Chapitre 1

Chaînes de Markov à espace d'états fini

1. Rappels du calcul des probabilités

1.1 Espace probabilisé

Un modèle probabiliste, appelé encore « espace probabilisé », est un triplet $(\Omega, \mathrm{A}, \mathrm{P})$, où :
- Ω est l'espace des réalisations possibles du phénomène aléatoire modélisé ;
- A est une tribu (ensemble de sous-ensembles de Ω contenant $\varnothing$, Ω, stable par l'union dénombrable et par le passage au complémentaire), les éléments de A étant appelés « évènements » ;
- et P est une probabilité, ou encore « loi de probabilité » (application de A dans $[0, 1]$ prenant 0 en $\varnothing$, 1 en Ω, et telle que l'image d'une union dénombrable d'ensembles deux à deux disjoints est la somme des images de ces ensembles).

Comment avoir une idée intuitive de la probabilité? Si, pour un évènement $A \in \mathrm{A}$, on a $P[A]$=0.3, que cela signifie-t-il? Une des possibilités de se forger une idée intuitive est de faire appel à la loi des grands nombres, selon laquelle "lorsque l'on effectue, de manière indépendante, n observations aléatoires dans Ω réalisées selon la même loi de probabilité P, la proportion des observations se trouvant dans A tend vers $P[A]$ lorsque n tend vers l'infini". On peut donc voir la probabilité comme étant une limite de quelque chose de bien concret.

1.2. Espérance mathématique

Rappelons que la mesure de Lebesgue "sur R " (c'est un abus de langage car cette mesure est définie sur la tribu borélienne de R , qui peut être définie – entre autres – comme étant la plus petite tribu, au sens de l'inclusion, contenant les intervalles de type $[a, b]$ associe à chaque intervalle $[a, b]$ sa longueur: $\mu([a, b]) = b - a$ (on montre que si une mesure est définie sur les ensembles de forme $[a, b]$, elle se prolonge de manière unique à la tribu borélienne). C'est une mesure de base, très intuitive, qui associe à chaque objet linéique sa longueur. On construit alors, en commençant par les fonctions en escalier, la théorie

de l'intégration : à une fonction $[a,b]$ de $\mathbb{R}$ dans $\mathbb{R}$ ayant les bonnes propriétés, on associe son intégrale $\int_{\mathbb{R}} f(t)dt$.

Lorsque l'on est amené à utiliser un ensemble discret E, fini ou non, on peut considérer la mesure de dénombrement, qui la deuxième mesure "de base" utilisée depuis la nuit des temps. La mesure de dénombrement ν sur E (là encore c'est un abus de langage car cette mesure est définie sur la tribu E constituée, dans le cas fini, de tous les sous-ensembles de E) associe à chaque $A \subset E$ fini son cardinal : $\nu[A] = \mathrm{Card}[A]$. La « mesure » d'un ensemble est donc simplement le nombre de ses éléments. Comme ci-dessus, on peut alors associer à une fonction f de E dans R son "intégrale par rapport à ν", qui est : $\sum_{e\in E} f(e)\nu[\{e\}] = \sum_{e\in E} f(e)$ (somme ou série).

Considérons $(\Omega, \mathrm{A}, \mathrm{P})$, un espace probabilisé. La probabilité P étant une mesure, on peut considérer, comme ci-dessus, des intégrales. Pour une fonction X de Ω dans R, son intégrale par rapport à P est dite "espérance" : $E[X] = \int_{\Omega} X dP$.

Il est important de noter que l'intégrale de la fonction indicatrice 1_A d'un ensemble A (qui est définie par $1_A(\omega) = 1$ si $\omega \in A$, et $1_A(\omega) = 0$ si $\omega \notin A$) par rapport à une mesure est la mesure de cet ensemble, nous avons donc $E[1_A] = P[A]$. L'espérance est à la probabilité ce que l'intégrale classique (dans R) est à la mesure de Lebesgue (pour un intervalle, sa longueur).

Les probabilités sont souvent définies par des "densités" par rapport à d'autres mesures. Une probabilité P "sur R" est donnée par une densité f par rapport à la mesure de Lebesgue μ lorsque $P[A] = \int_A f(t)dt$ (rappelons que ce type de densité a été introduit de manière intuitive dans le paragraphe 7 du chapitre 1). Pour E discret, une probabilité P "sur E" est donnée par une densité f par rapport à la mesure de dénombrement ν lorsque $P[A] = \sum_{e\in A} f(e)$, qui est l'intégrale de f sur A par rapport à ν. Notons que lorsque $P[\{e\}]$ existe pour tout $e \in E$, nous avons également $P[A] = \sum_{e\in A} P[\{e\}]$. Ainsi $P[\{e\}]$ peut être interprétée comme la valeur au point e de la densité de P par rapport à la mesure de dénombrement ν. Attention, cela n'est pas vrai dans le cas des probabilités P définies par des densités « sur R », on peut en effet montrer que dans ce cas $P[\{x\}] = 0$ pour tout $x \in \mathrm{R}$.

Lorsqu'il y a des densités (les deux cas ci-dessus), les espérances s'écrivent :

$$E[X] = \int_{\Omega} X dP = \int_{\mathrm{R}} x f(x) dt \ ; \qquad (1.1)$$

$$E[X] = \sum_{e\in\Omega} X(e)P[\{e\}] = \sum_{1\le i\le n} x_i f(x_i), \qquad (1.2)$$

où les $\{x_1,...,x_n\}$ est l'ensemble image de E par X. Les calculs des espérances se ramènent ainsi aux calculs classiques des intégrales, des séries, ou des sommes finies.
Notons dès à présent un fait pouvant apparaître comme troublant : la connaissance de la densité f de la loi d'une variable aléatoire $X:\Omega\to \mathrm{R}$ est suffisante pour calculer son espérance : on n'a besoin de connaître ni de l'ensemble de départ Ω, ni X en tant que fonction.
Soit un espace probabilisé (Ω, A, P) et X une variable aléatoire réelle (une fonction de Ω dans R « mesurable », voire Remarque 1 ci-après). Nous avons vu ci-dessus que la probabilité d'un événement pouvait être interprété comme « la fréquence de son apparition pour un grand nombre d'observations indépendantes ». Nous avons un résultat analogue pour les espérances : lorsqu'un phénomène modélisé par (Ω, A, P) se reproduit n fois de manière indépendante en donnant ω_1, …, ω_n, …, alors (cette convergence est appelée « loi des grands nombres ») :

$$\frac{X(\omega_1)+...+X(\omega_n)}{n}\xrightarrow[n\to+\infty]{} E[X] \qquad (1.3)$$

Ainsi l'espérance (qui n'est pas aléatoire) peut être vue comme la limite des moyennes arithmétiques des observations (qui sont aléatoires), lorsque le nombre d'observations tend vers l'infini. Remarquons qu'en prenant $X=1_A$ (fonction caractéristique de A), (1.3) donne l'interprétation de la probabilité de A mentionnée précédemment.

Remarque 1

Une fonction X de Ω dans R « mesurable » veut dire que l'image inverse de la tribu borélienne est une sous-tribu de A. Demander la mesurabilité est justifiée intuitivement : lorsque l'on étudie un phénomène aléatoire décrit par la réalisation de X il est naturel de vouloir pouvoir dire quelle est la probabilité pour que X tombe dans des intervalles $[a,b]$. Pour que $P[X\in[a,b]]=P[X^{-1}[a,b]]$ existe il faut que $X^{-1}[a,b]$ soit dans la tribu A (rappelons que P est définie sur A qui est une tribu sur Ω). Sachant que $X^{-1}[a,b]$ est dans A pour tout $[a,b]$ et sachant que les intervalles engendrent la tribu borélienne, on peut alors montrer que $X^{-1}(B)$ est dans A pour n'importe que élément B de la tribu borélienne.

La variance $Var[X]=E[(X-E[X])^2]$ est la deuxième caractéristique importante des variables aléatoires réelles. En appliquant (1.3) à $(X-E[X])^2$ au lieu de X on a

$$\frac{(X(\omega_1)-E[X])^2+...+(X(\omega_1)-E[X])^2}{n}\xrightarrow[n\to+\infty]{} Var[X]\ ; \qquad (1.4)$$

Ainsi $Var[X] = E[(X - E[X])^2]$ exprime, toujours en vertu de la loi des grands nombres, l'écart "moyen" des réalisations de la variable aléatoire X de leur valeur "moyenne" $E[X]$. Au plan intuitif, la variance mesure le « degré de stochasticité » d'un phénomène. Lorsque la variance est grande le phénomène est très aléatoire. Lorsqu'elle diminue il devient de plus en plus « déterministe ». Enfin, lorsque la variance est nulle l'aspect stochastique disparaît et X est une constante. Cette dernière propriété est à l'origine du fait que beaucoup de calculs faits de manière déterministe, en absence d'aspects aléatoires, apparaissent comme des cas particuliers des calculs fait en présence de l'aléatoire, cette dernière situation généralisant la première. On peut également noter que lorsque les variances sont petites le calcul déterministe, souvent bien plus simple que le calcul probabiliste, peut donner des résultats « approximatifs » satisfaisants.

1.3. Probabilité image et théorème de transfert

Soit un espace probabilisé (Ω, A, P) et X une fonction de Ω dans R. En munissant R de la tribu borélienne $\mathrm{B}(R)$ (qui est la plus petite - au sens de l'inclusion - tribu contenant les intervalles), X est dite mesurable, ou encore X est dite "variable aléatoire", lorsque pour tout B dans $\mathrm{B}(R)$ son image réciproque $X^{-1}[B]$ est dans A. La donnée de (Ω, A, P) et X mesurable définit alors une probabilité (notée P_X et appelée "loi de X") sur $\mathrm{B}(R)$ pa

$$P_X[B] = P[X^{-1}[B]] \tag{1.5}$$

De manière imagée, une variable aléatoire "transporte", ou "transfère" la probabilité de l'ensemble de départ sur l'ensemble de l'arrivée (P_X est également dite « probabilité image » de P par X).

En reprenant le cadre ci-dessus, considérons une fonction φ de R dans R mesurable (l'image réciproque d'un borélien est un borélien : pour $B \in \mathrm{B}(R)$, $\varphi^{-1}[B] \in \mathrm{B}(R)$). Nous avons alors le schéma suivant

$$(\Omega, \mathrm{A}, P) \xrightarrow{X} (R, \mathrm{B}(R), P_X) \xrightarrow{\varphi} (R, \mathrm{B}(R), P_{\varphi \circ X}) \tag{1.6}$$

et on peut se poser la question du calcul de l'espérance de la variable aléatoire $Y = \varphi o X$. Selon le théorème de transfert, nous avons :

$$E[\varphi o X] = \int_{\Omega} \varphi o X dP = \int_{R} \varphi dP_X \tag{1.7}$$

Le théorème concerne la deuxième égalité, la première étant la définition de l'espérance.

Le théorème de transfert est fondamental en théorie des probabilités. Il montre que l'on peut calculer $\int_{\Omega} \varphi o X dP$ ne connaissant ni Ω, ni A, ni P, ni même X - ce qui peut être quelque peu déroutant - la connaissance de la loi P_X de variable aléatoire X et de la fonction φ est suffisante. De manière imagée, le théorème de transfert permet de transférer les calculs d'un espace sur un autre, dans le but de les rendre possibles, ou plus simples. Notons que ce théorème généralise les résultats bien connus concernant les changements de variables pour le calcul d'intégrales.
Notons enfin que (1.7) donne (1.5) en prenant $\varphi = 1_B$. Par ailleurs, pour $\varphi = Id$ (identité), (1.7) donne $E[X]$ (si P_X admet une densité f par rapport à la mesure de Lebesgue, on retrouve bien (1.1)).

1.1.4. Conditionnement

Considérons deux variables aléatoires réelles X et Y modélisant deux phénomènes aléatoires. Nous avons :

$$(\Omega, \mathrm{A}, P) \xrightarrow{(X,Y)} (R^2, \mathrm{B}(R^2), P_{(X,Y)}), \tag{1.8}$$

où $B(R^2)$ est la tribu borélienne (la plus petite tribu - au sens de l'inclusion - contenant les rectangles). On peut alors se poser plusieurs questions, que nous commentons rapidement en se plaçant dans le cas particulier où $P_{(X,Y)}$ admet une densité $f_{(X,Y)}$ par rapport à la mesure de Lebesgue sur $\mathrm{B}(R^2)$ (on dira, par abus de langage, "mesure de Lebesgue sur R^2"), laquelle mesure associe à un ensemble sa surface (lorsque cela est possible, à savoir lorsque l'ensemble est dans la tribu borélienne ; rappelons cependant qu'il est difficile d'exhiber un ensemble qui n'y soit pas).
La première question concerne les "lois marginales" : si l'on ne regarde que X, comment décrire son comportement (comment trouver sa loi P_X)? La densité f_X de P_X par rapport à la mesure de Lebesgue sur R est donnée par

$$f_X(x) = \int_{\mathrm{R}} f_{(X,Y)}(x,y)dy \tag{1.9}$$

(formule analogue pour P_Y). De manière plus générale, cela signifie que la connaissance du comportement aléatoire de plusieurs variables (donnée par leur loi jointe, i.e. la loi du paquet) implique la connaissance du comportement aléatoire de tout sous-ensemble de ces variables (donnée par les lois jointes des sous-paquets). Attention, la réciproque n'est pas vraie en général. Ainsi P_X et P_Y ne déterminent pas $P_{(X,Y)}$ (cela n'est le cas que si X et Y sont indépendantes) et, plus généralement, les lois de tous les sous paquets de $X_1,\ldots,X_n$ de cardinal inférieur ou égal à $n-1$ ne déterminent pas, en général, la loi de

$(X_1,...,X_n)$. Ainsi, pour trois variables (X_1, X_2, X_3), les lois $P_{(X_1, X_2)}$, $P_{(X_1, X_3)}$, et $P_{(X_2, X_3)}$ ne déterminent pas $P_{(X_1, X_2, X_3)}$.

Supposons que Y est observable et X ne l'est pas. Que peut-on dire du comportement de X lorsque l'on a observé $Y = y$? La loi de X devient la loi conditionnelle dont la densité s'écrit :

$$f_X^{Y=y}(x) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{f_{(X,Y)}(x,y)}{\int_R f_{(X,Y)}(x,y)dx} \tag{1.10}$$

Cette loi conditionnelle permet de se faire une idée sur la manière dont les deux phénomènes aléatoires modélisés par X et Y sont liés.

L'espérance conditionnelle est l'espérance selon la loi conditionnelle. Pour deux variables X et Y, de loi conjointe de densité $f_{(X,Y)}(x,y)$ et de densités marginales $f_X(x)$, $f_Y(y)$, on note :

$$E[Y|X = x] = \int_R y f_Y^{X=x}(y)dy = \int_R y \frac{f_{(X,Y)}(x,y)}{f_X(x)} dy \tag{1.11}$$

avec formule analogue pour $E[X|Y = y]$.

Exemple 1

Supposons que $Y = y$ modélise le niveau de la bourse, la température dehors, ou encore la vitesse d'un véhicule, à l'instant présent t, et X modélise le même phénomène à l'instant futur $t + \Delta t$. Supposons que le phénomène évolue de manière stochastique mais continue. Lorsque Δt est "petit", connaître $Y = y$ c'est connaître X, le lien entre les deux observations est déterministe (la variance de la loi donnée par $f_X^{Y=y}$ sera « petite »). Dans de tels cas, nous n'avons pas besoin de la théorie des probabilités. Lorsque Δt est "grand", $Y = y$ n'aura aucune influence sur le comportement de X : les deux phénomènes sont "indépendants" (on a $f_{(X,Y)}(x,y) = f_X(x) f_Y(y)$, et donc $f_X^{Y=y} = f_X$: l'observation y de Y n'apporte pas d'information sur le comportement de X). Les cas intéressants sont les cas intermédiaires : Δt est trop grand pour que l'on puisse utiliser les démarches déterministes et il est suffisamment petit pour que $Y = y$ apporte une information significative sur le comportement de X (ce qui signifie que la densité $f_X^{Y=y}$ est "assez différente" de la densité f_X). Dans de telles situations, les méthodes statistiques de prévision peuvent présenter de qualités remarquables.

Notons que lorsque le couple (X,Y) prend ses valeurs dans un ensemble fini (ou dénombrable), les intégrale dans (1.9)-(1.11) sont remplacées par des sommes (ou séries). On peut également avoir un cas mixte, que nous verrons dans ce cours en classification Bayésienne, où l'une de deux variables (X par exemple) est discrète et l'autre continue.

Dans ce cas, (1.9) est une intégrale lorsque l'on "intègre" par rapport à y, mais est une somme (ou série) lorsque l'on "intègre" par rapport à x.

Le point important à noter est que toute l'information concernant les liens entre les phénomènes aléatoires modélisés par X et Y est contenue dans la loi du couple $P_{(X,Y)}$ (qui donne les deux lois marginales et les deux familles de lois conditionnelles). Réciproquement, $P_{(X,Y)}$ est donnée indifféremment par P_X et la famille $P_Y^{X=x}$, ou par P_Y et la famille $P_X^{Y=y}$.

Dans la pratique, on est souvent confronté au calcul de $E[\varphi(X,Y)]$, où φ est une fonction de R^2 dans R. On peut alors utiliser la formule générale suivante :

$$E[\varphi(X,Y)] = E[E[\varphi(X,Y)|Y]] \tag{1.12}$$

qui permet de calculer $E[\varphi(X,Y)]$ - qui est une intégrale "double" (deux variables) - par une succession de deux intégrales "simples" (ce qui est connu en théorie de l'intégration comme « théorème de Fubini »).
Lorsque la loi $P_{(X,Y)}$ admet une densité $f_{(X,Y)}$ on montre aisément (1.12) en écrivant

$$E[\varphi(X,Y)] = \int_{R^2} \varphi(x,y) f_{(X,Y)}(x,y)dxdy = \int_R [\int_R \varphi(x,y) f_{(X,Y)}(x,y)dx]dy =$$

$$\int_R [\int_R \varphi(x,y) f_Y(y) \frac{f_{(X,Y)}(x,y)}{f_Y(y)} dx]dy = \int_R f_Y(y)[\int_R \varphi(x,y) f_X^{Y=y}(x)dx]dy = E[E[\varphi(X,Y)|Y]]$$

La formule (1.12), comme le théorème de transfert, est fondamentale en probabilité-statistique. Son utilisation « intuitive » peut être la suivante. Supposons que l'on cherche l'espérance d'une fonction de plusieurs variables, et on se rend compte que l'on saurait trouver la solution si une partie de ces variables, notons leur ensemble Y, était connue. Cela signifie que l'on sait calculer $g(y) = E[\varphi(X,Y)|Y = y]$. La fonction g, étant donnée, il reste à calculer $E[g(Y)]$. On a transformé un problème en un autre, éventuellement plus facile à résoudre.

Par exemple, notons le cas particulier (1.12) suivant : en posant $\varphi(X,Y) = X$, on a pour tout couple de variables X, Y :

$$E[X] = E[E[X|Y]] \tag{1.13}$$

Ainsi, si l'on cherche calculer $E[h(X)]$. On peut on peut essayer de considérer un autre phénomène aléatoire dont l'observation rendrait le calcul (qui devient donc le calcul de l'espérance conditionnelle) possible. Cela est illustré par l'exemple suivant :

Exemple 2

Dans une promotion d'une école on note la proportion de 0.3 de filles et 0.7 de garçons. La taille des filles suit la loi gaussienne de moyenne 1.65 cm et celle des garçons suit la loi gaussienne de moyenne 1.75 cm. Quelle est l'espérance de la variable aléatoire « la taille de l'individu choisi au hasard dans la promotion » ? La réponse n'est pas immédiate car il y a deux phénomènes aléatoires simultanés : « fille ou garçon » et « taille ». Cependant, la réponse est immédiate si on fixe la variable « fille ou garçon ». On pose donc comme Y la variables « fille ou garçon », X la taille. On a donc $E[X|Y = \mathit{fille}] = 1.65$, et $E[X|Y = \mathit{garçon}] = 1.75$. En appliquant (1.13) on trouve donc la taille moyenne de la promotion :

$$E[X] = E[E[X|Y]] = 0.3E[X|Y = \mathit{fille}] + 0.7E[X|Y = \mathit{garçon}] = 0.3 \times 1.65 + 0.7 \times 1.75 = 1.72$$

Un autre type de problème couramment rencontré – y compris dans les sujets des différents examens – est le suivant. On se donne deux variables X, Y, la loi de X, et la loi de Y conditionnelle à X. Les diverses questions portent alors sur la loi de X conditionnelle à Y. On est donc amené à « inverser » le conditionnement : la loi de X et la loi de Y conditionnelle à X donnent la loi du couple (X,Y) suit, ce qui permet de calculer la loi de X conditionnelle à Y. Cela est illustré dans l'exemple suivant.

Exemple 3

Un élève prépare un examen de statistique et, après y avoir consacré une durée de temps x, il obtient la note y. En normalisant, on suppose que ces valeurs sont dans $[0,1]$. Le lien entre x et y n'étant pas déterministe, on le modélise en considérant un vecteur aléatoire (X,Y) à valeurs dans $[0,1]\times[0,1]$. Supposons que le temps de préparation suit la loi de probabilité de densité $f_X(x) = 2x$, et la note obtenue, après avoir travaillé pendant $X = x$, suit la loi de probabilité de densité $f_Y^{X=x}(y) = (2y-1)x + 1$. On peut alors poser différentes questions concernant la loi de X (temps de travail) sachant $Y = y$ (la note).

Par exemple, sachant $Y = y$, quelle est la probabilité p pour que l'élève ait travaille pendant plus de 0.5 ?

La loi du couple (X,Y) suit la loi de probabilité de densité $f_{(X,Y)}(x,y) = f_X(x) f_Y^{X=x}(y) = -2x^2 + 4x^2 y + 2x$, et la note obtenue $Y = y$ est une réalisation de Y qui suit alors la loi de probabilité de densité la $f_Y(y) = \int_0^1 f_{(X,Y)}(x,y)dx = \frac{4}{3}y + \frac{1}{3}$. Sachant la note $Y = y$ d'un élève donné, le professeur de statistique sait que le temps qu'il a consacré au travail suit une la loi de

probabilité de densité $f_X^{Y=y}(y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{6x[(2y-1)x+1]}{4y+1}$. Pour répondre à la question on intègre $f_X^{Y=y}$ sur [0.5,1] : $p = \int_{0.5}^{1} f_X^{Y=y}(x)dx$.

On peut noter que $f_X(x) = 2x$ est relativement optimiste car l'espérance (le temps moyen consacré à l'étude) est $\int_0^1 2x^2 dx = \frac{2}{3}$.

2. Chaînes de Markov à espace d'états fini et temps discret.

2.1 Introduction

On considère une collection infinie des variables aléatoires (un processus stochastique) $X = (X(t))_{t \in I}$, où I est un sous-ensemble de R et chaque X_t étant à valeurs dans un espace E (qui sera, dans la pratique, N ou R^n). Le processus X est markovien si pour tout instant u, pour une valeur $X(u) = x$ fixée les variables aléatoires $X(t), t > u$ sont indépendantes des variables aléatoires $X(s), s < u$. Un tel processus est aussi dit "sans mémoire": la loi de probabilité de son évolution après chaque instant dépend de son état à cet instant mais, *sachant son état à cet instant*, est indépendante de ses états antérieurs. Insistons sur le fait qu'il s'agit de l'indépendance conditionnellement à $X(u) = x$: sans conditionnement (on n'observe pas $X(u)$), les $X(s)$, $X(t)$ ne sont pas nécessairement indépendants (cela quelques soient $s < u < t$).

Exemple 1

Considérons $I = N$, et $X(n)$ modélisant le temps (au sens météorologique) : $X(n) = 1$ signifie "il fait beau" le jour n, et $X(n) = 0$ signifie "il ne fait pas beau". On peut imaginer que le temps qu'il fera demain dépend essentiellement du temps qu'il fait aujourd'hui : sachant le temps qu'il fait aujourd'hui le temps qu'il fera demain dépend peu du temps qu'il a fait hier, on peut donc considérer, du moins en première approximation, le processus $X = (X(m))_{m \in N}$ comme étant markovien. Supposons maintenant qu'il s'agit du temps dans un endroit lointain (nous ignorons le temps qu'il y fait aujourd'hui, mais nous connaissons le temps qu'il y a fait hier). Si l'on nous demande le temps qu'il y fera demain, nous tiendrons compte, bien entendu, du temps qu'il y a fait hier, ce qui signifie que $X(n+1)$ et $X(n-1)$ ne sont pas indépendants. Cet exemple illustre l'indépendance de $X(n+1)$ et $X(n-1)$ conditionnellement à $X(n)$, mais leur dépendance sans le conditionnement par $X(n)$.

Définition 1

Un processus stochastique réel $X = (X(t))_{t \in I}$ est appelé processus de Markov si pour tout sous-ensemble fini $t_1 < t_2 < ... < t_n$ de valeurs du paramètre t, la distribution de $X(t_n)$ conditionnelle à $X(t_1), X(t_2), ..., X(t_{n-1})$ dépend seulement de la valeur de $X(t_{n-1})$. Soit, en utilisant la fonction de répartition :

$$P[X(t_n) \le x_n / X(t_1) = x_1, X(t_2) = x_2, ..., X(t_{n-1}) = x_{n-1}] = P[X(t_n) \le x_n / X(t_{n-1}) = x_{n-1}] \qquad (2.2.1)$$

On dit aussi que connaissant le "présent" du processus, le "futur" est indépendant du "passé".

A un processus de Markov est donc associée la probabilité conditionnelle que l'état du système à un instant t appartienne à un sous-ensemble de l'espace des états, sachant qu'à un instant t_0 précédent l'état x du système est connu.
Cette probabilité conditionnelle est appelée "probabilité de transition" et notée

$$P[A,t/x,t_0] = P[X_t(\omega) \in A / X_{t_0}(\omega) = x] \qquad (2.2.2)$$

Un processus de Markov est dit "homogène" si $P[A,t/x,t_0]$ dépend seulement de $t-t_0$. L'homogénéité signifie que le mécanisme qui est à l'origine de la nature aléatoires des observations X_t (comme vent, houle, fluctuations économiques, rugosités de la route pour une voiture qui roule, réaction des organismes à un médicament, ...) n'évolue pas au cours du temps.
Les processus de Markov peuvent être classés en quatre catégories, suivant le caractère de l'espace des états et celui du temps. On dit que le processus est à temps "discret" ou "continu" selon que $I = N$ ou $I = R$. On dit que l'espace des états est continu lorsque chaque X_t prend ses valeurs dans R^n et qu'il est discret lorsque chaque X_t prend ses valeurs dans N.
Lorsque l'espace des états est discret on parle de chaînes de Markov, à temps discret ou continu selon le caractère de I.
Lorsque l'espace des états est continu on parle des processus de Markov, à temps discret ou continu selon le caractère de I. Pour espace des temps I on prendra $I = N$ pour les processus à temps discret et $I = R^+$ pour les processus à temps continu.
Nous limitons l'étude des processus markoviens à celle des chaînes de Markov à temps discret.

2.2 Chaînes de Markov à temps discret.

Une chaîne de Markov à temps discret est une suite de variables aléatoires $X_0, X_1, \ldots, X_n$ à valeurs dans un espace dénombrable Ω, fini ou infini, vérifiant :

$$P[X_m = x_m / X_0 = x_0, X_1 = x_1, \ldots, X_{m-1} = x_{m-1}] = P[X_m = x_m / X_{m-1} = x_{m-1}] \qquad (2.2.3)$$

Pour tout $n \in N$, la loi de $(X_0, X_1, \ldots, X_n)$ est alors donnée par les probabilités

$$p_j(0) = P[X_0 = x_j] \qquad (2.2.4)$$

et la famille des probabilités conditionnelles

$$p_{j,k}(m, m+1) = P[X_{m+1} = x_k / X_m = x_j] \qquad (2.2.5)$$

définies pour $0 \le m \le n-1$ et tous les états x_j, x_k.
En effet, en utilisant la markovianité définie par (2.2.1) m fois nous avons :

$$\begin{aligned}
&P[X_0 = x_0, X_1 = x_1, \ldots, X_{m-1} = x_{m-1}, X_m = x_m] = \\
&= P[X_m = x_m / X_0 = x_0, X_1 = x_1, \ldots, X_{m-1} = x_{m-1}] P[X_0 = x_0, X_1 = x_1, \ldots, X_{m-1} = x_{m-1}] = \\
&= P[X_m = x_m / X_{m-1} = x_{m-1}] P[X_0 = x_0, X_1 = x_1, \ldots, X_{m-1} = x_{m-1}] = \ldots = \\
&= P[X_m = x_m / X_{m-1} = x_{m-1}] P[X_{m-1} = x_{m-1} / X_{m-2} = x_{m-2}] \ldots P[X_1 = x_1 / X_0 = x_0] P[X_0 = x_0]
\end{aligned}$$

Notons alors que la donnée de la loi de la première variable X_0 et la donnée de la suite des probabilités conditionnelles (2.2.5) sont suffisant pour définir, en vertu du théorème de

Kolmogorov, une loi de probabilité unique, qui est la loi de la chaîne de Markov $X = (X_0, X_1, ..., X_n, ...)$, sur l'ensemble Ω^N.
Considérons, de façon plus générale

$$p_{j,k}(m,n) = P[X_n = x_k / X_m = x_j] \tag{2.2.6}$$

pour $0 \le m \le n$. Les probabilités conditionnelles ci-dessus sont dites "probabilités de transition" et les matrices

$$P(m,n) = \begin{bmatrix} p_{0,0}(m,n) & p_{0,1}(m,n) & \dots & p_{0,k}(m,n) & \dots \\ p_{1,0}(m,n) & p_{1,1}(m,n) & \dots & p_{1,k}(m,n) & \dots \\ \dots & \dots & \dots & \dots & \dots \\ p_{j,0}(m,n) & p_{j,1}(m,n) & \dots & p_{j,k}(m,n) & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{2.2.7}$$

matrices de transition, ou matrices stochastiques. Notons que ces matrices sont infinies pour le nombre infini (dénombrable) d'états; pour le nombre d'états fini (k par exemple) les matrices de transition sont finies :

$$P(m,n) = \begin{bmatrix} p_{0,0}(m,n) & p_{0,1}(m,n) & \dots & p_{0,k}(m,n) \\ p_{1,0}(m,n) & p_{1,1}(m,n) & \dots & p_{1,k}(m,n) \\ \dots & \dots & \dots & \dots \\ p_{k,0}(m,n) & p_{k,1}(m,n) & \dots & p_{k,k}(m,n) \end{bmatrix} \tag{2.2.8}$$

De telles matrices vérifient:

$$\begin{aligned} &\text{(i)} \quad p_{j,k}(m,n) \ge 0 \ \text{ pour tout couple } (j,k) \\ &\text{(ii)} \quad \sum_k p_{j,k}(m,n) = 1 \qquad \text{pour chaque } j \end{aligned} \tag{2.2.9}$$

On définit le produit de deux matrices infinies A et B comme la matrice C dont l'élément générique est

$$c_{i,k} = \sum_j a_{i,j} b_{j,k} \tag{2.2.10}$$

On peut alors montrer l'équation suivante, dite de Kolmogorov-Chapman:

$$\text{(iii)} \quad P(m,n) = P(m,u)P(u,n) \qquad \text{pour } 0 \le m < u < n \tag{2.2.11}$$

Réciproquement, étant donnée une famille de matrices $P(m,n)$ satisfaisant (i), (ii), (iii) il existe au moins une chaîne de Markov (X_n) dont les probabilités de transition sont les éléments des matrices $P(m,n)$.

Notons que pour connaître toutes les matrices $P(m,n)$ il suffit de connaître, en vertu de l'équation de Kolmogorov-Chapman, les matrices $P(m,n)$ pour $n-m=1$.

Définition 2.1

Une chaîne de Markov (X_n) est dite "homogène" si $p_{j,k}(n,m)$ ne dépend que de $n-m$. On pose alors

$$p_{j,k}(n) = P[X_{t+n} = x_k / X_t = x_j] \qquad (2.2.12)$$

$p_{j,k}(n)$, indépendante de t, est appelée probabilité de transition en n étapes. Une matrice de transition en une étape est simplement appelée, lorsqu'il n'y a pas de confusion possible, "matrice de transition".
Notons

$$P = P(1) = [p_{j,k}(1)] \qquad (2.2.13)$$

la matrice de transition (en une étape). En vertu de l'équation de Kolmogorov-Chapman la matrice de transition en n étapes $P(n)$ s'écrit

$$P(n) = P^n \qquad (2.2.14)$$

Par ailleurs, la loi $p(n)$ de X_n s'écrit

$$p(n) = p(0)P^n \qquad (2.2.15)$$

Finalement, dans le cas d'une chaîne de Markov homogène, la loi de tout $(X_0, X_1, ..., X_n)$ est donnée par $p(o)$ (loi de X_0) et la matrice de transition en une étape P.

Exemple 2.1

Reprenons l'exemple météorologique du paragraphe précèdent. Si nous sommes au mois de juillet, nous pouvons imaginer, à titre d'exemple, qu'il y ait trois chances sur quatre qu'il fasse beau demain s'il fait beau aujourd'hui, et une chance sur deux s'il ne fait pas beau aujourd'hui. La matrice de transition "demain sachant aujourd'hui" de notre chaîne est alors

$$P = \begin{bmatrix} 0.75 & 0.25 \\ 0.50 & 0.50 \end{bmatrix}$$

Notons que la matrice de transition "après-demain sachant aujourd'hui" est

$$P^2 = \begin{bmatrix} 0.75 & 0.25 \\ 0.50 & 0.50 \end{bmatrix} \begin{bmatrix} 0.75 & 0.25 \\ 0.50 & 0.50 \end{bmatrix} = \begin{bmatrix} \frac{11}{16} & \frac{5}{16} \\ \frac{5}{8} & \frac{3}{8} \end{bmatrix}$$

et donc la probabilité qu'il fasse beau après-demain sachant qu'il fait beau aujourd'hui est de $\frac{11}{16}$, et la probabilité qu'il fasse beau après-demain sachant qu'il ne fait pas beau est de $\frac{5}{8}$.
Nous pouvons supposer la chaîne homogène sur le mois de juillet, il serait cependant trop fort de la supposer homogène sur toute l'année.

2.3. Chaînes de Markov homogènes.

2.3.1 Classes transitoires et classes finales

Une chaîne de Markov est dite finie lorsque l'ensemble de ses états, que l'on notera S, est fini. Notons s le cardinal de S. La matrice de transition en une étape P est alors finie:

$$P = [p_{j,k}]_{\substack{1\le j\le s \\ 1\le k\le s}} \qquad (2.3.1)$$

.et $p(o)$ (loi de X_0) est une probabilité sur S.

Remarque 3.1

Le nombre limité de paramètres définissant les lois des $(X_0, X_1, ..., X_n)$, quel que soit n, constitue le principal avantage de cette modélisation. Si s est le cardinal de S ce nombre est de $s+s^2$, ce qui autorise généralement les traitements informatiques. Notons que sans l'hypothèse de markovianité la loi de $(X_0, X_1, ..., X_n)$ est une probabilité sur S^n, donnée donc par s^n paramètres. Dans les applications n peut être grand : en traitement d'images il est couramment de l'ordre de 256×256.

Remarque 3. 2

Notons que dans une chaîne de Markov toute les variables aléatoires X_i sont dépendantes. Ainsi, quel que soit le couple (m,n) des instants des observations les variables X_m, X_n sont dépendantes et donc leur loi conjointe n'est pas le produit de leurs lois marginales. Par contre si on a observé $X_k = x_k$ avec $m < k < n$ les variables X_m, X_n, dont la loi devient conditionnelle à $X_k = x_k$, sont indépendantes. Ainsi la loi du couple (X_m, X_n) conditionnelle à $X_k = x_k$ est le produit des lois de X_m, X_n respectivement conditionnelles à $X_k = x_k$.

On se propose d'étudier l'évolution au cours du temps d'une chaîne de Markov homogène, à nombre fini d'états, et à temps discret.

On pose $p_{j,j}(0) = 1$ pour tout j (la probabilité d'être dans l'état j à un instant donné sachant qu'à ce même instant on est dans l'état j est un) et $p_{i,j}(0) = 0$ si $i \neq j$.

Il est intéressant de classer les états de la chaîne suivant qu'il est, ou non, possible de passer d'un état donné à un autre état donné.

Définition 3.1

- Un état k est dit conséquent d'un état j, s'il existe un entier $N \geq 0$ tel que $p_{j,k}(N) > 0$; autrement dit si la probabilité de passer de l'état j à l'état k en N étapes est positive. On notera

$j \to k$. C'est une relation d'ordre faible (réflexive et transitive, la transitivité découlant de l'équation de Kolmogorov-Chapman).
- Deux états j et k sont dits communicants si j est conséquent de k et k est conséquent de j . On le notera $j \leftrightarrow k$.

La relation "$\leftrightarrow$" est une relation d'équivalence sur l'espace des états, on peut donc décomposer cet espace en classes d'équivalence. Soient $C \neq C'$ deux telles classes. Si $x \in C, x' \in C'$ alors trois cas peuvent se produire:

1. $x \to x'$ (alors x ne peut être conséquent de x'). On a $\forall i \in C,\ \forall i' \in C'\ \ i \to i'$ et on écrit $C \leq C'$
2. $x' \to x$ (alors x' ne peut être conséquent de x). On a $\forall i \in C \leq C',\ \forall i' \in C'\ \ i' \to i$ et on écrit $C' \leq C$
3. x et x' ne sont pas conséquents l'un de l'autre. Alors C et C' ne sont pas comparables.

On définit ainsi une relation d'ordre partiel (deux classes différentes ne sont pas nécessairement comparable) sur l'ensemble de classes.

Définition 3.2

Les éléments maximaux de la relation d'ordre "$\leq$" (dont le nombre peut être supérieur à un) sont appelés les classes finales ou ergodiques et les états de ces classes sont dits ergodiques ou essentiels.
Les autres états sont dits transitoires ou de passage.
Si une classe ergodique ne contient qu'un seul élément k, cet élément est dit "absorbant" et on a $p_{k,k} = 1$.
Si l'espace d'états d'une chaîne de Markov ne comprend qu'une seule classe d'équivalence la chaîne est dite irréductible.

Remarque 3.3

L'appellation "classes finales" sous-entend que le mobile finira fatalement dans une de telles classes. En effet, cette propriété peut être démontrée. De même, on peut montrer que le mobile quitte les classes "transitoires" (ou "de passage"), pour ne plus y revenir, avec la probabilité 1.

Exemple 3.1

Considérons les matrices de transition

$$A = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.8 & 0.1 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.8 & 0.2 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \quad C = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.8 & 0.2 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Notons $E = \{e_1, e_2, e_3\}$. On constate que la chaîne correspondant à A est irréductible (tous les éléments communiquent, dons une seule classe d'équivalence). Dans le cas de la chaîne correspondant à B il y a deux classes d'équivalence $C_1 = \{e_1, e_2\}$, $C_2 = \{e_3\}$ sans qu'elles soient

comparables. Dans le cas C on a toujours deux classes d'équivalence $C_1 = \{e_1, e_2\}$, $C_2 = \{e_3\}$ avec $C_1 \leq C_2$. La classe C_2 est donc ergodique.

2.4 Stationnarité et comportement asymptotique

Définition 2.4.3

Une chaîne $X = (X_n)_{n \in N}$ homogène est dite stationnaire si $p(n)$, loi de X_n, ne dépend pas de n

Étant donné que $p(n) = p(0)P^n$, où $p(0)$ est la loi de X_0 (loi initiale), la chaîne est ainsi stationnaire si et seulement si $p(0)$ est un vecteur propre à gauche de la matrice de transition P (un tel vecteur est dit une distribution stationnaire attachée à P).
Lorsque une chaîne n'est pas stationnaire, il est intéressant de chercher les conditions sur P sous lesquelles la distribution $p(n)$ admet une limite Π lorsque n tend vers l'infini. Si Π existe on doit avoir

$$\Pi = \lim_{n \to \infty} p(n) = \lim_{n \to \infty} p(n-1)P = \Pi P \qquad (2.4.2)$$

Donc Π vecteur ligne doit être un vecteur propre à gauche de la matrice P. Ainsi la limite éventuelle se trouve dans l'ensemble des vecteurs propres à gauche de la matrice P. Par ailleurs, des lois initiales $p(0)$ différentes peuvent donner des limites différentes. En effet, si P^n a une limite P^* lorsque n tend vers l'infini on a

$$\lim_{n \to \infty} p(n) = \lim_{n \to \infty} p(0)P^n = p(0)P^* \qquad (2.4.3)$$

Le cas intéressant, pour les applications, est celui ou cette limite existe et ne dépend pas de $p(0)$.

Définition 2.4.4

Lorsque $\Pi = \lim_{n \to \infty} p(n)$ existe et ne dépend pas de $p(0)$ la chaîne de Markov est dite régulière.

Étant donné la définition, l'étude de la régularité d'une chaîne de Markov peut se faire à partir de l'étude de la convergence de la suite des puissances d'une matrice stochastique. On peut en effet montrer que pour qu'une chaîne de Markov finie, homogène, à temps discret, soit régulière, il faut et il suffit que sa matrice de transition P n'admette que la valeur propre simple 1 comme valeur propre de module 1, toutes les autres valeurs propres étant de module strictement inférieur à 1. Cette propriété montre que dans une chaîne de Markov régulière le passé est "oublié à la vitesse exponentielle" (voir la Proposition 2.4.2. ci-après). Ce résultat est intéressant au plan intuitif; cependant, on l'utilise rarement dans la pratique pour démontrer la régularité d'une chaîne. Nous énonçons plus loin deux autres résultats, plus pratiques pour montrer la régularité d'une chaîne.
Intuitivement, lorsqu'un phénomène est modélisé par une chaîne de Markov, la régularité permet de s'affranchir de l'étude de la loi initiale.

Définition 2.4.5

Un état j est périodique de période $t>1$ si un retour à l'état j ne peut intervenir qu'aux étapes $t, 2t, 3t, \ldots$. Alors $p_{i,j}(n)=0$ si n n'est pas divisible par t. Si $t=1$ l'état est dit apériodique.
On a le résultat suivant

Proposition 2.4.1

1. Tous les états d'une classe ergodique ont la même période.
2. Une chaîne de Markov est régulière si et seulement si elle ne contient qu'une classe ergodique et cette classe est apériodique.

La deuxième partie de la proposition ci-dessus permet une reconnaissance rapide, à partir du graphe sagittal, de la régularité d'une chaîne de Markov. En effet, on peut procéder de la façon suivante :

(i) on reconnaît, à partir du graphe sagittal, l'existence et l'unicité d'une classe ergodique. Notons que cette condition est vérifiée pour une chaîne irréductible (une seule classe d'équivalence).

(ii) on montre l'apériodicité de la classe ergodique. Notons qu'il suffit pour cela de trouver, pour un élément donné de la classe, deux chemins qui partent de l'élément et y retournent en n_1 et n_2 coups respectivement, avec n_1, n_2 premiers entre eux.

Ainsi, sous certaines conditions, les lois marginales tendent vers une loi donnée. On peut alors se poser le problème de la vitesse de convergence. La proposition suivante (théorème ergodique) précise cette vitesse.

Définition 2.4.6

On appelle coefficient d'ergodicité

$$k(n_0)=1-\sup_{i,j}\sum_m \left|p_{im}(n_0)-p_{jm}(n_0)\right| \qquad (2.4.4)$$

Proposition 2.4.2

Si pour un certain n_0 on a $k(n_0)\rangle 0$ les limites $\lim_{n\to+\infty} p_j(n)=\Pi_j$ existent et forment une distribution stationnaire. De plus, on a pour toute distribution initiale $p(0)$:

$$\sup_{p(0)}\left|p_j(n)-\Pi_j\right| \le Ce^{-Dn} \qquad (2.4.5)$$

avec

$$C=\frac{1}{1-k(n_0)} \qquad (2.4.6)$$

$$D=\frac{1}{n_0}Ln\frac{1}{1-k(n_0)} \qquad (2.4.7)$$

Nous constatons ainsi une convergence rapide, de type exponentielle, vers Π.

Exemple 2.4.2

Reprenons l'exemple météorologique des paragraphes précédents. La matrice de transition de la chaîne étant $P = \begin{bmatrix} 0.75 & 0.25 \\ 0.50 & 0.50 \end{bmatrix}$, on montre aisément que la chaîne est régulière. Étant donné que $\Pi = (\frac{2}{3}, \frac{1}{3})$ est un vecteur "probabilité" propre à gauche de P, c'est aussi la distribution limite (en effet, cette limite existe et est indépendante de l'initialisation, or, lorsque l'on initialise avec Π, on obtient Π). On peut ainsi affirmer, "à la limite" donc "approximativement", qu'il fera beau deux jours sur trois (notons que cela n'apparaît pas immédiatement à partir de la matrice P). Pour $n_0 = 1$, le coefficient de d'ergodicité est $k(1) = 0.5$, ce qui donne $C = 2$ et $D = Log2$. Finalement

$$\sup_{p_i(0)} \left| p_j(n) - \Pi_j \right| \leq \left(\frac{1}{2} \right)^{n-1} \tag{2.4.8}$$

Ainsi, si nous ignorons le temps qu'il a fait les cinq derniers jours, le temps qu'il a fait il y a six jours ne peut modifier que modestement la probabilité $\Pi = (\frac{2}{3}, \frac{1}{3})$: l'écart entre cette probabilité et la probabilité conditionnelle au temps qu'il a fait il y a six jours est inférieur à $\frac{1}{2^5} \approx 0.03$.

2.5 Autre classification

Supposons que le système soit initialement dans un état j donné, et soit $f_j(n)$ la probabilité que le premier retour à l'état j ait lieu à la $n^{ième}$ étape. On a la relation:

$$p_{j,j}(n) = \sum_{m=1}^{n} f_j(m)\, p_{j,j}(n-m) \tag{2.5.1}$$

qui traduit le fait que pour être dans l'état j à la $n^{ième}$ étape, sachant que l'on part de l'état j, il faut et il suffit que l'on y soit revenu une première fois à un temps $m \leq n$ et que l'on revienne à cet état après $n-m$ étapes. Cette relation permet de calculer de proche en proche les $f_j(n)$ connaissant les $p_{i,j}(n)$.
La probabilité que le système retourne au moins une fois à l'état j est alors:

$$f_j = \sum_{n=1}^{\infty} f_j(n) \tag{2.5.2}$$

Définition 2.5.1

Si $f_j = 1$ l'état j est dit récurrent. Si $f_j < 1$ l'état j est dit transitoire.

Remarque 2.5.1

L'appellation "récurrent" sous-entend que l'état sera visité plus d'une fois. On peut en effet montrer que si la probabilité de retour à un état vaut 1 alors la probabilité de retour une infinité de fois à cet état vaut également 1. Autrement dit, la certitude d'un retour implique la certitude d'une infinité de retours. Intuitivement, cette propriété est due à l'absence de mémoire : à l'instant du premier retour le passé est oublié et le deuxième retour, qui devient le "premier", est certain.

De manière analogue, "transitoire" sous-entend que les retours ne dureront pas. On peut en effet montrer qu'un état transitoire est visité, avec la probabilité 1, un nombre fini de fois.

Proposition 2.5.1

1. Avec la probabilité 1, un état récurrent est visité une infinité de fois.
2. Avec la probabilité 1, un état transitoire est visité un nombre fini de fois.

Démonstration (facultatif)

1. Soit e un état récurrent et notons E l'événement "e est visité une infinité de fois". Par ailleurs, notons E_k l'événement "e est visité au moins k fois" et E_{kn} l'événement "le k ième retour à e s'effectue à l'instant n". Nous avons $E_{k+1} \subset E_k$ et $E = \bigcap_{k=1}^{+\infty} E_k$. Par conséquent , en appliquant le théorème de convergence monotone de Lebesgue à la suite des fonctions indicatrices des ensembles E_k on obtient : $P[E] = \lim_{k \to +\infty} P[E_k]$. Il suffit donc de montrer que $P[E_k] = 1$ pour tout k. Montrons le par récurrence. $P[E_1] = 1$ par définition d'un état récurrent, supposons $P[E_k] = 1$. Nous avons

$$P[E_{k+1}] = P[E_{k+1} \cap E_k] = P[E_{k+1} \cap (\bigcup_{n=k+1}^{+\infty} E_{kn})] = P[\bigcup_{n=k+1}^{+\infty} (E_{k+1} \cap E_{kn})] =$$
$$\sum_{n=k+1}^{+\infty} P[E_{k+1} \cap E_{kn}] = \sum_{n=k+1}^{+\infty} P[E_{k+1} / E_{kn}] P[E_{kn}] = \sum_{n=k+1}^{+\infty} P[E_{kn}] = P[E_k] = 1 \qquad (2.5.3)$$

car la probabilité $P[E_{k+1} / E_{kn}]$ est la probabilité de retour en e sachant que l'on s'y trouve, donc 1.

2. Dans le cas e transitoire les calculs restent valables à cette différence près que la probabilité $P[E_{k+1} / E_{kn}]$ vaut $\varepsilon \langle 1$. On obtient :

$$P[E_{k+1}] = \sum_{n=k+1}^{+\infty} P[E_{kn} / E_k] P[E_k] = P[E_k] \sum_{n=k+1}^{+\infty} P[E_{kn} / E_k] = P[E_k] \varepsilon \qquad (2.5.4)$$

Donc, par récurrence $P[E_k] = \varepsilon^k$. Il en résulte que $P[E] = \lim_{k \to +\infty} P[E_k] = \lim_{k \to +\infty} \varepsilon^k = 0$. La probabilité de l'infinité de retours étant nulle l'état e est visité un nombre fini de fois avec la probabilité 1.

Enfin, les approches par "classes d'équivalence" et les "probabilités de retour" sont liées par les résultats suivants :

Proposition 2.5.2

Supposons qu'il y a une seule classe ergodique.
Les états récurrents sont les états ergodiques définis par la relation d'ordre. Les états transitoires sont les états transitoires définis par la relation d'ordre.

Exercices

Exercice 1.

Étudier l'irréductibilité, la périodicité, et la régularité de la chaîne de Markov de matrice de transition

$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{7}{8} & \frac{1}{8} & 0 \end{bmatrix}$$

Exercice 2

On considère l'exemple des lancements d'une pièce de monnaie. Montrer que la probabilité de l'événement " 0 apparaît un nombre fini de fois" est nulle.

Exercice 3.

Un processus $X = (X_1, ..., X_n, ...)$ est stationnaire à l'ordre k si les lois de $(X_n, ..., X_{n+k-1})$ sont indépendantes de n. Il est fortement stationnaire s'il est stationnaire à l'ordre k pour tout k.

1. Montrer qu'un processus de Markov homogène est stationnaire à l'ordre 2 ssi il stationnaire à l'ordre 1 ;
2. Montrer qu'un processus de Markov stationnaire à l'ordre 2 est fortement stationnaire.

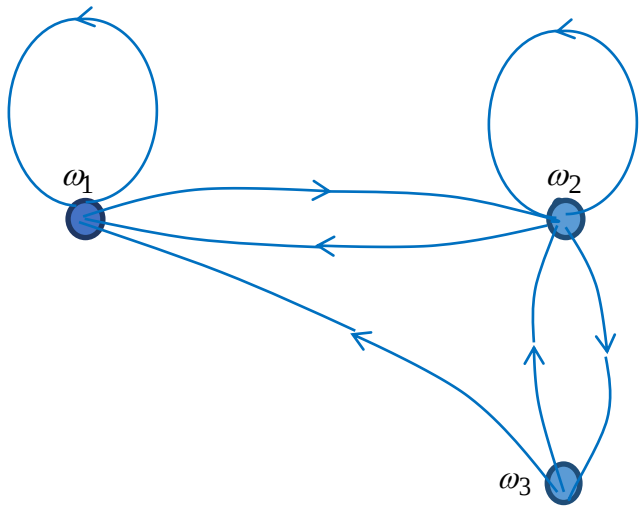
Solutions des exercices

Exercice 1.

Pour la chaîne de Markov de matrice de transition

$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{7}{8} & \frac{1}{8} & 0 \end{bmatrix}$$

Le diagramme sagittal est



On constate que la chaîne est irréductible (une seule classe d'équivalence).
La classe étant unique, elle est ergodique. Tous ses éléments ont donc la même périodicité. On constate sur e1 par exemple que la classe est apériodique.
La chaîne est donc régulière

Exercice 2

Il suffit de montrer que la probabilité de l'évènement complémentaire B^c "0 apparaît un nombre infini de fois" vaut 1. Selon la loi des grands nombres $\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow[n\to\infty]{} E[X_1] = p$, la convergence ayant lieu « presque surement », ce qui signifie que l'ensemble des trajectoires $x = (x_1, x_2, ..., x_n, ...) \in \{0,1\}^{N^*}$ où cela n'est pas vrai, qui sera noté F, est de probabilité nulle. Or l'événement B "0 apparaît un nombre fini de fois"

est inclus dans F (en effet, pour $x = (x_1, x_2, ..., x_n, ...) \in B$ on a $\frac{x_1 + x_2 + ... + x_n}{n} \xrightarrow[n\to\infty]{} 1 \neq p$), d'où le résultat .

Exercice 3.

Un processus $X = (X_1, ..., X_n, ...)$ est stationnaire à l'ordre k si les lois de $(X_n, ..., X_{n+k-1})$ sont indépendantes de n. Il est fortement stationnaire s'il est stationnaire à l'ordre k pour tout k.

1. $p(x_n, x_{n+1}) = p(x_n)p(x_{n+1}|x_n) = p(x_n)p(x_2|x_1)$ car $X = (X_1, ..., X_n, ...)$ homogène, donc $p(x_n, x_{n+1}) = p(x_n)p(x_2|x_1) = p(x_1, x_2) = p(x_1)p(x_2|x_1)$ ssi $p(x_n) = p(x_1)$;

2. $p(x_n, ..., x_{n+k-1}) = p(x_n)p(x_{n+1}|x_n)... p(x_{n+k-1}|x_{n+k-2}) =$
$= p(x_1)p(x_2|x_1)... p(x_{n-1}|x_{n-2}) = p(x_1, ..., x_k)$