$
- Si le drift $\mathbf{d}_M$ est bien défini, la marche est récurrente ssi $\mathbf{d}_M = 0$. Dans ce cas, $\lim_{n\to\infty} \frac{S_n}{n} = \mathbf{d}_S = 0$.
  Une légère perturbation des paramètres q_c conduit en général à de la transience. Le critère est global.
- Que se passe-t'il dans le cas où le drift n'est pas défini ?

Récurrence/transience

- Avec le double peigne, on construit des exemples de marches récurrentes et non symétriques sur $\mathbb{Z}$.
- Le comportement asymptotique de la marche est déterminé par les queues de distributions $\mathcal{T}^{\ell}(n) := \mathbb{P}(\tau^{\ell} \geq n)$
- Si le drift $\mathbf{d}_M$ est bien défini, la marche est récurrente ssi $\mathbf{d}_M = 0$. Dans ce cas, $\lim\limits_{n\to\infty} \dfrac{S_n}{n} = \mathbf{d}_S = 0$.
  Une légère perturbation des paramètres q_c conduit en général à de la transience. Le critère est global.
- Que se passe-t'il dans le cas où le drift n'est pas défini ?

Récurrence/transience

<table>
<tr><td></td><td colspan="2">$\Theta_u < \infty$</td><td colspan="2">$\Theta_u = \infty$</td></tr>
<tr><td rowspan="2">$\Theta_d < \infty$</td><td rowspan="2">recurrent
$\mathbf{d}_S = 0$</td><td>drifting $+\infty$
$\mathbf{d}_S > 0$</td><td colspan="2" rowspan="2">drifting $+\infty$</td></tr>
<tr><td>drifting $-\infty$
$\mathbf{d}_S < 0$</td></tr>
<tr><td rowspan="2">$\Theta_d = \infty$</td><td colspan="2" rowspan="2">drifting $-\infty$</td><td rowspan="2">recurrent
$J_{u|d} = J_{u|d} = \infty$</td><td>drifting $+\infty$
$\infty = J_{u|d} > J_{d|u}$</td></tr>
<tr><td>drifting $-\infty$
$\infty = J_{d|u} > J_{u|d}$</td></tr>
</table>

$$J_{\ell_1|\ell_2} := \sum_{n=1}^{\infty} \frac{n\mathbb{P}(\tau^{\ell_1} = n)}{\sum_{k=1}^{n} \mathbb{P}(\tau^{\ell_2} \geq k)}.$$

Preuve

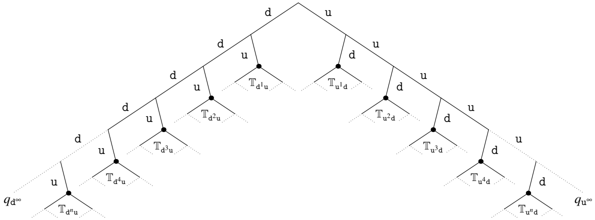
Récurrence/transience

- Contrairement au cas où le drift est bien défini, dans le cas où le drift n'est pas défini, la marche peut rester récurrente en modifiant les probabilités de changements de cap, tant que la perturbation reste asymptotiquement contrôlée. Critère asymptotique/global.
- Les marches persistantes (S_n) et $(\widetilde{S}_n)$ sont simultanément récurrentes ou transientes si leurs queues de distribution de longueur de saut sont du même ordre : pour $\ell \in \{u, d\}$,

$$\mathcal{T}_\ell(n) \asymp \widetilde{\mathcal{T}}_\ell(n).$$

Perturbations aléatoires

On peut perturber le double peigne par des petits arbres finis et sous des hypothèses faibles sur les probabilités de changement de cap, montrer que la marche persistante perturbée est de même nature que la marche non perturbée (pourtant on perd la propriété de renouvellement).



Merci !

Eléments de preuve

- Le comportement asymptotique en terme de récurrence/transience de (S_n) est le même que celui de (M_n).
- Dans le cas où le drift n'est pas défini, on introduit la marche

$$M_n^\xi := \sum_{k=1}^n J_k^\xi, \text{ avec } J_k^\xi = \xi_k \tau_k^u - (1-\xi_k)\tau_k^d,$$

  $(\xi_n)_{n\geq 0}$ i.i.d. Bernoulli de paramètre $p \in [0,1]$, indépendant des suite (τ_n^u) et (τ_n^d). $(M_n^\xi)_{n\geq 0}$ et $(M_n)_{n\geq 0}$ ont le même comportement asymptotique.
- On applique ensuite les critères de Erickson (1974) à la marche $(M_n^\xi)_{n\geq 0}$.