N$^{\text{O}}$ d'ordre :

THÈSE

présentée à

L'ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

PAR **Arnaud de LA FORTELLE**

POUR OBTENIR LE TITRE DE

DOCTEUR

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES

Contribution à la théorie des grandes déviations et applications

---

Directeur de thèse : **Guy FAYOLLE**

Soutenue le 16 novembre 2000 devant le jury formé de :

MM.	**N. BOULEAU**	Président
	F. COMETS	Rapporteur
	J. LEWIS	Rapporteur
	G. FAYOLLE	Examinateur
	V. MALYCHEV	Examinateur
	G. RUGET	Examinateur

À Nastia

Le calcul

Le calcul, c'est l'abîme. Ah ! tu sors de ta sphère,
Eh bien, tu seras seul. Homme, tâche de faire
Entrer dans l'infini quelque être que ce soit
De ceux que ta main touche et que ton regard voit ;
Nul ne le peut. La vie expire en perdant terre.
Chaque être a son milieu ; hors du bois la panthère
Meurt, et l'on voit tomber, sans essor, sans éclair,
Hors du feu l'étincelle et l'oiseau hors de l'air ;
Nulle forme ne vit loin du réel traînée ;
La vision terrestre à la terre est bornée ;
Le nuage lui-même, errant, volant, planant,
Allant d'un continent à l'autre continent,
S'il voyait l'absolu, serait pris de vertige ;
Sortir de l'horizon n'est permis qu'au prodige ;
L'homme le peut, étant le monstre en qui s'unit
Le miasme du nadir au rayon du zénith ;
Entre donc dans l'abstrait, dans l'obscur, dans l'énorme,
Renonce à la couleur et renonce à la forme ;
Soit ; mais pour soulever le voile, le linceul,
La robe de la pâle Isis, te voilà seul.
Tout est noir. C'est en vain que ta voix crie et nomme.
La nature, ce chien qui, fidèle, suit l'homme,
S'est arrêtée au seuil du gouffre avec effroi.

Regarde. La science exacte est devant toi...

Victor Hugo

Remerciements

Ce travail aurait été bien difficile — s'il eût jamais été achevé — sans l'accueil, l'aide et le conseils qui m'ont constamment été prodigués. Je voudrais donc remercier tous ceux qui ont éclairé ces années, pour ces attentions qui ont marqué mon esprit et mon cœur.

Ma pensée se tourne d'abord vers Guy Fayolle. Sa rencontre au D.E.A. de probabilités de Paris VI est de celles qui marquent. Sa confiance m'a laissé libre dans des recherches tâtonnantes ; les doutes menaient alors droit vers son bureau où sa disponibilité fut mise à rude épreuve, sans jamais vaciller ; ses conseils ont aplani bien des difficultés, ouvert de nouvelles perspectives et sa bonne humeur dissipait les nuages.

L'atmosphère décontractée de l'équipe est largement redevable à Jean-Marc, Franck, Christine, Vadim et Roudolf, avec lesquels les échanges ont été riches et agréables. Merci en particulier à Jean-Marc pour sa compétence en matière d'ordinateur et de logiciels — et pour sa psychologie des utilisateurs ; merci à Franck pour sa collaboration fructueuse ; merci à Danny pour son dévouement efficace.

Je suis reconnaissant à Nicolas Bouleau pour l'orientation qu'il a donnée à mes études et qui m'a conduit ici. Sa participation au jury représente de plus l'École nationale des ponts et chaussées et le ministère de l'Équipement et des transports qui m'ont soutenu, notamment financièrement, et qui continuent de manifester de l'intérêt pour ce travail de recherche.

Les rapporteurs, MM. Francis Comets et John Lewis ont bien voulu se charger de la lourde tâche de rapporteur. Leurs remarques constructives ont permis d'améliorer ce document. Qu'ils en soient remerciés. Merci également à MM. Vadim Malychev et Gabriel Ruget pour leur participation au jury.

Finalement, mes remerciements vont à tous ceux qui — à l'INRIA ou en dehors — ont contribué à enrichir ces années ; une pensée particulière s'adresse aux membres de l'équipe de tennis, dont la rencontre a débouché sur d'autres échanges que celui des balles jaunes.

Merci à tous d'avoir ensoleillé ces heures : la chaleur de votre rencontre ne s'éteindra jamais dans mon cœur.

Arnaud de La Fortelle

Avant-propos

« Le silence éternel de ces espaces infinis m'effraie. »
Blaise Pascal

« Ils traversent ainsi le noir illimité, ce frère du silence éternel. »
Charles Baudelaire

Imaginons les aveugles baudelairiens se déplaçant de façon erratique : voilà une illustration classique de marche aléatoire. Les mathématiciens se sont fixé la gageure de cerner le hasard, par exemple on peut délimiter un couloir que les aveugles ne quitteront presque jamais ; il encadre la trajectoire moyenne. La présente thèse se propose d'étudier ces situations improbables, quand l'évolution déraille vers l'espace infini qui est à l'extérieur de ce canal. Ces ténébreuses considérations, et les formules cabalistiques par lesquelles elles s'expriment, paraissent aussi effrayantes que mystérieuses, et il semble que mieux vaudrait les laisser au silence éternel. Mais enfin, puisqu'il faut être indulgent avec les érudits passionnés, quel sens ose-t-on donner à ces occultes hiéroglyphes ?

Je prends ici le parti d'expliquer en quelques lignes en quoi ce travail de thèse est digne d'intérêt, car il est fréquent de voir le regard se troubler chez des interlocuteurs peu prévenus des choses mathématiques, alors qu'un fil d'Ariane bien perceptible guide toute la réflexion.

Pour atteindre un état exceptionnel, un système emprunte la trajectoire de moindre information. Voilà résumée la thèse en une phrase, dans une formulation rappelant furieusement le principe de moindre action, dont on sait combien il est riche de conséquences. Cependant, pour bien concevoir ce qui s'énonce clairement, il faut définir les notions, comme celle d'*état exceptionnel*, de *système* ou d'*information*.

Ici s'explique le nom de la théorie : une grande déviation, c'est un écart à la normale qui mène à un état exceptionnel. Un système est décrit par l'évolution, selon une *dynamique*, d'un état représenté par des *coordonnées*.

Cette dynamique n'est pas déterministe, mais obéit aux règles des *probabilités*, qui lui dictent ainsi un comportement moyen. L'exception est ainsi définie comme un événement rare. Enfin, les distributions de probabilités permettent de quantifier l'information : mieux un évènement est cerné (et donc plus sa fréquence est faible), plus il fournit d'information.

Déjà le discours cède la place au formalisme mathématique, qui délimite précisément les objets et les opérations licites qui leurs sont associées même si parfois leur sens paraît absent. La différence avec une langue littéraire est que, si les déductions enrichissent dans les deux cas la signification originelle, les démonstrations formalisées n'offrent, elles, pas la moindre place au doute. En outre, la quantification qu'impose au départ la modélisation est conservée au fil des raisonnements si bien que non seulement le sens existe et est maintenu, mais il est encore précisé et concrétisé par une *méthode* qui dévoile de bonnes manières de *mesurer* les phénomènes.

Ces réflexions marquent l'organisation de ce mémoire. La première partie définit, du plus simple vers le plus complexe, les notions, leur représentation abstraite et leur signification. J'ai essayé de présenter en quoi elles me semblent naturelles. Les déductions qui découlent de ces structures ne font que reproduire le schéma de pensée de ce préambule, certes parfois dissimulé sous les notations. L'induction ensuite nous élève vers des sphères de plus haute abstraction, d'où l'on perçoit quelquefois mieux l'essence des phénomènes. Quoi qu'il arrive, il est apparu que certains calculs menaient à des conclusions surprenantes, qui nous font découvrir, ou préciser, les comportements des systèmes analysés. Cette première partie s'attache progressivement à ces explorations.

À coté des concepts et outils classiques — fonction des cumulants, transformée de Legendre, sous-additivité ou principe de contraction — nous développons particulièrement la technique du changement de mesure, les opérations sur les principes faibles et les techniques afférentes. Ces contributions personnelles sont réparties au fil des chapitres ; il s'agit des fonctions de Ruelle-Lanford (section 3.1.3), du changement de mesure tel qu'il est décrit à la section 3.2.1, la contraction pour les principes *faibles* (section 3.3.1) ; les chapitres 4, 5 et 6 sont surtout constitués d'apports personnels, notamment l'adaptation de ces méthodes au temps continu (section 4.2), à des processus non-markoviens (section 6.1.1) ou à des calculs plus précis (section 6.1.2).

De façon plus prospective, il semble que certains domaines soient à portée de main, bien que les démonstrations n'y soient pas rigoureuses : par exemple le lien fort entre fonctions harmoniques, grandes déviations et probabilités stationnaires. Ces considérations sont exposées à la section 4.4.3 et au chapitre 5. Enfin, les difficultés liées aux espaces d'états non-dénombrables sont mentionnées à la section 6.1.3.

La seconde partie est composée de trois articles soumis à publications. Deux sont essentiellement abstraits, mais les concepts, développés dans une grande généralité, n'y sont pas confinés. L'étude d'un modèle en troisième article, valable aussi bien en transport qu'en informatique ou en transmission de données, le prouve. On devrait, par la méthode proposée, parvenir à quantifier le mode de saturation de ces systèmes, et on en imagine sans peine les avantages pratiques, en conception ou pour le contrôle. Et la démarche est loin de se restreindre à cette seule application.

La structure linéaire imposée par l'exposition successive des notions reflète mal la densité des liens qui les unissent. De nombreuses remarques articulent le texte, soit en devançant des concepts d'un niveau plus abstrait, soit en offrant une réflexion personnelle sur des développements possibles. J'espère que cette tentative de m'exprimer clairement ne sera pas ruinée par le langage des mathématiques, car l'unité et la richesse de sens de la théorie des grandes déviations est vraiment un plaisir esthétique que j'espère ici partager.

Arnaud de La Fortelle

Table des matières

Première partie

Aperçu théorique

Chapitre 1

Introduction aux grandes déviations

Cette partie n'est qu'un aperçu de la théorie des grandes déviations. À partir des modèles les plus simples, les outils les plus couramment utilisés sont présentés, de façon à pouvoir plus aisément situer le travail de thèse. Le fil conducteur de ces pages est de montrer les point forts ou faibles de chaque méthode qui pourrait être employée dans les cas qui nous intéressent (réseaux et modèles *discrets* en général), aussi l'exposé n'est-il pas exhaustif. En particulier des branches entières de la théorie ne sont ici pas abordées, par exemple la finesse des topologies, car ces problèmes ne se posent pas dans le cadre que nous nous sommes fixé, mais de même les processus continus sont omis, dont le mouvement brownien, en dépit de leur immense intérêt, y compris dans le cadre des réseaux.

1.1 La notion de grande déviation

On considère une suite $\{X_i,\ i \geq 1\}$ de variables aléatoires prenant leurs valeurs sur un espace E. L'usage s'est établi de diviser les principes de grandes déviations (PGD) en trois niveaux (cf. [61]), qui correspondent à des propriétés statistiques de plus en plus fines :

– Le niveau 1, qui est l'étude des déviations de la *moyenne* ;

– le niveau 2, celui des déviations des *fréquences d'occurrence* ;

– le niveau 3, celui des déviations des *statistiques du processus.*

La hiérarchie des niveaux reflète l'ordre d'apparition des problèmes, mais n'épuise pas la variété des situations où un tel principe est étudié. Il est cependant significatif que les méthodes employées dans chaque cas soient similaires et parfois identiques. En effet tous les résultats peuvent se déduire d'un niveau supérieur, qui représente une description plus détaillée (c'est le principe de contraction), ce qui paraît naturel d'après l'adage, « qui peut le plus, peut le moins ». Inversement, on peut ramener les trois niveaux au seul premier cas en augmentant la taille de l'espace sur lequel les variables X_i sont définies. On voit donc que cette hiérarchie n'est pas aussi nette qu'on pourrait le supposer.

La présentation est ici ascendante dans la complexité. Les définitions et résultats sont établis pour chaque niveau, et on illustrera cette démarche qui consiste à appliquer un théorème simple sur un espace compliqué, qui donne souvent de meilleurs résultats qu'un théorème compliqué appliqué à un espace simple.

On suppose dans la suite, sauf mention explicite du contraire, que E est un *espace métrique complet séparable*, c'est-à-dire un espace polonais, dont la distance est notée d. Par exemple la droite réelle $\mathbb{R}$ est un espace polonais, ainsi que $\mathbb{R}^n$, ou encore tout ensemble dénombrable muni d'une distance (si nous pensons aux réseaux). Une des bonnes raisons de les choisir, c'est que la convergence d'une suite de mesures s'exprime au moyen des compacts. En outre, si l'espace E est polonais, alors l'ensemble des mesures de probabilité sur E l'est aussi. Or c'est ce type de grossissement qui permet de passer d'un niveau à l'autre. On conserve alors les bonnes propriétés des espaces polonais.

1.2 La moyenne empirique

Puisque la notion de moyenne nécessite que les variables $\{X_i,\ i \geq 1\}$ puissent être additionnées et multipliées par un scalaire, E est un espace vectoriel, à tout le moins, un sous-ensemble convexe d'un espace vectoriel.

Définition 1.1 *La moyenne empirique de la suite $\{X_i,\ i \geq 1\}$ est*

$$\hat{S}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{1.1}$$

Supposons brièvement que les variables $\{X_i,\ i \geq 1\}$ soient indépendantes et identiquement distribuées, avec une loi μ qui ait une certaine régularité. D'après la loi des grands nombres, la moyenne empirique converge vers l'espérance de μ, c'est-à-dire que

$$\lim_{n\to\infty} \hat{S}_n = \mathbb{E}\,[X_1] = \int_E x\mu(\,\mathrm{d}x), \tag{1.2}$$

au sens de la convergence presque sûre. La théorie des grandes déviations propose des moyens de calculer, sur une échelle exponentielle, la probabilité de converger vers d'autres valeurs. Plus précisément,

Définition 1.2 (Principe de grandes déviations) *On dit que $\hat{S}_n$ vérifie un* PGD *s'il existe une fonction $I : E \mapsto \mathbb{R}_+$ telle que*

(P1) I est semi-continue inférieurement ;

(P2) pour tout ouvert $O \subset E$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in O\right] \geq -\inf_{x\in O} I(x); \tag{1.3}$$

(P3) pour tout fermé $F \subset E$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in F\right] \leq -\inf_{x\in F} I(x). \tag{1.4}$$

I est appelée fonctionnelle d'action. On dit c'est une bonne *fonctionnelle d'action si ses niveaux $\{x \in E : I(x) \leq C\}$ sont compacts pour tout $C \geq 0$.*

On dit que $\hat{S}_n$ vérifie un principe faible *si les points (P1) et (P2) sont vrais, (P3) n'étant vérifié que pour les compacts.*

De façon heuristique, cela signifie que la probabilité d'être dans le voisinage d'un point $x \in E$ décroît exponentiellement vite et que la fonctionnelle d'action mesure cette vitesse :

$$\mathbb{P}\left[\hat{S}_n \simeq x\right] \simeq e^{-nI(x)}.$$

Il existe bien d'autres processus dont la moyenne converge, par exemple les chaînes de Markov ergodiques. Un principe de grandes déviations garde tout son sens dans ces contextes.

L'existence d'un tel principe est très courante ; on sait le montrer dans nombre de situations, mais c'est encore un sujet de recherche actif. Il s'agit, par exemple, d'exprimer les conditions de validité du principe fort à partir des donnés du problème, ou bien de l'unicité de la fonctionnelle d'action, et éventuellement son l'évaluation. La topologie elle-même peut intervenir : on connaît des cas où un principe est vrai dans une topologie et faux dans une autre.

1.3 La mesure empirique

Contrairement à la moyenne empirique, la mesure (ou distribution) empirique ne requiert aucune hypothèse de structure d'espace vectoriel pour E. L'espace $\mathcal{M}(E)$ des mesures finies sur E possède en effet une structure naturelle d'espace vectoriel. Nous considérerons souvent l'espace $\mathcal{M}_1(E)$ des mesures de probabilités sur E, qui est un ensemble convexe de $\mathcal{M}(E)$, et qui est polonais si E l'est, c'est-à-dire qui a toutes les bonnes propriétés.

Définition 1.3 *La distribution empirique de la suite* $\{X_i,\ i \geq 1\}$ *est*

$$L_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \in \mathcal{M}_1(E), \tag{1.5}$$

en notant δ_x *la mesure de probabilité concentrée au point* x.

Si les variables $\{X_i,\ i \geq 1\}$ sont indépendantes et identiquement distribuées de loi μ, alors la distribution empirique converge presque sûrement vers μ. Dans le cas des chaînes de Markov ergodiques, la convergence se fait vers la mesure stationnaire.

À ce niveau, la description est plus fine, puisque c'est la fréquence des occurrences de chaque état qui est mesurée. Pourtant la fonctionnelle d'action se révèle être une entropie, ou une quantité dérivée, dans les deux cas cités en exemple ci-dessus, ce qui permet des preuves plus aisément interprétables que dans le cas de la moyenne empirique.

1.4 La distribution empirique du processus

Définition 1.4 *La distribution empirique du processus* $\{X_i,\ i \geq 1\}$ *est*

$$L_n^\infty \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta^i X}, \tag{1.6}$$

en notant $X = (X_1, X_2, \dots)$ *le processus entier et* θ^i *le décalage* $\theta^i X = (X_{i+1}, X_{i+2}, \dots)$.

Cette distribution mesure le comportement global du processus. Nous ne l'étudierons pas (p. ex. voir [48]), mais une idée importante pour la suite s'en dégage : *plus la mesure est précise, mieux les propriété du processus s'expriment.* Par exemple, la mesure empirique suffit à *caractériser* une suite de variables i.i.d, mais pas une chaîne de Markov, pour laquelle il est nécessaire de mesurer les transitions.

1.5 Notations

Nous résumons ici les notations les plus couramment utilisées :

- E est l'espace d'état, le plus souvent dénombrable ;
- $\{X_k,\ k \geq 1\}$ est une chaîne de Markov sur E ;
- $P = (P_{xy})_{x,y \in E}$ est la matrice de transition de X_k ;
- Y_t est un processus de Markov sur E ;
- $R = (R_{xy})_{x,y \in E}$ est le générateur de Y_t ;
- T_n est le n–ième temps de saut de Y_t ;
- N_t est le nombre de sauts de Y_t avant la date t ;
- $\mathcal{M}_+(E)$ [resp. $\mathcal{M}_1(E)$] est l'ensemble des mesures positives [resp. de probabilité] sur E ;
- δ_x est la mesure de probabilité concentrée au point x (Dirac en x) ;
- on note souvent μ, ν, π des probabilités et f, g, h des fonctions ;
- $\mathcal{M}_s(E^2)$ est l'ensemble des mesures balancées sur $E^2 = E \times E$. Une mesure balancée $A \in \mathcal{M}_s(E^2)$ vérifie $A(E, .) \overset{\text{loi}}{=} A(., E)$ et on lui associe
  - $a_{xy} \overset{\text{def}}{=} A(\{(x,y)\})$ la mesure bidimensionnelle qui représente la fréquence des sauts de x à y ;
  - $a_x \overset{\text{def}}{=} A(E \times \{x\})$ l'unique projection qui représente la fréquence de passage en x ;
  - A_{xy} la loi conditionnelle (vérifiant $a_{xy} = a_x A_{xy}$) qui représente la matrice de transition ;
  - $\text{Supp}(A) \subset E^2$ le support de A.
- Γ est un graphe et $\mathcal{M}_s(\Gamma)$ représente l'ensemble des mesures balancées dont le support est inclus dans Γ. On note E^2_* l'ensemble des paires non-diagonales ((x, y) avec $x \neq y$) ;
- $L_n(\omega)$ ou $L_t(\omega)$ la mesure empirique ;
- $\hat{S}_n(\omega)$ la moyenne empirique ;
- $G_n(\omega)$ le générateur empirique ;

- $\mathcal{G}_s(E)$ [resp. $\mathcal{G}_s(\Gamma)$] l'ensemble des générateurs empiriques sur E [resp. sur $\Gamma \subset E_*^2$]. Pour $G = (\pi, A) \in \mathcal{G}_s(E)$, par unité de temps, on note
  - $A = (a_{xy}) \in \mathcal{M}_s(E_*^2)$: a_{xy} est le nombre de sauts de x à y ;
  - $\pi \in \mathcal{M}_1(E)$: π_x est le temps moyen de séjour en x ;
- $H(.\|.)$ est l'entropie relative, que l'on note H_d dans le cas de chaîne de Markov en temps *discret*, H_c en temps *continu* et I_p pour les processus de *Poisson* ;
- Λ est la fonction des cumulants d'une mesure et Λ^* sa transformée de Legendre ;
- M_n ou M_t est une martingale qui sert à définir un changement de mesure ;
- $\mathbb{P}$ est la probabilité initiale du processus X_n ou Y_t, et $\mathbb{E}$ l'espérance associée. $\mathbb{P}^*$ et $\mathbb{E}^*$ sont la probabilité et l'espérance modifiées ;
- O est un ensemble ouvert, F un fermé et K un compact.

Chapitre 2

Les variables indépendantes

On considère une suite $\{X_i,\ i \geq 1\}$ de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de loi μ, prenant leurs valeurs sur un espace E. Nous allons maintenant entrer davantage dans les concepts qui permettent de donner les conditions sous lesquelles les différents niveaux de principes de grandes déviations sont valides.

2.1 La moyenne empirique

La moyenne empirique satisfait *toujours* un principe faible.

Théorème 2.1 *Soit μ une mesure de probabilité sur un espace polonais E. Soit $\{X_i,\ i \geq 1\}$ une suite de variables i.i.d. de loi μ. Alors $\hat{S}_n$ vérifie un principe faible de grandes déviations avec la fonctionnelle Λ^*.*

La fonctionnelle Λ^* est définie immédiatement ci-dessous (Définitions 2.2 et 2.3). Nous allons cependant donner plusieurs techniques pour démontrer chaque point. Au fur et à mesure, d'autres concepts vont enrichir la signification de ce théorème, le renforcer et l'étendre.

2.1.1 La transformée de Legendre

Définition 2.2 (fonction des cumulants) *Soit μ une mesure de probabilité sur E, dont le dual est E^*. On appelle fonction des cumulants de μ la fonction $\Lambda :\ E^* \mapsto \bar{\mathbb{R}}$ définie par*

$$\Lambda(f) \stackrel{\text{def}}{=} \log \mathbb{E}\left[e^{\langle f, X_1 \rangle}\right] = \log \int_E e^{f(x)} \mu(\,\mathrm{d}x).$$

En fait, la fonction des cumulants recueille toute l'information utile aux grandes déviations. La fonctionnelle d'action est généralement exprimée comme une transformée de cette quantité.

Définition 2.3 (Transformée de Legendre) *On appelle transformée de Legendre de la fonction $\phi: \; E^* \mapsto \bar{\mathbb{R}}$, la fonction semi-continue inférieurement*[1] $\phi^*: \; E \mapsto \bar{\mathbb{R}}$ *définie par*

$$\phi^*(x) \stackrel{\text{def}}{=} \sup_{f \in E^*} \big\{ \langle f, x \rangle - \phi(f) \big\}. \tag{2.1}$$

La borne supérieure (1.4) peut être démontrée localement via la structure de la transformée de Legendre, en utilisant l'indépendance des variables X_i et la continuité des fonctions linéaires f. L'indépendance implique

$$\log \mathbb{E}\left[e^{n\langle f, \hat{S}_n\rangle}\right] = \log \mathbb{E}\left[\prod_{i=1}^{n} e^{\langle f, X_i\rangle}\right] = n \log \mathbb{E}\left[e^{\langle f, X_1\rangle}\right] = n\Lambda(f), \tag{2.2}$$

pour tout $f \in E^*$. Maintenant, considérons une petite boule $B(x, \varepsilon)$ de centre x et de rayon ε. Par une inégalité de type Chebychev, on a la minoration

$$\log \mathbb{E}\left[\mathbb{1}_{\{\hat{S}_n \in B(x,\varepsilon)\}} e^{n\langle f, \hat{S}_n\rangle}\right] \geq \log \mathbb{P}\left[\hat{S}_n \in B(x,\varepsilon)\right] + \inf_{y \in B(x,\varepsilon)} n\langle f, y\rangle.$$

En combinant cette inégalité avec (2.2), on obtient

$$\frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in B(x,\varepsilon)\right] \leq \sup_{f \in E^*} \left\{\Lambda(f) - \inf_{y \in B(x,\varepsilon)} \langle f, y\rangle\right\} = - \inf_{y \in B(x,\varepsilon)} \Lambda^*(y).$$

Sachant que Λ^* est semi-continue inférieurement, la borne inférieure dans la partie droite tend vers $\Lambda^*(x)$ quand ε devient petit ; le résultat local suivant est démontré : quel que soit $x \in E$, quel que soit $\delta > 0$, il existe $\varepsilon > 0$ tel que

$$\frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in B(x,\varepsilon)\right] \leq -\Lambda^*(x) + \delta. \tag{2.3}$$

Cette borne locale est souvent mise sous la forme suivante, légèrement plus faible que l'inégalité précédente.

$$\lim_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in B(x,\varepsilon)\right] \leq -\Lambda^*(x). \tag{2.4}$$

1. La transformée de Legendre est semi-continue inférieurement puisque c'est la borne supérieure de fonctions continues.

La borne inférieure ne peut se démontrer avec les mêmes arguments, car on a utilisé la majoration

$$\log \mathbb{E}\left[\mathbb{1}_{\{\hat{S}_n \in B(x,\varepsilon)\}} e^{n\langle f,\hat{S}_n\rangle}\right] \leq \log \mathbb{E}\left[e^{n\langle f,\hat{S}_n\rangle}\right] = n\Lambda(f).$$

Pour obtenir l'inégalité inverse il faudrait pouvoir montrer que la mesure correspondant à l'espérance *se concentre* au voisinage de x. C'est pourquoi on recoure souvent à un changement de mesure (voir section 2.2.4).

Remarque : Le point crucial de cette démonstration est que le théorème se démontre *localement*, au voisinage d'un point x où l'on veut estimer la fonctionnelle.

♦♦♦

2.1.2 Des bornes locales aux bornes globales

Supposons que l'on ait obtenu la borne inférieure, qui est le pendant de (2.4) :

$$\lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in B(x,\varepsilon)\right] \geq -\Lambda^*(x). \tag{2.5}$$

L'extension des bornes locales de type (2.4) aux bornes globales (1.3) et (1.4) est une affaire assez bien connue.

La borne inférieure se généralise sans peine. Considérons un ensemble ouvert $O \subset E$, $x \in O$ et un petit $\delta > 0$. Alors, en supposant que (2.5) soit valide, il existe une petite boule $B(x,\varepsilon)$ incluse dans O telle que son taux de décroissance soit inférieur à $I(x) + \delta$. Cela reste vrai pour O, puisqu'il contient $B(x,\varepsilon)$. Comme δ peut être choisi arbitrairement petit, le taux de décroissance de O est inférieur à $I(x)$, pour n'importe quel x appartenant à O. La borne inférieure sur les ouverts (1.3) est donc une conséquence de la borne locale inférieure.

L'extension de la borne locale supérieure (2.4) à la borne supérieure (1.4) pour tout compact ne requiert aucune condition, alors que l'extension à des fermés est beaucoup plus contraignante.

Considérons en effet un compact $K \subset E$ et fixons $\delta > 0$. Grâce à la borne locale (2.4), à chaque $x \in K$ on peut faire correspondre $\varepsilon_x > 0$ tel que (2.3) soit vérifié à l'infini (au sens de la limite supérieure). La réunion de toutes ces boules recouvrent K, et d'après la compacité de K, un sous-recouvrement fini peut en être extrait,

$$K \subset \bigcup_{i=1}^{p} B(x_i, \varepsilon_{x_i}).$$

Par la sous-additivité des probabilités, en minorant les taux de décroissance par le plus petit d'entre eux, en appliquant l'inégalité

$$\frac{1}{n}\log\left(\sum_{i=1}^{p} u_i\right) \leq \frac{1}{n}\left(\log p + \max_{i\leq p}\log u_i\right),$$

on obtient

$$\frac{1}{n}\log \mathbb{P}\left[\hat{S}_n \in K\right] \leq -\min_{i\leq p}\Lambda^*(x_i) + \delta \leq -\inf_{x\in K}\Lambda^*(x) + \delta.$$

Là encore, cette borne est valide pour tout $\delta > 0$, ce qui achève la démonstration du point (P3) pour les compacts.

2.1.3 La sous-additivité

La manipulation d'ensembles convexes conjuguée avec l'indépendance des variables X_i donne lieu à un outil très puissant pour la démonstration du principe faible de grandes déviations. Malheureusement, sa force est proportionnelle au nombre d'étapes que le raisonnement doit enchaîner. De plus, la fonctionnelle d'action n'est pas très explicite. Nous donnerons donc les grandes lignes de la méthode ici, en renvoyant aux excellents ouvrages de référence [40, 41] pour les preuves et davantage de détails.

Dans un premier temps, remarquons que si la moyenne empirique des X_i, pour i compris entre 1 et n, appartient à un ensemble convexe $C \subset E$, et si la moyenne des X_i entre $n+1$ et $n+m$ y appartient aussi, alors cela reste vrai de la moyenne des X_i entre 1 et $n+m$. Or, par indépendance, la loi des m derniers X_i est la même que celle des m premiers. Ce raisonnement s'écrit formellement

$$\mathbb{P}\left[\hat{S}_{n+m} \in C\right] \geq \mathbb{P}\left[\hat{S}_n \in C\right]\mathbb{P}\left[\hat{S}_m \in C\right]. \tag{2.6}$$

En posant maintenant

$$\phi(n) \stackrel{\text{def}}{=} -\log \mathbb{P}\left[\hat{S}_n \in C\right],$$

on constate que ϕ est sous-additive, ce qui signifie que

$$\phi(n+m) \leq \phi(n) + \phi(m), \qquad \forall n, m \geq 0.$$

Les fonctions sous-additives vérifient la bonne propriété de convergence suivante.

Lemme 2.4 *Soit une fonction* $\phi : \mathbb{N} \mapsto [0, +\infty]$ *sous-additive. S'il existe* $N \geq 0$ *tel que* $\phi(n) < \infty$ *pour tout* $n \geq N$*, alors*

$$\lim_{n\to\infty} \frac{\phi(n)}{n} = \inf_{n\geq N} \frac{\phi(n)}{n} < \infty.$$

Dans un deuxième temps, on montre que, soit la probabilité pour $\hat{S}_n$ d'appartenir à C est nulle pour tout n, soit elle est strictement positive au-delà d'un certain N, ce qui nous assure d'être dans les conditions d'utilisation du lemme 2.4. Pour cette raison, il est nécessaire de supposer que C est ouvert, et on utilise la propriété dite de *séparabilité* de E, ce qui est justifié puisque E est polonais. Finalement, le résultat est que

$$-\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in C\right] \stackrel{\text{def}}{=} J(C)$$

existe dans $[0, \infty]$, pour tout ensemble convexe

Dans un troisième temps, la fonctionnelle locale est construite. D'après la formule ci-dessus, il est clair que $J(C)$ est une fonction décroissante. La limite

$$\lim_{\varepsilon\to 0} J(B(x,\varepsilon)) \stackrel{\text{def}}{=} I(x), \qquad \forall x \in E$$

existe donc comme limite d'une suite monotone. La fonctionnelle d'action I est semi-continue inférieurement et convexe. En effet, si $I(x) > \alpha$, il existe $\varepsilon > 0$ tel que $J(B(x,\varepsilon)) > \alpha$, et pour chaque point $y \in B(x,\varepsilon)$, $I(y) \geq J(B(x,\varepsilon)) > \alpha$. Donc les ensembles $\{x : I(x) > \alpha\}$ sont ouverts pour tout α, ce qui caractérise la semi-continuité inférieure. Par ailleurs, de façon similaire à (2.6), l'inégalité

$$\mathbb{P}\left[\hat{S}_n \in C_1\right] \mathbb{P}\left[\hat{S}_m \in C_2\right] \leq \mathbb{P}\left[\hat{S}_{n+m} \in C\right],$$

avec $(n+m)C = nC_1 + mC_2$, où C, C_1 et C_2 sont convexes, entraîne la convexité de I.

Dans un quatrième temps on montre que $\hat{S}_n$ vérifie un principe faible de grandes déviations avec I pour fonctionnelle d'action. Remarquons d'abord que les deux bornes locales sont simultanément obtenues, en écrivant complètement la définition de I :

$$I(x) \stackrel{\text{def}}{=} -\lim_{\varepsilon\to 0} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in B(x,\varepsilon)\right]. \tag{2.7}$$

La semi-continuité inférieure de I vient d'être démontrée. La borne inférieure (1.3) est un corollaire de la définition de I, et la borne supérieure (1.4) pour les compacts se traite par recouvrement, comme on l'a vu au paragraphe 2.1.2.

Dans un cinquième temps, il faut montrer que la fonctionnelle d'action I est bien la transformée de Legendre Λ^*. Ce point n'est pas essentiel, puisque l'on peut se contenter d'un principe faible avec la fonctionnelle I, mais il est intéressant de noter que la fonctionnelle d'action n'est pas toujours unique, bien que dans le cas présent elle le soit. La démonstration de l'identification $I = \Lambda^*$ est assez longue, technique et les bénéfices n'en sont pas immédiats, aussi l'omettrons-nous (voir [40]).

> **Remarque :** La méthode de sous-additivité illustre ce point clef des grandes déviations qu'est le *comportement déviant*. Bien qu'on ne mesure la probabilité de $\hat{S}_n$ d'être au voisinage de x que pour n grand, on constate ici, de façon encore informelle, que c'est en fait toute la trajectoire $\{\hat{S}_m, m \leq n\}$ qui reste au voisinage de x. C'est donc un comportement global de la trajectoire, conditionnellement au fait que $\hat{S}_n \simeq x$.

♦♦♦

Le principe fort Dans $E = \mathbb{R}^d$, le dual est identifié à $\mathbb{R}^d$ lui-même via le produit scalaire. Dans ces conditions on peut montrer qu'une condition suffisante pour qu'un principe fort soit valide est que 0 soit dans l'intérieur du domaine où Λ est finie : c'est le théorème de Cramer dans $\mathbb{R}^d$ (qui est une conséquence du théorème 3.9). Cependant, nous allons voir que cette condition est superflue, en passant au niveau deux des grandes déviations, en appliquant la maxime : « Mieux vaut un résultat simple sur un espace compliqué qu'un résultat compliqué sur un espace simple. »

2.2 La mesure empirique

2.2.1 Réduction au niveau 1

Comme nous l'avons annoncé, le niveau 2 peut se réduire au niveau 1. C'est une occasion d'appliquer les résultats déjà acquis, avant de passer à des méthodes plus efficaces.

Puisque E est polonais, $\mathcal{M}_1(E)$ l'est aussi, par conséquent la suite de variables

$$Y_i \stackrel{\text{def}}{=} \delta_{X_i} \in \mathcal{M}_1(E)$$

est une suite i.i.d. sur un espace polonais. En appliquant le théorème 2.1, on en déduit que la moyenne empirique des Y_i, qui n'est autre que la mesure empirique des X_i, vérifie un PGD faible, avec la fonctionnelle

$$\sup_{f\in\mathcal{M}_1^*(E)} \left\{\langle f,\nu\rangle - \log\mathbb{E}\left[e^{\langle f,Y_1\rangle}\right]\right\} = \sup_{f\in\mathcal{C}_b(E)} \left\{\langle f,\nu\rangle - \log\mathbb{E}\left[e^{f(X_1)}\right]\right\}.$$

En effet, le dual $\mathcal{M}_1^*(E)$ de $\mathcal{M}_1(E)$ est l'ensemble $\mathcal{C}_b(E)$ des fonctions continues bornées sur E. Si E est fini, grâce au principe fort sur $\mathbb{R}^d$ pour la moyenne empirique, on peut prouver le principe fort pour la mesure empirique, ce qui est assez restreint comme champ d'application.

On voit néanmoins que la méthode de sous-additivité est remarquablement efficace, avec le léger bémol que la fonctionnelle d'action peut être mise sous une forme plus explicite. C'est ce que nous allons montrer maintenant.

2.2.2 L'entropie et l'information

En fait les grandes déviations s'expriment mieux au niveau 2 qu'au niveau 1, la raison profonde étant qu'une mesure plus détaillée exprime mieux les propriétés du processus, et qu'ici, la mesure empirique L_n (voir définition 1.3) *caractérise le processus* $\{X_i,\ i\geq 1\}$.

Théorème 2.5 *Soit μ une loi sur un espace polonais E. Soit $\{X_i,\ i\geq 1\}$ une suite de variables i.i.d. de loi μ. Alors L_n vérifie un* PGD *fort avec la bonne fonctionnelle d'action*

$$H(\nu\|\mu) \stackrel{\text{def}}{=} \begin{cases} \displaystyle\int_E \nu(\,\mathrm{d}x)\log\frac{\mathrm{d}\nu}{\mathrm{d}\mu} & \textit{si } \nu\ll\mu, \\ +\infty & \textit{sinon.} \end{cases} \tag{2.8}$$

$H(\nu\|\mu)$ est l'entropie de ν relativement à μ.

L'entropie est une passerelle entre, d'une part la théorie de l'information, et d'autre part la technique du changement de mesure, et donc les grandes déviations. On peut alors enrichir réciproquement les deux théories. Nous aurons tendance ici à privilégier le gain de sens qu'offre la théorie de l'information aux grandes déviations.

Interprétation en termes d'information La plupart des résultats de grandes déviations peuvent s'interpréter à la lumière de la théorie de l'information. Une excellente introduction à ce sujet, en lien avec les probabilités, est Rényi [121], que l'on peut approfondir par [3, 73].

Commençons par le cas d'une variable aléatoire X de loi μ. On définit une *information* comme la connaissance (ou la prédiction) d'un événement $X \in A$, à laquelle on associe une valeur, le *gain d'information*, par définition

$$I(A) \stackrel{\text{def}}{=} -\log \mathbb{P}\left[X \in A\right]. \tag{2.9}$$

On trouve un comportement raisonnable pour ce gain [2] : il est positif et si je fais deux fois la même prédiction, j'ai deux fois plus d'information et le gain est double (dans ce cas la loi est $\mu \star \mu$ et le logarithme transforme la multiplication en addition car les tirages sont indépendants).

On en déduit facilement l'information moyenne, ou *entropie*, liée à l'observation de la variable X de loi μ :

$$H(\mu) \stackrel{\text{def}}{=} -\sum_k \mu_k \log \mu_k.$$

Il faut noter que plus l'observation est précise (plus la loi μ_k a d'atomes), plus l'entropie est élevée.

Un autre concept est l'information *relative*. Lorsque l'information $X \in A$ est donnée, on peut se demander quelle est la nouvelle distribution ν. Il est naturel de poser $\nu(.) = \mu(.|A)$, c'est-à-dire que la mesure est uniformément répartie (selon μ) sur l'événement A. La donnée de ν est en fait la forme la plus générale et la plus précise que peut prendre une information.

> **Exemple :** Des bus sont numérotés de 1 à 10 ; il y a 30 véhicules numérotés 1 à 5 et 10 numérotés 6 à 10. Des bus sont verts : 10 pour les N^o 1 et 2, et 5 pour les N^o 5 et 6. Nous définissons μ comme la probabilité sur les numéros : $\mu_i = 3/20$ pour $1 \leq i \leq 5$ et $\mu_i = 1/20$ sinon. On tire un bus X au hasard. L'information « le bus est vert » revient à dire « Le bus est N^o 1 ou 2 avec fréquence 1/3 ou N^o 5 ou 6 avec fréquence 1/6 », ce qui est équivalent à la mesure ν, où $\nu_0 = \nu_1 = 1/3$ et $\nu_5 = \nu_6 = 1/6$, mais ne peut pas se mettre sous une forme du type « le bus appartient aux lignes 1, 2, 5 ou 6 ».

♦♦♦

Maintenant, connaissant la distribution μ, l'information associée à une mesure ν a la valeur $H(\nu\|\mu)$:

$$H(\nu\|\mu) \stackrel{\text{def}}{=} \int \nu(\,\mathrm{d}x) \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x).$$

2. Nous ne développerons pas ici une axiomatique, qui peut être trouvée dans les ouvrages précités. Par ailleurs, nous confondrons souvent dans la suite information et gain d'information.

C'est *l'entropie relative* de la probabilité ν par rapport à μ. Cette entropie jouit également de propriétés naturelles ; elle est positive et additive lorsque les variables sont indépendantes :

$$H\big(\nu_1 \otimes \nu_2 \| \mu_1 \otimes \mu_2\big) = H\big(\nu_1 \| \mu_1\big) + H\big(\nu_2 \| \mu_2\big).$$

Dans le cadre des grandes déviations, la mesure empirique L_n renseigne sur la distribution des n premiers tirages. Le gain d'information associé à[3] $L_n \simeq \nu$ (voir (2.9)), d'après le théorème 2.5, vaut à peu près $nH(\nu\|\mu)$; c'est la même quantité d'information que de se donner n fois l'information ν, pour chaque tirage $i \leq n$. Ceci vient par conséquent corroborer et préciser la vision de comportement déviant (voir la remarque de la page 14).

À cette interprétation locale s'ajoute une considération globale : l'information $L_n \in O$ est asymptotiquement équivalente à $L_n \simeq \nu$, où ν réalise le minimum de l'entropie sur O. C'est un principe de moindre information assez similaire au principe de moindre action : parmi toutes les mesures possibles, le processus choisit celle qui minimise l'information. C'est une formulation du principe de Gibbs en thermodynamique (par exemple, voir [24] et la remarque *loi conditionnelle*, p. 96).

2.2.3 La tension exponentielle

Après avoir formulé le principe faible, l'épithète lui-même suggère qu'il est possible de passer de faible à fort. La tension exponentielle est la condition nécessaire et suffisante[4] pour ce renforcement. L'idée directrice est d'assurer que ce sont bien les compacts qui portent toute la masse des probabilités (la *tension*) et ce sur une échelle exponentielle, d'où le nom de cette section. La définition et le théorème principal en ce domaine sont la définition 3.6 et le théorème 3.7.

La proposition suivante permet donc d'achever le renforcement du principe faible pour la mesure empirique.

Proposition 2.6 *La suite de mesures empiriques* $\{L_n, n \geq 1\}$ *est exponentiellement tendue.*

Preuve : Puisque E est polonais, la mesure μ est tendue. Donc, pour tout $\varepsilon_p > 0$, il existe un compact $K_p \subset E$ tel que

$$\mathbb{P}\left[X_1 \notin K_p\right] \leq \varepsilon_p.$$

3. C'est à dire que $L_n \in B(\nu, \varepsilon)$ pour $\varepsilon > 0$ petit.

4. La condition est toujours suffisante, mais est nécessaire seulement parce que l'espace E est polonais (voir section 3.1.4).

Cela permet de majorer la probabilité de sortir de K_p avec une fréquence plus élevée que $\eta_p > 0$:

$$\mathbb{P}\left[L_n(K_p^c) \geq \eta_p\right] \leq \sum_{i \geq n\eta_p} C_n^i \varepsilon_p^i \leq 2^n e^{n\eta_p \log \varepsilon_p} \leq e^{n(-p+\log 2)}.$$

La dernière inégalité est obtenue en choisissant des paramètres judicieux, de telle sorte que $\eta_p \log \varepsilon_p$ soit très négatif, ici $\varepsilon_p = e^{-p^2}$ et $\eta_p = 1/p$.

On introduit alors l'ensemble fermé (pour la topologie faible sur $\mathcal{M}_1(E)$)

$$\widetilde{K} \stackrel{\text{def}}{=} \bigcap_{p \geq N} \left\{\nu : \nu(K_p) \geq 1 - \eta_p\right\} \subset \mathcal{M}_1(E).$$

Cet ensemble de mesures est tendu puisque, pour tout p et toute mesure $\nu \in \widetilde{K}$, le compact K_p vérifie $\nu(K_p^c) \leq \eta_p$. C'est le critère de pré-compacité de Prokhorov. Comme $\widetilde{K}$ est fermé, il est donc compact. Finalement

$$\mathbb{P}\left[L_n \notin \widetilde{K}\right] \leq \sum_{p=N}^{\infty} \mathbb{P}\left[L_n \in \widetilde{K}_p^c \geq \eta_p\right] \leq \sum_{p=N}^{\infty} e^{n(-p+\log 2)} \leq e^{-n(N-1)},$$

si $n \geq 2$. Comme N est arbitrairement grand, (3.7) est vérifié, donc la suite $\{L_n, n \geq 1\}$ est exponentiellement tendue. ■

Remarque : Cette preuve repose en fait sur une opération fructueuse, *l'agrégation des états* en dehors des compacts K_p. On arrive, grâce à l'indépendance, à majorer la probabilité d'une fréquence strictement positive de cet état agrégé, et on en déduit la tension exponentielle. On peut cependant faire encore bien mieux et c'est le propos des *limites projectives* (*cf.* section 3.2.5).

♦♦♦

2.2.4 Le changement de mesure

Le changement de mesure permet, comme la sous-additivité, de prouver un PGD faible. Sa puissance réside en la construction explicite de la fonctionnelle d'action, l'*entropie* H. Cette dernière jouit d'excellentes propriétés et de plus s'interprète facilement (voir la section 2.2.2). Le sens qui est alors donné aux grandes déviations est un guide assez sûr permettant d'explorer des domaines où les techniques de preuve ne sont pas toujours mûres.

Soit $\nu \in \mathcal{M}_1(E)$, absolument continue par rapport à μ. Il existe donc une dérivée de Radon-Nikodym et on peut introduire

$$f \stackrel{\text{def}}{=} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

La martingale

$$M_n \stackrel{\text{def}}{=} \exp\left\{\sum_{i=1}^{n} f(X_i)\right\} = e^{n\langle L_n, f\rangle}$$

permet de définir la mesure « déformée » $\mathbb{P}^*$ sur le processus X_n par

$$\mathbb{P}^*\left[A\right] \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbb{1}_{\{A\}} M_n\right] = \int \mathbb{1}_{\{A\}} \nu(\,\mathrm{d}x_1) \ldots \nu(\,\mathrm{d}x_n).$$

Sous la loi $\mathbb{P}^*$, le processus X_n est donc i.i.d, de loi ν, et non plus μ. Ainsi, par définition

$$\mathbb{P}\left[L_n \in B(\nu, \varepsilon)\right] = \mathbb{E}^*\left[\mathbb{1}_{\{L_n \in B(\nu,\varepsilon)\}} e^{-n\langle L_n, f\rangle}\right]. \tag{2.10}$$

Nous allons supposer dans la suite que l'application $\pi \to \langle \pi, f\rangle$ est continue. Ce n'est pas toujours vrai, mais on peut faire tout comme, c'est le propos de [34] (repris chapitre II 2). Moyennant cette hypothèse,

$$\lim_{\varepsilon\to 0} \inf_{\pi\in B(\nu,\varepsilon)} \langle \pi, f\rangle = \lim_{\varepsilon\to 0} \sup_{\pi\in B(\nu,\varepsilon)} \langle \pi, f\rangle = \langle \nu, f\rangle = H(\nu\|\mu). \tag{2.11}$$

La dernière égalité découle de la définition de f. Puisque L_n converge p.s. vers ν sous $\mathbb{P}^*$, on déduit des égalités (2.10) et (2.11) :

$$\lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in B(\nu, \varepsilon)\right] = -H(\nu\|\mu), \tag{2.12}$$

$$\lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in B(\nu, \varepsilon)\right] = -H(\nu\|\mu). \tag{2.13}$$

On remarque la ressemblance avec l'équation (2.7) qui définit la fonctionnelle d'action pour la sous-additivité. On a ici deux limites, inférieure et supérieure, mais comme elles coïncident, les deux bornes locales sont bien prouvée. L'immense avantage, c'est que la fonctionnelle est explicite, c'est l'entropie $H(.\|\mu)$, et qu'elle jouit d'excellentes propriétés.

> **Remarque :** Une fois encore, cette méthode ne permet de prouver ni la tension exponentielle, ni le principe fort, seulement un PGD faible. Cependant, le fait de connaître la fonctionnelle d'action permet plus facilement de prouver la tension exponentielle, et donc le principe fort.
>
> Le plus naturel est d'utiliser les limites projectives, en choisissant une suite croissante de compacts K_p qui tends vers E, et en agrégeant les états de K_p^c en un unique point Δ. Pour chacun des espaces compacts $K_p \cup \{\Delta\}$, le projeté de X_n vérifie un PGD fort (puisque $K_p \cup \{\Delta\}$ est compact), et la fonctionnelle d'action est une approximation de $H(.\|\mu)$. On en déduit très aisément le principe fort. La méthode est à peine plus longue que la démonstration de la tension exponentielle du paragraphe 2.2.3, mais elle est bien plus intuitive.
>
> ♦♦♦

2.3 La distribution empirique du processus

Par souci d'exhaustivité, nous donnons ici sans preuve le PGD relatif à la distribution empirique du processus. Cependant, autant les applications du niveaux 1 et 2 sont nombreuses, autant celles du niveau 3 sont rares. Il est néanmoins satisfaisant de constater que les techniques usuelles sont ici encore efficientes.

Soit ν une probabilité sur $E^{\mathbb{N}}$. On note $\nu^{(k)} \in \mathcal{M}_1\big(E^k\big)$ sa projection sur les k premières composantes. On dit que ν est *invariante par décalage*, ou *balancée*, si

$$\nu^{(k)}\big(\,\mathrm{d}x_1, \ldots, \mathrm{d}x_{k-1}, E\big) \overset{\text{loi}}{=} \nu^{(k)}\big(E, \,\mathrm{d}x_1, \ldots, \mathrm{d}x_{k-1}\big), \quad \forall k \geq 2.$$

Théorème 2.7 *La distribution empirique L_n^{∞} du processus $\{X_i, \ i \geq 1\}$ vérifie un principe fort de grandes déviations dans $\mathcal{M}_1\big(E^{\mathbb{N}}\big)$ avec la bonne fonctionnelle d'action*

$$I(\nu) \overset{\text{def}}{=} \begin{cases} \sup\limits_{k \geq 2} H\left(\nu^{(k)} \| \nu^{(k-1)} \otimes \mu\right) & \textit{si } \nu \textit{ est balancée,} \\ \infty & \textit{sinon.} \end{cases}$$

Preuve : Comme le suggère la forme de la fonctionnelle d'action I, ce résultat est démontré par limites projectives (*cf.* théorème 3.15). ■

Chapitre 3

Théorie générale des grandes déviations

Le propos de ce chapitre est de donner un aperçu des propriétés globales, liées au principe même et non à des structures particulières (telles les variables i.i.d. ou les chaînes de Markov). L'essence des concepts comme la transformée de Legendre, la sous-additivité ou le changement de mesure, est ainsi séparé des preuves spécifiques que nous avons vues au chapitre 2.

Toutefois, la définition même du PGD peut prendre des formes très abstraites. Nous nous restreignons au cadre des espaces polonais — cadre suffisant pour toutes les applications présentées par la suite — en mentionnant parfois les structures minimales. Nous gardons ainsi une forme simple mais dans laquelle les idées ne sont pas dissimulées sous un taillis de notations.

Enfin, cette partie permet de dresser un état de l'art et de mieux situer les travaux de thèse. Il ne faut donc pas y voir une recension exhaustive des recherches sur les grandes déviations, mais plutôt une esquisse dans laquelle certains détails ont été plus intensément étudiés. Tout ce qui suit est donc bien connu, mais certaines sections exposent des développements personnels : il s'agit des fonctions de Ruelle-Lanford (section 3.1.3), du changement de mesure (section 3.2.1) et le principe de contraction (section 3.3.1).

3.1 Existence et unicité

3.1.1 Principe de grandes déviations

Définition 3.1 (PGD) *Soit* $\{X_n, n \geq 1\}$ *une suite de variables aléatoires sur un espace polonais* E*, et* $\mathbb{P}$ *la probabilité associée. On dit que* X_n *vérifie* *un* PGD *s'il existe une fonction $I: \ E \mapsto \mathbb{R}_+$ telle que*

(P1) I est semi-continue inférieurement ;

(P2) pour tout ouvert $O \subset E$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right] \geq -\inf_{x\in O} I(x); \tag{3.1}$$

(P3) pour tout fermé $F \subset E$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in F\right] \leq -\inf_{x\in F} I(x). \tag{3.2}$$

I est appelée fonctionnelle d'action. On dit c'est une bonne *fonctionnelle d'action si ses niveaux $\{x \in E: \ I(x) \leq C\}$ sont compacts pour tout $C \geq 0$.*

On dit que X_n vérifie un principe faible *si, les points (P1) et (P2) sont vrais, mais que (P3) n'est vérifié que pour les compacts.*

Dans la suite, O représentera toujours un ouvert, F un fermé et K un compact. Nous avons ici formulé le principe en temps discret. Pour le temps continu, il suffit de remplacer $n \geq 1$ par $t \in [0, \infty[$.

Notons que, dans le cadre des grandes déviations pour un *processus* X_n, seules les probabilités marginales $\mathbb{P}\left[X_n \in .\right]$ sont nécessaires, on n'a jamais besoin des lois jointes. C'est pourquoi le terme de « suite de variables aléatoires » et non de processus figure dans la définition 3.1. On ne fera désormais plus cette distinction.

Cependant, gardons à l'esprit que X_n peut être une variable beaucoup plus complexe qu'un simple réel, par exemple une mesure sur un espace dénombrable dans le cas du théorème de Sanov. Il s'ensuit que E peut être un très gros espace, avec une structure peu riche, et que pour cette raison nous devons garder une certaine généralité. Cependant, même dans ces cas l'espace est polonais, et c'est pourquoi nous ne présenterons pas les structures moins riches (espaces topologique et de Hausdorff, etc).

> **Remarque :** On introduit souvent le PGD par rapport à des *mesures* μ_ε indexées par un paramètre $\varepsilon > 0$. En toute généralité, on impose seulement à ces mesures d'être *positives* et *localement finies*. De plus, on introduit une notion de *changement d'échelle*, qui est une fonction $s: \ [0, \infty[\mapsto [0, \infty[$ croissante dont la limite en 0 est 0. Le principe s'exprime de la même manière, mais en remplaçant
>
> $$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in .\right] \quad \text{par} \quad \lim_{\varepsilon\to 0} s(\varepsilon) \log \mu_\varepsilon(.),$$

où le terme lim est soit la limite supérieure (lim sup), soit la limite inférieure (lim inf). On peut encore généraliser en remplaçant le logarithme par un changement d'échelle quelconque, comme dans la définition 5.23 — cette généralisation peut s'avérer utile : voir l'exemple p. 84. À condition de respecter quelques propriétés fondamentales, comme le « principe du terme le plus grand » (terme que nous reprenons de Lewis et Pfister [97]) et la croissance, tout ce qui s'ensuit reste valide, aussi garderons-nous la simplicité de la définition 3.1.

♦♦♦

3.1.2 Unicité

Lorsque l'espace est Hausdorff régulier[1] (c'est le cas des espaces polonais) toute fonction semi-continue inférieurement f satisfait une équation que l'on retrouvera sous différentes formes :

$$f(x) = \sup_{O \ni x} \inf_{y \in O} f(y). \tag{3.3}$$

Proposition 3.2 (Unicité) *Soit X_n une suite de variables aléatoires sur un espace polonais E vérifiant un* PGD *faible avec la fonctionnelle d'action I. Cette dernière est unique au moins dans les deux cas suivants :*

- *si X_n vérifie un principe fort ;*
- *si X_n vérifie un principe faible et que l'espace E est localement compact.*

Preuve : Soit x un point de E, et $\varepsilon > 0$. D'après (3.3), il existe un voisinage O_1 de x tel que $\inf_{y \in O_1} I(y) \geq I(x) - \varepsilon$. Le complémentaire O_1^c est fermé et il existe deux ouverts disjoints O_2 et O tels que $O_1^c \subset O_2$ et $x \in O$. On a donc construit un voisinage fermé $F = O_2^c$ de x, et $x \in O \subset F \subset O_1$. Si X_n vérifie un PGD *fort*, en appliquant les bornes de grandes déviations, on obtient

$$\begin{aligned} -I(x) \leq -\inf_{y \in O} I(y) &\leq \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right] \\ &\leq \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in F\right] \leq -\inf_{y \in F} I(y) \leq -I(x) + \varepsilon. \end{aligned}$$

Notons que, si E est localement compact et si X_n vérifie un PGD *faible*, F peut être choisi compact et la borne supérieure continue à s'appliquer. On

1. C'est donc la structure minimale sur E qui est nécessaire pour prouver la validité de la Proposition 3.2. C'est vrai aussi pour les résultats de la section 3.1.3.

déduit de ces inégalités que

$$-I(x) = \inf_{O \ni x} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right], \tag{3.4}$$

qui est une caractérisation de I, d'où l'unicité. ■

3.1.3 Fonctions de Ruelle-Lanford

Malgré le degré de généralité de la proposition 3.2, dans le cas du théorème de Sanov (PGD sur la mesure empirique L_n) sur un espace dénombrable, on ne peut en déduire l'unicité car il existe un PGD faible mais l'espace $\mathcal{M}_1(E)$ des probabilités sur E n'est pas localement compact. C'est la caractérisation (3.4) qui se généralise et offre la solution.

Définition 3.3 *On appelle fonctions de Ruelle-Lanford*

$$\underline{I}(x) \stackrel{\text{def}}{=} -\inf_{O \ni x} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right], \tag{3.5}$$

$$\overline{I}(x) \stackrel{\text{def}}{=} -\inf_{O \ni x} \sup_{K \subset O} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K\right]. \tag{3.6}$$

$\underline{I}$ et $\overline{I}$ sont respectivement les fonctions inférieure et supérieure de Ruelle-Lanford.

Cette présentation des fonctions de Ruelle-Lanford est un apport de la thèse (voir section II 3.4.2). Une fonction de Ruelle-Lanford a été introduite naturellement dans le cas de la sous-additivité où une unique limite existe pour tout ensemble convexe ouvert O,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right].$$

Elle a ensuite été dédoublée, notamment par Takahashi [131] et Orey [117]. Lewis et Pfister [97] notent brièvement la construction de (3.6), dans le cas où E n'est pas localement compact. Dans le cas contraire, la définition se simplifie car il n'est pas nécessaire de faire référence aux compacts.

Les fonctions de Ruelle-Lanford permettent de formuler simplement l'existence et l'unicité des PGD faibles (théorème 3.5). Mais dans un premier temps, prouvons quelques propriétés de ces fonctions.

Proposition 3.4 *Les fonctions de Ruelle-Lanford $\underline{I}$ et $\overline{I}$ sont positives et semi-continues inférieurement.*

Preuve : Par référence à (3.5), définissons la fonction $\underline{I}(O)$ par

$$\underline{I}(O) \stackrel{\mathrm{def}}{=} -\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right],$$

et de même $\overline{I}(O)$ par référence à (3.6). Ce sont des fonctions positives, donc $\underline{I}(x)$ et $\overline{I}(x)$ le sont aussi.

Elles sont, en outre, décroissantes : si $O_1 \subset O_2$, alors $\underline{I}(O_1) \geq \underline{I}(O_2)$. Par monotonie, $\underline{I}(O) \leq \underline{I}(x)$ pour tout $x \in O$. Prenons une suite $x_n \in E$ convergeant vers x. Pour tout $\varepsilon > 0$, par définition de $\underline{I}(x)$, il existe $O \ni x$ tel que $\underline{I}(O) \geq \underline{I}(x) - \varepsilon$. Pour n assez grand ($n \geq n_0$), x_n est inclus dans O et, par conséquent,

$$\underline{I}(x_n) \geq \underline{I}(O) \geq \underline{I}(x) - \varepsilon, \qquad \forall \varepsilon > 0.$$

Donc $x \to \underline{I}(x)$ est semi-continue inférieurement. $\overline{I}(x)$ l'est aussi grâce à la même propriété de décroissance de $\overline{I}(O)$ (définie en (3.6)). ■

> **Remarque :** En dépit des notations, l'inégalité $\underline{I} \leq \overline{I}$ n'est pas automatiquement acquise, comme le montre le théorème 3.5.

♦♦♦

Théorème 3.5 *Soit X_n une suite de variables aléatoires dans un espace polonais. X_n vérifie un* PGD *faible si, et seulement si, $\underline{I} \leq \overline{I}$. Toute fonction I, semi-continue inférieurement, telle que $\underline{I} \leq I \leq \overline{I}$, est une fonctionnelle d'action.*

La fonctionnelle d'action est unique si, et seulement si, $\underline{I} = \overline{I}$.

Preuve : La condition $\underline{I} \leq \overline{I}$ est nécessaire. Supposons en effet que X_n satisfasse un PGD avec la fonctionnelle I. Si $K \subset O$, alors, d'après le PGD faible,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K\right] \quad &\leq \quad -\inf_{x\in K} I(x) \\
\leq -\inf_{x\in O} I(x) \quad &\leq \quad \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right].
\end{aligned}$$

Donc, par (3.3) et par la définition 3.3, $\underline{I} \leq I \leq \overline{I}$ et l'unicité découle aisément de l'égalité $\underline{I} = \overline{I}$. Comme le montre le résultat d'existence ci-après, cette condition est également suffisante pour qu'il existe un PGD avec I comme fonctionnelle d'action, donc c'est une condition nécessaire et suffisante d'unicité.

Pour le résultat d'existence, prouvons que les bornes supérieure et inférieure sont toujours satisfaites avec les fonctionnelles respectives $\overline{I}$ et $\underline{I}$, même si l'inégalité $\underline{I} \leq \overline{I}$ n'est pas vérifiée. Par définition de $\underline{I}$ et la décroissance de $\underline{I}(O)$, nécessairement

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right] \geq -\underline{I}(x) \qquad \forall O \ni x,$$

d'où la borne inférieure.

Pour la borne supérieure, considérons un compact K. Pour tout $\varepsilon > 0$, par définition et décroissance, pour tout point $x \in E$, il existe $O_x \ni x$ tel que

$$\sup_{K\subset O} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K\right] \leq -\overline{I}(x) + \varepsilon.$$

Puisque $K \subset \bigcup_{x\in K} O_x$, on peut le recouvrir par une union finie de tels ouverts $\bigcup_{x\in X} O_x$, où $X \subset K$ est un ensemble fini. Muni de la topologie induite, K est un espace compact et une application du lemme d'Urysohn (voir [124, théorème 2.13]) permet de construire des compacts $K_x \subset O_x$ tels que K soit inclus dans $\bigcup_{x\in X} K_x$. On en déduit,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K\right] &\leq \limsup_{n\to\infty} \frac{1}{n} \log \sum_{x\in X} \sup_{K\subset O_x} \mathbb{P}\left[X_n \in K\right] \\
&\leq \max_{x\in X} \sup_{K\subset O_x} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K\right] \\
&\leq -\min_{x\in X} \overline{I}(x) + \varepsilon \\
&\leq -\inf_{x\in K} \overline{I}(x) + \varepsilon.
\end{aligned}$$

Les bornes supérieure et inférieure sont donc toujours satisfaites avec les fonctionnelles $\overline{I}$ et $\underline{I}$, respectivement, et par conséquent avec n'importe quelle fonction I semi-continue inférieurement, pourvu que $I \geq \underline{I}$ et $I \leq \overline{I}$.

L'existence d'une telle fonction I *satisfaisant les deux bornes* n'est toutefois possible que si $\underline{I} \leq \overline{I}$ et dans ce cas il en existe, par exemple $I = \underline{I}$. Le théorème est prouvé. ∎

Notons que si la technique de sous-additivité (voir section 3.2.3) est utilisable, notamment dans le cas de processus de renouvellement, alors on obtient directement que

$$I(x) \stackrel{\text{def}}{=} \sup_{O\ni x} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right]$$

est l'unique fonctionnelle d'action d'un PGD pour X_n.

Remarque : La majeure partie de la théorie classique des grandes déviations s'est construite autour de la notion de PGD fort, pour lequel les principaux résultats (existence, unicité, contraction...) ont été démontrés. Les preuves sont un petit peu plus techniques dans le cas des PGD faibles, mais en se servant des fonctions de Ruelle-Lanford, on y arrive de la même manière. Or beaucoup de systèmes classiques satisfont un PGD faible, mais non fort. En fait, les premiers théorèmes se situaient dans des espaces finis ou compacts dans lesquels, nécessairement, le PGD est fort. Nous avons consacré une partie de [32] (*cf.* section II 3.4) à étendre aux PGD faibles des résultats réservés jusqu'ici aux PGD forts.

♦♦♦

3.1.4 La tension exponentielle

Un des points délicats est de rendre fort un PGD faible. C'est le but de la tension exponentielle.

Définition 3.6 *Les lois des X_n sont exponentiellement tendues si, pour tout $C < \infty$, il existe $K \subset E$ tel que*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K^c\right] < -C. \tag{3.7}$$

On dira souvent plus brièvement que la suite X_n est exponentiellement tendue.

Théorème 3.7 *Si X_n vérifie un* PGD *faible et est exponentiellement tendue, alors elle vérifie un* PGD *fort et la fonctionnelle d'action est bonne.*

Supposons l'espace E localement compact ou polonais. Si X_n vérifie un PGD *fort avec la bonne fonctionnelle d'action I, alors X_n est exponentiellement tendue.*

La preuve est omise.

Remarque : Ces trois notions — l'intensité des PGD, *forts* ou *faibles*, la « magnitude[2] » de la fonctionnelle d'action, *bonne* ou *petite*, et la tension exponentielle — sont clairement liées, mais leurs relations précises sont encore un sujet de recherche.

2. Le terme de *bonne* fonctionnelle d'action possède désormais une valeur d'usage, mais pose le problème du substantif décrivant cette propriété et de son antonyme. Puisqu'une *bonne* fonctionnelle d'action est « grande » (ses niveaux sont compacts), nous choisissons *petit* comme antonyme et *magnitude* comme substantif.

Quelle que soit la « magnitude » de la fonctionnelle d'action, un PGD faible et la tension exponentielle impliquent un PGD fort. Le théorème 3.7 prouve, que pour certains espaces (au moins ceux qui sont localement compacts ou polonais), lorsqu'un PGD faible est valide, les deux propriétés, tension exponentielle *et* bonne fonctionnelle d'action, sont *équivalentes* à un PGD fort.

Certaines implications que l'on apprécierait sont fausses. Ainsi peut-on avoir un principe fort avec une *petite* fonctionnelle d'action (voir note 2 p. 27). Par exemple, si X_n est la variable uniforme sur $[0, n]$, alors X_n vérifie un PGD fort, avec la petite fonctionnelle d'action $I = 0$. On peut encore avoir un principe faible, mais pas fort, avec une bonne fonctionnelle d'action. La suite $X_n = n$ vérifie un PGD faible avec la bonne fonctionnelle d'action $I = \infty$, mais pas de PGD fort, sinon il existerait un point x tel que $I(x) = 0$.

C'est d'ailleurs une conjecture que, pour les chaînes de Markov à espace d'états dénombrable, pour lesquelles le principe faible est toujours valable avec une unique fonctionnelle d'action H, le PGD fort serait équivalent à ce que H soit bonne et que son minimum soit 0. Pour ce lien entre PGD faibles et forts, nous renvoyons aussi à la remarque « Limite projective et compactification » p. 39.

♦♦♦

3.2 Techniques de preuve d'existence

Les principales techniques permettant de prouver un PGD ont été introduites dans le chapitre 2. Certains outils conviennent plus particulièrement à l'une ou l'autre des bornes, et par conséquent, on peut être amené à utiliser, suivant les circonstances, plusieurs techniques simultanément. Le raffinement qui leur a été apporté et le champ des applications sont considérables, mais il n'y a guère d'autres outils que ceux mentionnés dans cette section. Nous omettons les outils combinatoires, qui sont extrêmement spécifiques.

3.2.1 Changement de mesure

C'est la méthode employée dans les premiers articles de Donsker et Varadhan [47] pour obtenir la borne inférieure, dans le cas markovien. Cette section est la formalisation de la démarche entreprise dans tous les articles de la deuxième partie.

Il s'agit probablement de l'approche la plus intuitive, dans le sens où l'on décrit quel est le *comportement déviant* du processus, qui est une information plus riche que les grandes déviations (ce serait le niveau 3 et nous limitons au niveau 2). Mais cela signifie aussi qu'on puisse décrire relativement simplement la loi du processus. Pour des variables i.i.d, c'est assez simple car le comportement déviant est aussi i.i.d ; pour une chaîne de Markov, il est markovien et on se doute que le même phénomène se reproduit pour de nombreuses classes de processus.

Cependant, à mesure que la description du processus déviant se complique, il devient de plus en plus difficile d'opérer un *changement de variable.* Le cas le plus général est actuellement celui des processus semi-markoviens (décrits en section 6.1.1) mais l'essentiel de la technique est déjà discernable dans le cas markovien (voir chapitre 4). On présente maintenant ce que donnerait le changement de mesure dans un cadre absolument général.

On suppose que le processus X_n appartient à une certaine classe de processus $\mathcal{G}$, dans laquelle chaque loi est entièrement décrite par un point $G \in \mathcal{G}$. La loi de X_n est notée R (c'est la loi de référence). On suppose aussi que $\mathcal{G}$ est muni d'une topologie qui le rende polonais.

Ensuite, il faut construire une statistique $G_n \in \mathcal{G}$, $\mathcal{F}_n$–mesurable, en notant $\mathcal{F}_n$ la filtration canonique engendrée par X_n. Cette statistique G_n représente toute l'information pertinente sur $\{X_i,\ i \leq n\}$ pour la classe considérée et elle converge vers R.

> **Exemple :** Si $X_n \in E$ sont des variables i.i.d, $\mathcal{G} = \mathcal{M}_1(E)$ est l'ensemble des probabilités sur E, et G_n est la mesure empirique L_n (voir chapitre 2). Si X_n est une chaîne de Markov, $\mathcal{G} = \mathcal{M}_s(E^2)$ est l'ensemble des mesures balancées et G_n est la mesure empirique des paires L_n (voir chapitre 4). Dans le cas du polling, $\mathcal{G}$ est un ensemble plus compliqué, décrivant les paramètres (arrivées, services, routage et temps de transfert) et G_n est une mesure empirique formée à partir de ces paramètres (voir chapitre II 4).
>
> ♦♦♦

Pour tout $G \in \mathcal{G}$, on doit construire une fonction $\phi_{R,G} : \mathcal{G} \to \mathbb{R}$ telle que

$$M_n = e^{\phi_{R,G}(G_n)} \tag{3.8}$$

soit une martingale. Le changement de mesure proprement dit consiste à définir la loi

$$\mathbb{P}^*[A_n] \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbb{1}_{\{A_n\}} M_n\right], \quad \forall A_n \in \mathcal{F}_n. \tag{3.9}$$

En outre, $\phi_{R,G}$ doit être construite de telle sorte que *la loi sous* $\mathbb{P}^*$ *du processus* X_n *appartienne à la classe de processus* $\mathcal{G}$ *et soit caractérisée par* G. On en déduit que G_n tend vers G en probabilité sous la loi $\mathbb{P}^*$. Pour un voisinage O de G,

$$\mathbb{P}\left[G_n \in O\right] = \mathbb{E}^*\left[\mathbb{I}_{\{G_n \in O\}} e^{-\phi_{R,G}(G_n)}\right]. \tag{3.10}$$

L'entropie de G relativement à R est définie par

$$H(G\|R) \stackrel{\text{def}}{=} \phi_{R,G}(G). \tag{3.11}$$

Supposons que $\phi_{R,G}$ soit continue. Puisque G_n tend vers G sous la loi $\mathbb{P}^*$, la probabilité pour G_n d'appartenir à O tend vers 1 et on en déduit les bornes locales des grandes déviations :

$$\inf_{O \ni G} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[G_n \in O\right] = -H(G\|R), \tag{3.12}$$

$$\inf_{O \ni G} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[G_n \in O\right] = -H(G\|R). \tag{3.13}$$

Le PGD faible s'en déduit classiquement.

Ce schéma de preuve est celui des variables i.i.d. (voir section 2.2.4) et des processus de Markov (voir section 4.2.2). Il est tout aussi efficace en temps discret, bien qu'une autre technique ait été utilisée (voir section 4.1). Nous avons simplement généralisé la procédure, ce qui appelle quelques commentaires. Les suppositions sont en effet nombreuses et la question de la concrétisation de la démarche se pose légitimement.

Dans les cas que nous avons traités (variables i.i.d, chaînes de Markov en temps discret et continu), l'entropie est explicite et significative ; elle révèle un rapport étroit avec la théorie de l'information (voir sections 2.2.2, 4.1.2 et 4.2.3). Cette entropie se décompose en une somme d'informations élémentaires qui correspondent à autant de modifications aboutissant finalement à la modification globale du processus, c'est-à-dire à la construction de la fonction $\phi_{R,G}$, et donc de la martingale et du changement de mesure ayant les propriétés requises. À partir de ce point, le reste découle naturellement.

Nous avons supposé que $\phi_{R,G}$ était continue pour tout $G \in \mathcal{G}$. Le cas des processus de Markov (section 4.2.2) montre que ce n'est pas nécessaire ; si l'ensemble des G tels que $\phi_{R,G}$ est continue est dense dans $\mathcal{G}$, alors on peut procéder exactement comme si $\phi_{R,G}$ l'était toujours. Cependant, la preuve de cette propriété repose sur une étude fine de l'entropie et c'est là où son caractère explicite et sa décomposition concourent à la réussite de la démarche.

En outre n'avons-nous pas encore mentionné la construction de la statistique G_n. Dans les cas généraux, par exemple les variables i.i.d. ou les chaînes de Markov, il est relativement aisée de la trouver, en recourant aux mesures empiriques ; mais dans des cas plus spécifiques, comme les réseaux à polling (voir [35] ou le chapitre II 4), c'est une partie déterminante du travail que de bien choisir la représentation $\mathcal{G}$; ces exemples d'applications sont de surcroît toujours markoviens, grâce à une description *ad hoc*, et nous ne savons guère ce qui remplacerait les mesures empiriques pour des processus plus généraux. Insistons sur le fait que la correspondance entre les éléments de $\mathcal{G}$ et les lois des processus doit être *bijective*. Si la description, via G_n, est trop détaillée, la classe risque d'être trop vaste et de bonnes propriétés peuvent être perdues ; à l'opposé, si cette description n'est pas assez précise, certains changements de mesure ne pourront être opérés, et les bornes locales ne pourront être dérivées en tout point de $\mathcal{G}$.

> **Exemple :** Dans le cas du polling (voir chapitre II 4), on aurait pu mesurer plus précisément l'évolution du processus ; une idée naturelle, par analogie avec les chaînes de Markov, est de mesurer la fréquence de toutes les *transitions* et la durée moyenne de chacune d'entre elle — au lieu de mesurer la fréquence et la durée de *passage* en chaque point. Cependant on obtient dans ce cas une classe de polling beaucoup plus large que « prévu », où les paramètres (durée de services, temps de transfert) dépendent des points de départ *et d'arrivée* de la transition.

♦♦♦

De la construction de la statistique comme une mesure, dans les cas précités, dérivent encore de bonnes propriétés. Tout d'abord l'espace $\mathcal{G}$ est naturellement un espace polonais. Les problèmes topologiques disparaissent donc. Ensuite, si la statistique est une fonctionnelle additive du processus, l'entropie est convexe (*cf.* section 3.2.3). En fait il suffit que l'on puisse écrire à la limite l'équation (3.18) avec G_{n+m}, pour n et m grands, ce qui ne requiert qu'une forme de sous-additivité de G_n.

Enfin, si l'information G_n complète ne nous intéresse pas, mais seulement une partie L_n, ce que nous formalisons par $L_n = f(G_n)$, alors il faut dans une étape supplémentaire contracter le PGD faible, ce qui exige quelques conditions supplémentaires (théorème II 3.18). C'est ce qui est fait pour les polling (voir chapitre II 4), où seules les longueurs des queues sont intéressantes.

Nous avons beaucoup insisté, malgré les hypothèses, sur cette méthode car c'est sur elle qu'a porté une grande partie du travail de la thèse, et qu'il est donc ici important de bien la situer, à la fois dans sa généralité, ses restrictions et les applications possibles. Quant à ce dernier point, outre les théorèmes classiques (contraction, lemme de Varadhan : voir [32, 34] ou les chapitres II) 2 et II) 3, un premier axe de recherche est d'établir un PGD trajectoriel fort pour des réseaux (voir [35] ou Ch. II 4 pour les réseaux à polling), et d'en déduire les queues de la distribution stationnaire. Une deuxième possibilité est de calculer des équivalents exacts, et non des équivalents logarithmiques comme dans les grandes déviations (voir section 6.1.2). La troisième voie explorée au chapitre 5 consiste à transformer directement le PGD faible en PGD fort sur un espace étendu. En fait, au-delà de l'utilisation de la même technique, ces trois voies ne sont pas divergentes et elles comportent des similarités que nous essayons d'expliciter et d'utiliser.

3.2.2 Analyse convexe

C'est une généralisation de la transformée de Legendre (voir section 2.1.1). Cet axe de preuve a été largement exploré [5, 9, 41, 61, 70, 129]. L'analyse convexe est en effet d'autant plus intéressante que l'on peut montrer que, si un PGD est valide et si la fonctionnelle d'action est convexe (voir proposition 3.10), alors nécessairement c'est la transformée de Legendre de la fonction des cumulants. Ces idées sont très liées avec le principe de Laplace et son extension dans le cadre des grandes déviations qui est la base du lemme de Varadhan (théorème 3.20).

On considère une suite de variables X_n appartenant à un espace polonais E. On note E^* son dual, c'est-à-dire l'ensemble des fonctions linéaires continues bornées de E dans $\mathbb{R}$. Pour une fonction $f : E \mapsto \mathbb{R}$, on note $f^* : E^* \mapsto \mathbb{R}$ sa transformée de Fenchel-Legendre (voir définition 2.3). En combinant la définition 2.2 et l'équation (2.2), on introduit la fonction des cumulants limite, si elle existe :

$$\Lambda(f) \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle f, X_n\rangle}\right] \in [-\infty, \infty], \qquad \forall f \in E^*. \tag{3.14}$$

Lemme 3.8 *Si $\Lambda(f)$ existe pour tout $f \in E^*$, alors Λ^* est une fonctionnelle d'action convexe et pour tout ensemble compact $K \subset E$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in K\right] \leq - \inf_{x\in K} \Lambda^*(x).$$

Preuve : La convexité découle de la définition de Λ^* comme transformée de Fenchel-Legendre. De plus $\Lambda(0) = 0$, et donc $\Lambda^* \geq 0$. Par ailleurs Λ^* est la borne supérieure de fonctions continues (voir (2.1)), et par conséquence Λ^* est semi-continue inférieurement.

Pour la borne supérieure, on procède comme pour les équations (2.3) et (2.4). Considérons un ouvert O ; on obtient la minoration

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\hat{S}_n \in O\right] \leq - \inf_{y\in O} \Lambda^*(y).$$

Sachant que Λ^* est semi-continue inférieurement, la borne inférieure dans la partie droite tend vers $\Lambda^*(x)$ quand $O \ni x$ devient petit (voir équation (3.3)); le résultat local suivant est démontré :

$$\inf_{O \ni x} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right] \leq -\Lambda^*(x), \quad \forall x \in E. \tag{3.15}$$

Cette borne locale s'étend aux compacts (voir section 2.1.2). ■

Dans le cas suivant, particulier mais intéressant, on obtient non seulement la borne inférieure, mais aussi le PGD fort. On note $\mathcal{D}_\Lambda$ le domaine où Λ est finie, et $\mathcal{D}_\Lambda^o$ son intérieur.

Théorème 3.9 (Gärtner-Ellis) *On suppose que $E = \mathbb{R}^d$. Si Λ est partout définie, si $0 \in \mathcal{D}_\Lambda^o$, si elle est différentiable sur $\mathcal{D}_\Lambda^o$ et que la différentielle tend vers l'infini vers le bord de $\mathcal{D}_\Lambda^o$, alors X_n vérifie un* PGD *fort avec la bonne fonctionnelle d'action convexe Λ^*.*

Preuve : La preuve de la tension exponentielle est en fait assez facile. C'est l'obtention de la borne inférieure qui pose le plus de difficulté, car aucun changement de mesure n'est vraiment naturel. Une analyse convexe poussée est utilisée, pour laquelle nous renvoyons à l'excellent ouvrage [38, p. 45]. ■

Nous donnons maintenant sans preuve un théorème qui corrobore, si besoin était, l'intérêt de l'analyse convexe [38, théorème 4.5.10].

Proposition 3.10 *Supposons que X_n vérifie un* PGD *avec la bonne fonctionnelle d'action I, et que*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle f, X_n\rangle}\right] < \infty \qquad \forall f \in E^*.$$

Alors Λ existe, est finie et satisfait à

$$\Lambda(f) = \sup_{x\in E} \left\{\langle f, x\rangle - I(x)\right\}, \quad \forall f \in E^*.$$

De plus Λ^* est la régularisation affine[3] *de I, et est donc égale à I si cette dernière est convexe.*

3. C'est-à-dire le maximum des fonctions convexes s.c.i. inférieures à I ou de façon équivalente l'enveloppe supérieure des fonctions affines continues, inférieures à I en tout point, d'où le nom.

3.2.3 Sous-additivité

Lorsque les variables pour lesquelles on souhaite prouver un PGD sont des fonctionnelles additives d'un processus assez régulier, la méthode de sous-additivité permet d'établir l'existence d'un PGD. C'est une méthode ancienne (en lien avec les fonctions de Ruelle-Lanford, voir [95, 125, 126]) mais puissante, qui a connu un fort développement ces dernières années ([9] puis, dans le contexte des chaînes de Markov [28, 31, 43, 87]).

On considère une fonctionnelle additive L_n construite sur le processus X_n qui se régénère, c'est-à-dire que lorsque X_n repasse par un état x_0 (au temps τ), la loi du processus à venir $\{X_n,\ n \geq \tau\}$ est la même que celle de $\{X_n,\ n \geq 0\}$. Cette technique est aisément adaptée aux chaînes de Markov, par exemple. Une seconde manière de procéder est de supposer une certaine uniformité dans la loi par rapport à la condition initiale.

On suppose que L_n appartient à un espace polonais[4] $(F, \mathcal{F})$ et qu'il existe une fonction $f:\ E \mapsto F$ telle que

$$L_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Il existe deux possibilités pour utiliser la sous-additivité. La première si la loi $\mathbb{P}_x$ du processus partant de x est assez uniforme en x ; on introduit la quantité

$$\mu_n(A) \stackrel{\text{def}}{=} \inf_{x \in E} \mathbb{P}_x \left[L_n \in A\right], \quad \forall A \in \mathcal{F}; \tag{3.16}$$

La deuxième utilise la mesure

$$\mu_n(A) \stackrel{\text{def}}{=} \mathbb{P}_{x_0} \left[L_n \in A,\ X_n = x_0\right], \quad \forall A \in \mathcal{F}. \tag{3.17}$$

Dans ces deux cas, la mesure μ_n est sous-stochastique.

Lemme 3.11 *Pour tout ensemble convexe $C \in \mathcal{F}$, $\mu_n(C)$ est sur-multiplicative et donc $-\log \mu_n(C)$ est sous-additive.*

Preuve : On vérifie, exactement comme à la section 2.1.3 que

$$\mu_{n+m}(C) \geq \mu_n(C)\mu_m(C),$$

4. Le minimum requis est que F soit un vectoriel topologique $(F, \mathcal{F})$ localement convexe et de Hausdorff.

qui est l'équivalent de l'équation (2.6). Cette inégalité est obtenue en écrivant

$$L_{n+m} = \frac{n}{n+m} L_n + \frac{m}{n+m} \left(\frac{1}{m} \sum_{i=n+1}^{n+m} f(X_i) \right). \tag{3.18}$$

L'uniformité (3.16) de μ_n dans le premier cas et la propriété de renouvellement de X_n dans le deuxième (3.17) permettent, par conditionnement, de séparer les deux termes de droite et d'identifier le second terme entre parenthèse à L_m. Puisque C est convexe, l'appartenance de L_n et L_m à C implique que L_{n+m} y appartient aussi. Le lemme est ainsi prouvé. ∎

La suite de la construction de la fonctionnelle d'action est absolument similaire à celle donnée dans la section 2.1.3. Nous résumons ce travail dans la proposition 3.12, en soulignant la propriété de convexité, fort utile dans les applications. Exceptionnellement, nous exprimons un PGD relatif à une suite de mesures, pour condenser la formulation.

Proposition 3.12 *Soit L_n est une fonctionnelle additive sur un processus X_n. La mesure μ_n définie en (3.16) vérifie un* PGD *faible avec une fonctionnelle d'action convexe. Il en est de même pour μ_n définie en (3.17), à condition de supposer de plus que X_n est un processus de renouvellement.*

Remarque : Notons que dans le cas (3.16), on obtient un PGD faible, *uniforme en les conditions initiales.* C'est un résultat très fort et général. Cependant, le lemme 2.4 nuance ce jugement, car il est tout à fait possible que μ_n soit uniformément nulle, auquel cas le PGD est trivial. C'est ce qui se passe dans bon nombre d'applications.

En fait, c'est par l'intermédiaire de (3.17) que des progrès ont été accomplis dans les cas non-homogènes. On montre (par exemple dans [43]) comment étendre la borne supérieure aux fermés convexes, puis aux fermés. Néanmoins, cela fait appel à l'identification de la fonctionnelle d'action par des moyens [42] beaucoup plus sophistiqués, et il parait encore impossible de se débarrasser du conditionnement $\{X_n = x_0\}$. Ce terme, semble-t-il, impose que la chaîne reste au voisinage de x_0, et comme on le verra au chapitre 5, cela exclut des comportements qui emmènent la chaîne à l'infini.

♦♦♦

3.2.4 Méthodes d'approximation

Dans un certain nombre de situations, il est plus simple de démontrer un PGD pour un processus *voisin*. Si les deux processus sont suffisamment proches, ils satisfont — ou non — toujours *simultanément* un PGD. Le but de cette section est de donner un critère à cette notion de proximité. Nous ne donnerons aucune démonstration, car cette méthode ne sera pas utilisée dans la suite, mais il faut savoir qu'elle existe pour un certain nombre d'applications, où l'on perturbe un processus simple pour en étudier un autre.

On se place sur un espace métrique (E, d), où l'on considère deux suites de variables aléatoires X_n et Y_n, qui suivent les lois $\mathbb{P}_x$ et $\mathbb{P}_y$. On introduit une variante de *couplage*, c'est-à-dire une loi jointe $\mathbb{P}$ sur (X_n, Y_n), dont les marginales sont $\mathbb{P}_x$ et $\mathbb{P}_y$:

$$\mathbb{P}\left[X_n \in A\right] = \mathbb{P}_x\left[X_n \in A\right], \quad \text{et} \quad \mathbb{P}\left[Y_n \in A\right] = \mathbb{P}_y\left[Y_n \in A\right].$$

Pour mesurer la distance entre X_n et Y_n, on définit la diagonale, que l'on suppose mesurable,

$$\Delta_\varepsilon \stackrel{\text{def}}{=} \left\{(x, y) \in E^2 : \ d(x, y) \le \varepsilon\right\}.$$

Définition 3.13 *X_n et Y_n sont exponentiellement équivalentes si,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[(X_n, Y_n) \in \Delta_\varepsilon^c\right] = -\infty, \quad \forall \varepsilon > 0. \tag{3.19}$$

Cette notion d'équivalence exponentielle permet de faire passer le PGD d'une variable à l'autre.

Théorème 3.14 *Si X_n vérifie un* PGD *fort avec une bonne fonctionnelle d'action et si Y_n est exponentiellement équivalente à X_n, alors Y_n vérifie un* PGD *fort avec la même fonctionnelle d'action.*

> **Remarque :** On voit l'analogie entre la condition (3.19) et la tension exponentielle. Cette condition traduit un couplage extrêmement fort entre les variables aléatoires. Comme il n'est pas toujours commode de construire un tel couplage, on peut remplacer cette condition par une série de conditions similaires, mais avec une inégalité inférieure et une série tendant vers $-\infty$ à droite, dans (3.19).
>
> Enfin, il est possible d'adapter les conditions et le théorème pour inclure des approximations dans le principe de contraction, en remplaçant l'unique fonction f par une suite f_m. On remarque d'ailleurs que, comme pour le principe de contraction, un *principe fort avec une bonne fonctionnelle d'action* est requis ici (voir remarque page 41).

♦♦♦

3.2.5 Les limites projectives

Pour formaliser l'agrégation des états, les limites projectives, recourent à la notion de suite d'espaces E_n et de projections $\phi_{n,p}$ entre espaces. On peut en effet considérer que les espaces E_n représentent des partitions de plus en plus fines et que la projection $\phi_{n,p}$ explicite les points de E_n qui se rassemblent en un seul point de E_p.

Nous utilisons une variante de cette technique pour démontrer le PGD faible pour les chaînes de Markov en temps discret (voir chapitre II 2). Le point fort de la méthode est de dériver un PGD sur un espace infini, à partir d'une suite de PGD sur des espaces *finis*, pour lesquels la démonstration est simple. Le revers, c'est que l'on doit bien contrôler les variations de la suite de fonctionnelles d'action.

Formellement, soit $\{(E_n, \mathcal{E}_n), n \geq 1\}$ une suite d'espaces topologiques, et $\{\phi_{n,p}, n \geq p \geq 1\}$ des surjections[5] de E_n sur E_p. On les appelle projections si elles sont continues et vérifient la condition de transitivité :

$$\phi_{n,p} \circ \phi_{p,q} = \phi_{n,q}, \qquad \forall n \geq p \geq q.$$

La limite projective $E = \lim E_n$ est l'ensemble des suites $x = (x_1, x_2, \dots)$ telles que $x_n \in E_n$ pour tout $n \geq 1$ et $\phi_{n,p}(x_n) = x_p$ pour tout $n \geq p \geq 1$. La topologie est induite par la topologie de l'espace produit $E_1 \times E_2 \times \dots$, dont l'ensemble $\{\phi_n^{-1}(O_n),\ O_n \text{ ouvert de } E_n\}$ est une base. On note par ϕ_n la projection canonique de E dans E_n. La limite projective des fermés s'exprime simplement :

$$F = \lim_n F_n = \bigcap_{n \geq 1} \phi_n^{-1}(F_n).$$

Dans ce formalisme, l'agrégation des états donne lieu au théorème suivant.

Théorème 3.15 *Soit E la limite projective des espaces E_p munis des projections $\phi_{n,p}$ et X_n une variable aléatoire sur E. Si pour tout p, $\phi_p(X_n)$ vérifie un* PGD *fort avec la bonne fonctionnelle d'action I_p, alors c'est encore vrai pour X_n avec*

$$I(x) \stackrel{\text{def}}{=} \sup_{p \geq 1} I_p(\phi_p(x)), \qquad \forall x \in E.$$

5. L'hypothèse de surjectivité ne réduit pas la généralité. On peut choisir des fonctions, et réduire les espaces pour qu'elles deviennent des surjections, mais une surjection correspond bien à la vision d'agrégation : deux points $x, y \in E_n$ sont agrégés dans E_p si $\phi_{n,p}(x) = \phi_{n,p}(y)$.

Preuve : I est positive et semi-continue inférieurement car elle est la limite de telles fonctions. Le principe de contraction (théorème 3.16) implique que les fonctionnelles I_p sont croissantes :

$$I_p(x_p) = \inf_{x_n \in \phi_{n,p}^{-1}(x_p)} I_n(x_n).$$

Ceci prouve également que les niveaux $\{x_p \in E_p : \; I_p(x_p) \leq C\}$ des I_p sont décroissants, donc que le niveau correspondant de I en est la limite projective. Il est compact, puisque le produit d'espaces compacts est compact (théorème de Tychonoff). Ainsi I est une bonne fonctionnelle d'action.

Considérons un ensemble ouvert $O \subset E$ et un point $x \in O$. Comme $\{\phi_p^{-1}(O_p), \; O_p$ ouvert de $E_p\}$ est une base d'ouverts, il existe un ouvert O_p de E_p tel que $\phi_p(x)$ appartienne à O_p et que $\phi_p^{-1}(O_p)$ soit inclus dans O. On en déduit

$$-I_p(\phi_p(x)) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[\phi_p(X_n) \in O_p\right] \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in O\right],$$

et par conséquent la borne inférieure est vérifiée.

Soit $F \subset E$ un ensemble fermé. Par transitivité des projections, la limite limite projective des ensembles fermés $\overline{\phi_p(F)}$ existe, et puisque c'est un ensemble fermé,

$$F \subset \lim \overline{\phi_p(F)} \subset \overline{F} = F.$$

Puisque les niveaux de I sont eux-mêmes des limites projectives,

$$F \cap \{x : \; I(x) \leq C\} = \lim \left(\overline{\phi_p(F)} \cap \{x_p : \; I_p(x_p) \leq C\} \right).$$

Tous ces ensembles sont compacts et la limite projective de compacts non-vides n'est pas vide ; donc, si $I(x) > C$ pour tout $x \in F$, puisqu'alors l'intersection de F et de $\{x : \; I(x) \leq C\}$ est vide, il existe p tel que $I_p(x_p) > C$ pour tout $x_p \in \overline{\phi_p(F)}$. On en déduit que

$$\sup_{p\geq 1} \inf_{x_p \in \overline{\phi_p(F)}} I_p(x_p) \geq \inf_{x\in F} I(x).$$

D'autre part, $x \in F$, implique que $\phi_p(x)$ appartienne à $\overline{\phi_p(F)}$. Comme cette implication est valide pour tout $p \geq 1$, la borne supérieure pour $\phi_p(X_n)$ donne

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[X_n \in F\right] \leq -\sup_{p\geq 1} \inf_{x_p \in \overline{\phi_p(F)}} I_p(x_p) \leq -\inf_{x\in F} I(x).$$

Cette borne supérieure achève la preuve du PGD fort pour X_n. ■

Ce théorème est confectionné sur mesure pour les variables indépendantes. En effet, si $\{X_i,\ i \geq 1\}$ est une suite de variables indépendantes de loi μ et si $\phi:\ E \mapsto E'$ est une application mesurable, alors les variables aléatoires $\phi(X_i)$ sont aussi indépendantes, de loi $\mu \circ \phi^{-1}$. Or, si E' est compact, le principe faible de grandes déviations pour la mesure ou la distribution empirique est automatiquement fort, puisque tous les fermés sont compacts.

Limite projective et compactification : Un espace polonais est une union dénombrable de compacts[6] :

$$E = \bigcup_{n \geq 1} K_n.$$

On peut supposer, sans restreindre la généralité, que ces compacts sont emboîtés, au sens où $K_n \subset K_{n+1}$. Une suite de projections est définie sur $E_p \stackrel{\text{def}}{=} K_p \cup \{\Delta\}$ en posant :

$$\phi_{n,p}(x) \stackrel{\text{def}}{=} \begin{cases} x & \text{si } x \in K_p, \\ \Delta & \text{si } x \notin K_p. \end{cases}$$

Le point Δ s'interprète comme l'agrégation des points extérieurs à K_p. Dans ce cas, la limite projective est bien définie, mais l'ensemble limite est $E \cup \{\Delta\}$ où Δ est le point à l'infini d'Alexandroff, et la topologie est celle de la compactification du même nom. C'est cohérent, puisque la limite projective d'ensembles compacts (les E_p le sont) est compacte d'après le théorème de Tychonoff.

La tension exponentielle (voir section 3.1.4 et la remarque p. 18) n'est alors que la démonstration du fait que la fonctionnelle I de grandes déviations, qui existe par le théorème 3.15, est infinie en Δ. On peut ainsi considérer les ensembles $F \setminus \Delta$ pour les grandes déviations, qui sont les ensembles de la topologie initiale sur E. Plus précisément, pour la mesure empirique, la fonctionnelle $I(\nu)$ est infinie dès que Δ appartient au support de $\nu \in \mathcal{M}_1(E \cup \{\Delta\})$, et on peut en fait se restreindre à $\nu \in \mathcal{M}_1(E)$.

Finalement, cette présentation s'interprète ainsi : lorsqu'un principe fort est valide, l'espace peut être compactifié en conservant le PGD et avec $I(\Delta) = \infty$. Un principe faible (et seulement faible) serait l'inverse : $I(\Delta) < \infty$. La question est alors de savoir si on peut compactifier l'espace (et donc traiter les ensembles fermés) en choisissant un bonne valeur pour $I(\Delta)$. La réponse est en général négative. La compactification d'Alexandroff n'est pas la seule possible, et il faut

6. C'est la base de la démonstration du théorème qui stipule que toute mesure est tendue sur un espace polonais.

en choisir une qui respecte la structure du problème. Une conjecture, crédible dans ce contexte, est que c'est la compactification de Martin qui est la « bonne » compactification pour les chaînes de Markov (c'est le propos du chapitre 5).

♦♦♦

3.3 Propriétés générales

Les deux propriétés suivantes, le principe de contraction et le lemme de Varadhan sont d'une très grande importance pratique. Grâce au premier, un PGD pour X_n implique un PGD pour toute image $f(X_n)$, avec une fonctionnelle d'action presque explicite. Le second conduit à de nombreux résultats, p. ex. en mécanique statistique.

3.3.1 Le principe de contraction

Le principe de contraction est à nos yeux le cœur de cette maxime que nous avons fait nôtre, que mieux vaut un résultat simple sur un espace compliqué qu'un résultat compliqué sur un espace simple : on prouve « naturellement » un PGD pour une fonctionnelle compliquée du processus (p. ex, voir la section 3.2.1) et on le contracte pour obtenir le PGD pour une fonctionnelle plus simple, pour laquelle une démonstration directe équivaudrait souvent à une contraction déguisée.

Théorème 3.16 (Contraction) *Soit $(E, \mathcal{E})$ et $(F, \mathcal{F})$ deux espaces polonais*[7] *et $f : E \mapsto F$ une fonction continue. Si $I : E \mapsto \mathbb{R}_+$ est une bonne fonctionnelle d'action, alors*

$$J(y) \stackrel{\text{def}}{=} \inf_{x \in f^{-1}(y)} I(x), \qquad \forall y \in F \tag{3.20}$$

l'est aussi. Si de plus $X_n \in E$ vérifie un PGD fort avec I, alors $f(X_n)$ vérifie aussi un PGD fort, avec J.

Preuve : Soit $C < \infty$. Puisque les niveaux de I sont compacts, la borne inférieure est atteinte dans (3.20). On en déduit que les niveaux de J sont l'image des niveaux de I :

$$\big\{y \in F : J(y) \leq C\big\} = f\Big(\big\{x \in E : I(x) \leq C\big\}\Big).$$

7. Il faut au minimum que $(E, \mathcal{E})$ et $(F, \mathcal{F})$ soient des espaces de Hausdorff réguliers.

Comme f est continue, ces niveaux sont compacts, et puisqu'ils sont fermés, J est semi-continue inférieurement. Finalement J est une bonne fonctionnelle d'action.

Par construction, pour tout ensemble $A \in \mathcal{F}$,

$$\inf_{y \in A} J(y) = \inf_{x \in f^{-1}(A)} I(x).$$

Puisque f est continue, si A est ouvert ou fermé, $f^{-1}(A)$ l'est aussi, d'où on déduit le PGD fort pour $f(X_n)$ avec la bonne fonctionnelle d'action J. ■

> **Remarque :** On le voit dans cette preuve, la seule difficulté est de prouver la semi-continuité inférieure de J. Dans ce contexte classique, les conditions sont drastiques puisqu'il faut une fonction f *continue* et un PGD *fort avec une bonne fonctionnelle d'action* pour que le PGD passe à $f(X_n)$.
>
> Il est naturel d'exiger la continuité pour f. En revanche, et c'est un des buts de [32] (voir aussi la section II 3.4.3), on peut formuler autrement les conditions pour obtenir que J soit semi-continue inférieurement. Il n'est notamment pas nécessaire d'avoir un principe fort ni une bonne fonctionnelle d'action, mais il suffit du PGD faible et d'une version partielle de la tension exponentielle, pour obtenir la borne supérieure sur les fermés, ainsi qu'une certaine compacité de la fonctionnelle d'action I, pour assurer la semi-continuité inférieure de J (ce sont les hypothèses (H1) et (H2) de la section II 3.4). On obtient dans ce cas un PGD faible pour $f(X_n)$.

♦♦♦

Il est assez étonnant que peu de travaux aient été consacrés à l'étude de la transmission des propriétés de la fonctionnelle d'action par contraction. C'est une tâche à laquelle nous nous sommes attelé dans [32] (repris au chapitre II 3), et dont nous exprimons la substance ici.

Nous nous plaçons maintenant sous les hypothèses (H1) et (H2) de la section II 3.4, qui expriment les conditions sous lesquelles un PGD faible se contracte. On suppose dans la suite que f est continue, que X_n vérifie un PGD avec la fonctionnelle d'action I, et de même $f(X_n)$ avec J.

Proposition 3.17 *Si I est une fonction de Ruelle-Lanford pour X_n, alors J l'est aussi pour $f(X_n)$.*

Proposition 3.18 *Supposons que $F = f(E)$. Si B est un ensemble dense dans E, alors $f(B)$ est dense dans F. Si I est la plus petite fonction semi-continue inférieurement qui coïncide avec I sur B, alors J est la plus petite fonction semi-continue inférieurement qui coïncide avec J sur $f(B)$.*

Proposition 3.19 *Si I est convexe et si f est affine, alors J est convexe.*

> **La convexité :** Il est essentiel, en vue des applications notamment, mais théoriquement aussi, de noter que l'*optimisation* exprimant la fonctionnelle du processus contracté est parfois un *programme convexe.* C'est le cas sous les hypothèses de la proposition 3.19. La complexité algorithmique se réduit alors drastiquement (voir la remarque « Complexité algorithmique » page 66). On peut aussi prouver certaines propriétés supplémentaires (voir proposition II 4.21) qui, dans le cas des PGD faibles, sont quelquefois déterminantes.

♦♦♦

3.3.2 Le lemme de Varadhan

Le lemme de Varadhan est une extension de la méthode de Laplace, qui indique naturellement comment obtenir un développement d'une intégrale du type

$$\int e^{nf(x)}\,\mathrm{d}x.$$

Supposons que la fonction f soit assez régulière et ait un unique maximum, que l'on décale en 0. Alors, en développant au voisinage de 0, $f'(0) = 0$ et $f''(0) < 0$ donc

$$\int e^{nf(x)}\,\mathrm{d}x \approx e^{nf(0)} \int e^{nf''(0)\frac{x^2}{2}}\,\mathrm{d}x = \sqrt{\frac{2\pi}{n|f''(0)|}}e^{nf(0)}. \tag{3.21}$$

La théorie des grandes déviations complique un peu le problème en considérant une mesure dont les variations sont localement exponentielles, mais sans prendre en compte le coefficient en $1/\sqrt{n}$.

Théorème 3.20 *Si X_n vérifie un* PGD *avec la bonne fonctionnelle d'action I, si $f:\ E \mapsto \mathbb{R}$ est une fonction continue qui remplit la condition*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\mathbb{1}_{\{f(X_n)\geq C\}} e^{nf(X_n)}\right] = -\infty, \tag{3.22}$$

alors

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{nf(X_n)}\right] = \sup_{x\in E} \big\{f(x) - I(x)\big\}. \tag{3.23}$$

Preuve : L'intuition de la preuve est la suivante : le taux de décroissance de la probabilité que X_n soit dans un voisinage de x est $-I(x)$, mais l'espérance croît avec un taux $f(x)$; on obtient donc le taux local $f(x) - I(x)$. On prouve en fait (3.23) localement ; la condition (3.22) joue le rôle d'une tension exponentielle et permet d'étendre la borne supérieure aux fermés. ■

Remarque : La méthode de Laplace (voir l'équation (3.21)) peut encore s'appliquer dans un certain nombre de situations où l'on souhaite obtenir un équivalent exact (et non pas l'équivalent logarithmique que fournit les grandes déviations). Dans le cas des variables i.i.d, voir p. ex. [14] ou la remarque p. 93. La théorie est cependant rendue difficile, car les processus évoluent naturellement dans des espaces *infinis*. Dans ce cadre, la dérivée seconde doit être remplacée par le hessien, mais on ne peut pas exprimer de densité par rapport à une mesure uniforme, avec un terme gaussien provenant de la dérivée seconde. Il faut donc montrer directement la convergence vers un processus gaussien, ce qui revient à démontrer un théorème central limite. C'est ce que nous esquissons dans la section 6.1.2.

♦♦♦

Chapitre 4

Grandes déviations pour les chaînes de Markov

Point n'est besoin d'insister sur l'importance des chaînes de Markov dans des applications de nature très variée. Les principaux résultats ont été obtenus dans ce cadre, mais on peut les étendre à des processus plus généraux. Toutefois, plutôt que de généraliser les conditions d'applications, nous avons préféré obtenir davantage d'applications, parfois extrêmement simples, convaincu que les méthodes se perfectionneront au fil du temps. En effet, certaines formules sont si naturelles et ont une forme si générale qu'il paraîtrait étrange qu'elles ne s'appliquent qu'elles ne puissent avoir une portée plus large.

Ce chapitre est essentiellement une présentation des résultats de la thèse obtenus dans ce contexte, les sections 4.1–4.3 reprenant les idées des chapitres II 2–II 4. Les applications présentées à la section 4.4 sont des approfondissements des méthodes employées. Le passage du transitoire au stationnaire présenté à la section 4.4.3 est un aspect connu (voir p. ex. [83] dans un cas particulier) mais la présentation que nous donnons semble nouvelle. Ce n'est qu'une esquisse sans preuve, mais la méthode permettrait de résoudre de nouveaux problèmes (p. ex. la décroissance des queues de distribution dans certains systèmes à polling).

4.1 Le théorème de Sanov

L'étude du théorème de Sanov, dans sa version sur les paires, a occupé une place importante dans ce travail de thèse, pour plusieurs raisons. La principale est que la mesure empirique des paires L_n correspond au générateur empirique tel qu'il est défini à la section 3.2.1 : il existe une bijection entre de telles mesures L_n et les générateurs ergodiques. La méthode de changement de mesure permet de parcourir l'ensemble de ces générateurs, d'obtenir en tout point des bornes locales et de les réunir en un PGD faible. Une mesure plus restreinte n'aurait pas permis de parcourir tout l'ensemble et d'obtenir des bornes locales en tout point ; une mesure plus fine aurait conduit à un changement de mesure tel que le processus modifié n'aurait pas conservé la structure markovienne initiale. C'est en ce sens que nous pouvons dire que la mesure empirique des paires est une description exacte des déviations qui conserve la structure markovienne par changement de mesure (voir section 3.2.1). Si nous avions eu affaire à des processus plus compliqués (par exemple semi-markoviens), nous aurions dû employer une version de PGD qui, tout en ressemblant encore au théorème de Sanov, n'en aurait pas moins substantiellement différée (voir section 6.1.1).

Outre les raisons techniques (*cf.* section 2.2.4) qui peuvent faire préférer la méthode de changement de mesure, qui a aussi ses propres limites, ce sont des motifs plutôt théoriques, voire même intuitifs qui nous ont guidés. D'une part la notion d'information suggère une description complète du processus considéré. D'autre part, son synonyme d'entropie évoque irrésistiblement le troisième principe de la thermodynamique, (ou ce que nous avons appelé « principe de moindre information » en avant-propos).

On conçoit donc aisément sous cet angle que les PGD devraient être la règle et qu'il faudrait des propriétés surprenantes pour que l'information liée à un processus ne représente pas — ou mal — ce dernier. Notre but est de *démontrer* la validité de ce propos.

Les preuves apportent cependant plus qu'une simple confirmation de notre intuition. Le générateur empirique, par exemple, est introduit au départ comme un intermédiaire de calcul commode, mais certains encadrements signalent des développements inattendus et il acquiert une place centrale dans nos constructions. Il semble en effet que cette variante du changement de mesure, non seulement s'interprète et se généralise, mais encore approfondissent la connaissance de certains comportements.

4.1.1 La mesure des transitions

L'entropie relative, dans le cas des chaînes de Markov, possède les mêmes propriétés que pour les variables i.i.d. et s'interprète semblablement. Cependant, puisque la structure de chaîne de Markov est plus complexe que celle de variable i.i.d, la mesure empirique L_n l'est également. On caractérise une chaîne de Markov par ses *transitions* et c'est pourquoi on mesure leur fréquence :

$$L_n \stackrel{\text{def}}{=} \frac{1}{n}\left(\sum_{i=1}^{n-1}\delta_{X_iX_{i+1}} + \delta_{X_nX_1}\right) \in \mathcal{M}_1(E^2), \tag{4.1}$$

où δ_{xy} représente la probabilité concentrée sur la paire (x, y). La dernière transition (X_n, X_1) n'est pas indispensable, mais elle fait revenir la chaîne en X_1, ce qui confère à L_n la bonne propriété d'être *balancée*, ce qui signifie que le flux entrant est égal, pour chaque point de E, au flux sortant.

On dit que $A = (a_{xy}) \in \mathcal{M}_1(E^2)$, mesure sur les paires, est balancée si

$$\sum_{y\in E} a_{yx} = \sum_{y\in E} a_{xy} \stackrel{\text{def}}{=} a_x, \qquad \forall x \in E. \tag{4.2}$$

Soit $\mathcal{M}_s(E^2)$ l'ensemble des mesures balancée ; il s'avère être le bon outil pour les chaînes de Markov : a_{xy} représente la fréquence des transitions de x vers y et a_x celle des passages en x. La correspondance entre les matrices de transitions P *ergodiques* et les mesures balancées A est donnée par

$$a_{xy} = \pi_x P_{xy}, \qquad \forall x, y \in E,$$

où π_x est la probabilité stationnaire du point $x \in E$. Il ressort que a_x est bien égal à la mesure invariante π_x. On note $A_{xy} = a_{xy}/a_x$ la probabilité conditionnelle (égale de fait à P_{xy}).

Dans la suite on notera souvent $A = P$, ce qui signifie seulement que les probabilités conditionnelles sont égales : $A_{xy} = P_{xy}$. Contrairement à ce que l'égalité suggère, une telle mesure balancée n'existe pas nécessairement et il n'y a pas unicité[1].

Notons que cette restriction aux chaînes *ergodiques* est naturelle, puisque l'on compte des *fréquences stationnaires* a_{xy}. La mesure d'un comportement transient doit se faire autrement et est beaucoup plus difficile (voir chapitre 5) sans une hypothèse d'uniformité des transitions (comme dans le polling : voir II section 4.3) permettant d'utiliser des analogues de l'ergodicité. En revanche, rien ne s'oppose à ce que la chaîne étudiée soit transiente : dans ce cas, l'analyse est limitée aux comportements ergodiques. On ne peut donc espérer avoir un principe fort dans cette situation.

Puisque la loi du processus X_n « interdit » les transitions de probabilité nulle, nous sommes amenés à restreindre les mesures balancées. Un ensemble de paires $(x, y) \in E^2$ est un *graphe*, que nous noterons Γ, et $\mathcal{M}_s(\Gamma)$ est le

1. La non-unicité est liée au problème d'*irréductibilité* de la chaîne, que l'on rencontrera ultérieurement.

sous-ensemble des mesures balancée $\mathcal{M}_s(E^2)$ dont le support est inclus dans $\Gamma \subset E^2$:

$$A \in \mathcal{M}_s(\Gamma) \iff a_{xy} = 0 \quad \forall (x, y) \notin \Gamma.$$

Par souci de simplicité, nous noterons également P pour le graphe de P, c'est à dire les transitions (x, y) telles que $P_{xy} > 0$.

4.1.2 L'entropie

L'entropie d'une mesure balancée A, au lieu d'être calculée relativement à une mesure, l'est par rapport à une matrice de transition P qui, elle, peut être transiente. Il est plus simple de la supposer irréductible bien que ce ne soit pas strictement nécessaire. L'entropie est définie comme

$$H_d(A\|P) \stackrel{\text{def}}{=} \sum_{x,y \in E} a_{xy} \log \frac{A_{xy}}{P_{xy}}. \tag{4.3}$$

On reconnaît l'information, telle qu'elle a été présentée à la section 2.2.2 ; en notant A_x et P_x les probabilités de transition respectives partant de x :

$$H_d(A\|P) \stackrel{\text{def}}{=} \sum_{x,y \in E} a_x H(A_x \| P_x). \tag{4.4}$$

L'interprétation est directe : le gain d'information apporté par la dynamique A est égal à la moyenne (pondérée par a_x) du gain d'information lié à chaque saut — $H(A_x\|P_x)$ en partant de x. L'entropie est bien exactement le *gain moyen d'information*. Ses propriétés sont nombreuses (voir l'appendice II 2.B) et nous retiendrons les plus utiles dans notre cadre :

Propriétés *L'entropie $H_d(A\|P)$ est*

(i) positive ; nulle si, et seulement si $A = P$;

(ii) convexe et semi-continue inférieurement comme fonction de A ;

(iii) continue sur $\mathcal{M}_s(\Gamma)$, si Γ est un sous-graphe fini de P ;

(iv) pour toute mesure $A \in \mathcal{M}_s(P)$, il existe une suite $A(n)$ de mesures à support fini, qui convergent vers A et dont l'entropie converge vers $H_d(A\|P)$:

$$\lim_{n\to\infty} A(n) = A \quad \textit{et} \quad \lim_{n\to\infty} H_d(A(n)\|P) = H_d(A\|P).$$

Soulignons que ce dernier point *(iv)* n'est pas complètement évident mais est capital pour toutes les démonstrations — aussi bien dans le cadre du temps discret que du temps continu — et pour les applications. On peut aussi voir cette propriété de la manière suivante : H_d est la plus petite fonction semi-continue inférieurement qui vaut $H_d(A\|P)$ sur les mesures à support fini. Cela évite d'avoir à utiliser des fonctions régularisées semi-inférieurement, opération qui apparaît fréquemment lorsqu'on ne connaît pas l'expression de la fonctionnelle d'action (voir la section II 1.3 ou plus précisément la remarque 2 p. 172).

4.1.3 Résultats

Le chapitre II 2 contient des résultats plus détaillés, ainsi que les démonstrations et quelques remarques supplémentaires. Nous condensons en un seul les théorèmes II 2.2 et II 2.12 :

Théorème *Soit X une chaîne de Markov irréductible, de matrice de transition P. La mesure empirique des paires L_n vérifie un* PGD *faible avec la fonctionnelle d'action H_d. La propriété reste vraie si la chaîne est réductible, alors c'est encore vrai avec la fonctionnelle d'action identique à H_d sur un ensemble $\mathcal{M} \subset \mathcal{M}_s(E^2)$ et infinie ailleurs.*

Ce théorème souligne, comme à la section 2.2.2, la notion de comportement déviant (voir la remarque de la page 14) et le principe de moindre information. Contrairement à ce que nous avions souligné jusqu'ici, la démonstration proposée n'est pas fondée sur un changement de mesure, mais sur une méthode similaire aux limites projectives.

On montre dans un premier temps le PGD pour des chaînes à espace d'états *fini*, ce qui est classique. Puis, grâce au point *(iv)* des propriétés (p. 48), on construit des approximations d'une mesure dont le support est *a priori* infini. En remarquant que seule une version *locale* de la propriété de Markov suffit à démontrer des bornes de grandes déviations, on étend le PGD des espaces finis aux espaces dénombrables. Toutefois, la « magnitude » du PGD décroît : le PGD obtenu est faible et on montre que dans certains cas (en fait dans beaucoup de modèles usuels) on ne peut espérer un principe fort.

> **Remarque :** On connaît des conditions suffisantes pour obtenir un PGD fort, par exemple les conditions uniformes suivantes [40] : il existe des entiers $M \geq 1$, N et $0 < l \leq N$ tels que, pour tous les états x, y,
>
> $$P^{(l)}(y, .) \leq \frac{M}{N} \sum_{m=1}^{N} P^{(m)}(x, .). \tag{4.5}$$

Cette condition est bien adaptée à des chaînes sur des espaces compacts, ou bien telles que les transitions la ramènent uniformément sur un compact (de façon semblable à la condition de Döblin). Dans les applications, notamment lorsque les sauts sont bornés, (4.5) n'est pas vérifiée. C'est une des raisons qui nous a incité à développer la théorie des PGD faibles.

♦♦♦

4.2 Temps continu

Dans le cas du temps continu, une méthode similaire à celle employée en temps discret permettrait d'obtenir le théorème de Sanov, mais la démonstration est finalement plus simple en choisissant la méthode de changement de mesure. Pour le reste, il n'y a pas de différence substantielle de nature avec le théorème de Sanov en temps discret. Nous insisterons donc sur le caractère général de la démarche.

4.2.1 Le générateur empirique

Dans le cas du temps discret, une chaîne de Markov est caractérisée par sa distribution initiale ν et sa matrice de transition P. En temps continu, un processus markovien de saut est caractérisé aussi par sa distribution initiale ν et par son générateur $R = (R_{xy})$, où R_{xy} mesure l'intensité des sauts de x vers y. L'analogue de la mesure empirique des paires L_n est donc le *générateur empirique* G_t, que l'on définit de la manière suivante :

- T_n est le n-ième temps de saut (avec $T_0 = 0$) ;
- $N_t \stackrel{\text{def}}{=} \sum_{n\geq 0} \mathbb{1}_{\{T_n \leq t\}}$ est le nombre de sauts au temps t ($N_0 = 1$) ;
- $X_{n+1} \stackrel{\text{def}}{=} Y_{T_n}$ est la chaîne incluse ($Y_0 = X_1$),
- L_t est la mesure empirique de Y_t et $\hat{L}_n$ la mesure empirique des paires du processus inclus X_n,

$$\begin{aligned} L_t &\stackrel{\text{def}}{=} \frac{1}{t}\int_0^t \delta_{Y_s}\,\mathrm{d}s, \\ \hat{L}_n &\stackrel{\text{def}}{=} \frac{1}{n}\left(\sum_{i=1}^{n-1} \delta_{X_i X_{i+1}} + \delta_{X_n X_1}\right); \end{aligned}$$

Alors, le générateur empirique s'écrit

$$G_t \stackrel{\text{def}}{=} \left(L_t, \frac{N_t}{t} \hat{L}_{N_t} \right). \tag{4.6}$$

L'analogie se poursuit avec l'ensemble $\mathcal{G}_s$ des générateurs empiriques :

$$\mathcal{G}_s \stackrel{\text{def}}{=} \mathcal{M}_1(E) \times \mathcal{M}_s(E_*^2).$$

E_*^2 est l'ensemble des paires (x, y) où $x \neq y$. Si on remplace E_*^2 par un sous-graphe Γ, on définit $\mathcal{G}_s(\Gamma)$, l'ensemble des générateurs empiriques dont les transitions sont dans Γ. On notera systématiquement dans la suite $G = (\pi, A)$ où $\pi \in \mathcal{M}_1(E)$ et $A \in \mathcal{M}_s(E_*^2)$. La forme habituelle L_{xy} des générateurs se retrouve grâce à la correspondance

$$L_{xy} = \frac{a_{xy}}{\pi_x} \quad \text{si } x \neq y \qquad \text{et} \quad L_{xx} = -\frac{a_x}{\pi_x}. \tag{4.7}$$

La construction de $\mathcal{G}_s$ est absolument analogue à celle des mesures balancées $\mathcal{M}_s$, pour la bonne raison que les structures (addition, distance, topologie) s'en déduisent. Ces détails sont présentés plus avant dans la section II 3.2.1.

4.2.2 Changement de mesure

Pour obtenir la borne inférieure, dans le cas du temps discret, on introduit un changement de mesure défini par la martingale

$$M_n \stackrel{\text{def}}{=} \exp \sum_{i=1}^{n-1} h(X_i, X_{i+1}), \quad \text{où } h(i,j) = \log \frac{A_{ij}}{P_{ij}}.$$

On procède de même pour le temps continu, avec une martingale à peine plus compliquée :

$$M_t \stackrel{\text{def}}{=} \exp \left\{ \sum_{i=1}^{N_t - 1} h_A(X_i, X_{i+1}) - \int_0^t \Lambda(Y_s)\, \mathrm{d}s \right\},$$

avec la nouvelle fonction

$$h_A(x,y) \stackrel{\text{def}}{=} \begin{cases} \log(L_{xy}/R_{xy}) \in \mathbb{R} \cup \{-\infty\} & \text{si } \ x \in \operatorname{Supp}(G), \\ 0 & \text{sinon}, \end{cases}$$

où L est déduite de $G = (\pi, A)$ par les relations (4.7), Λ étant le compensateur

$$\Lambda(x) \stackrel{\text{def}}{=} \sum_{y \neq x} R_{xy}(e^{h_A(x,y)} - 1).$$

De façon similaire l'équation (3.9) de la section 3.2.1, une nouvelle mesure $\mathbb{P}^*$ est définie sous laquelle le processus Y_t a pour générateur L_{xy} sur le support de G (qui est absorbant) et R_{xy} en dehors. L'entropie apparaît en introduisant la fonctionnelle linéaire $\phi_{R,G}: \mathcal{G}_s \mapsto \mathbb{R}$:

$$\phi_{R,G}(\pi', A') \stackrel{\text{def}}{=} \sum_{x,y \in E} a'_{xy} h_A(x,y) - \sum_{x \in E} \pi'_x \Lambda(x).$$

On obtient alors deux relations importantes, analogues aux équations (3.8) et (3.11) :

$$\phi_{R,G}(G) = H_c(G \| R), \tag{4.8}$$
$$\log M_t \simeq t\phi_{R,G}(G_t). \tag{4.9}$$

La section II 3.3 montre que l'approximation (4.9) peut de fait être traitée comme une égalité. Grâce à la continuité de $\phi_{R,G}$ et aux égalités (4.8) et (4.9), un développement de M_t^{-1} autour du point G (comme en (3.10)) aboutit aux inégalités locales de grandes déviations (similaires à (3.12) et (3.13)).

On obtient en fait une borne inférieure un peu plus forte qui permet de déduire le principal résultat (théorème II 3.10) :

Théorème *Soit Y_t une processus de Markov irréductible, non-explosif, de générateur R sur un espace dénombrable E. Alors l'entropie $H_c(.\|R)$ est la fonction de Ruelle-Lanford pour le générateur empirique G_t.*

Grâce à une version particulière de la tension exponentielle, on en déduit aussi (théorème II 3.12) :

Théorème *Soit Y_t un processus de Markov irréductible, non-explosif, de générateur R sur un espace dénombrable E. Alors le générateur empirique G_t vérifie un* PGD *faible. Si E est fini, le* PGD *est fort.*

4.2.3 Entropie

Définition 4.1 *Soit R un générateur et $G = (\pi, A)$ un générateur empirique. Soit L_{xy} le générateur défini par les relations (4.7). L'entropie de G relativement à R est*

$$H_c(G \| R) \stackrel{\text{def}}{=} \sum_{x \neq y \in E} \pi_x I_p(L_{xy} \| R_{xy}),$$
$$\text{où } I_p(\nu \| \lambda) \stackrel{\text{def}}{=} \nu \log \frac{\nu}{\lambda} - \nu + \lambda.$$

Comme dans le cas des chaînes de Markov, cette entropie s'interprète aisément en terme d'information. On peut même lui associer deux variantes. Mais tout d'abord, considérons la fonction I_p.

$I_p(\nu\|\lambda)$ est le gain d'information lorsque l'intensité d'un processus de Poisson est changée de λ en ν, qui est lié à la fonctionnelle d'action $I_e(.\|x)$ des variables i.i.d. exponentielles de moyenne x par la relation

$$I_p(\nu\|\lambda) = \nu I_e(\nu^{-1}\|\lambda^{-1}) = \nu k(\lambda\nu^{-1}), \tag{4.10}$$

où $k(u) = u - 1 - \log u$ est la fonctionnelle d'action pour des variables exponentielles de moyenne 1. L'explication de la première égalité est la suivante : les temps d'inter-arrivées sont distribués exponentiellement, leur moyenne est déformée de λ^{-1} en ν^{-1} ; le « coût » de chaque distorsion est de $I_e(\nu^{-1}\|\lambda^{-1})$ et cet événement a lieu ν fois par unité de temps. En multipliant le coût des saut I_e par le nombre ν^{-1} de sauts par unité de temps, on obtient le coût par unité de temps, à savoir I_p.

Cette arithmétique très simple explique aussi bien la formule plus compliquée de l'entropie H_c. Le coût (par unité de temps) de la déformation de R_{xy} en L_{xy} est $I_p(L_{xy}\|R_{xy})$; mais il ne faut le comptabiliser que lorsque le processus est en x ; or la fraction de temps passée en x est π_x, donc le coût moyen est exactement H_c.

La meilleure preuve de l'efficience de ces calculs est de déduire une autre forme de l'entropie. On peut en effet considérer que Y_t est décrite par sa chaîne incluse et ses temps de saut. Pour chaque saut en x, sans considération du point y atteint, on modifie l'intensité de saut de R_x en L_x ; le nombre de passages en x par unité de temps est a_x, ce qui donne un coût

$$\sum_{x\in E} a_x I_e\big(L_x^{-1}\|R_x^{-1}\big)$$

pour la modification des vitesses de saut. La modification de la chaîne incluse coûte $H_d(A/|A|\|P)$ par saut, mais il y a $|A| = \sum_x a_x$ sauts par unité de temps, ce qui donne le coût $H_d(A\|P)$ par unité de temps. Finalement on obtient, en additionnant ces deux composantes, on obtient

$$H_c(G\|R) = H_d(A\|P) + \sum_{x\in E} a_x I_e(L_x^{-1}\|R_x^{-1}).$$

On vérifie sans difficulté l'égalité avec la définition 4.1.

4.2.4 Opérations sur les principes faibles

Bien que le changement de mesure de la section 4.2.2 soit une généralisation assez directe du PGD pour le temps discret [32], une formalisation plus poussée aboutit à une extension des méthodes. Dans un premier temps, la démonstration de la borne inférieure est un peu plus explicite dans [32] que dans [34] ; de plus, l'abandon des processus locaux, en dépit de l'intérêt de cette voie, conduit à une unification des concepts. C'est à partir de ces considérations que la section 3.2.1 a été construite. De plus, les équations (4.8)–(4.9) sont bien plus signifiantes qu'elles n'y paraissent. En effet, nous n'avons pas eu recours à la structure très spéciale de la fonctionnelle $\phi_{R,G}$, qui notamment est linéaire. La section 6.1.2 montre qu'on peut en tirer beaucoup plus qu'un PGD.

Dans un deuxième temps, la déduction du théorème de Sanov (unidimensionnel) par contraction requiert, dans le cadre du temps continu, un développement des propriétés des PGD faibles. Il semble que la propriété de Ruelle-Lanford soit la bonne notion, à partir de laquelle la plupart des théorèmes se déduisent sous une forme simple. Nous avons déjà vu le théorème d'existence et d'unicité (théorème 3.5). Sous des hypothèses, certes plus contraignantes que dans le cas de PGD forts, le principe de contraction ou le lemme de Varadhan sont encore valides. La contraction étant l'opération qui nous intéresse le plus, on montre également qu'elle préserve d'importantes propriétés.

De cette manière on parvient à dériver le théorème de Sanov en temps continu pour un espace dénombrable (théorème II 3.25) et à démontrer l'unicité, la convexité et l'exacte semi-continuité de la fonctionnelle, ce qui est plus ardu qu'en temps discret :

Théorème *Soit Y_t un processus de Markov irréductible sur un espace dénombrable E, dont le générateur R est borné, c'est-à-dire que les intensités R_x sont bornée par $\bar{r}$. Alors*

$$I_0(\pi \| R) \stackrel{\text{def}}{=} \inf_{G=(\pi,A)} H_c(G \| R) \tag{4.11}$$

est la fonction de Ruelle-Lanford de la mesure empirique L_t. Elle est unique, convexe et c'est la plus petite fonction semi-continue inférieurement qui soit égale à I_0 sur les mesures $\pi \in \mathcal{M}_1(E)$ irréductibles[2] à support fini.

La difficulté intrinsèque au temps continu est qu'il peut y avoir un nombre infini de sauts en temps fini. Il faut donc s'assurer qu'un tel événement ne vient pas perturber les grandes déviations. Le générateur empirique, en ce qu'il mesure les intensités et donc le nombre de sauts, circonscrit ce phénomène. Il est intéressant en effet de remarquer que la méthode proposée

2. Une mesure $\pi \in \mathcal{M}_1(E)$ est dite *irréductible* si son support est connexe dans le graphe de R.

reste efficace, même pour un processus explosif (qui peut sauter infiniment souvent en temps fini). Dans ce cas, on restreint en fait l'analyse aux trajectoires dont le nombre de transitions est fini. On n'obtient donc qu'une information partielle et c'est finalement pourquoi un PGD n'est valable que pour des processus non-explosifs.

Les hypothèses auxquelles nous avons recouru ci-dessus pour la contraction sont même trop fortes ; en effet, borner l'intensité du générateur est plus que suffisant pour garantir un nombre fini de sauts en temps fini : par exemple la file $M/M/\infty$ a des intensités de sauts non-bornées mais ne saute presque jamais un nombre infini de fois sur une durée finie. On peut imaginer que la contraction — c'est-à-dire le théorème ci-dessus — est valide pour tous les processus non-explosifs.

4.3 PGD trajectoriel

En se concentrant maintenant nettement plus sur des problèmes concrets, on s'aperçoit de l'insuffisance de la restriction de l'analyse à des trajectoires ergodiques (c'est bien ce que sont les mesures balancées ou les générateurs empiriques). Prenons par exemple l'archétype du réseau qu'est la file $M/M/1$. On imagine bien qu'il soit possible, par des arguments de grandes déviations, de donner le taux de décroissance de la probabilité pour la queue de se remplir à vitesse constante. Ce serait une méthode pour approcher la probabilité qu'une file d'attente, initialement vide, contienne cent clients au bout d'une heure, ce que l'on souhaite être un événement rare !

L'étude des propriétés asymptotiques des systèmes en considérant une petite perturbation (théorie de Freidlin-Ventzel', voir [67, 6]) ou bien en renormalisant le processus (c'est le cas pour les réseaux, voir p. ex. [128]). Les modèles utilisés actuellement ($M/M/1$, réseaux de Jackson, réseaux à polling, etc.) jouissent d'excellentes propriété d'homogénéité, qui conduisent naturellement à des changements d'échelle. On peut facilement imaginer des modèles plus réalistes — et plus compliqués — mais l'absence de propriétés fortes rend l'analyse difficile. Nous préférons donc focaliser notre attention sur des modèles simples, afin d'en dégager de bonnes propriétés. La suite présente cette démarche sur un exemple (voir la figure 4.2), mais les démonstrations peuvent aisément être adaptées à d'autres modèles (voir la section II 4.6).

4.3.1 Changement d'échelle

Le but de l'analyse trajectorielle est de calculer la probabilité que le processus suive une trajectoire donnée. De façon similaire à la loi des grands

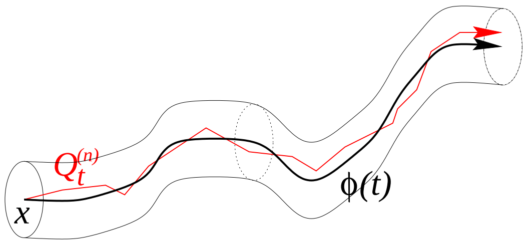

Fig. 4.1: *Un* PGD *trajectoriel mesure la probabilité pour un processus renormalisé $Q_t^{(n)}$ de suivre une trajectoire $\phi(t)$.*

nombres, les *limites fluides* montrent, en considérant un processus convenablement normalisé, que l'aléatoire disparaît. Par exemple, les atomes d'une substance radioactive se désintègrent au hasard, mais sur un échantillon assez grand on observe juste un *flux* radioactif *constant*.

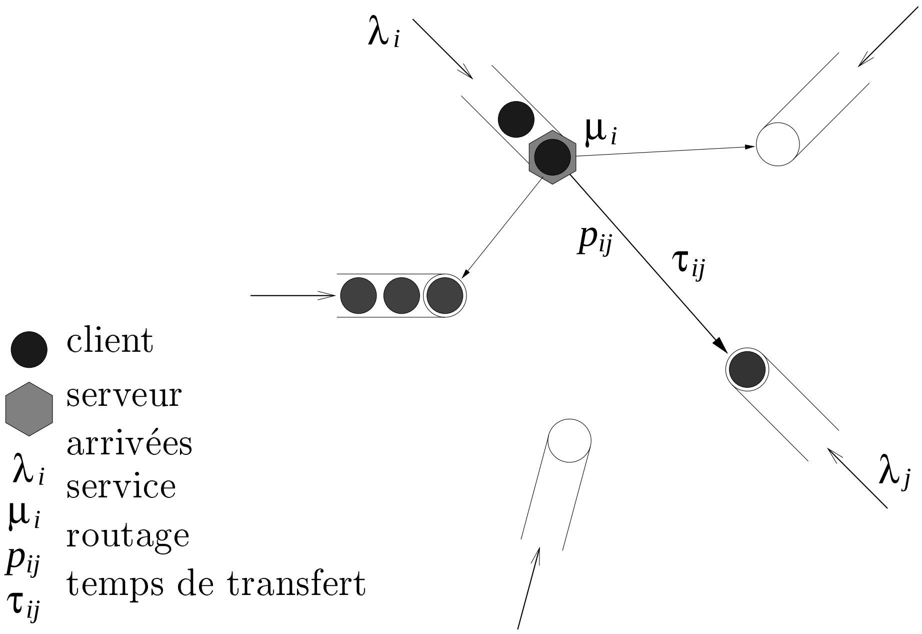

Fig. 4.2: *Un réseau à polling, pour lequel un* PGD *trajectoriel a été obtenu.*

Prenons le cas d'un réseau à polling (voir la figure 4.2, l'analyse proposée en [35] ou le chapitre II 4). On souhaite étudier le nombre de clients $Q(t) = \big(q_1(t), \dots, q_N(t)\big)$ en attente à chaque queue. On introduit un changement d'échelle et, au lieu du processus $Q(t)$, on s'intéresse à $Q_t^{(n)} \stackrel{\text{def}}{=} n^{-1}Q(nt)$ et à la probabilité de suivre une trajectoire $\phi(t)$ dans $\mathbb{R}_+^N$ (voir la figure 4.1).

Dans l'espace constitué de toutes les trajectoires, un PGD est valide (théorème II 4.2), dont une version simplifiée est énoncée ci-dessous.

Théorème *La suite* $\{n^{-1}Q(nt),\ n \geq 1\}$ *satisfait un* PGD *dans l'espace des fonctions càdlàg* $D\big([0,T],\mathbb{R}_+^N\big)$ *nulles en 0, avec la bonne fonctionnelle d'action*[3] $I_T(.)$ *:*

$$I_T(\varphi) = \begin{cases} \displaystyle\int_0^T L\left(\varphi(t),\dot\varphi(t)\right)dt, & \text{si } \varphi \text{ est absolument continue,} \\ +\infty, & \text{sinon.} \end{cases}$$

Le PGD *s'énonce classiquement : pour tout* $T > 0$,

1. *pour tout fermé* F *de* $D\big([0,T],\mathbb{R}_+^N\big)$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[n^{-1}Q(nt) \in F\right] \leq - \inf_{\phi\in F} I_T(\phi);$$

2. *pour tout ouvert* O *de* $D\big([0,T],\mathbb{R}_+^N\big)$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[n^{-1}Q(nt) \in O\right] \geq - \inf_{\phi\in O} I_T(\phi);$$

4.3.2 Entropie

Bien que non explicite dans la formulation du théorème ci-dessus, l'entropie est au cœur de la démonstration. On utilise en effet un changement de mesure qui conduit au résultat grâce aux bonnes propriétés de l'entropie associée. Nous ne les détaillerons pas ici, mais il est intéressant d'écrire cette entropie, dans la mesure où elle est très intuitive :

$$H(G\|R) \stackrel{\text{def}}{=} \sum_{i\in\mathcal{S}} \Big(I_p(\tilde\lambda_i\|\lambda_i) + \pi_i I_p(\tilde\mu_i\|\mu_i)\Big) + \sum_{i,j\in\mathcal{S}} \pi_{ij} I_p(\tilde\tau_{ij}^{-1}\|\tau_{ij}^{-1}) + H_d(\overline{A}\|P).$$

Il s'agit de l'entropie d'un polling G relativement au polling de référence R. Succintement, le générateur du processus est $R = (\lambda_i, \mu_i, \tau_{ij}, P_{ij})$, dont les composantes indiquent l'intensité des arrivées λ_i et des services μ_i, les temps moyens de déplacement τ_{ij} et la matrice de routage P_{ij} (voir la figure 4.2 p. 56). Quant à G, c'est le générateur empirique du polling — le comportement déviant que l'on veut mesurer — qui admet deux représentations $G = (\tilde\lambda_i, \tilde\mu_i, \tilde\tau_{ij}, \tilde P_{ij})$ ou $G = (A, \pi, D)$, comme à la section 4.2.1, liées par des

3. La fonctionnelle d'action fait intervenir une *fonctionnelle locale* $L(x,y)$. On la note L comme un lagrangien, ce qui n'a donc rien à voir avec le L_t des mesures empiriques.

relations similaires à (4.7). $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})$ s'interprète comme R et A, π et D sont respectivement le nombre de transitions par unité de temps (A), la fraction de temps que le serveur passe dans un site donné (π) et la vitesse d'accroissement des files d'attente (D). La matrice $\overline{A}$ est une sous-matrice de A qui ne compte que les déplacements du serveur.

L'entropie s'interprète finalement classiquement : on déforme les arrivée — pour un « coût » $I_p(\tilde{\lambda}_i \| \lambda_i)$ — les temps de service — $I_p(\tilde{\mu}_i \| \mu_i)$ — et les temps de déplacement — $I_p(\tilde{\tau}_{ij}^{-1} \| \tau_{ij}^{-1})$ — chacun en fonction de la fraction de temps durant laquelle ces événements sont possibles (respectivement 1, π_i et π_{ij}) ; on y ajoute la modification du routage — qui donne $H_d(\overline{A} \| P)$.

Comme le générateur empirique G est de *dimension finie*, l'entropie H possèdent d'excellentes propriété : elle est positive, finie, continue et convexe. Elle n'est nulle que si $G = R$ et possède des niveaux compacts (proposition II 4.13). Ces propriétés permettent d'établir que la *fonctionnelle locale* L est obtenue à partir de H par une contraction, qui est de fait une *projection*. Les études générales sur les processus de Markov en temps continu (section 4.2 ou chapitre II 3) montrent que les propriétés passent à la fonctionnelle contractée. L est donc positive, finie, continue et convexe avec des niveaux compacts.

De plus, on en déduit que la définition de L par une équation analogue à (4.11) constitue en réalité un programme convexe. D'un côté c'est un avantage pour l'analyse numérique : la complexité algorithmique d'un programme convexe est raisonnable et le calcul de L devrait être rapide ; d'autre part on parvient à montrer d'autres propriété de L qui conduisent à la forme intégrale de I_T, ce qui n'était qu'une conjoncture (voir [55] ou [56]).

4.4 Applications

Mieux sans doute que le mot applications, le mot « implications » conviendrait pour décrire les sous-produits des PGD, qui ne sont pas matériels. Nous en recensons trois types, qui correspondent aux différentes facettes de la notion de PGD : la statistique, la simulation et les probabilités.

4.4.1 Statistique

L'approche statistique propose d'évaluer, à partir d'une ou de plusieurs réalisation d'un processus, des caractéristiques intrinsèques, comme le générateur. Deux cas sont particulièrement intéressants : les estimateurs et les tests.

Pour un estimateur, le premier critère est qu'il converge bien vers la quantité évaluée, mais, parmi toutes les variantes possibles, on souhaite caractériser l'estimateur optimal en un certain sens. Une méthode classique est celle du maximum de vraisemblance. On construit une variable aléatoire qui dépend du générateur R — c'est ici le paramètre des modèles statistiques — et on cherche, pour une réalisation donnée ω_0, quel est le générateur R qui donne la probabilité la plus importante pour cet événement. Sous des hypothèses assez générales, pour un processus de Markov sur un espace fini, Albert [1] propose exactement le générateur empirique G_t défini en (4.6).

Le lien avec les grandes déviations est assez net : on cherche à maximiser, relativement au générateur R, la vraisemblance $V(R)$, c'est-à-dire le logarithme de la probabilité que le générateur G_t soit égal à la réalisation[4] $G_t(\omega_0)$:

$$V(R) \stackrel{\text{def}}{=} \log \mathbb{P}_R\left[G_t = G_t(\omega_0)\right].$$

Or, si t est assez grand, les grandes déviations impliquent que

$$V(R) \simeq -tH(A\|R),$$

où on a posé $A \stackrel{\text{def}}{=} G_t(\omega_0)$. Le calcul de $V(R)$ se simplifie substantiellement et, partant, le maximum de vraisemblance.

Un test doit typiquement trancher entre deux possibilités, par exemple si le générateur R vaut R_0 ou R_1. On définit deux types d'erreur. Pour les « bons » tests (consistants), les deux erreurs tendent vers 0, le plus souvent exponentiellement rapidement. C'est cette vitesse de convergence qui détermine l'optimalité du test choisi et qui est naturellement en lien avec les grandes déviations.

Ce point de vue, initié par Bahadur [7], est exposé simplement dans [71, chapitre 4] dans le cas de variables i.i.d. et les articles de Hornik [74, 75] pour les processus de Markov. Ces considérations de vitesse de convergence sont assez anciennes et c'est pourquoi le premier article sur le PGD du générateur empirique (bien qu'il ne soit pas nommé ainsi) provient du point de vue statistique [76].

4.4.2 Simulation

Concrètement, il est très fréquent de ne pouvoir calculer explicitement certaines probabilités, ou bien les calculs pour y parvenir sont trop longs. Dans ce cas la simulation est une bonne solution. Prenons le cas du calcul de

4. Bien sûr cette probabilité est nulle, mais en considérant un petit voisinage, ou des densités, on contourne cette restriction.

la fréquence d'un événement (par exemple une probabilité de panne). Lorsque cette quantité est trop petite (ou proche de 1), la simulation donne (presque) toujours le même résultat : une estimation égale à 0.

Sachant qu'il faudrait d'autant plus de tirages que la probabilité est faible (par exemple, 10^{10} tirages pour une fréquence de 10^{-9}), une solution consiste à faire dévier le processus vers l'événement intéressant. C'est *l'échantillonnage avec poids* (ou *importance sampling*). C'est exactement la méthode du changement de mesure, car cela revient à écrire

$$\mathbb{P}\left[G_t \in O\right] = \mathbb{E}^*\left[M_t^{-1}\mathbb{1}_{\{G_t \in O\}}\right],$$

où $\mathbb{P}^*$ est la loi du processus modifié (de générateur G). On prend comme estimateur (sans biais) de $\mathbb{P}\left[G_t \in O\right]$, si on considère N tirages ω_i de processus de loi $\mathbb{P}^*$, la quantité

$$\frac{1}{N}\sum_{i=1}^{N} M_t^{-1}(\omega_i)\mathbb{1}_{\{G_t(\omega_i) \in O\}}.$$

On se rend compte l'optimisation de ces calculs nécessite de choisir un nouveau générateur G, duquel on déduit le changement de mesure. On se doute que G est optimal quand il réalise le minimum de l'entropie sur O. C'est le but des articles [10, 22, 92], qui se limitent toutefois à l'estimation par simulation de quantités dérivées de la mesure empirique, soit $L_t \in O$, soit $\langle f, L_t\rangle \geq C^{\text{ste}}$.

Les changements de mesure que nous avons introduits élargissent l'éventail des possibilités, à la fois en pouvant estimer des quantités liées au générateur G_t, et non seulement la mesure empirique L_t, mais aussi en montrant comment procéder sur des modèles plus sophistiqués, comme les réseaux à polling, avec des changements de mesure adéquats.

4.4.3 Probabilité stationnaire : un exemple de calcul asymptotique

L'analyse des systèmes stochastiques, dans bon nombre de cas (réseaux de télécommunications, transports, etc.), permet de classifier la *stabilité* : c'est l'ergodicité. Cependant, mis à part quelques cas particuliers (comme les réseaux BCMP), on ne sait pas calculer la distribution stationnaire. Or la *performance* (temps moyen d'attente, longueur moyenne des queues...) des systèmes stables est principalement déterminée par cette distribution. Toute indication sur leur comportement est donc bienvenue.

Les PGD trajectoriels (voir section 4.3) sont justement liés aux probabilités stationnaires (voir, p. ex. [17]) ; ils permettent de décrire la décroissance à l'infini de ces distributions. Cependant rares sont les exemple où l'on a pu expliciter cette relation. Le cas intéressant et instructif de marches dans $\mathbb{Z}_+^2$ est entièrement traité dans [83] (voir aussi [18]). Pour éviter des notations trop lourdes, nous considérerons désormais le cas du polling à N files et 1 serveur (voir la section 4.3 ou le chapitre II 4). Dans ce cas, on peut montrer qu'il existe un temps T_0 et un chemin ϕ_0 optimaux entre deux points x et y de $\mathbb{R}_+^N$:

$$I_{T_0}(\phi_0) = \inf_{T\geq 0} \inf_{\phi} \{I_T(\phi) : \quad \phi(0) = x,\ \phi(T) = y\} \stackrel{\text{def}}{=} \mathcal{I}(x, y). \tag{4.12}$$

Les queues de distributions sont liées à $\mathcal{I}$ par

$$\lim_{n\to\infty} \frac{1}{n} \log \pi([nx]) = -\mathcal{I}(0, x), \tag{4.13}$$

où $\pi(x)$ est la probabilité stationnaire d'être en x.

Cette démarche se heurte à plusieurs pierres d'achoppement. En effet, s'il est relativement aisé de prouver (4.12), la limite (4.13) est beaucoup plus difficile. Encore n'est-ce là qu'une difficulté minime à côté du problème que représente le calcul explicite de la trajectoire optimale ϕ_0 (au moins une, s'il n'y a pas unicité). La faible dimension permet de limiter les possibilités dans le cas de $\mathbb{Z}_+^2$: les trajectoires optimales sont linéaires par morceaux, avec au plus deux segments. En dimension N il n'en en plus de même, des trajectoires « en spirale » avec un nombre infini de segments sont envisageables. De plus, dans le cas $N = 2$, on peut faire appel à l'analyse complexe [63, 94], ce qui n'est plus possible dans les mêmes termes pour $N \geq 3$.

Polling Dans la suite, nous nous attachons à l'exemple particulier — mais que nous espérons générique — du réseau à polling (voir figure 4.2). Les notations sont reprises du chapitre II 4.

Λ représente une *face* (voir la section II 4.1), qui est un sous-ensemble de $\{1, .., N\}$ décrivant les files où des clients attendent. Cette notion de face est importante pour les descriptions asymptotiques et on sait [64] que des phénomènes très compliqués (« scattering ») peuvent se produire pour des dimensions $N \geq 4$. Sans en préjuger de la réussite, nous proposons une méthode susceptible de diminuer la complexité de cette question d'optimisation.

Fonctions harmoniques Nous constatons tout d'abord que les trajectoires optimales linéaires sont associées à des fonctions presque harmoniques, au sens suivant. Si on écrit l'entropie sous la forme

$$H(G\|R) = \sum_{i\in\mathcal{S}} \Big(I_p(a_i+D_i\|\lambda_i)+I_p(a_i\|\pi_i\mu_i)\Big)+\sum_{i,j\in\mathcal{S}} I_p(a_{ij}\|\pi_{ij}\tau_{ij}^{-1})+H_d(\overline{A}\|P),$$

le programme convexe (voir définition II 4.15) qui définit la fonctionnelle locale L est

$$L(\Lambda, D) \stackrel{\text{def}}{=} \inf_{G \in f_\Lambda^{-1}(D)} H(G\|R), \qquad \forall D \in \mathbb{R}^\Lambda, \tag{4.14}$$

s'écrit assez simplement : calculer $\inf_{A,\pi} H(G\|R)$ sous les contraintes

$$\begin{aligned}
\overline{a}_i \stackrel{\text{def}}{=} \sum_{j\in\mathcal{S}} a_{ji} &= \sum_{j\in\mathcal{S}} a_{ij}, && \forall i \in \mathcal{S}, && (4.15)\\
D_i = 0, \quad \text{et} \quad 0 \le a_i &\le \overline{a}_i, && \forall i \in \Lambda^c, && (4.16)\\
0 \le a_i &= \overline{a}_i, && \forall i \in \Lambda, && (4.17)\\
\tilde{\lambda}_i \stackrel{\text{def}}{=} a_i + D_i &\ge 0, && \forall i \in \Lambda, && (4.18)\\
\sum_{s\in\mathcal{S}_0} \pi_s &= 1. && && (4.19)
\end{aligned}$$

Tab. 4.1: *Contraintes et multiplicateurs de Lagrange, pour le programme convexe défini par (4.14).*

contrainte	multiplicateur	domaine	convention
$\sum_j a_{ij} - \sum_j a_{ji} = 0$	ϕ'_i	$\forall i \neq 1$	$\phi'_1 = 0$
$\sum_j a_{ij} - a_i \ge 0$	ϕ_i	$\forall i \in \Lambda$	$\phi_i = 0 \ \forall i \in \Lambda^c$
$\sum_s \pi_s = 1$	$-\theta$		

Supposons dans un premier temps que les contraintes (4.16) sont strictes et introduisons les multiplicateurs de Lagrange, selon le tableau 4.1. Supposons que le minimum soit atteint en un point $(A, \pi, D) \equiv (\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})$ où H est différentiable, alors on obtient

$$\begin{aligned}
\tilde{\lambda}_i \tilde{\mu}_i &= \lambda_i \mu_i e^{\phi_i}, && (4.20)\\
\tilde{P}_{ij} \tilde{\tau}_{ij}^{-1} &= P_{ij} \tau_{ij}^{-1} e^{-\phi_i - \phi'_i + \phi'_j}, && (4.21)\\
\tilde{\mu}_i &= \mu_i - \theta, && (4.22)\\
\tilde{\tau}_{ij}^{-1} &= \tau_{ij}^{-1} - \theta. && (4.23)
\end{aligned}$$

On construit de cette manière une fonction C–harmonique f de la forme

$$f(x,s) \;=\; e^{\langle \alpha, x\rangle} h(s) > 0,$$

$$\text{avec} \quad \begin{cases} \alpha_i & \stackrel{\text{def}}{=} \log \dfrac{\tilde{\lambda}_i}{\lambda_i}, \\ h(i) & \stackrel{\text{def}}{=} e^{\phi'_i}, \\ h(ij) & \stackrel{\text{def}}{=} \dfrac{e^{\phi'_i}}{\tau_{ij}\theta + 1}. \end{cases}$$

On vérifie en effet que f est C–harmonique :

$$Rf = Cf, \qquad \text{avec } C \stackrel{\text{def}}{=} \sum_{k\in\mathcal{S}} \lambda_k \left(e^{\alpha_k} - 1\right) + \theta.$$

En fait f est C–harmonique pour le système à polling *localisé* $(Q_t^\Lambda, S_t^\Lambda)$. Elle est donc C–harmonique si $x_i > 0$, pour tout $i \in \Lambda$, mais elle n'est pas nécessairement C–harmonique pour les autres points. De plus, si on ne suppose plus $\phi_i = 0$ pour tout $i \in \Lambda^c$, l'inégalité (4.16) implique que $\phi_i \le 0$ et on en déduit que f est C–surharmonique sur le domaine où $x_i > 0$, pour tout $i \in \Lambda$.

Maintenant, l'optimisation de la trajectoire conduit à considérer, pour des raisons techniques, non pas directement le minimum de $L(\Lambda, D)$, mais celui de $L(\Lambda, D)/\|D\|$. Grâce à cette contrainte supplémentaire, on obtient des fonctions harmoniques car $C = 0$. Notons cependant que ce problème d'optimisation n'est pas convexe, mais on peut montrer qu'il admet au moins une solution.

Écrivons que $D = d\vec{u}$, où $\vec{u} \in \mathbb{R}^\Lambda$ est un vecteur unitaire. Annuler la dérivée par rapport à d de $L(\Lambda, d\vec{u})/d$ nous procure l'équation

$$\frac{1}{d}\frac{\partial H}{\partial d} - \frac{H}{d^2} = 0 \iff d \sum_{i\in\mathcal{S}} u_i \log \frac{\tilde{\lambda}_i}{\lambda_i} - H = \langle \alpha, D\rangle - H(G\|R) = 0.$$

Finalement, un calcul au point optimal G, utilisant les égalités (4.20)–(4.23), montre que $H(G\|R) = \langle \alpha, D\rangle - R(\alpha) - \theta$, avec $R(\alpha) \stackrel{\text{def}}{=} \sum_{i\in\mathcal{S}} \left(\tilde{\lambda}_i - \lambda_i\right)$. Donc ceci implique

$$0 = \langle \alpha, D\rangle - H(G\|R) = R(\alpha) + \theta = C,$$

En résumé, pour chaque direction $\vec{u} \in \mathbb{R}^\Lambda$, à toute solution du problème d'optimisation

$$\inf_{d>0} \frac{L(\Lambda, d\vec{u})}{d} = \inf_{\substack{G\in f_\Lambda^{-1}(d\vec{u}) \\ d>0}} \frac{H(G\|R)}{d}$$

on fait correspondre une fonction $f(x,s) = e^{\langle \alpha,x\rangle} h(s) > 0$, harmonique en dehors des frontières $x_i = 0$ pour $i \in \Lambda$.

> **Remarque :** Ce lien entre trajectoire optimale et fonction harmonique vient renforcer ce sentiment qu'il existe un lien fort et général entre la frontière de Martin et les grandes déviations. La fonction f, qui correspond à la meilleure manière de saturer le réseau dans la direction $\vec{u}$, *pour le processus localisé*, semble en effet correspondre à un point minimal de la frontière de Martin. Une approche analytique le montre bien dans le cas particulier de $\mathbb{Z}_+^2$ [83, 94] et c'est ce qui nous a incité à approfondir ce lien dans le chapitre 5.

♦♦♦

Trajectoire optimale Ces fonctions harmoniques permettent à leur tour de construire des trajectoires optimales *rectiligne* pour des modèles localisés auxquels il manque des frontières. Il reste à prendre en compte les effets aux frontières[5]. Pour cela nous utilisons la représentation suivante de la mesure invariante. Notons $X_t = (Q_t, S_t)$ l'état du système, $R(z)$ l'intensité de saut au point z et τ_A le temps de retour en A :

$$\begin{cases} \tau_1 \stackrel{\text{def}}{=} \inf\{t > 0 : \ X_t \neq X_0\}, \\ \tau_A \stackrel{\text{def}}{=} \inf\{t > \tau_1 : \ X_t \in A\}. \end{cases}$$

Alors, pour tout $y \in \mathbb{Z}_+^N \times \mathcal{S}_0$ et tout ensemble $A \subset \mathbb{Z}_+^N \times \mathcal{S}_0$,

$$\pi(y) = \sum_{z\in A} \pi(z) R(z) \mathbb{E}_z \left[\int_0^{\tau_A} \mathbb{1}_{\{X_t = y\}} \, \mathrm{d}t \right]. \tag{4.24}$$

Choisissons maintenant $(q,s) \notin A$. En conditionnant l'espérance de (4.24) par le passage en (q,s), on obtient

$$\begin{aligned} \pi(q,s) &= \sum_{z\in A} \pi(z) R(z) \mathbb{P}_z \left[\exists t < \tau_A : \ X_t = (q,s)\right] \mathbb{E}_{(q,s)} \left[\int_0^{\tau_A} \mathbb{1}_{\{X_t = (q,s)\}} \, \mathrm{d}t \right] \\ &= \sum_{z\in A} \pi(z) R(z) \mathbb{P}_z \left[X_{\tau_{A\cup(q,s)}} = (q,s) \right] T(q,s), \end{aligned}$$

5. On appelle *effet frontière* l'attractivité d'une frontière, c'est-à-dire qu'il est plus avantageux en terme de « coût » de passer par cette frontière plutôt que dans son voisinage. C'est à cause des effets frontières que la trajectoire optimale n'est pas toujours une droite. La question de savoir quand cela se produit est une des difficultés essentielles des grandes déviations sur les réseaux.

en notant $T(q,s) \stackrel{\text{def}}{=} \mathbb{E}_{(q,s)}\left[\int_0^{\tau_A} \mathbb{1}_{\{X_t=(q,s)\}}\,\mathrm{d}t\right]$.

Si A représente une union de faces (de type $x_i = 0$), on peut faire intervenir les fonctions qui sont harmoniques en dehors de A. Si f est une telle fonction, alors,

$$\begin{aligned}
\mathbb{E}_z\left[f(X_{\tau_{A\cup(q,s)}})\right] &= \mathbb{E}_z\left[f(X_{\tau_1})\right] = \frac{Rf(z)}{R(z)} + f(z) \\
= &\sum_{y\in A} f(z)\mathbb{P}_z\left[X_{\tau_{A\cup(q,s)}} = y\right] + f(q,s)\mathbb{P}_z\left[X_{\tau_{A\cup(q,s)}} = (q,s)\right].
\end{aligned}$$

Puisque $f(q,s) = e^{\langle\alpha,q\rangle}h(s)$ est positive,

$$\mathbb{P}_z\left[X_{\tau_{A\cup(q,s)}} = (q,s)\right] \le \frac{e^{-\langle\alpha,q\rangle}}{h(s)}\left(\frac{Rf(z)}{R(z)} + f(z)\right).$$

Finalement

$$\pi(q,s) \le T(q,s)\frac{e^{-\langle\alpha,q\rangle}}{h(s)}\sum_{z\in A}\pi(z)\big(Rf(z) + R(z)f(z)\big). \tag{4.25}$$

Ceci n'est valable que si f et Rf sont π–intégrables sur A, ce qui revient à la condition que f le soit. Si ce n'était pas le cas, cela signifierait que la trajectoire pour aller de 0 à (nq,s) (en renormalisant) en passant par A est plus avantageuse que la ligne droite : il y a donc un effet frontière.

L'étude des effets frontières est ramenée à une question d'intégrabilité et même à une comparaison entre taux, puisque toutes les valeurs décroissent exponentiellement. Remarquons cependant que la solution n'est pas encore rendue immédiate, car il faut connaître le comportement de π sur A (sauf si f est bornée).

On peut néanmoins réduire de nouveau la complexité en ne cherchant que le majorant des taux de décroissance, c'est-à-dire le plus grand nombre $\alpha_{\max}$ tel que

$$\lim_{n\to\infty}\frac{1}{n}\log\pi([nx]) \le -\alpha_{\max}\|x\| \qquad \forall x \in \mathbb{R}_+^N.$$

La dernière étape serait de montrer que, si la meilleure façon de saturer le système (à savoir de sortir de la boule $\|x\| \le 1$) est de partir en ligne droite, ce qui est notamment toujours le cas dans $\mathbb{Z}_+^2$, alors le meilleur exposant est

$$\alpha_{\max} = \inf_{\Lambda\neq\emptyset}\inf_{G\in\mathcal{G}_s^\Lambda}\frac{H(G\|R)}{\|D\|}. \tag{4.26}$$

On peut conjecturer que la minimisation (4.26) est souvent la solution au problème de saturation « optimale » d'un réseau. Il est certain que (4.26) ne dépend pas du caractère markovien, puisque Anantharam (voir [2], qui dépasse largement l'utilisation que nous en faisons ici) prouve dans le cas de la file $GI/G/1$ qu'elle se sature en ligne droite, avec un comportement déviant donné par une équation similaire à (4.26) — pourvu que les temps d'arrivée et de service possèdent des moments exponentiels. La file $GI/G/1$ est cependant unidimensionnelle et les problèmes de faces n'existent pas : cet exemple montre que l'optimisation (4.26) est pertinente.

Complexité : Bien que l'optimisation (4.26) soit compliquée (il y a 2^N faces Λ), on arriverait à obtenir des informations sur la probabilité stationnaire, qui est solution d'un système infini-dimensionnel, grâce à un système fini-dimensionnel. Ce système n'est pas quelconque, car, on l'a vu, $G \to H(G\|R)$ est convexe, ce qui simplifie certains calculs. Encore n'avons-nous pas fait intervenir la structure particulière du réseau à polling et il n'est pas interdit de penser qu'on pourrait rendre polynômiale la complexité de (4.26). Pour ce qui est des algorithmes, il faut citer Majewski ([101] et ses travaux ultérieurs), qui travaille sous des hypothèses plus restrictives (les processus sont obtenus par transformation de Skorokhod) .

♦♦♦

Chapitre 5

La frontière de Martin et les grandes déviations

Nous avons estimé au moment des limites projectives (voir remarque « Limite projective et compactification » p. 39) que l'extension de faible à fort d'un PGD pourrait se faire par la voie d'une compactification, qui respecterait la structure probabiliste du processus stochastique considéré. Dans le cas des chaînes de Markov, une telle opération a été proposée depuis longtemps : c'est la compactification de Martin. Elle est plus compliquée que la celle d'Alexandroff, dans le sens où on doit ajouter non pas un, mais en général plusieurs points. Ces points limites forment la *frontière de Martin.*

Cette construction, introduite par Martin [102], a pour origine la description des fonctions harmoniques dans des domaines continus (telles que leur laplacien soit nul sur un ouvert de $\mathbb{R}^d$). Par ailleurs, Kakutani [89, 90] explicita le lien entre le mouvement brownien et les fonctions harmoniques (détaillé dans l'ouvrage de référence [50]), qui a ensuite été généralisé aux chaînes de Markov par Hunt [77, 78, 79]. C'est la théorie du potentiel, dont les concepts physiques (charge, potentiel, capacité, énergie, etc.) et leurs homologues probabilistes sont bien décrits dans [91, p. 169]. La synthèse de ces deux concepts, fut peu après proposée par Doob [49] et Hunt [80] : c'est la frontière de Martin pour les chaînes de Markov, telle que nous allons la décrire.

Le lien entre cette frontière et les grandes déviations ont déjà été notés (voir [94]) car certains calculs montrent des similarités, mais la nature de ce lien n'a pas été explicitée. En se servant de la remarque p. 39 comme guide, nous espérons en donner ici une interprétation plausible. La section 5.1 est un rappel de la construction et des propriétés de la frontière de Martin ; la section 5.2 est une contribution personnelle (sauf la proposition 5.22 qui est classique). C'est une version adaptée de la méthode du changement de mesure (voir section 3.2.1).

Nous ne sommes malheureusement pas en mesure de montrer un PGD, mais seulement des bornes supérieure et inférieure qui ne coïncident pas nécessairement. Cependant une conclusion au moins est sûre : s'il existe une trajectoire (correspondant à un point de la frontière de Martin) qui est d'entropie ni nulle, ni infinie, alors il ne peut y avoir de principe fort. Peut-être ce résultat permet-il d'approcher une condition nécessaire et suffisante pour qu'une chaîne de Markov admette un PGD *fort* ? Un exemple (p. 84) vient illustrer le lien entre frontière de Martin, grandes déviations et comportement des probabilités stationnaires.

5.1 La frontière de Martin

Puisque notre propos n'est pas l'étude en elle-même de la frontière de Martin, nous en proposons une construction et exposons les principales propriétés sans en donner les démonstrations. On se référera pour cela aux excellents ouvrages [91, 122].

5.1.1 Construction de la frontière de Martin

Fonction de Green Soit E un espace dénombrable et X_n une chaîne de Markov homogène sur E de noyau de transition $P = (p(x,y))$ et de distribution initiale ν. On notera $\mathbb{P}$ et $\mathbb{E}$ respectivement, la probabilité et l'espérance liées à X_n ; $\mathbb{P}_x$ et $\mathbb{E}_x$ lorsque la distribution initiale est concentrée en x, i.e. $\nu = \delta_x$. Les itérés $P^{(n)} = (p^{(n)}(x,y))$ du noyau de transition sont donnés par

$$\begin{cases} p^{(0)}(x,y) \stackrel{\text{def}}{=} \mathbb{1}_{\{x=y\}}, \\ p^{(n)}(x,y) \stackrel{\text{def}}{=} \displaystyle\sum_{z\in E} p^{(n-1)}(x,z)p(z,y), \quad \text{si } n \geq 1. \end{cases}$$

Définition 5.1 *La fonction de Green associée à P est définie par*

$$G(x,y) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} p^{(n)}(x,y) = \mathbb{E}_x \left[\sum_{n=0}^{\infty} \mathbb{1}_{\{X_n=y\}} \right].$$

La fonction de Green compte le nombre moyen de passages en y, lorsque la chaîne part de x. La transience est justement caractérisée par un nombre moyen de passages *fini* dans chaque état.

Définition 5.2 *Une chaîne de Markov est transiente si $G(x,y) < \infty$ pour tout $x, y \in E$. Si $G(x,y) = \infty$ pour tout $x, y \in E$, elle est dite récurrente.*

On s'intéresse désormais aux chaînes *transientes*. En définissant T_y comme étant le temps de premier passage en y, avec la convention usuelle que $T_y = \infty$ si la chaîne ne passe jamais en y, la fonction de Green prend la forme :

$$G(x,y) = \mathbb{P}_x\left[T_y < \infty\right] G(y,y). \tag{5.1}$$

Nous notons $x \rightsquigarrow y$ si $G(x,y) > 0$.

Mesure de référence La construction de la frontière de Martin dépend assez fortement du choix d'une mesure de référence η sur E. En choisissant $\eta \stackrel{\text{def}}{=} \delta_x$ pour $x \in E$ quelconque, les frontières de Martin obtenues sont parfois isomorphes mais ce n'est pas le cas général. Dans la suite, nous convenons que $\eta \stackrel{\text{def}}{=} \nu$ est la distribution initiale.

Définition 5.3 *Le noyau de Martin est*

$$K(x,y) \stackrel{\text{def}}{=} \frac{G(x,y)}{\nu G(y)}.$$

> **Remarque :** On constate dans la définition 5.3 que le noyau de Martin est mal défini si $\nu G(y)$ est nul ou infini, ce qui souligne deux problèmes.
>
> En premier lieu, *l'atteignabilité* de y, à partir de la mesure de référence : $\nu G(y) > 0$ si, et seulement si, à partir d'un point x du support de ν, $x \rightsquigarrow y$. En restreignant ν à δ_0, par exemple, on réduit l'espace effectif[1] de E à $E_0 = \{x \in E : \ 0 \rightsquigarrow x\}$. La solution de ce problème serait que E tout entier soit le support de ν.
>
> Grâce à la majoration $G(x,y) \leq G(y,y)$ déduite de (5.1), on a $\nu G(y) \leq G(y,y) < \infty$ et donc *l'intégrabilité* de $G(.,y)$ par rapport à ν. Cependant, on verra dans la suite qu'on ne construit dans ce cas-là que les fonctions ν–intégrables. Or il n'est pas sûr qu'il existe une mesure dont le support soit E et telle que toute fonction surharmonique finie soit ν–intégrable.
>
> ♦♦♦

Compactification On définit la distance

$$d_M(y,y') \stackrel{\text{def}}{=} \sum_{x\in E} w_x \nu G(x) |K(x,y) - K(x,y')|, \tag{5.2}$$

1. Cette remarque est développée dans [91, p. 338].

en choisissant $w_x > 0$ tel que $\sum_x w_x / \mathbb{P}[T_x < \infty] < \infty$. Prenons par exemple

$$w_x \stackrel{\text{def}}{=} 2^{-N(x)} \mathbb{P}[T_x < \infty],$$

où $N: E \mapsto \{1, 2, \ldots\}$ est une bijection. Dans ce cas, on a $d_M(y, y') \leq 1$. On caractérise alors les suites de Cauchy.

Proposition 5.4 *Une suite $\{y_n,\ n \geq 0\}$ est de Cauchy dans (E, d_M) si, et seulement si, les suites de réels $\{K(x, y_n),\ n \geq 0\}$ sont de Cauchy pour tout $x \in E$.*

Cette proposition prouve que la complétion E^* de Cauchy de E ne dépend pas des poids w_x choisis. On peut voir aussi que si $\{y_n, n \geq 0\}$ est de Cauchy, alors soit y_n est constante au-delà d'un certain n_0, soit $N(y_n)$ tend vers l'infini, c'est-à-dire que la suite sort de tout compact et en ce sens « tend vers l'infini ». Ce sont ces limites infinies qui constituent la *frontière de Martin*.

Définition 5.5 *La frontière de Martin ∂E est l'ensemble des points-limites pour la complétion de Cauchy de l'espace métrique (E, d_M). La compactification de Martin est $E^* = E \cup \partial E$.*

D'après la proposition 5.4, $K(x, .)$ est uniformément continue sur E et s'étend donc de manière unique en une fonction continue sur E^* encore notée $K(x, y)$, pour $y \in \partial E$. La proposition 5.4 implique aussi que $K(x, y) = K(x, y')$ pour tout $x \in E$ si, et seulement si, y et y' sont égaux dans E^*. Comme E est dénombrable et E^* complet, E^* est bien compact, ainsi que le suggère la définition 5.5. De plus, par continuité de $K(x, .)$ sur E^*, la distance d_M s'étend à E^* par la même équation (5.2). La topologie sur E^* est induite par la distance d_M.

Proposition 5.6 *Le compactifié de Martin (E^*, d_M) est un espace compact.*

5.1.2 Fonctions harmoniques et surharmoniques

Définition 5.7 *Une fonction $h: E \mapsto \mathbb{R}$ est surharmonique (ou excessive) si*

$$Ph \leq h \iff \sum_{y \in E} p(x, y) h(y) \leq h(x), \quad \forall x \in E.$$

La fonction h est harmonique (ou invariante[2]) s'il y a égalité. Elle est strictement surharmonique en x si $Ph(x) < h(x)$.

2. Nous adoptons le vocabulaire de la théorie classique du potentiel. Les synonymes entre parenthèses sont surtout employés lorsque la théorie du potentiel est construite directement à partir de la notion de noyau (voir [36]).

Proposition 5.8 *Si g_1 et g_2 sont [sur]harmoniques, α_1 et α_2 positifs, alors $\alpha_1 g_1 + \alpha_1 g_1$ est [sur]harmonique. Si g_n est une suite de fonctions surharmoniques, alors* $\liminf_n g_n$ *est surharmonique.*

Remarquons que $G(., y)$ est surharmonique, en fait harmonique en tout point, sauf en y où elle est strictement surharmonique. Cette propriété reste vraie pour le noyau de Martin, dérivé de celui de Green par renormalisation. Maintenant, si le point y est à l'infini, c'est-à-dire si $y \in \partial E$, on devine que $K(., y)$ est harmonique. Ce n'est pas tout à fait exact, mais on construit ainsi toutes les fonctions harmoniques et surharmoniques, et de manière unique. Pour montrer la force des théorèmes, nous devons d'abord aborder les propriétés classiques des fonctions surharmoniques.

Définition 5.9 *Soit f une fonction positive et $g = Gf$. Alors f est la charge de g et g est le potentiel de f. En outre le potentiel g est positif et il est dit fini s'il est fini en tout point de E.*

On peut considérer que $G(., x)$ est le potentiel induit par une charge unitaire située en x. Dans ce cas, g est le potentiel créé par une distribution de charges f sur E. Il s'ensuit que les potentiels sont surharmoniques. En fait, ils *caractérisent* la partie strictement surharmonique d'une fonction surharmonique. On ne considérera ici ni charges négatives, ni charges ou potentiels définis par une mesure.

Théorème 5.10 *Une fonction surharmonique g est un potentiel si, et seulement si,* $\lim_{n\to\infty} P^{(n)}g$ *ne prend que les valeurs 0 ou ∞. Si g est un potentiel fini, alors $(I - P)g$ est son unique charge.*

Théorème 5.11 (Riesz) *Toute fonction surharmonique g est somme d'une fonction harmonique h et d'un potentiel Gf. Si Pg est finie, alors cette décomposition est unique, $f = (I - P)g$ et $h =$* $\lim_{n\to\infty} P^{(n)}g$.

Voici le lien avec la frontière de Martin, qui éclaire la décomposition de Riesz (théorème 5.11).

Théorème 5.12 *Pour toute fonction g surharmonique et ν–intégrable, il existe une mesure finie μ_g sur E^* telle que*

$$g(.) = \int_{E^*} K(., y)\mu_g(\mathrm{d}y).$$

C'est la mesure spectrale *de g. De plus,*

$$\mu_g(y) = \nu G(y)(I - P)g(y).$$

5.1.3 Frontière de Martin minimale

Bien que la décomposition de Riesz (théorème 5.11) soit unique, la mesure spectrale définie au théorème 5.12 ne l'est pas nécessairement. Il se peut en effet que des fonctions harmoniques en plusieurs points de la frontière ne soient pas indépendantes. Il se trouve que la frontière de Martin contient des surprises et la lecture de [51] est édifiante, outre l'intérêt intrinsèque de cet article. Par exemple, un point de E peut aussi être un point limite, c-à-d appartenir à ∂E. La notion de *frontière minimale* élimine tous ces artefacts.

Définition 5.13 *Une fonction surharmonique g est dite minimale si, les seules décomposition $g = g_1 + g_2$ en fonctions g_1 et g_2 surharmoniques sont celles où $g_1 = \alpha_1 g$ et $g_2 = \alpha_2 g$ avec α_1 et α_2 positifs.*

Il s'ensuit qu'une fonction harmonique minimale h est telle que si une autre fonction harmonique h_1 lui est inférieure, alors c'est un multiple de h :

$$h_1 \leq h \implies h_1 = \alpha h, \quad \text{avec } 0 \leq \alpha \leq 1.$$

Définition 5.14 *Un point $x \in \partial E$ est dit minimal (ou plus proprement, extrême) si $K(.,x)$ est harmonique minimale. L'ensemble $B \subset \partial E$ de ces points est appelé frontière minimale de Martin.*

Proposition 5.15 *Pour chaque point limite $x \in \partial E$, $K(.,x)$ est surharmonique et normalisée, c'est-à-dire que $\nu K(x) = 1$. Les ensemble ∂E et B sont des boréliens sur la tribu $\mathcal{B}\big(E^*, d_M\big)$.*

Proposition 5.16 *Les seules fonctions minimales sont $K(.,x)$ pour x appartenant à $E \cup B$.*

Théorème 5.17 *Toute fonction g surharmonique et ν–intégrable se décompose de manière unique sous la forme*

$$g(.) = \int_{E \cup B} K(.,y)\mu_g(\,\mathrm{d}y), \tag{5.3}$$

où le support de μ_g est inclus dans $E \cup B$ et μ_g est finie. Réciproquement, pour de telles mesures μ_g, la fonction g définie par (5.3) est surharmonique. g est harmonique si, et seulement si, $\mu_g(E) = 0$.

5.1.4 Convergence vers la frontière de Martin

La frontière de Martin permet d'étudier le comportement à l'infini de la chaîne de Markov X_n, ce qui nous rapproche de notre propos, à savoir le comportement asymptotique de X_n.

Théorème 5.18 *La chaîne X_n possède $\mathbb{P}$–presque sûrement une limite dans la frontière de Martin, notée X_∞.*

En dehors de l'ensemble de mesure nulle sur lequel X_∞ n'est pas défini, $\omega \to X_\infty(\omega)$ est une application mesurable sur ∂E. On définit alors la mesure μ_1 induite par $\mathbb{P}$: pour tout borélien A de E^*,

$$\mu_1(A) \overset{\text{def}}{=} \mathbb{P}\left[X_\infty \in A\right].$$

C'est une mesure de probabilité concentrée sur la frontière, car $\mu_1(\partial E) = 1$ et $\mu_1(E) = 0$. Remarquons que μ_1 est la mesure spectrale de la fonction harmonique 1 : les notations sont bien cohérentes.

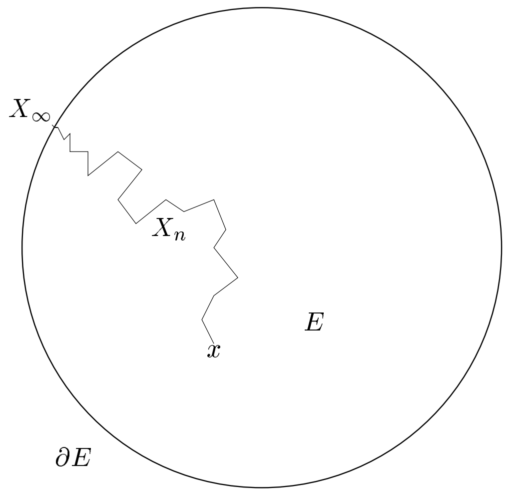

Fig. 5.1: *La trajectoire partant de x converge vers un point-limite $X_\infty \in \partial E$.*

Constatons tout d'abord que X_n converge nécessairement dans le support de μ_1. C'est l'analogue du théorème de Fatou, dans la théorie classique.

Théorème 5.19 *Soit h une fonction harmonique et $\mu_h = \phi\mu_1 + \mu_s$ la décomposition de Lebesgue de sa mesure spectrale par rapport à μ_1, alors*

$$\lim_{n\to\infty} h(X_n) = \phi(X_\infty) \quad p.s..$$

Voici l'analogue du théorème classique, qui permet d'obtenir les fonctions harmoniques via les probabilités et qui mène à la simulation.

Théorème 5.20 *Soit ϕ une fonction μ_1–intégrable sur ∂E et*

$$h(.) = \int_B K(., y)\phi(y)\mu_1(\,\mathrm{d}y). \tag{5.4}$$

Alors

$$\mathbb{E}_x\left[\phi(X_\infty)\right] = h(x). \tag{5.5}$$

Inversement, soit h une fonction harmonique bornée, alors il existe une fonction bornée ϕ sur B telle que (5.4) et (5.5) soient vérifiées. C'est la dérivée de Radon-Nikodym de μ_h par rapport à μ_1.

5.1.5 Exemples de frontières de Martin

Peu de travaux ont été consacrés au calcul de la frontière de Martin. On connaît les frontières de marches aléatoires homogènes : leur étude a commencé avec $\mathbb{Z}^d$ [51, 113] et s'est poursuivie sur des groupes généraux (*cf.* [114, 132, 88] et leurs références). Quelques autres modèles ont été analysés : les urnes de Polya [13], certains processus de branchement [118], les marches avec frontières, dans $\mathbb{Z}_+ \times \mathbb{Z}^d$ ou $\mathbb{Z}_+^2$ [94].

Marches aléatoires sur $\mathbb{Z}^d$ On suppose que la chaîne est irréductible et que la dérive n'est pas nulle. Dans ce cas, la frontière de Martin est représentée par les directions de $\mathbb{R}^d$ (les vecteurs unitaires) et chaque point est minimal.

$$y_n \to \vec{u} \iff \|y_n\| \to \infty \text{ et } \frac{y_n}{\|y_n\|} \to \vec{u}.$$

C'est un exemple très simple car les fonctions harmoniques sont des exponentielles et les h–processus (voir définition 5.21) sont eux-mêmes des marches aléatoires. Toutes les quantités sont ici explicites.

Marches aléatoires sur les arbres On considère ici un arbre binaire. Chaque état est représenté par une suite finie de 0 et de 1. Les transitions sont homogènes : $p(s, s0) = p(s, s1) = p$. Si $s \neq \emptyset$, on peut enlever le dernier symbole de la chaîne s avec probabilité q (voir figure 5.2). La frontière de Martin est alors homéomorphe au segment $[0, 1]$ des suites infinies de 0 et 1 muni de la distance $d(x, y) = \sup_i 2^{-i}|x_i - y_i|$. C'est un cas qui se déduit de la marche sur les groupes et, encore une fois, tout est explicite.

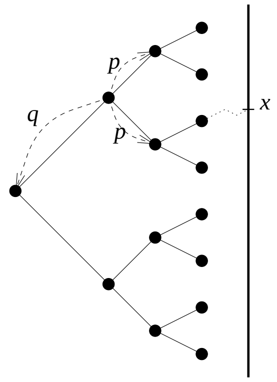

Fig. 5.2: *Frontière de Martin sur un arbre binaire.*

Marches aléatoires sur $\mathbb{Z}_+^2$ Il s'agit de « marches simples », dont les paramètres sont les probabilités de sauts : une seule coordonnée peut changer à chaque saut, de ± 1 sur chaque *face* (l'origine, les deux axes et l'intérieur). Plusieurs configurations sont décrites dans [94]. Dans l'une d'elles, il existe (voir figure 5.3, cas (1)) une direction γ qui partage en trois catégories les points limites :

- Si la direction asymptotique est strictement inférieure à γ, la trajectoire converge vers un point B minimal ;
- si la direction asymptotique est strictement supérieure à γ, la trajectoire converge vers un point A minimal ;
- si la direction asymptotique est égale à γ et que $y_n - \|y_n\|\gamma$ converge dans $\mathbb{R} \cup \{-\infty, \infty\}$, la trajectoire y_n converge vers un point mixte, combinaison convexe de A et de B.

Dans une autre configuration des paramètres (voir figure 5.3, cas (2)), la direction γ partage en deux catégories les points limites :

- Si la direction asymptotique est inférieure ou égale à γ, la trajectoire converge vers un unique point B minimal ;
- si la direction asymptotique est supérieure à γ, la trajectoire converge vers un point minimal pour chaque direction.

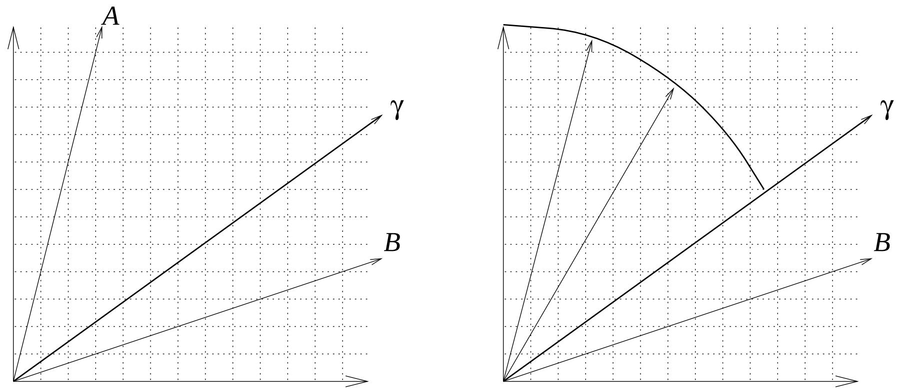

FIG. 5.3: *Frontières de Martin sur $\mathbb{Z}_+^2$. Cas (1) à gauche et (2) à droite.*

5.2 Grandes déviations et frontière de Martin

Il se trouve que la frontière de Martin décrit les différentes directions où une chaîne de Markov X_n peut diverger, mais il manque à cette description un terme de vitesse. Pour y remédier, ce n'est pas la frontière de la chaîne X_n que nous allons étudier, mais de la chaîne (n, X_n), que nous appellerons la chaîne (et la frontière) en espace-temps. Dans ce contexte, *rien n'empêche de considérer des chaînes non-homogènes en temps* ; cette généralisation est cependant coûteuse en notations supplémentaires, que nous épargnerons au lecteur.

5.2.1 Changement de mesure

Définition 5.21 *Soit h une fonction harmonique positive, non identiquement nulle sur le support de ν. Le h–processus X_n^h est la chaîne de Markov de noyau de transition $P^h = \big(p^h(x,y)\big)$*

$$p^h(x,y) \stackrel{\text{def}}{=} \frac{p(x,y)h(y)}{h(x)}, \quad si\ h(x) > 0.$$

Le h–processus n'est défini que sur l'ensemble $E^h \stackrel{\text{def}}{=} \{x \in E : \ h(x) > 0\}$. Si h est surharmonique, alors P^h est sous-markovien.

Le h–processus peut être interprété comme un changement de mesure qui amène la chaîne X_n sur d'autres points de la frontière minimale. Dans le cas où P^h est sous-markovien, cela signifie que le processus peut mourir en temps fini. Pour que la théorie s'applique convenablement, il faut faire le lien entre les frontières de Martin de X_n et X_n^h. La construction particulière ici s'inspire de [91, Definition 10–23 et suite].

Tout d'abord il convient de bien choisir la mesure de référence — qui est aussi la distribution initiale. Nous prenons

$$\nu^h(y) \stackrel{\text{def}}{=} Z^{-1}\nu(y)h(y), \quad \text{avec} \quad Z \stackrel{\text{def}}{=} \sum_y \nu(y)h(y),$$

en supposant que h est ν–intégrable. On vérifie alors que

$$G^h(x,y) = \frac{G(x,y)h(y)}{h(x)}$$

et donc que

$$K^h(x,y) \stackrel{\text{def}}{=} \frac{G^h(x,y)}{\nu^h G^h(y)} = K(x,y)\frac{Z}{h(x)},$$

où G^h et K^h sont les noyaux de Green et de Martin correspondant au h-processus. On définit de même la distance de Martin (voir (5.2)) et, finalement, la frontière de Martin de X_n^h est la restriction de celle de X_n à E^h, d'après la proposition suivante.

Proposition 5.22 *L'application identique de (E^h, d_M^h) sur (E, d_M) est une isométrie : $d_M^h(x,y) = d_M(x,y)$ pour tous $x, y \in E^h$. Si $y \in E^h$,*

1. *$K^h(.,y)$ est normalisé si, et seulement si, $K(.,y)$ l'est ;*
2. *$K^h(.,y)$ est P^h–harmonique si, et seulement si, $K(.,y)$ est P–harmonique ;*
3. *$K^h(.,y)$ est P^h–minimale si, et seulement si, $K(.,y)$ est P–minimale.*

Donc $B^h = B \cap \partial E^h$.

De plus, les décompositions se correspondent : si μ_h est la mesure spectrale de h selon P, c'est l'équivalent pour P^h de la mesure μ_1 :

$$\int_{B^h} K^h(x,y)\mu_h(\mathrm{d}y) = 1, \quad \forall x \in E^h.$$

Finalement, ce changement de mesure conserve suffisamment la structure de la chaîne de Markov, notamment la frontière de Martin, pour qu'il soit envisageable d'appliquer la méthode générale de la section 3.2.1 ; la section 5.2.2 offre l'heuristique qui sous-tend cette démarche et la suivante (section 5.2.3) donne des résultats plus quantitatifs, mêmes s'ils ne sont pas tout à fait complets.

5.2.2 Chaîne en espace-temps

Posons $\tilde{E} \stackrel{\text{def}}{=} \mathbb{N} \times E$. La chaîne en espace-temps est la chaîne de Markov $(n, X_n) \in \tilde{E}$ dont la matrice de transition est

$$p(n,x;m,y) \stackrel{\text{def}}{=} \begin{cases} p(x,y), & \text{si } m = n+1, \\ 0 & \text{sinon.} \end{cases}$$

Le noyau de Green est très simple :

$$G(n,x;m,y) = \mathbb{P}\left[X_m = y | X_n = x\right] = p^{(m-n)}(x,y), \quad \forall m \geq n.$$

On définit la frontière de Martin en espace-temps $\partial\tilde{E}$ et l'ensemble des points minimaux $\tilde{B}$ comme à la section 5.1.1. Les frontières en espace et en espace-temps sont très liées. Moltchanov [108, 109, 110] a en effet montré que, dans un certain nombre de cas, les points minimaux $f(n,x)$ en espace-temps se décomposent comme $f(n,x) = h(x)\lambda^{-n}$ avec $Ph = \lambda h$ (on dit dans ce cas que h est λ-*harmonique*). Or on construit — pour les fonctions λ–harmoniques aussi bien que 1-harmoniques — une frontière de Martin à partir du noyau de Green

$$G(x,y,z) \stackrel{\text{def}}{=} \sum_{n\geq 0} p^{(n)}(x,y) z^n, \qquad \forall z \in \big(0, 1/\rho(P)\big),$$

où $\rho(P)$ est le rayon spectral de P. Le paramètre z s'interprète comme l'accélération de la chaîne si $z \leq 1$, ou la décélération si $z \geq 1$. À la limite, si $z = 1/\rho(P)$, la chaîne est presque arrêtée : soit elle part très lentement à l'infini, soit elle ne part pas à l'infini (récurrence).

Illustrons sur un exemple l'*accélération*. Si $0 \leq z \leq 1$, définissons une nouvelle chaîne Y_n, de matrice de transitions zP sous-stochastique : la probabilité que le processus meurt à chaque étape est donc $1 - z$. Conditionnellement au fait qu'elle n'est pas tuée, la chaîne Y_n se déplace plus vite que X_n. Par exemple, si T_y et T_y^z désignent les temps respectifs de première atteinte de y, alors

$$\mathbb{E}_x\left[T_y^z | T_y^z < \infty\right] \leq \mathbb{E}_x\left[T_y | T_y < \infty\right], \quad \forall z \leq 1. \tag{5.6}$$

Posons $\pi_n \stackrel{\text{def}}{=} \mathbb{P}_x\left[X_n = y\right]$, alors

$$\begin{aligned} \phi(z) &\stackrel{\text{def}}{=} \mathbb{E}_x\left[T_y^z | T_y^z < \infty\right] = \frac{\sum_{n\geq 0} n\pi_n z^n}{\sum_{n\geq 0} \pi_n z^n}, \\ \phi'(z) &= \frac{1}{z}\left(\frac{\sum_{n\geq 0} n^2\pi_n z^n}{\sum_{n\geq 0} \pi_n z^n} - \left(\frac{\sum_{n\geq 0} n\pi_n z^n}{\sum_{n\geq 0} \pi_n z^n}\right)^2\right) = \frac{1}{z}\operatorname{Var}(\pi^z) \geq 0, \end{aligned}$$

où $\pi_n^z \stackrel{\text{def}}{=} \pi_n z^n / \sum_n \pi_n z^n$. Donc, par croissance, $\phi(z) \leq \phi(1)$, pour $z \leq 1$, d'où le résultat (5.6).

Cependant (5.6) n'est vraie que pour $y \in E$. On voudrait que cette heuristique reste vraie pour les points frontières : si $z \leq 1$ et $y \in \partial E$, alors Y_n converge plus vite que X_n vers y. Puisque $G(x, y, \lambda^{-1})$ génère les fonctions λ–harmoniques, cette heuristique peut s'exprimer sous une forme plus précise : si y est un point de la frontière de Martin correspondant à la fonction λ–harmonique h avec $\lambda \geq 1$ et si $y_n \in E$ est une suite convergeant vers y, alors y est atteint plus vite sous P^h que sous P.

Ici commencent les premières difficultés. L'heuristique ci-dessus ne mentionne qu'une seule frontière, ce qui signifie que les frontières de Martin sont homéomorphes pour différents $z \neq 1/\rho(P)$, c'est-à-dire que y_n converge. Cette propriété, appelée *stabilité*, a été formalisée et étudiée récemment par Picardello et Wöss [119, 120]. Elle est vraie pour certaines classes de chaînes de Markov (les marches aléatoires homogènes), mais est malheureusement fausse dans beaucoup d'autres cas — les marches avec frontières, p. ex. dans $\mathbb{Z}_+ \times \mathbb{Z}$ (voir [83]).

5.2.3 Bornes de grandes déviations

Définition 5.23 *On appelle changement d'échelle une suite de fonctions continues croissantes $\phi_n : \mathbb{R}_+ \mapsto [-\infty, 0]$, vérifiant le « principe du terme le plus grand », c'est-à-dire que pour toute paire de suites x_n, y_n bornées*

$$\limsup_{n\to\infty} \phi_n(x_n + y_n) = \max\left\{ \limsup_{n\to\infty} \phi_n(x_n), \limsup_{n\to\infty} \phi_n(y_n) \right\}.$$

On impose de plus la condition de normalisation $\phi_n(x) = 0$, pour tout n et pour tout $x \geq 1$.

Remarque : La condition de normalisation est arbitraire. Elle est choisie afin que l'entropie minimum soit 0 — c'est celle qui correspond à une probabilité 1 et à $x = 1$ dans la définition 5.23. Quant à la condition pour $x \geq 1$, compte tenu de la croissance des ϕ_n, cela équivaut, lorsqu'on majore une probabilité par une fonction f, à prendre la borne $f \wedge 1$ (p. ex. dans la borne (5.7)). Ce n'est pas une obligation théorique, mais cela nous rapproche du changement d'échelle habituel, qui est $\phi_n(.) = n^{-1} \log(.)$.

Nous aurions pu conserver ce changement d'échelle, qui est le plus courant, mais il est intéressant de constater que deux changements d'échelle différents peuvent conduire à deux PGD distincts, et que la

« magnitude » (forte ou faible) d'un PGD peut varier suivant le changement d'échelle choisi (voir l'exemple de la p. 84).

♦♦♦

Proposition 5.24 *Si $\phi_n(x_n)$ converge, alors $\phi_n(ax_n)$ converge aussi, pour tout $a > 0$, vers la même limite.*

Preuve : En considérant des sous-suites convergentes, il est possible d'écrire

$$\lim_{n\to\infty} \phi_n(x_n + x_n) = \lim_{n\to\infty} \phi_n(x_n).$$

En itérant cette égalité, on obtient

$$\lim_{n\to\infty} \phi_n(2^p x_n) = \lim_{n\to\infty} \phi_n(x_n), \quad \forall p \in \mathbb{Z},$$

et en encadrant $2^p \leq a \leq 2^q$, on obtient le résultat, grâce à la croissance des fonctions ϕ_n. ∎

Nous allons travailler sur des ensembles *en espace-temps*. La section au temps n de l'ensemble $O \subset \tilde{E}^*$ est $O_n \subset E$, c'est-à-dire que $y \in O_n$ si $(n, y) \in O$. Par ailleurs, nous supposons que la mesure de référence ν — qui est la distribution initiale du processus — est concentrée au temps 0 : nous la noterons donc ν_x (et non $\nu_{(0,x)}$) pour $x \in E$.

Proposition 5.25 *Pour tout ensemble $F \subset \tilde{E}^*$ et toute fonction f surharmonique et ν-intégrable, X_n vérifie la borne supérieure :*

$$\limsup_{n\to\infty} \phi_n\big(\mathbb{P}\left[X_n \in F_n\right]\big) \leq \limsup_{n\to\infty} \sup_{y\in F_n} \phi_n\left(f(n,y)^{-1}\right). \tag{5.7}$$

Dans la suite, l'écriture $f \in O$ signifiera que f est une fonction surharmonique ν-intégrable dont le support est dans O. Si $O \subset \tilde{E}^*$ est un ouvert et $f \in O \cap \tilde{B}$, alors X_n vérifie la borne inférieure :

$$\liminf_{n\to\infty} \phi_n\big(\mathbb{P}\left[X_n \in O_n\right]\big) \geq \liminf_{n\to\infty} \inf_{y\in O_n} \phi_n\left(f(n,y)^{-1}\right). \tag{5.8}$$

Preuve : Soit f une fonction harmonique en espace-temps, normalisée (c-à-d $Z = \sum_x \nu_x f(0,x) = 1$). Le calcul du f–processus correspondant donne

$$\mathbb{P}^f\left[X_n = y\right] = \sum_{x\in E} \nu_x^f p^f(0,x;n,y) = \mathbb{P}\left[X_n = y\right] f(n,y). \tag{5.9}$$

En convenant que $0/0 = 0$, on déduit de (5.9)

$$\begin{aligned}
\phi_n\left(\mathbb{P}\left[X_n \in F_n\right]\right) &= \phi_n\left(\sum_{y \in F_n} \mathbb{P}^f\left[X_n = y\right] f(n,y)^{-1}\right) \\
&\leq \phi_n\left(\mathbb{P}^f\left[X_n \in F_n\right] \sup_{y \in F_n} f(n,y)^{-1}\right) \qquad (5.10)\\
&\leq \sup_{y \in F_n} \phi_n\left(f(n,y)^{-1}\right).
\end{aligned}$$

En effet la probabilité $\mathbb{P}^f\left[X_n \in F_n\right]$ est inférieure à 1 et les fonction ϕ_n sont croissantes. À partir de cette borne, on déduit aisément (5.7).

Si la fonction n'est pas normalisée, puisqu'elle est ν-intégrable, les inégalités comprennent un terme multiplicatif $\nu f < \infty$ et d'après la proposition 5.24, ce terme disparaît dans (5.10) en passant à la limite.

La borne inférieure s'obtient en retournant la majoration (5.10) :

$$\phi_n\left(\mathbb{P}\left[X_n \in O_n\right]\right) \geq \phi_n\left(\mathbb{P}^f\left[X_n \in O_n\right] \inf_{y \in O_n} f(n,y)^{-1}\right).$$

Si $f \in O \cap \tilde{B}$, alors, d'après le théorème 5.18 et la proposition 5.22, $\mathbb{P}^f\left[X_n \in O_n\right]$ tend vers 1. Grâce aux propriétés du changement d'échelle, on en déduit

$$\begin{aligned}
\liminf_{n\to\infty} \phi_n\left(\mathbb{P}\left[X_n \in O_n\right]\right) &\geq \liminf_{n\to\infty} \phi_n\left(\inf_{y \in O_n} f(n,y)^{-1}\right) \\
&= \liminf_{n\to\infty} \inf_{y \in O_n} \phi_n\left(f(n,y)^{-1}\right).
\end{aligned}$$

La borne inférieure (5.8) est démontrée ■

Bien qu'aucune des deux bornes ne soit vraiment bonne — dans le sens où elle coïnciderait *toujours* avec le membre gauche dans les inégalités (5.8) et (5.7) lorsqu'une limite existe — elles peuvent être raffinées pour obtenir un résultat de grandes déviations.

Une partie de cette construction est inspirée de Lewis et Pfister, à ceci près que nous souhaitons « expliciter » la fonction ; nous renvoyons à leur article [97] pour cette approche.

Définition 5.26 *Soit f une fonction surharmonique, dont le support est S^f. On note $B(f,\varepsilon)$ la boule ouverte :*

$$B(f,\varepsilon) \stackrel{\text{def}}{=} \left\{x \in \tilde{E}^* : d_M(x, S^f) < \varepsilon\right\}.$$

La section au temps n est notée $B_n(f,\varepsilon)$. On définit les fonctions $\underline{I}$ et $\overline{I}$ par :

$$\underline{I}(f) \stackrel{\text{def}}{=} \lim_{\varepsilon\to 0}\liminf_{n\to\infty}\inf_{y\in B_n(f,\varepsilon)} \phi_n\left(f(n,y)^{-1}\right), \tag{5.11}$$

$$\overline{I}(f) \stackrel{\text{def}}{=} \lim_{\varepsilon\to 0}\limsup_{n\to\infty}\sup_{y\in B_n(f,\varepsilon)} \phi_n\left(f(n,y)^{-1}\right); \tag{5.12}$$

et puis successivement, en notant génériquement I au lieu de $\underline{I}$ et $\overline{I}$:

$$I(O) \stackrel{\text{def}}{=} \sup_{f\in O\cap\tilde{B}} I(f), \qquad \forall O\subset \tilde{E}^* \text{ ouvert}, \tag{5.13}$$

$$I(x) \stackrel{\text{def}}{=} \inf_{O\ni x} I(O), \qquad \forall x\in\tilde{E}^*. \tag{5.14}$$

Proposition 5.27 *Les fonctions $\underline{I}$ et $\overline{I}$, notées génériquement I, possèdent les propriétés suivantes.*

La fonction $f\to I(f)$ est négative.

La fonction $O\to I(O)$ est négative et croissante :

$$\begin{cases} I(O)\le 0, \\ I(O_1)\le I(O_2), \quad \forall O_1\subset O_2. \end{cases}$$

La fonction $x\to I(x)$ est négative et semi-continue supérieurement :

$$\begin{cases} I(x)\le 0, \\ \limsup_{y\to x} I(y)\le I(x). \end{cases}$$

Preuve : La fonction $f\to I(f)$ est négative car ϕ_n l'est.

La fonction $O\to I(O)$ est négative car $f\to I(f)$ l'est. Elle est croissante par construction car le supremum croît avec l'ensemble sur lequel on le calcule.

Enfin $x\to I(x)$ est négative car $O\to I(O)$ l'est. Montrons que la semi-continuité supérieure est la conséquence de la croissance de $O\to I(O)$. En effet, par croissance, $I(O)\ge I(x)$ pour tout $x\in O$. Prenons une suite $x_n\in\tilde{E}^*$ convergeant vers x et soit $\varepsilon>0$. Par définition de $I(x)$, il existe $\eta>0$ tel que $I(B(x,\eta))\le I(x)+\varepsilon$. Pour n assez grand ($n\ge n_0$), x_n est inclus dans $B(x,\eta)$ et par conséquent

$$I(x_n)\le I(B(x,\eta))\le I(x)+\varepsilon, \qquad \forall\varepsilon>0.$$

Ainsi $x\to I(x)$ est semi-continue supérieurement. ∎

Théorème 5.28 *Pour tout changement d'échelle ϕ_n, $\phi_n\big(\mathbb{P}\,[X_n \in .]\,\big)$ vérifie les bornes de grandes déviations dans l'espace de Martin en espace-temps avec les bonnes fonctionnelles d'action $\overline{H}(x) \stackrel{\text{def}}{=} -\underline{I}(x)$ et $\underline{H}(x) \stackrel{\text{def}}{=} -\overline{I}(x)$.*

Soit O un ouvert et F un fermé de $\tilde{E}^*$. Alors

$$\liminf_{n\to\infty} \phi_n\big(\mathbb{P}\,[X_n \in O_n]\,\big) \;\geq\; -\inf_{x\in O} \overline{H}(x), \tag{5.15}$$

$$\limsup_{n\to\infty} \phi_n\big(\mathbb{P}\,[X_n \in F_n]\,\big) \;\leq\; -\inf_{x\in F} \underline{H}(x). \tag{5.16}$$

Preuve : $\underline{H}$ et $\overline{H}$ sont de bonnes fonctionnelles d'actions car, étant semi-continues inférieurement, leurs niveaux sont fermés et donc compacts (car $\tilde{E}^*$ est compact et métrique).

La borne supérieure (5.16) se déduit directement de la construction de $\underline{H}$. Soit $x \in \tilde{E}^*$ et $\varepsilon > 0$. Alors il existe η_x tel que $\overline{I}\big(B(x,\eta_x)\big) \leq \overline{I}(x) + \varepsilon$ et, pour toute fonction harmonique $f \in B(x,\eta_x)$,

$$\limsup_{n\to\infty} \sup_{y\in B_n(x,2\eta_x)} \phi_n\left(f(n,y)^{-1}\right) \leq \overline{I}\big(B(x,\eta_x)\big) \leq \overline{I}(x) + \varepsilon.$$

D'après la borne (5.7), on en déduit

$$\limsup_{n\to\infty} \phi_n\big(\mathbb{P}\,[X_n \in B_n(x,2\eta_x)]\,\big) \leq \overline{I}(x) + \varepsilon.$$

Puisque F est fermé — donc compact — on peut le recouvrir par un nombre fini de telles boules $B(x,2\eta_x)$. De manière classique, en utilisant la sous-additivité des probabilités et le principe du terme le plus grand, on en déduit la borne supérieure.

La borne inférieure se déduit de manière similaire, ponctuellement. ■

Chaque fois que $\underline{H}(x) = \overline{H}(x)$ pour tout $x \in X \subset \tilde{E}^*$, un principe de grandes déviations local sur X découle des inégalités précédentes. On retrouve en particulier les résultats connus sur les marches aléatoires homogènes dans $\mathbb{Z}^d$ ou sur les arbres. Lorsqu'il y a des frontières (comme dans $\mathbb{Z}^d_+$), il se peut qu'en certains points apparaissent des effets de frontière et que $\underline{H}(x) < \overline{H}(x)$. Dans ce cas il faudrait poursuivre le travail, à savoir introduire des subdivisions dans $B(f,\varepsilon)$:

$$B(f,\alpha,\varepsilon) \stackrel{\text{def}}{=} \left\{(n,y) \in B_n(f,\varepsilon) : \left|\phi_n\big(f(n,y)^{-1}\big) - \alpha\right| < \varepsilon\right\}.$$

Or $\mathbb{P}^f$ charge $B(f,\varepsilon)$, dans le sens où $\lim \mathbb{P}^f[X_n \in B_n(f,\varepsilon)] = 1$; il existe donc un ensemble A de α tels que la limite $\lim \mathbb{P}^f[X_n \in B_n(f,\alpha,\varepsilon)]$ existe et soit strictement positive, quel que soit $\varepsilon > 0$. On définit alors

$$I(f) \stackrel{\text{def}}{=} \sup_{\alpha\in A} \alpha.$$

C'est l'entropie H déduite de cette fonction I qui devrait donner la bonne fonctionnelle d'action. Il est facile de vérifier que $\underline{I} \leq I \leq \overline{I}$. La démonstration de tels résultats, dans le cas général, n'est pas très aisée, mais ces définitions donnent un bon fil directeur pour des démonstrations particulières. On peut émettre la conjoncture que A est réduit à un seul élément, ou mieux encore, que $H = \overline{H}$. Quoi qu'il arrive, le lien entre grandes déviations et frontière de Martin est réel et assez direct. Nous espérons que ce chapitre, même sans complètement démontrer tous les résultats, suscite la conviction chez le lecteur.

Conséquences Tout d'abord, il ne peut y avoir de principe *fort* de grandes déviations sur E que si $\underline{H}(x) = \overline{H}(x) = \infty$ pour tout point de la frontière, excepté l'unique point correspondant à $f = 1$, qui correspond aux suites stationnaires $\{(n, y), n \geq 1\}$ avec y fixe, et dont l'entropie est nulle.

Ceci reste vrai pour le théorème de Sanov, dont la généralisation, devrait aussi faire appel à la frontière de Martin. Nous avons ici une condition qui semble *suffisante* pour le PGD fort. Il est donc envisageable, à partir de ces considérations, d'imaginer un élargissement de la condition uniforme (4.5), suffisante pour démontrer un principe fort mais non nécessaire.

Exemple : Prenons la chaîne suivante, dérivée de la $M/M/\infty$

$$p(n, n+1) = 1/(n+1) \quad \text{et} \quad p(n, n-1) = n/(n+1), \qquad \forall n \geq 0.$$

La frontière de Martin — sans tenir compte de la cyclicité — est paramétrée par $\lambda \geq 1$, et est définie par les fonctions λ–harmoniques

$$\begin{cases} h_1(n) = 1 & \text{si } \lambda = 1, \\ h_\lambda(n) = n! \lambda^n e^{o(n)} & \text{si } \lambda > 1. \end{cases}$$

Pour $\lambda > 1$ les chaînes obtenues par changement de mesure (*cf.* définition 5.21) sont

$$p_\lambda(n, n+1) = 1 - p_\lambda(n, n-1) \quad \text{et} \quad p_\lambda(n, n-1) = \left(n\lambda^2\right)^{-1} + o\left(n^{-1}\right).$$

On conjecture que les les trajectoires (n, x_n) correspondant aux points λ de la frontière de Martin sont

$$\begin{cases} x_n \leq n - \log n + o(\log n) & \text{si } \lambda = 1, \\ x_n = n - \lambda^{-2} \log n + o(\log n) & \text{si } 1 < \lambda < \infty, \\ x_n = n + o(\log n) & \text{si } \lambda = \infty. \end{cases}$$

où ∞ dépend du choix de la mesure de référence — p. ex. $\infty = (0,0)$ si $\nu = \delta_{(0,0)}$. Ces trajectoires ne diffèrent donc que très peu du mouvement déterministe $x_n = n$: remarquons même qu'elles ne diffèrent que par un terme de *deuxième* ordre.

Le changement d'échelle le mieux adapté à la structure des fonctions λ–harmoniques est

$$\phi_n(x) = \frac{1}{n}\frac{\log x}{\log|\log x|} \quad \text{si} \quad 0 < x < 1 \qquad \text{et } 0 \text{ sinon.}$$

qui est équivalent à la forme plus classique

$$\phi_n(x) = \frac{1}{n \log n} \log x.$$

Cette autre formule est équivalente à la première, en paramétrant $(n\log n, n\log n)$ les trajectoires (n, x_n) convergeant vers la frontière de Martin, puis en calculant l'entropie selon les définitions (5.11)–(5.12) en remplaçant $\phi_n\big(h_\lambda(n,x_n)^{-1}\big)$ par $\phi_{n\log n}\big(h_\lambda(n\log n, n\log n)^{-1}\big)$. On voit bien en quoi sont liés le changement d'échelle et la paramétrisation des trajectoires convergentes au sens de Martin.

L'entropie définie au théorème 5.28 est univaluée pour $\lambda > 1$ car $\overline{H}(\lambda) = \underline{H}(\lambda) = 1$; en revanche on a $\underline{H}(1) = 0$ mais $\overline{H}(1) = 1$. En introduisant des subdivisions dans $B(1,\varepsilon)$ (comme indiqué ci-dessus), on conjecture que la « bonne » valeur est 0 ; les trajectoires correspondantes sont celles qui sont *stables*, quand x_n est borné. Une autre entropie serait définie de manière semblable en prenant le changement d'échelle classique : $\phi_n(x) = (\log x)/n$. L'entropie vaudrait alors 0 pour $\lambda = 1$ et serait infinie ailleurs.

En utilisant les techniques de la section 4.4.3 avec le « bon » changement d'échelle, on déduit que les probabilités stationnaires doivent décroître en $\exp\{-x\log x(1+o(1))\}$. Le calcul montre que c'est juste, puisque $\pi_x = 1/(x!e)$. On ne peut rien déduire de tel avec le changement de mesure usuel. Dans ce dernier cas en revanche, puisque l'entropie est partout infinie — sauf pour les trajectoires stables — on conjecture que le théorème de Sanov est fort. En fait, on peut le montrer directement par des encadrements.

Cet exemple pourrait se transposer facilement au temps continu. Cependant, ce cadre réserve des difficultés théoriques. Dans le cas de la $M/M/\infty$, les trajectoires correspondant aux points de la frontière de Martin sont des fonctions exponentielles du temps et les processus modifiés ne sont pas explosifs ; mais dans un exemple *plus* ergodique — l'intensité de service au point n vaut n^2 et les arrivées sont toujours d'intensité λ — les processus modifiés sont explosifs. Ce genre de difficultés rend la construction théorique plus lourde, mais les idées sont en tout point similaires à celles que cet exemple met en valeur : *une unité de structure entre la frontière de Martin, les grandes déviations et la décroissance des probabilités stationnaires.*

♦♦♦

Chapitre 6

Extensions et conclusion

6.1 Extensions

Mous présentons dans cette section trois possibilités d'élargissement de la technique de changement de mesure : la prise en compte de distributions générales et non pas exponentielles comme dans le cas markovien (section 6.1.1) ; le calcul de développements asymptotiques plus précis que l'équivalent logarithmique fourni par la théorie des grandes déviations (section 6.1.2) ; enfin la généralisation à des espaces continus (section 6.1.3). Les domaines abordés ici ne sont pas vierges, mais toutes ces considérations sont personnelles.

6.1.1 Processus semi-markoviens

Beaucoup de modèles peuvent être mis sous une forme plus générale, lorsqu'on remplace les hypothèses de temps de service ou d'inter-arrivées de lois exponentielles par des distributions quelconques.

Poursuivons avec le cas générique du polling à N files et un serveur (II 4, ou I 4.3). Prenons des temps de service de loi $\mu_i(\,\mathrm{d}t)$, des temps de déplacement de loi $\tau_{ij}(\,\mathrm{d}t)$, et des inter-arrivées de loi $\lambda_i(\,\mathrm{d}t)$; le routage reste P_{ij}. On notera $R = (\lambda_i, \mu_i, \tau_{ij}, P_{ij})$ les paramètres de ce modèle.

En notant λ_i, μ_i les intensités des lois correspondantes, et τ_{ij} la moyenne, les conditions d'ergodicité restent les mêmes qu'au théorème II 4.4.

Nous continuons à appliquer la maxime *mieux vaut appliquer un théorème simple sur un espace compliqué qu'un théorème compliqué sur un espace simple.* Pour cela nous allons enrichir le processus semi-markovien (Q_t, S_t) pour qu'il devienne markovien sur un espace plus compliqué. Nous pourrons alors appliquer un changement de mesure et en déduire un PGD. Une contraction nous ramènera ensuite sur l'espace d'origine. Notons que c'est déjà ce que nous avons fait : Q_t n'est pas markovien mais (Q_t, S_t) l'est, donc on obtient un PGD pour la paire, puis on contracte ([35] ou II 4).

On introduit donc les temps résiduels d'arrivées Z_t^i (associé à la file i), et le temps résiduel de service (ou de déplacement) Y_t. Finalement

$$X_t \stackrel{\text{def}}{=} \big(Q_t, Z_t, S_t, Y_t\big)$$

est markovien. C'est exactement le cas décrit en [68], lorsqu'il y a un seul serveur et que la politique est 1-limitée. L'évolution est la suivante ; si tous les temps résiduels sont positifs, ils décroissent linéairement, le reste du système étant inchangé :

$$\forall 0 \leq s < \min\{Z_t^i, Y_t\} \qquad \begin{cases} Z_{t+s}^i = Z_t^i - s, \\ Y_{t+s} = Y_t - s, \\ Q_{t+s} = Q_t, \\ S_{t+s} = S_t. \end{cases}$$

Sinon, un saut intervient ; prenons par exemple, le cas du serveur quittant la file i (et donc la file i n'était pas vide, $Q_{t^-}^i \geq 1$ mais $Y_{t^-} = 0$) :

$$\mathbb{E}\left[f(X_t) \middle| X_{t^-} = (q, z, i, 0)\right] = \sum_{j \in \mathcal{S}} P_{ij} \int_0^\infty \tau_{ij}(\,\mathrm{d}y) f(q - e_i, z, ij, y).$$

Remarque : La suite, par souci de concision, n'est pas entièrement développée. La démarche est absolument la même qu'à la section II 4 : il faudrait considérer des processus localisés, avec des changements de mesures qui conservent la stabilité du système, puis des contractions partielles. Pour éviter de reproduire ces considérations, ces questions sont sous-entendues ici. De plus, on suppose assez de régularité sur les lois λ_i, μ_i et τ_{ij} pour procéder à toutes les opérations qui suivent (et notamment des moments exponentiels suffisamment réguliers pour que le théorème de Cramer s'applique).

♦♦♦

Générateur empirique Les notations sont essentiellement reprises du chapitre II 4 ; p. ex. $\mathcal{S}_0$ est l'ensemble des positions du serveurs. On introduit

- T_n [resp. T_n^i] le n–ième temps de saut pour le serveur [resp. les arrivées sur la file i] ; on convient que $T_0 = T_0^i = 0$;
- $\Delta T_n \stackrel{\text{def}}{=} Y_{T_n}$ [resp. $\Delta T_n^i \stackrel{\text{def}}{=} Z_{T_n^i}^i$] la n–ième durée inter-saut pour le serveur [resp. inter-arrivées sur la file i] ; les nombres ΔT_0 et ΔT_0^i sont des données du problème ;

– N_t [resp. N_t^i] le nombre de sauts sur $[0,t]$ pour le serveur [resp. les arrivées sur la file i] ;

– S_n la chaîne incluse ($S_n \stackrel{\text{def}}{=} S_{T_n}$) ;

– le générateur empirique $G_t \stackrel{\text{def}}{=} (L_t, D_t)$:

$$\begin{aligned} L_t &\stackrel{\text{def}}{=} \frac{1}{t}\sum_{n=1}^{N_t} \delta(S_n, \Delta T_n) \in \mathcal{M}_+(\mathcal{S}_0 \times \mathbb{R}_+), \\ D_t^i &\stackrel{\text{def}}{=} \frac{1}{t}\sum_{n=1}^{N_t^i} \delta(\Delta T_n^i) \in \mathcal{M}_+(\mathbb{R}_+). \end{aligned}$$

De façon similaire à la section II 3.2.1, le générateur empirique $G_t \in \mathcal{G}$ tend vers un ensemble limite $\mathcal{G}_s$ de « générateurs balancés », qui décrivent les contraintes moyennes du système, notamment que, pour chaque file, le serveur entre autant de fois qu'il sort d'une file i ou bien qu'il ne peut y avoir plus de départs que d'arrivées sur les files non-saturées. Dans l'ensemble limite $\mathcal{G}_s$ les éléments $G = (A, D)$ s'interprètent comme

$$\begin{cases} \widetilde{\lambda}_i(\,\mathrm{d}t) & \stackrel{\text{def}}{=} D^i(\,\mathrm{d}t)/d^i, \\ \widetilde{\mu}_i(\,\mathrm{d}t) & \stackrel{\text{def}}{=} A(i,\,\mathrm{d}t)/a(i), \\ \widetilde{\tau}_{ij}(\,\mathrm{d}t) & \stackrel{\text{def}}{=} A(ij,\,\mathrm{d}t)/a(ij), \\ \widetilde{P}_{ij} & \stackrel{\text{def}}{=} a(ij)/\sum_j a(ij), \end{cases}$$

en posant les constantes de normalisation (de façon à travailler sur des probabilités)

$$d^i \stackrel{\text{def}}{=} D^i(\mathbb{R}_+) \quad \text{et} \quad a(s) \stackrel{\text{def}}{=} A(s, \mathbb{R}_+).$$

Changement de mesure On définit ensuite les fonctionnelles

$$\begin{cases} f_1^i(.) & \stackrel{\text{def}}{=} \log \dfrac{\mathrm{d}\widetilde{\lambda}_i}{\mathrm{d}\lambda_i}, \\ f_2(i,.) & \stackrel{\text{def}}{=} \begin{cases} \log \dfrac{\mathrm{d}\widetilde{\mu}_i}{\mathrm{d}\mu_i}, & \text{si } s = i, \\ \log \dfrac{\mathrm{d}\widetilde{\tau}_{ij}}{\mathrm{d}\tau_{ij}}, & \text{si } s = ij, \end{cases} \\ f_3(ij) & \stackrel{\text{def}}{=} \log \dfrac{\widetilde{P}_{ij}}{P_{ij}}, \end{cases}$$

et finalement une fonctionnelle linéaire $\phi_{R,G}$ sur les générateurs empiriques :

$$\langle G_t, \phi_{R,G}\rangle \stackrel{\text{def}}{=} \sum_{i\in\mathcal{S}} \int D_t^i(\mathrm{d}u) f_1^i(u) + \sum_{s\in\mathcal{S}_0} \int L_t(s,\mathrm{d}u) f_2(s,u) + \sum_{i,j\in\mathcal{S}} L_t(ij,\mathbb{R}_+) f_3(ij).$$

On en dérive une martingale et le changement de mesure convenable

$$M_t \stackrel{\text{def}}{=} e^{t\langle G_t, \phi_{R,G}\rangle}.$$

On vérifie que sous la nouvelle probabilité $\mathbb{P}^*$, le processus X_t est un polling de paramètres $G = (\widetilde{\lambda}_i, \widetilde{\mu}_i, \widetilde{\tau}_{ij}, \widetilde{P}_{ij})$. On pose $\widetilde{A} \stackrel{\text{def}}{=} \big(a(ij)\big)$ et on introduit l'entropie[1] relative

$$\begin{aligned} H(G\|R) \stackrel{\text{def}}{=} & \sum_{i\in\mathcal{S}} d^i H(\widetilde{\lambda}_i\|\lambda_i) + a(i) H(\widetilde{\mu}_i\|\mu_i) \\ & + \sum_{i,j\in\mathcal{S}} a(ij) H(\widetilde{\tau}_{ij}\|\tau_{ij}) + H_d(\widetilde{A}\|P). \end{aligned} \qquad (6.1)$$

Une fois de plus cette entropie est exactement la somme des coûts liés aux modifications introduites et pouvait se deviner avant tout calcul. Elle est liée au changement de mesure par

$$H(G\|R) = \langle G, \phi_{R,G}\rangle.$$

Or [68] montre que G_t, sous $\mathbb{P}^*$, converge en probabilité vers G ; de plus H possède de bonnes propriétés de continuité. On en déduit que $G_t \in \mathcal{G}$ vérifie un PGD avec la bonne fonctionnelle d'action qui vaut H sur $\mathcal{G}_s$ et l'infini en dehors.

Contraction Un inconvénient de travailler sur un espace immense, et assurément $\mathcal{G}_s$ l'est, est de ne pas aisément pouvoir implémenter les calculs si la théorie n'aboutit pas à une formule simple. Or, si comme à la section 4.4.3 on s'intéresse à la manière de saturer le réseaux et au régime stationnaire, il faudrait disposer d'une représentation de la probabilité de partir dans une certaine direction. Même si on ne s'intéresse qu'à la simulation (section 4.4.2), il faut pouvoir choisir convenablement le changement de mesure.

En fait, la démarche la plus commode est de réduire G_t de manière à se ramener exactement aux termes du chapitre II 4. La suite est identique, avec une fonctionnelle $\bar{H}$ modifiée.

1. On note identiquement H pour l'entropie relative entre générateurs ou entre distributions de probabilité, dans la mesure où il n'y a pas de confusion possible, en considérant les objets sur lesquels s'applique l'entropie.

On note $\bar{G} = (\bar{\lambda}_i, \bar{\mu}_i, \bar{\tau}_{ij}, \bar{P}_{ij})$ un générateur contracté, à la manière de la section II 4.3.2. On s'intéresse donc au problème de minimisation :

$$\bar{H}(\bar{G}\|R) \stackrel{\text{def}}{=} \inf H(G\|R) \quad \text{avec} \left\{ \begin{array}{rcl} \displaystyle\int_0^\infty t\widetilde{\lambda}_i(\,\mathrm{d}t) & = & \bar{\lambda}_i^{-1}, \\ \displaystyle\int_0^\infty t\widetilde{\mu}_i(\,\mathrm{d}t) & = & \bar{\mu}_i^{-1}, \\ \displaystyle\int_0^\infty t\widetilde{\tau}_{ij}(\,\mathrm{d}t) & = & \bar{\tau}_{ij}, \\ \widetilde{P}_{ij} & = & \bar{P}_{ij}. \end{array} \right.$$

Le « miracle » c'est que cette minimisation est résolue, dans la mesure où on connaît la solution du problème de moyenne empirique pour des variables i.i.d. (voir section 2.1). Le théorème 2.1 donne la forme de l'entropie contractée. En notant $\Lambda^*(.\|\nu)$ la transformée de Legendre de la fonction des cumulant de la mesure ν sur $\mathbb{R}$,

$$\begin{aligned} \bar{H}(\bar{G}\|R) \;=\; & \sum_{i\in\mathcal{S}} \Lambda^*(\bar{\lambda}_i^{-1}\|\lambda_i) + a(i)\Lambda^*(\bar{\mu}_i^{-1}\|\mu_i) \\ & + \sum_{i,j\in\mathcal{S}} a(ij)\Lambda^*(\bar{\tau}_{ij}\|\tau_{ij}) + H_d(\bar{A}\|P). \end{aligned} \tag{6.2}$$

En remarquant que, si $\bar{G} = (\bar{\lambda}_i, \bar{\mu}_i, \bar{\tau}_{ij}, \bar{P}_{ij})$ est représenté de manière équivalente par $\bar{G} = (\bar{A}, \bar{\pi}, \bar{D})$, alors on retrouve les relations (4.24)–(4.27) du chapitre II 4 :

$$\left\{ \begin{array}{rcll} \bar{a}_s & = & a(s), & \\ \bar{\pi}_i & = & \int tA(i,\,\mathrm{d}t) & = \bar{a}_i/\bar{\mu}_i, \\ \bar{\pi}_{ij} & = & \int tA(ij,\,\mathrm{d}t) & = \bar{a}_{ij}\bar{\tau}_{ij}, \\ \bar{D}_i & = & \int D^i(\,\mathrm{d}t) - \int A(i,\,\mathrm{d}t) & = \bar{\lambda}_i - \bar{a}_i, \\ \bar{P}_{ij} & = & a(ij)/\sum_j a(ij) & = \bar{a}_{ij}/\sum_j \bar{a}_{ij}. \end{array} \right.$$

Finalement, on poursuit l'étude du polling, de façon similaire au modèle markovien, en considérant l'entropie

$$\begin{aligned} \bar{H}(\bar{G}\|R) \;=\; & \sum_{i\in\mathcal{S}} \bar{\lambda}_i\Lambda^*(\bar{\lambda}_i^{-1}\|\lambda_i) + \bar{\pi}_i\bar{\mu}_i\Lambda^*(\bar{\mu}_i^{-1}\|\mu_i) \\ & + \sum_{i,j\in\mathcal{S}} \bar{\pi}_{ij}\bar{\tau}_{ij}^{-1}\Lambda^*(\bar{\tau}_{ij}\|\tau_{ij}) + H_d(\bar{A}\|P). \end{aligned} \tag{6.3}$$

Notons la cohérence avec la définition II 4.12, en posant

$$I(\bar{\nu}\|\nu) \stackrel{\text{def}}{=} \bar{\nu}\Lambda^*(\bar{\nu}^{-1}\|\nu).$$

C'est aussi exactement le raisonnement que nous avons fait pour déduire la forme de I_p, pour des sauts poissonniens, à partir de variables exponentielles (voir la relation (4.10) et son interprétation).

6.1.2 Développements asymptotiques

Nous allons voir comment la structure de l'entropie permet de pousser plus loin certains calculs asymptotiques. La question n'est pas neuve puisque dans les années soixante, on a montré que la décroissance de la probabilité pour la moyenne empirique $\hat{S}_n$ d'appartenir à un ensemble O était en $O\big(n^{-1/2}\big)e^{-nH}$, pour des variables i.i.d. sur $\mathbb{R}$ [8] et pour des chaînes de Markov à espace d'états fini [106]. Plus récemment, une analyse fine dans $\mathbb{R}^d$ utilisant les techniques des grandes déviations [111, 84] a montré que l'on pouvait calculer l'exposant exact du terme n^γ, avec $-d/2 \leq \gamma \leq -1/2$, et que la décroissance était en $O(1)n^{-\gamma}e^{-nH}$. Nous souhaitons poursuivre ici ces développements asymptotiques, notamment en explicitant les constantes, par des moyens qui ne dépendent que peu du type de problème envisagé. De plus on considérera des fonctionnelles additives $\hat{S}_n$ plus générales que la moyenne.

Le cadre est celui du chapitre 2 : X_i est une suite de variables i.i.d. de loi μ, L_n est leur mesure empirique et $\hat{S}_n$ la moyenne empirique. Supposons de plus que μ soit une distribution continue assez régulière. Dans ce cas, le PGD fort implique que

$$\log \mathbb{P}\left[\hat{S}_n \geq x_0\right] \sim -n\Lambda^*(x_0) = -nH(\nu\|\mu),$$

pour une certaine mesure ν (qui réalise le minimum de la contraction de L_n en $\langle L_n, f\rangle = \hat{S}_n$ avec $f(t) = t$). Nous allons montrer qu'on peut calculer des quantités plus précises, du type

$$\mathbb{P}\left[\frac{a}{n} \leq \hat{S}_n - x_0 \leq \frac{b}{n}\right] \sim \frac{e^{-nH(\nu\|\mu)}}{\sqrt{2\pi n \operatorname{Var}(\nu)}} \int_a^b e^{-\alpha u}\,\mathrm{d}u. \tag{6.4}$$

Cette analyse repose sur une identité remarquable satisfaite par l'entropie (lorsque les termes ont un sens) :

$$H(\nu + \mathrm{d}\nu\|\mu) = H(\nu\|\mu) + \nabla H(\nu\|\mu).\,\mathrm{d}\nu + H(\nu + \mathrm{d}\nu\|\nu), \tag{6.5}$$

où μ et ν sont deux mesures de $\mathcal{M}_1(E)$ et $\mathrm{d}\nu$ un élément de l'espace vectoriel tangent à $\mathcal{M}_1(E)$ en ν, que nous noterons $\mathcal{T}(\nu)$; c'est l'espace des mesures η, signées, finies, dont la masse vaut 0, absolument continues à ν et dont la dérivée de Radon-Nikodym $\mathrm{d}\eta/\mathrm{d}\nu$ est bornée.

L'équation (6.5) constitue, d'une certaine mesure, un *développement* de $H(\nu + \mathrm{d}\nu\|\mu)$, avec le terme constant $H(\nu\|\mu)$, la dérivée $\nabla H(\nu\|\mu).\mathrm{d}\nu$ et les termes d'ordre supérieur à deux $H(\nu + \mathrm{d}\nu\|\nu)$. De façon heuristique, la suite consiste en ce développement de H, c'est-à-dire en analysant la *concentration* de L_n au voisinage de la mesure optimale ν.

Soit $\nu \in M_1(E)$ absolument continue relativement à μ. On prouve que H est deux fois dérivable en ν, que ses différentielles sont des applications *continues* définies par

$$\begin{aligned} \nabla H(\nu\|\mu).\theta &= \int \theta(\mathrm{d}x) \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \int \nu(\mathrm{d}x) \frac{\mathrm{d}\theta}{\mathrm{d}\nu} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}, && (6.6) \\ \nabla^2 H(\nu).\theta.\eta &= \int \eta(\mathrm{d}x) \frac{\mathrm{d}\theta}{\mathrm{d}\nu} = \int \nu(\mathrm{d}x) \frac{\mathrm{d}\eta}{\mathrm{d}\nu} \frac{\mathrm{d}\theta}{\mathrm{d}\nu}, && (6.7) \end{aligned}$$

où $\theta, \eta \in \mathcal{T}(\nu)$. On constate bien sous cette forme la convexité de H, car l'opérateur quadratique $\nabla^2 H(\nu)$ est positif, correspondant au produit scalaire et à la norme canoniques sur les fonctions appartenant à $L_2(\nu)$. En outre $\nabla^2 H(\nu)$ est *inversible*. Ce qui est remarquable, c'est que $\nabla^2 H(\nu)$ *ne dépende que de* ν*, et pas de* μ — on peut déjà le remarquer sur (6.5).

On peut généraliser les situations en considérant une fonctionnelle réelle f et en définissant $\hat{S}_n \stackrel{\text{def}}{=} \langle L_n, f\rangle$. On suppose que f n'est pas constante, cas suffisamment particulier pour qu'on puisse le résoudre à part. Les différentielles peuvent aussi être représentées par des fonctions de $L_1(\nu)$:

$$f \equiv \nabla H(\nu\|\mu) \iff \langle f, \theta\rangle = \nabla H(\nu\|\mu).\theta, \quad \forall \theta \in \mathcal{T}(\nu).$$

On peut prouver que la différentielle en ν est définie à une constante près, c'est-à-dire que $f \equiv g$ si, et seulement si, $f = g + C^{\text{ste}}$ ν–p.s.

Considérons une mesure $\nu \in \mathcal{M}_1(E)$ absolument continue relativement à μ ; on lui associe (comme à la section 2.2.4) la fonctionnelle

$$\phi_{\mu,\nu} \stackrel{\text{def}}{=} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \equiv \nabla H(\nu\|\mu), \tag{6.8}$$

qui vérifie $\langle \nu, \phi_{\mu,\nu}\rangle = H(\nu\|\mu)$, puis définissons la martingale M_n par

$$\log M_n \stackrel{\text{def}}{=} n\langle L_n, \phi_{\mu,\nu}\rangle = nH(\nu\|\mu) + n\nabla H(\nu\|\mu).(L_n - \nu).$$

Le changement de mesure intervient très naturellement, car pour $O \subset \mathcal{M}_1(E)$ quelconque,

$$\mathbb{P}\left[L_n \in O\right] = e^{-nH(\nu\|\mu)} \mathbb{E}^* \left[e^{-n\nabla H(\nu\|\mu).(L_n - \nu)} \mathbb{1}_{\{L_n \in O\}}\right]. \tag{6.9}$$

Méthode de Laplace : Un problème très voisin, qui donne des résultats similaires, mais bien distincts de ceux-ci, est de considérer dans (6.9) une fonction $\Phi(L_n)$ *régulière* (plusieurs fois différentiable), au lieu de l'indicatrice $\mathbb{I}_{\{L_n \in O\}}$, et de calculer $\mathbb{E}\left[e^{n\Phi(L_n)}\right]$ (voir [14]). On peut alors montrer que l'asymptotique est en $Ce^{nC'}\big(1+o(1)\big)$. Notre approche diffère d'une part en ce qu'elle montre le lien entre les différentielles de l'entropie et la solution, et d'autre part par les effets de la discontinuité, qui induit des termes en n^γ, voire d'autres perturbations (*cf.* remarque « Discrétisation » p. 95).

♦♦♦

On choisit le domaine O de la forme $\hat{S}_n \in [x_0 + an^{-1}, x_0 + bn^{-1}]$, soit

$$O = \big\{a \leq n\langle f, L_n - \nu\rangle \leq b\big\} \qquad \text{avec } -\infty < a \leq b \leq \infty.$$

Sans perte de généralité, on peut supposer que $\langle f, \mu\rangle \leq x_0$. Dans un premier temps on montre que le minimum de l'entropie sur $\{\nu : \langle f, \nu\rangle \geq x_0\}$ est atteint sur la frontière $\langle f, \nu\rangle = x_0$, que H y est différentiable et que

$$\nabla H(\nu\|\mu) \parallel f \quad \text{et} \quad \langle \nabla H(\nu\|\mu), f\rangle \geq 0.$$

Cela signifie qu'il existe α tel que $\nabla H(\nu\|\mu) \equiv \alpha f$ (car f n'est pas constante).

Dans un deuxième temps on montre que le minimum de l'entropie sur O (qui dépend de n) est atteint en une mesure ν_n qui converge vers ν et telle que $H(\nu_n\|\nu) = \sigma^2/2n$, où

$$\sigma^2 \stackrel{\text{def}}{=} \big\langle \nabla^2 H(\nu)^{-1}.f, f\big\rangle = \int f^2 \, \mathrm{d}\nu. \tag{6.10}$$

Or, on sait que $\sqrt{n}\langle f, L_n - \nu\rangle$ vérifie un théorème central limite (TCL) sous $\mathbb{P}^*$. Le calcul précédent permet de montrer que la variance est exactement σ^2.

En appliquant le TCL *local* (voir [66, Ch. XVI 2] ou [81, Theorem 4.2.1] pour les variables discretes), on en déduit que $n\langle f, L_n - \nu\rangle$, qui est voisine de 0, est une variable de densité

$$n\langle f, L_n - \nu\rangle = \frac{\mathrm{d}u}{\sqrt{2\pi n\sigma^2}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{6.11}$$

où $\mathrm{d}u$ est la mesure de Lebesgue (uniforme sur $\mathbb{R}$). Or, puisque $\nabla H(\nu\|\mu) \equiv \alpha f$, on a $\nabla H(\nu).(L_n - \nu) = \alpha\langle f, L_n - \nu\rangle$ et donc, en combinant (6.11) et (6.9), on obtient le développement (6.4). En effet, dans le cas de $f(t) = t$, σ^2 est exactement la variance de ν.

$$\mathbb{P}\left[a \leq n\langle f, L_n - \nu\rangle \leq b\right] \sim e^{-nH(\nu\|\mu)} \int_a^b e^{-\alpha u} \frac{\mathrm{d}u}{\sqrt{2\pi n\sigma^2}}.$$

Discrétisation : De toutes les considérations précédentes, dont on peut légitimement penser qu'elles requièrent des conditions restrictives sur les données du problème, μ, f, x_0, a, b, une analyse plus poussée montre que c'est (6.11) qui est le plus sujet à caution.

Il semble que (6.11) ne soit à modifier que dans seul le cas *très particulier* où f est « arithmétique », à savoir qu'il existe $k > 0$ et k' tels que $f(x) \in k\mathbb{Z} + k'$ pour μ–presque tout $x \in E$. On se rend compte en effet dans ce cas que $n\hat{S}_n$ appartient nécessairement à $k\mathbb{Z} + nk'$. L'approximation (6.11) est encore valable, mais au lieu d'être la mesure de Lebesgue sur $\mathbb{R}$, $\mathrm{d}u$ vaut k fois la mesure de comptage sur $k\mathbb{Z} + nk'$, ce qui est l'équivalent discret de la mesure de Lebesgue. A cette restriction près la suite reste identique.

Les termes constants dans les formules asymptotiques deviennent cependant oscillants. La probabilité ne décroît donc pas exactement en $Cn^{-\gamma}e^{-nH}$. Considérons les deux cas de figure, avec $f(t) = t$, $a = 0$ et $b = \infty$. En posant $h_0(n)$ le décalage de la grille $k\mathbb{Z} + nk'$ par rapport à $k\mathbb{Z} + n\langle \nu, f\rangle$:

$$h_0(n) \stackrel{\text{def}}{=} k\left(\left\lceil \frac{n}{k}\big(\langle \nu, f\rangle - k'\big)\right\rceil - \frac{n}{k}\big(\langle \nu, f\rangle - k'\big)\right) \in [0, k[,$$

on obtient, dans le cas non-arithmétique,

$$\mathbb{P}\left[\hat{S}_n \geq x_0\right] \sim \frac{e^{-nH(\nu\|\mu)}}{\sqrt{2\pi n\sigma^2}}\frac{1}{\alpha}, \tag{6.12}$$

ou bien, dans le cas arithmétique,

$$\mathbb{P}\left[\hat{S}_n \geq x_0\right] \sim \frac{e^{-nH(\nu\|\mu)}}{\sqrt{2\pi n\sigma^2}}\frac{ke^{-\alpha h_0(n)}}{(1 - e^{-k\alpha})}. \tag{6.13}$$

L'influence de la discrétisation apparaît donc nettement à ce niveau, par le biais de $h_0(n) \in [0, k[$ et elle engendre des difficultés pour l'obtention de développements plus précis. On remarque par ailleurs que si k tend vers 0, alors (6.12) se déduit de (6.13).

♦♦♦

On peut généraliser la méthode à des parties de $\mathbb{R}$ plus générales qu'un segment, mais aussi à un nombre de contraintes plus grand. Considérons une fonctionnelle dans $\mathbb{R}^p$, $f = (f_1, \ldots, f_p)$ — qui définit $\hat{S}_n \stackrel{\text{def}}{=} \langle f, L_n\rangle$ — et l'ensemble

$$O \stackrel{\text{def}}{=} \left\{n(\hat{S}_n - x_0) \in K \subset \mathbb{R}^p\right\}.$$

Sous des hypothèses qui expriment que toutes les contraintes sont actives, on montre qu'il existe $\alpha \in \mathbb{R}^p$ tel que $\nabla H(\nu\|\mu) \equiv \sum_i \alpha_i f_i = \alpha^t f$. En reproduisant les calculs précédents, si μ et f sont assez régulières, on obtient

$$\mathbb{P}\left[n(\hat{S}_n - x_0) \in K\right] \sim \frac{e^{-nH(\nu\|\mu)}}{(2\pi n)^{p/2}\sigma} \int_K e^{-\langle \alpha, u\rangle}\,\mathrm{d}u, \tag{6.14}$$

avec

$$\sigma^2 \stackrel{\text{def}}{=} |\det \Sigma| \quad \text{et} \quad \Sigma_{ij} \stackrel{\text{def}}{=} \left\langle \nabla^2 H(\nu)^{-1}.f_i, f_j\right\rangle = \int (f_i(x) - \bar{f}_i)(f_j(x) - \bar{f}_j)\,\mathrm{d}\nu.$$

Unification des échelles Outre le calculs de développements asymptotiques, on peut déduire plusieurs conclusions de ces caclculs. En premier lieu, le changement de mesure permet de *calculer la variance*. Nous avons en effet supposé que $\sqrt{n}\langle f, L_n - \mu\rangle$ satisfaisait un TCL et nous en avons déduit la variance, à partir de développements locaux sur un problème d'optimisation.

Le second point est d'avoir montré que ceci restait vrai pour n'importe quelle dimension : on en déduit que, localement autour de μ, $\langle f, L_n - \mu\rangle$ tend en loi dans $\mathcal{T}(\mu)$ vers un processus gaussien centré dont l'inverse[2] de la covariance est $\nabla^2 H(\mu)$.

Le troisième point est que nous n'avons à aucun moment considéré la structure de l'espace E. Le résultat est donc valable pour toute suite de variables i.i.d. En fait, on peut montrer que toute la construction théorique est encore valable pour des chaînes de Markov. En utilisant le hessien exactement comme dans l'équation (6.10), on obtient dans ce cas la variance sous la forme [29, équation (2.2)] (qui est utilisée dans le cadre des *moyennes* déviations : voir aussi [30]). La conclusion, en particulier les développements asymptotiques et l'identification de la variance, reste valide ; malheureusement, le hessien $\nabla^2 H$ n'est pas toujours inversible dans le cas markovien et on doit donc recourrir à des hypothèses restrictives. Quoi qu'il advienne, il y a ici une jolie voie pour unifier petites (TCL), moyennes et grandes déviations.

Les deux exemples qui suivent montrent comment quelques propriétés peuvent simplement être calculées par ces méthodes. Notons toutefois que, dans le cas des variables i.i.d, une simplification supplémentaire intervient car, à partir de la caractérisation de l'équivalence $\alpha f \equiv \nabla H(\nu\|\mu)$ et de (6.8), on en déduit que $\phi_{\mu,\nu} = \alpha f + \beta$ et finalement ν est donnée par le système :

$$\nu = \mu e^{\alpha f + \beta} \quad \text{avec} \quad \begin{cases} \langle \nu, \mathbb{1}\rangle = 1, \\ \langle \nu, f\rangle = x_0. \end{cases} \tag{6.15}$$

2. Remarquons qu'ici $\nabla^2 H(\mu)$ peut être identifiée à l'opérateur identité, qui est égal à son inverse, mais ce n'est pas toujours le cas.

Loi conditionnelle : À partir de l'équation (6.14) on peut exhiber diverses lois conditionnelles. Nous prenons comme exemple la condition $\{\hat{S}_n \geq x_0\}$ (avec $\hat{S}_n = \langle L_n, f\rangle$), mais on peut facilement en imaginer d'autres.

La loi conditionnelle la plus simple est celle de $\hat{S}_n$, conditionnellement à $\{\hat{S}_n \geq x_0\}$. En effet, en utilisant la définition de la probabilité conditionnelle, pour des événements de probabilité non-nulle, on obtient

$$\begin{aligned}\mathbb{P}\left[n(\hat{S}_n - x_0) \in [a,b] \mid \hat{S}_n \geq x_0\right] &= \frac{\mathbb{P}\left[n(\hat{S}_n - x_0) \in [a,b] \cap \mathbb{R}_+\right]}{\mathbb{P}\left[\hat{S}_n \geq x_0\right]} \\ &\stackrel{n\to\infty}{\longrightarrow} \alpha \int_a^b \mathbb{1}_{\{u\geq 0\}} e^{-\alpha u}\,\mathrm{d}u.\end{aligned}$$

On en déduit que, conditionnellement à $\{\hat{S}_n \geq x_0\}$, $n(\hat{S}_n - x_0)$ tends vers une loi exponentielle de paramètre α. On voit que la moyenne et la variance conditionnelles de $\hat{S}_n$ sont $x_0 + 1/n\alpha$ et $1/(n\alpha)^2$. La distribution conditionnelle est vraiment très ramassée au voisinage de x_0.

La plus connue des lois conditionnelles est décrite par le *principe de conditionnement de Gibbs*, qui stipule que, conditionnellement à un événement rare sur la moyenne (ici $\{\langle L_n, f\rangle \geq x_0\}$), les k premières variables $X_1, \ldots, X_k$ sont indépendantes et suivent la loi ν qui minimise l'entropie $H(\nu\|\mu)$ sur l'ensemble $\{\langle \nu, f\rangle \geq x_0\}$. Grâce aux développements asymptotiques, ce principe est facilement démontré. Pour la simplicité, nous le démontrons pour $k = 1$. Tout d'abord

$$\mathbb{P}\left[X_1 \in O_1 \mid \hat{S}_n \geq x_0\right] = \int_{O_1} \mu(\,\mathrm{d}x_1) \frac{\mathbb{P}\left[\hat{S}_n \geq x_0 \mid X_1 = x_1\right]}{\mathbb{P}\left[\hat{S}_n \geq x_0\right]}.$$

Posons $L_{n-1}^{(2)}$ pour la mesure empirique des termes de 2 à n et développons nL_n en $(n-1)L_{n-1}^{(2)} + \delta_{X_1}$. En se servant du développement (6.4) et des relations (6.15), il est aisé de voir que le rapport à l'intérieur de l'intégrale converge μ–presque sûrement vers :

$$\begin{aligned} &= \frac{\mathbb{P}\left[(n-1)\langle f, L_{n-1}^{(2)} - \nu\rangle \geq x_0 - f(x_1)\right]}{\mathbb{P}\left[\hat{S}_n \geq x_0\right]} \\ &\sim e^{H(\nu\|\mu)} \sqrt{\frac{n}{n-1}}\,\alpha \int_{x_0 - f(x_1)}^{\infty} e^{-\alpha u}\,\mathrm{d}u \\ &\stackrel{n\to\infty}{\longrightarrow} e^{\alpha x_0 + \beta} e^{-\alpha(x_0 - f(x_1))} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x_1).\end{aligned}$$

Donc X_1 est distribué selon ν. On constate que, conditionnellement, les X_i ne sont pas tous indépendants, de loin s'en faut, car si c'était le cas, l'ordre de grandeur de la variance de $\hat{S}_n$ serait $1/n$, alors que nous avons vu qu'il est $1/n^2$. Il existe donc bien une corrélation, mais pour de petits échantillons, les X_i sont indépendants de loi ν. Ici on voit que $k = o(n)$ est nécessaire et suffisant pour assurer l'indépendance (voir aussi [37, 39] et [103] pour les chaînes de Markov). Il est intéressant de voir que [39, condition (2.16)] est ici le conditionnement par rapport à *une seule* fonction f (comparer avec (6.14)). Si le conditionnement est du type $\hat{S}_n \in x_0 + K$, où $K \subset \mathbb{R}^p$ est un cône, alors on retrouve par les mêmes calculs le principe de Gibbs. Enfin, le développement multidimensionnel (6.14) permet en outre de calculer la loi conditionnelle de variables $\hat{S}'_n = \langle g, L_n \rangle$.

♦♦♦

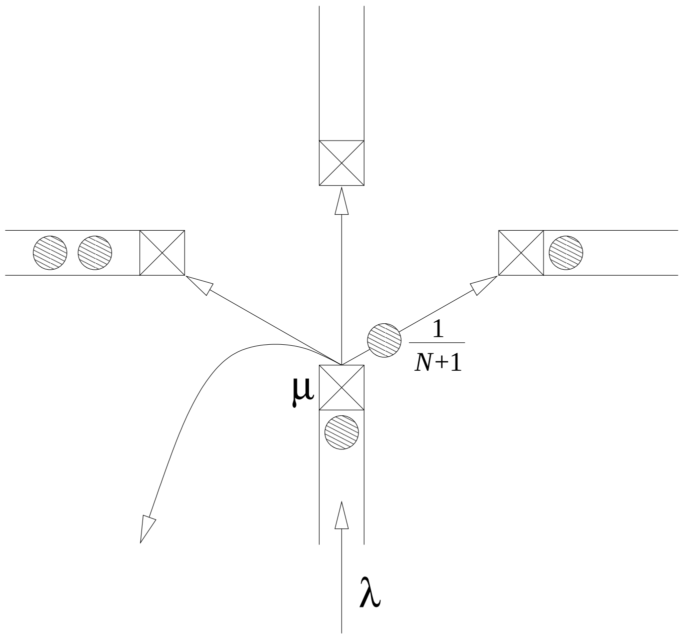

Fig. 6.1: *Réseau de Jackson à N files complètement connecté: après un service, un client choisit une queue avec probabilité* $1/(N+1)$ *ou bien quitte le système avec cette même probabilité. À chaque file, l'intensité d'arrivées [resp. de services] est* λ *[resp.* μ*].*

Corrélation dans les réseaux : Nous considérons ici un réseau de Jackson, décrit à la figure 6.1. Le but est de mesurer d'une certaine manière la corrélation entre les queues, notamment de voir la différence avec N files d'attentes $M/M/1$ indépendantes. On note la charge $\rho \stackrel{\text{def}}{=} \lambda/(N+1)\mu$ pour le réseau de Jackson et $\rho \stackrel{\text{def}}{=} \lambda/\mu$ pour les $M/M/1$. Les probabilités stationnaires sont connues dans les deux cas et sont identiques : une distribution à forme produit en $(1-\rho)\rho^n$ sur chaque file.

Puisque l'entropie du réseau fait intervenir la fraction de temps où le serveur est occupé, on peut calculer la matrice de covariance de ces périodes d'activité : on ne mesure pas directement la covariance de la longueur des queues, mais on obtient une information sur la corrélation entre les services. En notant a_{ij} la fréquence des transitions de i vers j — avec $i = 0$ [reps. $j = 0$] pour une arrivée [resp. un départ] — et π_i le taux d'utilisation du serveur i, l'entropie d'un réseau de Jackson s'écrit

$$H = \sum_{i,j=0}^{N} I_p\big(a_{ij} \| \lambda_{ij}\big) \quad \text{avec} \quad \begin{cases} \lambda_{0i} = \lambda_i, \\ \lambda_{ij} = \pi_i \mu_i P_{ij} & \text{sinon,} \end{cases}$$

en prenant conventionnellement $I_p(a_{00} \| \lambda_{00}) = 0$.

Le calcul du hessien $\nabla^2 H$ est aisé, au point où l'entropie est nulle — ici $a_{ij} = \lambda$ et $\pi_i = 1-\rho$. On cherche à minimiser la forme quadratique obtenue (de variables a'_{ij} et π'_i), sous les contraintes usuelles — la matrice A doit être balancée, sauf pour $i = 0$ — afin d'éliminer les variables a'_{ij}. En utilisant ici la symétrie du problème, on trouve que la matrice de covariance Σ est

$$\Sigma_{ii} = \frac{4\rho}{\mu} \quad \text{et} \quad \Sigma_{ij} = \left(\frac{N-3}{N-1}\right)\frac{\rho}{\mu} \quad \text{si} \quad i \neq j.$$

Notons que cette méthode est exacte en N. À titre de comparaison, la variance est moitié moindre pour les $M/M/1$ et la covariance est nulle. L'effet du réseau est particulièrement perceptible en considérant le temps de service *total* des N queues. Sa variance vaut $N(N+1)\rho/\mu$ contre $2N\rho/\mu$ pour les $M/M/1$ indépendantes. Lorsque le réseau est grand $N \to \infty$, il se comporte ici de manière quasi-coordonnée : les périodes de service sont presque synchrones.

♦♦♦

6.1.3 Espace d'états continu

Les grandes déviations sont valides sur un espace d'états *continu* aussi bien que *dénombrable*. Aussi est-il naturel de se demander si la technique que nous avons utilisée (voir section 3.2.1) ne peut pas s'y étendre. Une part importante de la recherche effectuée (voir II 1) s'appuyait sur un résultat de continuité (propositions II 2.19 et II 2.22). La décomposition en cycles des mesures balancées (proposition II 2.22) n'a pas d'équivalent *intéressant* sur un espace continu E. Il s'ensuit que la propriété d'exacte semi-continuité sur les mesures à support fini (proposition II 2.19) ne tient pas non plus. Ce ne sont pas les preuves qui manquent, ce sont les propositions elles-mêmes qui sont fausses, telles qu'elles sont énoncées. Nous ne traiterons ici que du temps discret, mais les processus de sauts à temps continu ne sont pas substantiellement plus compliqués.

Changement de mesure Il est important d'insister sur ce fait que nous cherchons à étendre la technique du changement de mesure, exactement dans les mêmes termes qu'au chapitre 4 (et plus particulièrement la section 4.2). On remplace formellement E par un espace continu, les sommations par des intégrations et les rapports par des dérivées de Radon-Nikodym. Par exemple, une mesure balancée se définit toujours (voir équation (4.2)) comme une mesure sur E^2 avec

$$A(E,.) \stackrel{\text{loi}}{=} A(.,E) \stackrel{\text{loi}}{=} a(.). \tag{6.16}$$

L'entropie s'exprime comme

$$H(A\|P) \stackrel{\text{def}}{=} \int_{E^2} A(\,\mathrm{d}x,\,\mathrm{d}y) \log \frac{\mathrm{d}A(x,.)}{\mathrm{d}P(x,.)}(y),$$

si $A(x,.) \ll P(x,.)$ pour a–presque tout x.

La martingale s'exprime comme

$$M_n \stackrel{\text{def}}{=} e^{\langle L_n, \phi_{P,A} \rangle}$$

avec L_n la mesure empirique des paires (voir l'équation (4.1)) et

$$\phi_{P,A}(x,y) \stackrel{\text{def}}{=} \log \frac{\mathrm{d}A(x,.)}{\mathrm{d}P(x,.)}(y),$$

et la fonctionnelle $B \to \langle B, \phi_{P,A} \rangle$ est telle que $\langle A, \phi_{P,A} \rangle = H(A\|P)$.

Notations Nous avons défini les mesures balancées $A \in \mathcal{M}_s(E^2)$. En temps discret s'y ajoute la condition de normalisation $A(E,E) = 1$: la mesure a est donc une probabilité (voir (6.16)). La probabilité conditionnelle $A(x,\,\mathrm{d}y)$ est définie par $A(x,O) = \frac{\mathrm{d}A(.,O)}{\mathrm{d}a}(x)$ (car $A(.,O) \ll a$ pour tout borélien O). On vérifie que $aA = a$, c'est-à-dire que a est bien une probabilité invariante du noyau de transition $A(x,\,\mathrm{d}y)$.

L'analogue des mesures balancées sur un graphe $\mathcal{M}_s(\Gamma)$ se définit par rapport à un noyau de transition $\Gamma(x,\,\mathrm{d}y)$.

$$A \in \mathcal{M}_s(\Gamma) \iff A \ll a \otimes \Gamma.$$

On peut définir alors la densité, soit par le noyau $A(x,.)$, soit par la mesure A, mais elles sont égales dans le sens où :

$$\frac{\mathrm{d}A(x,.)}{\mathrm{d}\Gamma(x,.)}(y) \stackrel{\text{def}}{=} \frac{\mathrm{d}A}{\mathrm{d}(a\otimes\Gamma)}(x,y) \quad a\otimes\Gamma\text{–p.s} \quad (\text{et donc } A\text{–p.s}).$$

Cycles La notion de cycle n'est pas tout à fait identique pour des espaces continus ou discrets. En effet, les cycles minimaux (dans un sens à préciser) peuvent fort bien être infinis dans le premier cas. La « bonne » définition est de considérer la mesure empirique $L(C)$, et non C lui-même. L'analogie avec les espaces discrets est alors la suivante : un cycle est un *flux déterministe ergodique*, c'est-à-dire que $L(C)$ est un flux (une mesure balancée) qui indique en tout point du cycle C quel est l'unique suivant (flux déterministe) ; il est ergodique car en partant de tout point de C, la chaîne construite à partir de $L(C)$ est ergodique. On peut montrer la caractérisation suivante de ces deux aspects, mesure minimale ou flux déterministe ergodique.

Définition 6.1 *Une mesure balancée $A \in \mathcal{M}_s(E^2)$ est :*

- *minimale si $B \le A \implies B = \alpha A, \quad \forall B \in \mathcal{M}_s(E^2)$;*
- *déterministe s'il existe $f:\ E \mapsto E$ telle que $A(x,.) = \delta_{f(x)}(.)$, a–p.s ;*
- *ergodique si, sous la loi A, $\mathbb{P}_x[L_n \to A] = 1$, a–p.s.*

Proposition 6.2 *Une mesure $A \in \mathcal{M}_s(E^2)$ est minimale si, et seulement si, elle est déterministe et ergodique. On l'appelle cycle si de plus elle est normalisée : $A(E,E) = 1$.*

Exemple : Pour se fixer les idées, un cycle typique sur $E = [0,1[$ est

$$A \stackrel{\text{def}}{=} \int_0^1 \delta_{(x,f(x))}\,\mathrm{d}x \in \mathcal{M}_s([0,1[^2).$$

où $f(x) = x + x_0 \mod 1$, avec x_0 irrationnel.

♦♦♦

En s'inspirant des remarques sur les génératrices des cônes (juste avant la proposition II 2.22, p. 136) on peut alors décomposer les mesures balancées sur les cycles (voir encore [20, Ch. II 7], et particulièrement la proposition 5).

Décomposition et continuité Une décomposition sur les cycles ne permet pas de conclure au PGD. Deux types de continuité sont nécessaires et leur relation éclaire sous un jour nouveau les travaux sur les espaces d'états discrets. Tout d'abord la fonctionnelle linéaire $B \to \langle B, \phi_{P,A} \rangle$ doit être continue. Il faut donc trouver un ensemble $\mathcal{B}$ dense dans $\mathcal{M}_s(E^2)$, tel que $\phi_{P,A} = \log \frac{\mathrm{d}A}{\mathrm{d}P}$ soit continue sur les mesures $A \in \mathcal{B}$ et H doit être exactement semi-continue inférieurement sur $\mathcal{B}$. Pour un espace E discret, les deux conditions sont satisfaites en contraignant le support de A à être fini ; la transcription de fini en compact est fausse : si le support de A est compact, $\phi_{P,A}$ n'est pas nécessairement continue. Il faudrait construire des approximations $A^{(\varepsilon)} \in \mathcal{M}_s(E^2)$ telles que

$$-\frac{1}{\varepsilon} \leq \phi^{(\varepsilon)}_{P,A} \equiv \log \frac{\mathrm{d}A^{(\varepsilon)}}{\mathrm{d}P} \leq \frac{1}{\varepsilon} \quad \text{et} \quad A^{(\varepsilon)} \overset{\varepsilon \to 0}{\longrightarrow} A.$$

Les problèmes techniques posés par l'étude de telles fonctions $\phi_{P,A}(x, y)$ sont multiples. De plus nous n'avons pas parlé de l'irréductibilité de la chaîne — propriété plus complexe que dans le cas des espaces discrets — et encore moins de la régularité des probabilités de transitions $P(x, .)$. Finalement, il est probable que sous des hypothèses suffisamment contraignantes, on puisse étendre le changement de mesure aux espaces continus, mais il n'est pas clair que ce soit une méthode techniquement plus élégante que la variante de limite projective utilisée au chapitre II 2 (ou [34]).

6.2 Conclusion

Nous espérons avoir convaincu le lecteur au moins des points suivants. Le processus choisit le chemin de moindre information lors d'écarts importants au comportement moyen : cette idée directrice simple et significative, se décline de multiples manières. L'entropie associée au PGD, non seulement quantifie l'information et donne l'interprétation précédente des résultats, mais encore sa forme est prédite aisément. C'est un premier point, la démarche entreprise ici se généralise bien, à partir de réflexions sur les variables i.i.d. et les chaînes de Markov, on parvient à conjecturer assez facilement et précisément les résultats analogues pour des situations soit plus particulières (p. ex. les réseaux à polling), soit plus générales (p. ex. les processus semi-markoviens).

Après cette étape d'induction, il faut revenir à la déduction. Les preuves utilisent ici principalement la technique du changement de mesure, bien que ce ne soit pas la seule. Elle repose sur des propriétés fines de l'entropie, notamment de continuité, qui conduisent à des concepts probabilistes forts (fonctions de Ruelle-Lanford), et leurs conséquences en terme de grandes déviations (PGD faibles, contraction). Comme ces propriétés ne sont pas l'apanage exclusif du modèle que nous avons étudié, en s'inspirant toujours de la notion d'information, on en déduit des résultats assez variés, sur les réseaux, markoviens ou non. L'étude générale du changement de mesure, par exemple via la frontière de Martin, ou sur les espaces continus, devrait permettre d'approfondir quelles sont exactement les propriétés requises.

Nous jugeons esthétique cette technique en ce qu'elle montre bien le rapport entre les probabilités et l'entropie. En effet, cette dernière est évocatrice de phénomènes probabilistes, mais en outre sa dérivée et son hessien sont aussi signifiants : la martingale qui définit le changement de mesure est directement liée au gradient et le théorème central limite au hessien. À ce stade, il est remarquable de constater à quel point la mise en forme rigoureuse de l'idée générale qui nous a servi de fil d'Ariane se rehausse elle-même : l'information, c'est-à-dire l'entropie, « guide » effectivement les déviations des processus considérés, mais elle le fait sur plusieurs échelles différentes !

Il faut bien admettre que les résultats obtenus dépassent nos espérances initiales. Mais n'est-ce pas cet espoir qui nous motive, lorsque l'on se lance à travers ces espaces infinis ? Cette unité découverte au détour de nos recherches est une facette de cette étonnante et magnifique adéquation de la pensée aux objets, et pour notre plus grand bonheur il reste assurément bien des surprises dans cette voie.

« UN COUP DE DÉS
JAMAIS
MÊME LANCÉ DANS DES CIRCONSTANCES ÉTERNELLES
N'ABOLIRA
LE HASARD »
STÉPHANE MALLARMÉ

Deuxième partie

Articles

Chapitre 1

Introduction aux articles

Dans cette partie, nous présentons les travaux achevés de cette thèse. Les articles qui suivent sont disponibles comme rapports de recherche [34, 32, 35], et ont été soumis à publication ; c'est pourquoi ils ont été peu retouchés ; les modifications devraient rendre l'ensemble plus cohérent et concernent essentiellement les notations.

L'ordre d'exposition est chronologique, mais reflète une évolution naturelle des concepts. Nous n'insisterons pas ici sur la spécificité de ces travaux par rapport à l'état de l'art (voir la partie I, et plus spécifiquement les sections I 4.1, I 4.2 et I 4.3), mais plutôt sur la continuité des développements.

1.1 PGD pour les chaînes de Markov en temps discret

L'obtention de l'existence d'un PGD *faible* n'est pas nouvelle, aussi n'est-ce pas le seul but de la méthode proposée. Tout d'abord il s'agissait de montrer que l'entropie est toujours l'*unique* fonctionnelle d'action, y compris dans le cas réductible. Ensuite ces résultats, qui sont valables pour la mesure empirique des paires, le restent pour la mesure empirique simple (théorème de Sanov).

Ce sont donc trois axes qui se dégagent de cet article : premièrement l'unicité de la fonctionnelle d'action dans le cas de PGD faibles ; deuxièmement les opérations qui leurs sont associées, notamment la contraction. Ces deux premiers points se fondent sur un troisième résultat plus important qu'il n'y paraît, que sont les propriétés de continuité de l'entropie.

Dans cette perspective, la décomposition d'une mesure balancée à partir de cycles (proposition 2.22) est capitale en ce sens qu'on en déduit que l'entropie est complètement calculable, par continuité, à partir des *mesures* *irréductibles à support fini* (proposition 2.19), qui est dense dans l'ensemble des mesures. Cette proposition est démontrée dans le cadre général du graphe de la matrice de transition P, qui impose des restrictions sur les possibilités de construire une suite d'approximations $A(n)$ d'une mesure A : une simple troncation ne suffit pas, car il faut que $A(n)$ soit proche de A tout en restant balancée et dans le graphe de P (voir, par exemple, la proposition 2.21).

Ce premier résultat de continuité sur l'entropie permet d'en utiliser un second, lors du changement de mesure. En effet, pour une mesure A à support fini Γ, l'application $A' \to \sum_{x,y} a'_{xy} h(x,y)$ (où $h(x,y) \stackrel{\text{def}}{=} \log(A_{xy}/P_{xy})$) est continue. On en déduit la borne inférieure par (2.6) et (2.7), la continuité prouvant que le terme (2.6) converge vers $H(A\|P)$ et l'irréductibilité que (2.7) est négligeable. Le changement de mesure n'est donc valide que pour les *mesures irréductibles à support fini*. Heureusement, par la continuité de l'entropie (proposition 2.19), la borne inférieure est étendue à toutes les mesures.

La borne supérieure locale ne fait appel qu'à la semi-continuité inférieure de l'entropie et à une propriété *locale* (voir la définition 2.4) qui permet d'utiliser une méthode semblable aux limites projectives. Le PGD faible est déduit des deux bornes, inférieure et supérieure.

La volonté de rattacher ces résultats à la théorie existante, qui est essentiellement dirigée vers le théorème de Sanov unidimensionnel, nous a incité à prouver d'une part un principe de contraction faible (théorème 2.10), qui prouve que le théorème de Sanov faible est toujours valide, d'autre part que les fonctionnelles d'actions que l'on trouve dans la littérature sont toutes égales. Ici se trouve une subtilité logique : on prouve que la fonctionnelle d'action est unique, mais on ne peut en déduire immédiatement l'égalité de toutes les expressions ; en effet, suivant le type de méthode utilisée et les propriétés de la matrice de transition P, elles servent parfois à ne démontrer qu'une seule borne (et non un PGD). Ce ne sont donc pas *a priori* des fonctionnelles d'action, mais seulement des majorants ou des minorants.

Finalement, le traitement du cas réductible exhibe ce phénomène intéressant, que l'entropie n'est pas convexe (*cf.* théorème 2.12), *uniquement du fait que le domaine sur lequel elle est finie n'est pas convexe*. On voit d'ailleurs bien les cas où ce domaine est convexe, et donc l'entropie aussi : si, et seulement si, les classes d'irréductibilité E_i forment un ensemble complètement ordonné [1]. Le second phénomène qui est mis en évidence est le lien entre l'homogénéité spatiale et l'homogénéité temporelle des trajectoires déviantes. On retrouve cette considération dès qu'il faut traiter des processus avec frontières (notamment le cas du polling : voir la section 1.3). Finalement,

1. En fait, dans ce cas, *toutes* les mesures balancées sont « sérielles » (*cf.* définition 2.13).

cette remarque est développée dans le cadre général des frontières de Martin (*cf.* chapitre I 5), où l'on constate que c'est la frontière en *espace-temps* qui importe : dans cette perspective, l'homogènéité en espace équivaut à celle en temps.

1.2 PGD pour les chaînes de Markov en temps continu

Le second article poursuit les développements du premier en ce qui concerne le changement de mesure, la continuité et les opérations sur les PGD faibles. La notion fondamentale est celle de *générateur empirique* (*cf.* définition 3.2). Elle généralise naturellement pour le temps continu la mesure empirique des paires ; elle conduit logiquement à une entropie entre générateurs et à une forme bien précise de changement de mesure.

On prouve donc la propriété de continuité de l'entropie (*cf.* proposition 3.9), ainsi que sa convexité. Cependant, au lieu de recourir à la notion de processus localement markovien, nous insistons sur celle de fonction de Ruelle-Lanford (*cf.* définition 3.15). La démonstration de la borne inférieure reste sensiblement la même, alors que pour la borne supérieure, puisque la continuité de la fonctionnelle linéaire $\phi_{R,G}$ (3.10) n'est plus assurée, on doit employer la convexité de l'entropie.

Le résultat est cependant nettement plus fort que le PGD faible, qui s'en déduit, puisque l'on obtient automatiquement l'unicité de la fonctionnelle d'action (*cf.* section 3.4.2). Par ailleurs, la contraction requiert la définition d'une *fonction auxiliaire* (voir section 3.4.3). À partir de là, on constate que cette notion est féconde, puisque l'on obtient en outre le passage de bonnes propriétés par contraction : les propriétés de continuité, de convexité et de Ruelle-Lanford. Cette analyse des PGD faibles entraîne la validité du théorème de Sanov unidimensionnel pour les processus à intensité bornée (mais on peut relâcher cette condition) ; en outre, la fonctionnelle est unique, convexe, possède de bonnes propriétés de continuité et s'exprime comme un programme convexe.

Cet article ne se contente donc pas d'élargir au temps continu une technique provenant du temps discret, il montre que la notion de PGD faible ne restreint pas les opérations qui sont applicables à un PGD ; de plus il montre comment le changement de mesure peut être mis en œuvre pour obtenir des résultats plus puissants que le PGD. Ce sont ces considérations qui mènent aux développements asymptotiques (voir la section I 6.1.2). Par ailleurs, l'idée de procéder à un changement de mesure sur un espace plus grand que l'espace d'origine, puis de procéder à une contraction pour obtenir le théorème de Sanov (ou tout autre PGD), avec une fonctionnelle solution d'un programme convexe (et donc algorithmiquement simple à calculer) conduit à l'article suivant sur le polling, et aux développements sur des processus non-markoviens (voir la section I 6.1.1).

1.3 PGD pour les systèmes à polling

L'existence des grandes déviations pour des modèles, comme les systèmes à polling, n'est pas vraiment un enjeu théorique, mais il nous a paru que c'était un bon terrain d'application pour la méthode de changement de mesure. Notons tout d'abord que, dans ce modèle, le PGD ne découle pas strictement des principes généraux (voir [56]). En outre, la méthode employée [56] débouche sur une expression de la fonctionnelle d'action locale $L(x, D)$ sous la forme de « limite de limites ». Dans le cas bien étudié des réseaux de Jackson, le calcul de la fonctionnelle locale est algorithmiquement complexe, alors qu'il est réduit à un programme convexe dans notre approche.

La spécificité du problème oblige à une écriture plus complexe, avec des modèles locaux, par exemple. Néanmoins, on met la main ainsi sur des problèmes cruciaux de la théorie des réseaux, qui sont les processus à statistiques discontinues, c'est-à-dire ceux dont l'espace d'états possède des frontières (typiquement $\mathbb{Z}_+^N$). L'étude des PGD pour les marches homogènes dans $\mathbb{Z}_+^N$ est encore difficile — il semble que des phénomènes nouveaux apparaissent — mais grâce à cette étude, on voit une condition nécessaire pour l'application du changement de mesure, qui est que les limites fluides (ou les seconds champs de vecteurs) doivent être *uniques*, et de préférence explicites pour pouvoir conduire un calcul.

Finalement, les propriétés de la fonctionnelle locale $L(x, D)$ sont suffisamment bonnes, pour que la fonctionnelle d'action du PGD trajectoriel puisse se mettre sous la forme d'une intégrale d'action (4.4), ce qui est une avancée théorique. En outre, une réflexion sur les conditions exactes d'application des différentes étape du raisonnement montre que non seulement le PGD garde la même forme avec des temps de services, d'arrivées et de déplacements généraux (voir la section I 6.1.1), mais que cela reste encore valide pour des politiques plus générales et un nombre quelconque de serveurs ! Finalement, la pierre d'achoppement, actuellement, dans la poursuite de l'évaluation de ces systèmes, est l'optimisation des trajectoires (voir la section I 4.4.3), dans la perspective d'obtenir les queues de la distribution stationnaire, lorsqu'elle existe.

Chapitre 2

Large Deviation Principle for Markov Chains in Discrete Time

Arnaud de La Fortelle & Guy Fayolle
October 1999

Abstract

Let E be a denumerable state space, X be an homogeneous Markov chain on E with kernel P. Then the chain X verifies a *weak* Sanov's theorem, i.e. a weak large deviation principle holds for the law of the pair empirical measure. In our opinion this is an improvement with respect to the existing literature, insofar as the LDP in the Markov case often requires either the finiteness of E, or strong uniformity conditions, which important classes of chains do not verify (e.g. classical queueing networks with bounded jumps). Moreover this LDP holds for *any* discrete state space Markov chain, possibly non ergodic.

The result is obtained by a new method, allowing to extend the LDP from a finite state space setting to a denumerable one, somehow like a the projective limit approach. The analysis presented here gives some by-products, among which an analogue of Varadhan's integral lemma and, under restrictive conditions, a contraction principle leading directly to a weak Sanov's theorem for the one–dimensional empirical measure.

keywords Large deviations, Markov chain, countable, pair empirical measure, Sanov, entropy, information, cycle.

2.1 Introduction and main ideas

Large deviations theory is a powerful tool to analyze the behaviour of stochastic processes in the long run, off the Central Limit Theorem regime. Estimates have been used in number of applications, some general results being specialized in various ways to fit each case.

The goal of our study is to propose a proof of a LDP for Markov chains defined on a countable space. This general framework includes most of the classical queueing networks with bounded jumps and Markovian evolution.

Assumptions, techniques and bounds When the state space is finite, results are well known and the rate function is expressed as an *entropy function*, the so-called *Kullback–Leibler information* [93]. In fact, the link between large deviations and entropy was frequently used [61, 15, 23, 104], and earlier works on information theory already emphasized the equivalence of entropy to a *mean information gain* [121, 69]. This interpretation often proved very useful [99, 98, 73, 105, 16] and is briefly recalled in Appendix 2.B.1.
Since its infancy [47], the standard theory (e.g. [41, 38, 40]), interested in Markov chains defined on Polish spaces, does impose a pretty strong uniformity condition of the following type:

There exist integers $M \geq 1$ and $N, 0 < l \leq N$ such that, for all states x, y,

$$P^{(l)}(y,.) \leq \frac{M}{N} \sum_{m=1}^{N} P^{(m)}(x,.)\,, \tag{2.1}$$

where $P^{(m)}(x,.)$ denotes the m–step transition probability, given the initial state x. Alas, this condition is very restrictive. As a rule, the M/M/1 queue, Jackson networks and most of the discrete event systems with downward-bounded jumps *do not satisfy it.* Although refinements were proposed [62, 28, 133, 100], they still do not suffice for our purpose. General bounds have been obtained by considering kernels of the form

$$K_\xi(x, A) = \int_A e^{\langle \xi, f(y) \rangle} P(x,\, \mathrm{d}y),$$

for all bounded continuous functions f and all continuous linear functionals ξ (see [27, 28, 42, 31]). In this respect, it is worth quoting the study in [31], which nicely brings out the fact that the sole irreducibility is necessary for the lower bound to hold (even for an unbounded functional f).
In [43], the authors consider the quantity $\mathbb{P}\left[L_n \in B(\mu, \varepsilon), X_n \in C\right]$, where $B(\mu, \varepsilon)$ and C denote respectively a ball and a *small* set. Then, by means of

regeneration arguments, it is shown that the sets $B(\mu, \varepsilon)$ are subject to an exponential decay, whence general lower and upper bounds follow; but the problem of getting rid of the set C still remains.
At that moment, it is important to point out the existence of strong technical subtleties, which concern the coincidence of the rate functions for the upper and lower bounds. This point is illustrated in Section 2.3.5.

In network models, the families of probability measures of interest are rarely exponentially tight, so that, in many applications, a strong LDP does not hold. Therefore, we chose to concentrate on the weak LDP. Our approach (improved since the first release [33]) allows to get the upper bound with the *same* rate function for both compact and open sets. Moreover proofs are short, constructive, and they extend to special situations, e.g. reducible chains, continuous time, etc.

Our goal is to analyze the following object (see [48] and [60, 61, 62]).

Definition 2.1 (pair empirical measure) *The pair empirical measure is defined by*

$$L_n(\omega) \stackrel{\text{def}}{=} \frac{1}{n}\left(\sum_{i=1}^{n-1} \delta_{X_i(\omega),X_{i+1}(\omega)} + \delta_{X_n(\omega),X_1(\omega)}\right) \in \mathcal{M}_s(E^2),$$

where $\mathcal{M}_s(E^2)$ is the set of balanced *measures (see Definition 2.15). The last transition*[1] *(X_n, X_1) will be referred to as a "ghost transition".*

The organization of the paper is streamlined. Starting from the finite state space LDP (Section 2.2.1), an extension to *locally Markov processes* is obtained (Section 2.2.2), which, by using continuity properties of the rate function (Appendix 2.B), leads to the weak Sanov's theorem (Sections 2.2.3–2.2.4).

Theorem 2.2 (Generalized Sanov's theorem) *Let X be an irreducible Markov chain with kernel P. Then the pair empirical measure L_n satisfies a weak LDP with rate function $H_d(.\|P)$, i.e. for all open sets $O \subset \mathcal{M}_s(E^2)$ and all compact sets $K \subset \mathcal{M}_s(E^2)$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in O\right] \geq - \inf_{A\in O} H_d(A\|P), \tag{2.2}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \leq - \inf_{A\in K} H_d(A\|P), \tag{2.3}$$

[1]This transition is added for L_n to stay in $\mathcal{M}_s(E^2)$ which is the "natural" set-up for the LDP. When Sanov's theorem is derived from the LDP for the single empirical measure, the ghost transition is chosen to be (X_0, X_1) with X_0 arbitrary. Then $L_n \in \mathcal{M}_1(E^2)$, but the functional is the same and is infinite outside $\mathcal{M}_s(E^2)$).

where the entropy function H_d *(see Appendix 2.B) is defined by*

$$H_d(A\|P) \stackrel{\text{def}}{=} \sum_{i,j\in E} a_{ij} \log\left(\frac{a_{ij}}{a_i P_{ij}}\right).$$

This LDP holds without any restriction on the initial distribution. In particular, no uniformity condition is imposed.

For the upper bound (2.3), the analysis of the entropy function H_d (i.e. the *rate function*) in Section 2.B shows important differences, that change the nature of the LDP. Indeed, in the countable state space case, H_d is lower semi-continuous and the level sets

$$\Psi(\alpha) \stackrel{\text{def}}{=} \left\{A \in \mathcal{M}_s(E^2) : H_d(A\|P) \leq \alpha\right\}$$

are *not necessarily compact*. This means at once a *full* LDP cannot be achieved, but we prove nevertheless that a *weak* LDP holds.

The main idea to get the lower bound (2.2) (see Section 2.2.4) is to construct successive finite approximations of a measure A, remarking that the lower bound is easily proved for finite support measures. Nonetheless, convergence problems might occur, since H_d is not continuous. To circumvent them, we analyze in Appendix 2.C the link between balanced measures and cycles, and the continuity of the entropy is discussed in Appendix 2.B.

2.2 Sanov's theorem

2.2.1 The finite case

Large deviation principles in a finite state space have been proved quite a while ago, since famous Sanov's theorem [127] was first published in 1957, in Russian. Convexity and eigenvalues properties are the major tools to get the result [107], but combinatorial estimates are sufficient [25]. In our opinion, good books presenting methods and major references are [41, 40]. The rate function H_d, sometimes called the *Kullback–Leibler divergence*, became a classical quantity, since its appearance in information theory in 1951 [93, 69],

Theorem 2.3 (Sanov's theorem) *Let* X *be an irreducible Markov chain on a finite state space* E. *The pair empirical measure* L_n *satisfies a LDP with the good rate function* $H_d(.\|P)$.

Proof : Only the main lines will be presented.

Let $F \subset \mathcal{M}_s(E^2)$ be closed. The upper bound relies on the following combinatorial inequalities (see [33])

$$\mathbb{P}[L_n = A] \le e^{\varepsilon_n} \prod_{i,j \in E} \frac{(na_i)!}{(na_{ij})!} P_{ij}^{na_{ij}} \le e^{-nH_d(A\|P)+\varepsilon_n}, \tag{2.4}$$

where $\varepsilon_n = o(n)$ and is independant of A. Roughly, there is a multinomial number of paths (but not exactly) with the same probability (depending on the initial point). Since there are at most $(n+1)^{|E|^2}$ empirical measures L_n of length n, we have

$$\mathbb{P}[L_n \in F] = \sum_{A \in F} \mathbb{P}[L_n = A] \le (n+1)^{|E|^2} \exp\{-n \inf_{A \in F} H_d(A\|P) + \varepsilon_n\},$$

and hence the result.

Let $O \subset \mathcal{M}_s(E^2)$ be open, $A \in O$ with $H_d(A\|P) < \infty$, so that $A \in \mathcal{M}_s(\Gamma)$, where Γ is the graph associated to P, remarking that these assumptions are sufficient, as far as the infimum of the lower bound is concerned.
One will use a classical exponential change of measure, via the martingale

$$M_n \stackrel{\text{def}}{=} \exp \sum_{i=1}^{n-1} h(X_i, X_{i+1}), \quad \text{where } h(i,j) = \log \frac{A_{ij}}{P_{ij}}. \tag{2.5}$$

So, M_n defines a new probability $\mathbb{P}^*$, under which X is a Markov chain with transition matrix A (the function h is finite on $\mathrm{Supp}(A) \subset \Gamma$). Denoting $\mathrm{Supp}(A)$ by Γ', we have, for any open neighborhood $\mathcal{V}_A \subset O$ of A,

$$\begin{aligned} \frac{1}{n} \log \mathbb{P}[L_n \in \mathcal{V}_A] &= \frac{1}{n} \log \mathbb{E}^* \left[\mathbb{1}_{\{L_n \in \mathcal{V}_A\}} M_n^{-1} \right] \\ &\ge - \sup_{A' \in \mathcal{V}_A \cap \mathcal{M}_s(\Gamma')} \sum_{(i,j) \in \Gamma'} a'_{ij} h(i,j) \qquad (2.6) \\ &\quad + \frac{1}{n} \log \mathbb{E}^* \left[\mathbb{1}_{\{L_n \in \mathcal{V}_A\}} e^{-h(X_n, X_1)} \right]. \qquad (2.7) \end{aligned}$$

The term in (2.7) tends to 0, since X_n is ergodic, (so that L_n converges to A in probability), and h is bounded on $\mathcal{M}_s(\Gamma')$. When $\mathcal{V}_A$ is sufficiently close to A, the term in (2.6) does not differ substantially from $H_d(A\|P)$, due to the continuity of the mapping

$$A' \to \sum_{(i,j) \in \Gamma'} a'_{ij} h(i,j).$$

Since $\mathbb{P}[L_n \in O] \ge \mathbb{P}[L_n \in \mathcal{V}_A]$ for all A, the lower bound is proved. ∎

Actually A might be not irreducible, in which case $\mathbb{P}^*$ does not have necessarily a unique stationary measure. Thus one must first prove the lower bound for any irreducible A, and then extend the result to any A, by a continuity argument, using Proposition 2.21 and Proposition 2.18.

M_n is only defined on the set $\{x \in E: \ a_x > 0\}$ (also denoted by Supp(A)), which is absorbing under $\mathbb{P}^*$. This would be enough to conclude, provided that the initial distribution ν satisfies $\nu\big(\mathrm{Supp}(A)\big) > 0$, but this condition is not true in general. Therefore, one chooses h to be null outside Supp(A), so that M_n is defined everywhere.
Under $\mathbb{P}^*$, X behaves like P outside Supp(A). From the irreducibility of P, there exists n_0 with $\mathbb{P}^*\left[X_{n_0} \in \mathrm{Supp}(A)\right] > 0$. It suffices now to consider the chain for $n \geq n_0$, so that the bounds still hold, as the finite part $(X_n, n \leq n_0)$ does not play any role.
In fact, the way of defining X under $\mathbb{P}^*$ outside the support of A is not crucial, so that the next section is in some sense quite natural.

2.2.2 Extension to locally Markov processes

Here one proposes to extend Theorem 2.3 to a countable state space. This leads first to make a slight generalization of Theorem 2.3 to the case of *substochastic* transition matrices.

Definition 2.4 (Locally Markov process) *A process X_n is said to be locally Markovian (and homogeneous) on $E' \subset E$, with transition matrix P, if*

$$\mathbb{P}\left[X_{n+1} = y | \mathcal{F}_n \cap \{X_n = x\}\right] = P_{xy}, \quad \forall x, y \in E', \ \forall n \geq 0, \tag{2.8}$$

where $\mathcal{F}_n$ is the natural filtration of X_n. For the sake of consistency, one requires the probability of ever reaching E' to be strictly positive.

For exemple, we shall consider in the following a Markov chain X_n on a denumerable state space E. Let $E' \subset E$ be a finite set and define a new process Y_n on $E' \cup \{\Delta\}$ by $Y_n = X_n$ if $X_n \in E'$ and $Y_n = \Delta$ otherwise. This is an *aggregated* process and, naturally, it is locally Markovian on E'.
Note that there is no restriction on the initial probability distribution, so that $\mathbb{P}\left[X_n \in E'\right]$ may behave arbitrarily.[2]

[2]For instance, if X is a random walk on $\mathbb{Z}$ and E' is finite, a heavy tailed initial probability distribution disturbs a LDP for $\mathbb{P}\left[X_n \in E'\right]$. Fortunately this is not the case for the empirical measure.

Using the combinatorics of (2.4) for the upper bound, taking the value 1 for the unknown probabilities, and the above change of measure for the lower bound, one easily obtains the following extension of Theorem 2.3.

Theorem 2.5 *Let X be a locally Markovian process on a finite set E'. The pair empirical measure L_n satisfies the LDP lower bound on $\mathcal{M}_s(E'^2)$ and, restricted to E', satisfies the LDP upper bound, both with the good rate function $H_{E'}(.\|P)$. More formally, for all open $O \subset \mathcal{M}_s(E'^2)$ and all closed $F \subset t_{E'}(\mathcal{M}_s(E^2))$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in O\right] \geq - \inf_{A \in O} H_{E'}(A\|P), \tag{2.9}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[t_{E'}(L_n) \in F\right] \leq - \inf_{A \in F} H_{E'}(A\|P), \tag{2.10}$$

where $t_{E'}(A) \stackrel{\text{def}}{=} (a_{ij})_{i,j\in E'}$ is the truncation on E' and $H_{E'}$ the associated entropy function (see Definition 2.16).

Although Theorem 2.5 is not a LDP, it looks very much like a kind of *projective limit*. Nonetheless, the projection of a Markov chain onto some subset is no longer a Markov chain, and consequently the projective limit method should undergo ad hoc modifications. Our approach is qualitatively different, since a *full* LDP for finite state spaces becomes in general a *weak* LDP for the limiting countable state space.

The set $t_{E'}(\mathcal{M}_s(E^2))$ is *not equal* to $\mathcal{M}_s(E'^2)$, nor to the sets of all measures on E'^2 with mass less than 1. The above formulation is quite natural, and avoids tedious problems related to the existence of paths entering or leaving each state a fixed number of times.

2.2.3 Upper bound

The heart of the extension consists of three properties. First the compacity of the set of measures, second the fact that a Markov chain is locally Markovian on any finite set, and third the lower semi-continuity of the rate function H_d.

Theorem 2.6 *Let X be a Markov chain on a countable state space E, with transition matrix P. The pair empirical measure L_n satisfies the LDP upper bound on compact sets with the rate function $H_d(.\|P)$, that is, for all compact sets $K \subset \mathcal{M}_s(E^2)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \leq - \inf_{A \in K} H_d(A\|P). \tag{2.11}$$

Proof : Let $x < \inf_K H_d(.\|P)$ be fixed. Then there exists a finite subset $E' \subset E$ such that

$$\inf_{A \in K} H_{E'}(A\|P) \geq x. \tag{2.12}$$

Suppose this is not true. Then, for all finite E_n, there exists $A(n) \in K$ such that $H_{E_n}(A(n)\|P) < x$. Choosing $E_n = \{1, \ldots, n\}$, this writes

$$\liminf_{n \to \infty} H_{E_n}(A(n)\|P) \leq x.$$

Since K is compact, there exists a subsequence $A(n)$ converging to A in K. Of course $t_{E_n}(A(n))$ also converges to A, because E_n increases to E. Applying Definition 2.16, we have

$$H_{E_n}(A(n)\|P) = H_d(t_{E_n}(A(n))\|P).$$

By Lemma 2.17, H_d is lower semi-continuous, so that

$$H_d(A\|P) \leq \liminf_{n \to \infty} H_d(t_{E_n}(A(n))\|P) \leq x.$$

But this contradicts $x < \inf_K H_d(.\|P)$. Hence, (2.12) is true for some finite subset $E' \subset E$. As X is obviously locally Markovian on E', Theorem 2.5 yields

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[t_{E'}(L_n) \in t_{E'}(K)\right] \leq - \inf_{A \in t_{E'}(K)} H_{E'}(A\|P).$$

Now

$$\inf_{A \in t_{E'}(K)} H_{E'}(A\|P) = \inf_{A \in K} H_{E'}(t_{E'}(A)\|P) = \inf_{A \in K} H_{E'}(A\|P) \geq x.$$

Since $L_n \in K$ implies that $t_{E'}(L_n) \in t_{E'}(K)$, we obtain

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \leq -x,$$

which this is true for any $x < \inf_K H_d(.\|P)$. Theorem 2.6 is proved. ∎

2.2.4 Lower bound

The extension for the lower bound is simple, by using continuity properties of H_d.

Theorem 2.7 *Let X be a Markov chain on a countable state space E, with transition matrix P. The pair empirical measure L_n satisfies the LDP lower bound with the rate function $H_d(.\|P)$, that is, for all open sets $O \subset \mathcal{M}_s(E^2)$,*

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in O\right] \geq - \inf_{A \in O} H_d(A\|P). \tag{2.13}$$

Proof : Let E' be a finite subset of E and $A \in \mathcal{M}_s(E'^2) \cap O$. Since O is open, there exists an open neighborhood $\mathcal{V}_A \subset \mathcal{M}_s(E'^2)$ of A included in O. Applying Theorem 2.5 for such a measure, one has

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in \mathcal{V}_A\right] \geq -H_{E'}(A\|P).$$

But $H_{E'}(A\|P) = H_d(A\|P)$ and $\mathcal{V}_A \subset O$, so that the lower bound is satisfied for any measure $A \in O$ having a finite support. This reads, for such A,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in O\right] \geq -H_d(A\|P). \tag{2.14}$$

Let us fix now $A \in O$. By Proposition 2.19, there exist measures $A(m)$ with finite support, converging to A, and ensuring also the convergence of $H_d(A(m)\|P)$ to $H_d(A\|P)$. Applying (2.14), with A replaced by $A(m)$, results in

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in O\right] \geq -H_d(A(m)\|P) \overset{m\to\infty}{\longrightarrow} -H_d(A\|P).$$

Thus (2.14) is valid for any measure $A \in O$ and the proof of Theorem 2.7 is concluded. ∎

Remark 1 : The change of measure works in fact even for an infinite state space. In this case, problems encountered have a twofold aspect. First, the method works only for irreducible A: no matter which approach is taken, the study of continuity can never be avoided, if the goal is to extend the lower bound to all $A \in O$. Secondly, when $\mathrm{Supp}(A)$ is countable, the supremum in (2.6) may be infinite (because h itself can be unbounded), and therefore it is necessary to place some restrictions on A.

♦♦♦

2.3 Links and transformations of weak LDP

2.3.1 Another derivation via Ruelle–Lanford functions

In a beautiful paper [97], J. Lewis and C. Pfister analyze, in a very broad setting, large deviation phenomena from a thermodynamical point of view. It is interesting to note that their *vague* LDP corresponds exactly to our[3] *weak* LDP.

[3]The terminology [97] follows O'Brien and Vervaat [116], while we use Deuschel and Strook [41] vocabulary.

In this respect, a pragmatic question arises: could the weak LDP be obtained in a more direct way by means of Ruelle–Lanford functions? More precisely, assume one shows $H_d(A\|P)$ is a Ruelle–Lanford function: then [97, Theorem 3.1] yields immediately the weak LDP. This is true, but (in agreement with the law saying that miracles are rare events!), the proof is in essence not simpler as the one given in Sections 2.2.3 and 2.2.4. Indeed Theorem 2.2 can be derived without the notion of local Markov processes, but at the expense of a deeper analysis of the change of measure. For the sake of completeness, we prove the following result.

Theorem 2.8 *The entropy is a Ruelle-Lanford function for the empirical measure L_n.*

Proof : The notation is as in Theorem 2.3, but here E is infinite. When $A \in \mathcal{M}_s(\Gamma)$ is an irreducible measure with a finite support, the line of argument which leads to the lower bound still holds, but (2.6) also yields the upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in \mathcal{V}_A\right] \le - \inf_{A' \in \mathcal{V}_A \cap \mathcal{M}_s(\Gamma')} \sum_{(i,j)\in\Gamma'} a'_{ij} h(i,j). \tag{2.15}$$

By continuity on $\mathcal{M}_s(\Gamma')$ of the mapping $A \to \sum_{i,j} a_{ij} h(i,j)$, for all $\varepsilon > 0$, there exists a neighborhood $\mathcal{V}_A$, such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in \mathcal{V}_A\right] \le -H_d(A\|P) + \varepsilon. \tag{2.16}$$

Let now A be arbitrary, $\mathcal{V}_A$ a neighborhood of A, and $K \subset \mathcal{V}_A$. In Propositions 2.21 and 2.22, one shows the set of irreducible measures with finite support to be dense in $\mathcal{M}_s(\Gamma)$. Consider all such $A' \in \mathcal{V}_A$ and their corresponding neighborhoods $\mathcal{V}_{A'}$, subject to (2.16). Then $\mathcal{V}_A$ and K can be covered by a collection of sets $\mathcal{V}_{A'}$, from which, since K is compact, we can extract a finite covering such that $K \subset \bigcup_{A'\in C(\varepsilon)} \mathcal{V}_{A'}$. Hence, by (2.16),

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \le - \inf_{A'\in C(\varepsilon)} H_d(A'\|P) + \varepsilon. \tag{2.17}$$

Therefore, by the lower semi-continuity of H_d and (2.17),

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \le -H_d(A\|P) + 2\varepsilon, \tag{2.18}$$

and the local upper bound is proved.

Using the lower semi-continuity of H_d, Proposition 2.19 and the density of irreducible measures with finite support, on can prove in a similar way a local lower bound for any measure A, exactly the reverse inequality of (2.16). Finally, $H_d(A\|P)$ is a Ruelle–Lanford function for $\mathbb{P}[L_n \in .]$ in the Hausdorff topological space $\mathcal{M}_s(E^2)$, so that, by [97, Theorem 3.1], the weak Sanov's theorem is proved. ■

Comparison As asserted at the beginning of this subsection, the two proofs have roughly the same complexity.

- In Section 2.2, one needs a detailed analysis of finite substochastic transition matrices, which form the locally Markov processes, and the result is a corollary of the finite state space LDP.
- The method proposed here is an alternative to that of Section 2.2. Since $\mathcal{M}_s(E^2)$ might be not locally compact, obtaining the upper bound is not really simpler than the LDP upper bound. Also, though further properties are explicitly used (e.g. continuity, density irreducible measures with finite support), no more work is needed because they are indispensable for the lower bound.

The choice of the method should be made according to the respective complexity of the bounds (2.16) and (2.10).

2.3.2 Varadhan's lemma and contraction principle

The next corollary[4], which is the analog of Varadhan's integral lemma, is given in [97].

Theorem 2.9 *Let $g : \mathcal{M}_s(E^2) \to \mathbb{R}$ be a continuous function. Then, for all open sets O and compact sets K,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\mathbb{I}_{\{L_n \in O\}} e^{ng(L_n)}\right] \geq - \inf_{A\in O}\{H_d(A\|P) - g(A)\}, \qquad (2.19)$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\mathbb{I}_{\{L_n \in K\}} e^{ng(L_n)}\right] \leq - \inf_{A\in K}\{H_d(A\|P) - g(A)\}. \qquad (2.20)$$

Hereafter, we prove a somewhat restricted contraction principle, which however contains interesting by-products (see Theorem 2.11).

[4]This could also be deduced easily from the change of measure. In fact the proof already relies on the fact that M_n^{-1} is a continuous functional of L_n.

Theorem 2.10 (contraction) *Let $f : \mathcal{M}_s(E^2) \to F$ be a continuous functional such that $f^{-1}(K)$ is compact, for any compact set $K \subset F$. Then $\mathbb{P}\left[f(L_n) \in .\right]$ satisfies a weak LDP with the rate function J.*

$$J(y) \stackrel{\text{def}}{=} \inf_{A \in f^{-1}(y)} H_d(A\|P) \tag{2.21}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[f(L_n) \in O\right] \geq -\inf_{\mu \in O} J(y), \tag{2.22}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[f(L_n) \in K\right] \leq -\inf_{\mu \in K} J(y). \tag{2.23}$$

Proof : The positivity of J, like equations (2.22)–(2.23), are immediate. The only difficulty concerns the lower semi–continuity of J.
Letting $\mu(n) \stackrel{n\to\infty}{\longrightarrow} \mu$, the set

$$K(n_0) = \{\mu(n)\}_{n \geq n_0} \cup \{\mu\}$$

is compact, and so is $f^{-1}\big(K(n_0)\big)$. By a finite covering argument and the lower semi–continuity of H_d, is it not difficult to prove the inequalities

$$\begin{aligned} \liminf_{n\to\infty} J(\mu(n)) &= \liminf_{n\to\infty} \inf_{A \in f^{-1}(\mu(n))} H_d(A\|P) \\ &\geq \inf_{A \in f^{-1}(\mu)} H_d(A\|P) = J(\mu). \end{aligned}$$

The contraction theorem is proved. ■

Instantiating for example $f(A) = A(E,.)$, one can checks in this case that all assumptions of Theorem 2.10 are satisfied: hence, the contraction is valid and there is a LDP for the one–dimensional empirical measure. The function $\phi : \mathcal{M}_s(E^2) \mapsto \mathcal{M}_1(E)$ will henceforth denote the projection $A \to A(E,.)$.

Theorem 2.11 *Let X be an irreducible Markov chain with kernel P. Then the empirical measure satisfies a weak LDP with the rate function*

$$I_0(\mu\|P) \stackrel{\text{def}}{=} \inf_{\phi(A)=\mu} H_d(A\|P). \tag{2.24}$$

The identification of I_0 in (2.24) does not require any particular condition (see Section 2.3.5), so that the weak Sanov's theorem in the one–dimensional case is really an easy consequence of Theorem 2.2.

2.3.3 Links with a full LDP

There is no way of improving our results toward a full LDP, as shown in the remark in Appendix 2.B.1. We emit the conjecture that the problem of existence of a full LDP reduces to the analysis of the entropy function, and this would give a real interest to the weak LDP. Unfortunately, even in our framework, the *goodness* of the rate function does not imply a full LDP with this good rate function. It is possible that $\inf_A H_d(A\|P) > 0$, in which case a full LDP cannot hold, H_d being nevertheless a good rate function.
Then, some directions may be explored.

1 When is the rate function H_d a *good* rate function?

2 What are the precise relationships between goodness, exponential tightness, and fullness of the LDP?

3 To which extent is a full LDP necessary?

Partial answers to the first question do exist, e.g. uniformity conditions which can be found in [38, 62, 100, 133]. A general *characterization* of goodness seems difficult to obtain, but perhaps some more specialized results might include bounded jump networks. In fact, for almost all open networks, there is no full LDP, due to the strong homogeneity.

The second question is open. It is known that the exponential tightness implies a full LDP holds with good rate function H_d (see [38, p. 8]). One can conjecture the following: if H_d is good *and* $\inf_A H_d(A\|P) = 0$, then the family $\{P^{(n)}, n \geq 0\}$ is exponentially tight. Section 2.3.2 gives a partial answer to the third point.

2.3.4 Further extensions

The LDP proved in the preceding sections extends directly to k-tuples empirical measures, to process level, or by considering continuous time or larger state spaces (e.g. Polish spaces).
On the other hand, the entropy function contains some surprises. Actually, when the chain is transient, the global infimum of H_d can be strictly positive. This infimum is related to the convergence parameter of P

$$\inf H_d(.\|P) = -\log \mathcal{R}(P),$$

where

$$\mathcal{R}(P) = \inf_\lambda \{\lambda : Ph = \lambda h, \text{ with } h \geq 0\},$$

and this leads naturally to harmonic analysis.

On the other hand, extension of the LDP upper bound to closed sets faces a serious problem. It appears to be closely related to boundary theory, and more precisely to space-time Martin boundary, that is again to harmonic functions. In order to keep the LDP formulation in terms of minima, the entropy function should be extended to the (minimal) Martin boundary, by means of an expression of the form

$$H_d(h) \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{t_n} \log h(x_n, t_n), \tag{2.25}$$

where h is a minimal harmonic function and $(x_n, t_n) \stackrel{n\to\infty}{\longrightarrow} h$ for the Martin compactification. The difficulty is that the limit does not exist, although (2.25) makes sense in some concrete situations.

2.3.5 Identification of the rate function

There are several ways of representing the rate function in Sanov's theorem for the one-dimensional empirical measures

$$L_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Actually, the formulations depend on the spaces of $E \mapsto \mathbb{R}$ functions usually considered: $C_b(E)$ (bounded continuous functions), $L_b(E)$ (bounded Lipschitz functions), and also $U(E)$, the space of continuous bounded functions h with $\inf h > 0$. Note that in discrete spaces there is no continuity matter: e.g. all functions are continuous. The rate functions are given in terms of transforms of the quantities

$$\begin{aligned}
K_f(u)(x) &\stackrel{\text{def}}{=} \int P(x,\, \mathrm{d}y) e^{f(y)} u(y), \\
T_f(u)(x) &\stackrel{\text{def}}{=} e^{f(x)} \int P(x,\, \mathrm{d}y) u(y), \\
\Lambda(f) &\stackrel{\text{def}}{=} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E} \left[\exp \sum_{i=1}^{n} f(X_i) \right],
\end{aligned}$$

which are related by the relation

$$\mathbb{E}_x \left[\exp\Big(\sum_{i=1}^{n} f(X_i)\Big) u(X_n) \right] = e^{f(x)} K_f^{(n-1)}(u)(x) = T_f^{(n-1)}(ue^f)(x). \tag{2.26}$$

An almost exhaustive list of the main rate functions is given below.

$$\begin{cases} I_0(\mu) = \inf_{\phi(A)=\mu} H_d(A\|P), \\ I_1(\mu) = \sup_{h\in U(E)} \sum_{i\in E} \mu_i \log\left(\frac{h(i)}{Ph(i)}\right), \\ I_2(\mu) = \sup_{f\in C_b(E)} \left[\langle f,\mu\rangle - \Lambda(f)\right], \\ I_3(\mu) = \sup_{f\in L_b(E)} \left[\langle f,\mu\rangle + \log \mathcal{R}(K_f)\right], \\ I_4(\mu) = \sup_{f\in L_b(E)} \left[\langle f,\mu\rangle - \log \rho(T_f)\right]. \end{cases}$$

where ρ denotes the spectral radius and $\mathcal{R}$ the convergence parameter.

- I_0 is the rate function obtained in Theorem 2.11, and is known to be the same both for the upper and lower bounds.
- I_1 is convenient for computational purposes. It often gives the final form of the result, although it does not appear at first sight in the proofs.
- I_2 is a classical mean of identifying the good rate function whenever the uniformity condition (2.1) holds (see [41, 38]). Unfortunately, when working on $\inf_{x\in E} \mathbb{P}_x[L_n \in .]$, the subadditivity method cannot be transposed into a general setting.
- As mentioned earlier, the regeneration argument used in [43, 31]) for the quantity $\mathbb{P}[L_n \in B(\mu,\varepsilon), X_n \in C]$ yields I_3. Getting rid of C is easily done for the lower bound, but is a stumbling block for the upper bound. I_3 also appears in [42] to determine the rate function when there is a full LDP.
- In [27], the inequality $\Lambda(f) \le \rho(K_f)$ allows to derive I_4 for the upper bound, but under very stringent conditions regarding the uniformity (see (2.1)).

In the very interesting paper [43], all the above expressions, but for I_0, are shown to be equal, assuming only (roughly speaking) P is irreducible. However, since the goal of the authors is not really to prove a weak LDP, they obtain an upper bound of the form

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[L_n \in K, X_n \in C] \le -\inf_{\mu\in K} I_3(\mu),$$

for all closed convex sets K, C being typically a non-empty compact set. Moreover, the above upper bound is only proved for exponentially tight sequences $\mathbb{P}\left[L_n \in .\right]$. Theorem 2.2 shows that I_0 is in fact valid for both bounds.

In our opinion, an interesting part of our study is to prove the rate function to be the best possible. As a matter of fact, the rate function is unique when a LDP holds[5]. If one aims at getting only one bound, the question is open to decide whether the rate function is optimal or not.

Remark : Clearly, I_0 is equal to the other quantities. For finite support measures μ, an optimization problem, with the $h(i)$ as Lagrange's multipliers, yields $I_0(\mu) = I_1(\mu)$. Now, by the continuity property in Proposition 2.19, it is not difficult to check that I_0 is the least lower semi-continuous function such that $I_1(\mu) = I_0(\mu)$, for all such μ.

Hence $I_0 \leq I_1$, which can also be deduced from regeneration arguments. The reverse inequality is less immediate. Let us denote by L_n^1 [resp. L_n^2] the empirical [resp. pair empirical] measure.

Starting from $\mathbb{P}\left[L_n^2 \in B(A,\varepsilon), X_n = x\right]$, for any fixed x with $a_x > 0$, one carries out the lower bound proof by extending the result from finite support irreducible to any A. The rate function is again $H_d(A\|P)$, since, from the ergodicity $\mathbb{P}^*\left[L_n^2 \in B(A,\varepsilon), X_n = x\right] \stackrel{n\to\infty}{\longrightarrow} a_x$.

Fixing $\mu \in \mathcal{M}_1(E)$, we have, for any $\phi(A) = \mu$,

$$L_n^2 \in B(A,\varepsilon) \implies L_n^1 \in B(\mu,\varepsilon).$$

Hence,

$$\mathbb{P}\left[L_n^2 \in B(A,\varepsilon), X_n = x\right] \leq \mathbb{P}\left[L_n^1 \in B(\mu,\varepsilon), X_n = x\right],$$

which gives

$$H_d(A\|P) \geq I_1(\mu), \quad \forall A : \phi(A) = \mu.$$

Thus $\inf_{\phi(A)=\mu} H_d(A\|P) = I_0(\mu) \geq I_1(\mu)$, which is the expected result.

Note that the machinery has to be re-run, in order to show that the rate function I_1 is not too large, which was a priori not obvious for measures with infinite support.

♦♦♦

[5]The *usual* uniqueness argument does not apply here, since the LDP is weak and $\mathcal{M}_s(E^2)$ is not locally compact. The uniqueness is derived by the inequalities (2.6) and (2.15)

Appendix

2.A The reducible case

For the sake of completeness, we give in this section a LDP for reducible kernels. No concept is really new, nor is the result surprising, but there is a dense notational thicket hiding the simple idea stated now: instead of evolving ad æternam with the transitions $A = (a_{ij})$, under the twisted probability $\mathbb{P}^*$ the chain behaves for a long time like $A(1)$, then like $A(2)$, and so on, where the $A(k)$ correspond to different classes of irreducibility.
In fact, in this case, the bounds are obtained by a change of measure which is *not time-homogeneous*, and the subset $\mathcal{M}$ (see Definition 2.13) is not convex nor is the rate function, thus rendering the convex analysis ineffectual.

Theorem 2.12 (reducible case) *Let X be a Markov chain with kernel P and let $\mathcal{M}$ be the set of serial measures relatively to P (see Definition 2.13). The pair empirical measure of X satisfies a weak large deviation principle, with the rate function*

$$H_{|\mathcal{M}}(A\|P) \stackrel{\text{def}}{=} \begin{cases} H_d(A\|P), & \textit{for } \ A \in \mathcal{M}; \\ \infty, & \textit{otherwise.} \end{cases} \tag{2.27}$$

2.A.1 Irreducible classes and associated objects

Throughout this section, P will be a reducible transition matrix. When $P^{(n)}(x,y) > 0$, there exits a path of length n, of positive probability, connecting the points x and y, and we shall write $x \rightsquigarrow y$. Two points belong to the same class of irreducibility if, and only if, $x \rightsquigarrow y$ and $y \rightsquigarrow x$.
The state space is partitioned into classes of irreducibility $E = \bigcup_{i\in\mathcal{I}} E_i$, the graph of P on E_i being denoted by Γ_i. Then there is the following classical partial ordering $\succ$ between the classes: $E_j \succ E_i$ if there exist $x \in E_i$ and $y \in E_j$, such that $x \rightsquigarrow y$. Thus $x \rightsquigarrow y$ for all $x \in E_i$ and $y \in E_j$. Eventually, the points reachable from $x \in E_i$ are all the sets E_j with $E_j \succ E_i$. The forthcoming definition is a formal description of the "suitable" measures.

Definition 2.13 (serial measures) *A* serial *measure A relatively to P is a balanced measure*

$$A = \sum u_i A(i), \ \ \textit{with } A(i) \in \mathcal{M}_s(\Gamma_i) \ \ \textit{and } E_{i+1} \succ E_i, \ \forall i.$$

In addition it is required that $\nu\big(\{x \in E : \ E_1 \succ x\}\big) > 0$, where ν is the initial distribution of the chain. The set of serial measures will be denoted by $\mathcal{M}$.

Lemma 2.14 *Serial measures are characterized by the equivalence*

$$A \in \mathcal{M} \iff \begin{cases} \exists A(m) \overset{m\to\infty}{\longrightarrow} A, \\ \mathbb{P}\left[L_{n(m)} = A(m)\right] > 0. \end{cases} \tag{2.28}$$

Proof : The sufficient condition $\Longleftarrow$
Let i, j be such that $a_{ij} > 0$. This implies $a_{ij}(m) > 0$ for large m , so that $i \rightsquigarrow j$ and also $j \rightsquigarrow i$. Therefore A can be split into irreducible classes, and one can write

$$A = \sum u_i A(i), \quad \text{with } A(i) \in \mathcal{M}_s(\Gamma_i).$$

Take now $j_0 \in E_j$ and $k_0 \in E_k$, with $a_{j_0} a_{k_0} > 0$, so that $u_j u_k > 0$. Consequently, for m sufficiently large, we have $a_{j_0}(m) a_{k_0}(m) > 0$, and a path crosses j_0 and k_0 with a positive probability. So, either $E_j \succ E_k$ or $E_k \succ E_j$, which means that $\succ$ is a total order relation on the sets E_i for which $u_i > 0$. Consequently, A is serial.

The Necessary condition $\Longrightarrow$
Assume first A has a finite support and construct the sequence $A(m)$. The decomposition of A as a serial measure is finite and so is its decomposition into cycles (see[6] Proposition 2.22). Hence $A = \sum_{i=1}^{p} u_i L(C_i)$, and cycles can be ordered so that $C_{i+1} \succ C_i$. There exist paths $T_{i,i+1} = (x_{n_i}, \ldots, x_{n_{i+1}})$, with positive probability, such that $x_{n_i} \in C_i$, for $i \geq 1$, and $\nu(x_{n_0}) > 0$. Then, taking integer approximations $\alpha_i(m)$ of mu_i, modified by the length of C_i, the path

$$T(m) = T_{01} C_1^{\alpha_1(m)} T_{12} C_2^{\alpha_2(m)} T_{23} \ldots C_p^{\alpha_p(m)}$$

has also a positive probability. Finally $A(m) \overset{\text{def}}{=} L(T(m))$ converges to A when m grows and, by construction, $L_{n(m)} = A(m)$.
In the case A has an infinite support, there exists a sequence of serial $A(n)$ with finite support converging to A. For each $A(n)$, the above paragraph shows there exists a sequence $A(n,m)$ converging to $A(n)$, with positive probability.
Finally, letting $A'(n) \overset{\text{def}}{=} A(n, m_n)$, with m_n such that

$$\|A'(n) - A(n)\| \leq 1/n,$$

one sees that the sequence $A'(n)$ also converges to A, and the proof is concluded. ∎

[6]Note that Proposition 2.22 does not postulate irreducibility.

2.A.2 Lower bound

Theorem *For any open set $O \subset \mathcal{M}_s(E^2)$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in O\right] \geq - \inf_{A \in O \cap \mathcal{M}} H_d(A\|P). \tag{2.29}$$

Proof : The argument is like in the irreducible case, but classes have to be separated. Let $A \in O \cap \mathcal{M}$ be a measure with finite support, so that the decomposition $A = \sum_{i=1}^{p} u_i A(i)$ is finite. Define

$$h_k(i,j) \stackrel{\text{def}}{=} \begin{cases} \log \frac{A_{ij}(k)}{P_{ij}}, & \text{if } i,j \in \operatorname{Supp}(A(k)), \\ 0 & \text{otherwise.} \end{cases} \tag{2.30}$$

Fix a large integer N, take $n \leq N$, and let

$$t_0 = 0 \quad \text{and} \quad t_i = \lfloor N(u_1 + \cdots + u_i) \rfloor \wedge n.$$

The martingale M_n of (2.5) is replaced by a sequence of martingales

$$M_n^{(N)} \stackrel{\text{def}}{=} \exp\left\{ \sum_{k=1}^{p} \sum_{i=t_{k-1}+1}^{t_k} h_k(X_i, X_{i+1}) \right\}. \tag{2.31}$$

By the definition of t_k, the summation is taken over the n first terms at most. It is easily checked that $M_n^{(N)}$ is a martingale, for $n \leq N$, defining a new measure $\mathbb{P}^*$. Under $\mathbb{P}^*$, X is a Markov chain, which is no longer time homogeneous: it behaves like $A(i)$ on $\operatorname{Supp}(A(i))$ and like P outside, for a duration $[t_{i-1}+1, t_i]$, for all $i \geq 1$.
For each k, we introduce the partial empirical measures

$$L_n^k \stackrel{\text{def}}{=} \frac{1}{t_k} \left(\sum_{i=t_{k-1}+1}^{t_k - 1} \delta_{X_i X_{i+1}} + \delta_{X_{t_k} X_{t_{k-1}+1}} \right) \in \mathcal{M}_s(\Gamma_k). \tag{2.32}$$

Choose an open neighborhood $\mathcal{V}_A \subset O$ of A. For n sufficiently large, the ghost transitions are small enough, since their total mass being less than k/n. Hence, there exist open neighborhoods $\mathcal{V}_A(k)$ of $A(k)$, with

$$L_n^k \in \mathcal{V}_A(k), \ \forall k \implies L_n \in \mathcal{V}_A. \tag{2.33}$$

Combining (2.6) and (2.33), yields, when $N = n$,

$$
\begin{aligned}
\frac{1}{n}\log \mathbb{P}\left[L_n \in \mathcal{V}_A\right] &\geq \frac{1}{n}\log \mathbb{E}^*\left[\prod_{k=1}^{p} \mathbb{1}_{\{L_n^k \in \mathcal{V}_A(k)\}} \left(M_n^{(n)}\right)^{-1}\right] \\
&\geq \sum_{k=1}^{p} -\frac{t_k}{n} \sup_{A' \in \mathcal{V}_A(k) \cap \mathcal{M}_s(\Gamma_k')} \sum_{(i,j)\in\Gamma_k'} a'_{ij} h_k(i,j) \qquad (2.34)\\
&+ \frac{1}{n}\log \mathbb{E}^*\left[\prod_{k=1}^{p} \mathbb{1}_{\{L_n^k \in \mathcal{V}_A(k)\}} \exp\bigl(h_k(X_{t_k}, X_{t_{k-1}+1})\bigr)\right], \qquad (2.35)
\end{aligned}
$$

where Γ_k' reads both for the graph and the support of $A(k)$. By continuity, the term in (2.34) gives $u_k H_d(A(k)\|P)$, and hence the sum tends to $-H_d(A\|P)$, for $n \to \infty$ and $\mathcal{V}_A$ small enough.
The expectation appearing in (2.35) is not that complicated. The k-th functional indeed depends only on X_i, for $i \in [t_{k-1}+1, t_k]$, and X behaves like $A(k)$ on this time interval. The chain has been split into p pieces, to each of which the result of the irreducible case can be applied. But the chain must hit once Γ_k', and only after this moment it behaves like $A(k)$. Remembering that $\mathbb{P}^*$ depends on n, it is not obvious whether the probability of ever reaching Γ_k' under $\mathbb{P}^*$ remains uniformly bounded on the sets of interest.

One will show by induction that there exist n_k and $b_k > 0$, such that

$$\mathbb{P}^*\left[X_{t_{l-1}+n_l} \in \Gamma_l',\ \forall l \leq k\right] \geq b_k > 0, \qquad (2.36)$$

independently of n. Since $\nu(\{x \in E:\ E_1 \succ x\}) > 0$, the assertion is true for $k = 1$. Assume it is true for k. Then

$$\mathbb{P}^*\left[X_{t_{l-1}+n_l} \in \Gamma_l',\ \forall l \leq k \text{ and } X_{t_k} \in \Gamma_k'\right] \geq b_k, \qquad (2.37)$$

as, under $\mathbb{P}^*$, Γ_k' is absorbing during the time interval $[t_k+1, t_{k+1}]$. Choose n_{k+1} such that

$$\inf_{x\in\Gamma_k'} \mathbb{P}^*\left[X_{t_k+n_{k+1}} \in \Gamma_{k+1}' | X_{t_k} = x\right] = c_k > 0. \qquad (2.38)$$

This can be done since $E_{k+1} \succ E_k$ and Γ_k' is finite. Thus, applying the Markov property to equations (2.37)–(2.38),

$$\mathbb{P}^*\left[X_{t_{l-1}+n_l} \in \Gamma_l',\ \forall l \leq k+1\right] \geq b_k c_k = b_{k+1} > 0, \qquad (2.39)$$

and (2.36) holds for $k+1$.

Now, one will use the ergodic theorem on the interval $[t_{k-1}+n_k, t_k]$. Since $n_1+\cdots+n_p$ is constant, independently of n, it is easy to modify the partial empirical measures (see (2.32)) to ensure finally

$$\liminf_{n\to\infty} \mathbb{E}^* \left[\prod_{k=1}^{p} \mathbb{1}_{\{L_n^k \in \mathcal{V}_A(k)\}} e^{h_k(X_{t_k}, X_{t_{k-1}+n_k})} \right] \geq b_p \prod_{k=1}^{p} \inf_{(i,j)\in\Gamma'_k} e^{h_k(i,j)} > 0,$$

because Γ'_k is finite, so that the infima are strictly positive. Therefore the (modified) sum of line (2.35) tends to 0 when $n \to \infty$.

For measures with infinite supports, approximations and continuity arguments (Proposition 2.19 still holds for reducible kernels) yield the result along the same way as in the irreducible case. ∎

Remark : Note that the proof would have been more difficult in taking directly infinite support measures, for uniformity reasons.

The above construction can be related to the control and cost structure described in [56], but we do not need here to deal with all control policies, since it is possible to choose directly the optimal one. However the aims are somewhat different and the change of measure allows to obtain the rate function in a closed form, instead of an expression as a "limit of limit".

♦♦♦

2.A.3 Upper bound

Theorem *For any compact set $K \subset \mathcal{M}_s(E^2)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \leq - \inf_{A \in K\cap\mathcal{M}} H_d(A\|P). \tag{2.40}$$

Proof : It is much simpler. When A is not serial, by Lemma 2.14,

$$A \notin \mathcal{M} \implies \exists \mathcal{V}_A,\ \exists n_0 : \forall n \geq n_0, \quad \mathbb{P}\left[L_n \in \mathcal{V}_A\right] = 0,$$

where $\mathcal{V}_A$ is an open neighborhood of A. Then $\mathbb{P}[L_n \in K \setminus \mathcal{V}_A]$ is equal to $\mathbb{P}[L_n \in K]$, for n large enough, and the upper bound becomes

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[L_n \in K\right] \leq - \inf_{A' \in K\setminus\mathcal{V}_A} H_d(A'\|P).$$

This can be achieved by means of finite number of neighborhoods. Using the lower semi-continuity of H_d and the closure of $K \cap \mathcal{M}$, we get

$$\inf \left\{ \inf_{A' \in K \setminus \cup_{i=1}^{p} \mathcal{V}_{A_i}} H_d(A'\|P) \ ,\ A_1, \ldots, A_p \notin \mathcal{M} \right\} = \inf_{A' \in K\cap\mathcal{M}} H_d(A'\|P),$$

which gives the upper bound (2.40). ∎

2.B Properties of the entropy function

The entropy function lies in the heart of the LDP, and we need to analyze some of its properties to derive the large deviation assertion.

2.B.1 Entropy and information

Definition 2.15 (balanced measures) *The set $M_s(E^2)$ of balanced measures is the set of measures on E^2, with both identical 1–dimensional projections*

$$A \in M_s(E^2) \iff A(E,.) = A(.,E) \in \mathcal{M}_1(E). \tag{2.41}$$

Let Γ be a graph on E. $\mathcal{M}_s(\Gamma) \subset \mathcal{M}_s(E^2)$ is the set of balanced measures with support included in Γ.

For $A \in M_s(E^2)$, let us put

- $a_{ij} \stackrel{\text{def}}{=} A(\{(i,j)\})$ the 2–dimensional law;
- $a_i \stackrel{\text{def}}{=} A(\{i\} \times E)$, the unique 1–dimensional projection;
- $A_{ij} \stackrel{\text{def}}{=} A(E \times \{j\} | \{i\} \times E)$, the conditional law, and hence $a_{ij} = a_i A_{ij}$.

Definition 2.16 (relative entropy) *Let $A \in M_s(E^2)$. The relative entropy of A with respect to P is defined by*

$$H_d(A\|P) \stackrel{\text{def}}{=} \sum_{i,j\in E} a_{ij} \log\left(\frac{a_{ij}}{a_i P_{ij}}\right), \tag{2.42}$$

with the usual convention

$$0\log(0/0) = 0, \quad 1/0 = +\infty, \quad 0\log 0 = 0.$$

This definition extends in a natural way to finite positive measures A and substochastic matrices P. $H_{E'}$ is the restriction of the entropy H_d to E', i.e. taking the indices $i, j \in E'$ in (2.42).

All results, and the next as well, do have an information theory significance. The reader is referred to Rényi [121], which provides an excellent introduction containing the link between information theory and probability. For a more complete study, see [3], and Hamming [73].

Recalling that the *information* brought by a Bernoulli trial Y is $-\log \mathbb{P}[Y]$, one sees $-n^{-1}\log \mathbb{P}[L_n \in F]$ as the mean information brought by the knowledge of L_n being in F. By the way, we think the expression "relative entropy" not to be as illuminating as *information gain*. Indeed the relative entropy is the *mean information gain*:

$$H_d(A\|P) = \sum_{i\in E} a_i \sum_{j\in E} A_{ij} \log\left(\frac{A_{ij}}{P_{ij}}\right) = \mathbb{E}_A\left[H(A_{X_2}\|P_{X_1})\right], \tag{2.43}$$

where $H(A_{X_2}\|P_{X_1}) \stackrel{\text{def}}{=} \sum_j A_{ij}\log(A_{ij}/P_{ij})$ is the information gain[7] for X_2 conditioning on the event $\{X_1 = i\}$ and the expectation is taken for X_1 distributed with the law (a_i) (the stationnary law for A_{ij}). Thus $H_d(A\|P)$ is the mean information gain for each sample, under the dynamics A.

Lemma 2.17 *The relative entropy $H_d(A\|P)$ is positive, convex and lower semi-continuous with respect to A. It is null if, and only if, one can write $A = P$ as a Markov kernel, i.e. $A_{ij} = P_{ij}$, $\forall i, j \in E$.*

Proof : Positivity, lower semi-continuity and convexity are immediate from equation (2.43).
As for the convexity, let $A, A' \in \mathcal{M}_s(E^2)$, $x, x' \geq 0$ with $x + x' = 1$. Then $B = xA + x'A'$ is also balanced with

$$b_i = xa_i + x'a_i', \quad B_{ij} = (xa_{ij} + x'a_{ij}')/(xa_i + x'a_i'),$$

so that

$$\begin{aligned} xH_d(A\|P) + x'H_d(A'\|P) &= \sum_{i\in E} xa_iH(A_i\|P_i) + x'a_i'H(A_i'\|P_i) \\ &\geq \sum_{i\in E} H(xa_iA_i + x'a_i'A_i'\|(xa_i + x'a_i')P_i) \\ &= \sum_{i\in E} b_iH(B_i\|P_i) = H_d(B\|P), \end{aligned}$$

by the convexity and the 1–homogeneity of H. ■

We will state now an important *negative* property. Let

$$\Psi(K) \stackrel{\text{def}}{=} \left\{A \in \mathcal{M}_s(E^2) : H_d(A\|P) \leq K\right\}$$

denote the level set of the entropy function.

The closed level set $\Psi(K)$ is not necessarily compact.

[7]Here H is the relative entropy between measures in $\mathcal{M}_1(E)$.

Proof : Consider the irreducible ergodic kernel P defined on $\mathbb{N}$ by

$$P_{01} = 1, \qquad \text{and} \qquad P_{ij} = \begin{cases} 3/8 & \text{if} \quad j = i-1, \\ 1/2 & \text{if} \quad j = i, \\ 1/8 & \text{if} \quad j = i+1, \end{cases} \qquad \text{for } i \geq 1.$$

Then, for each $q > 0$,

$$H_d(\delta_{qq} \| P) = -\log P_{qq} = \log 2.$$

Hence, the set $\{\delta_{qq} : q \geq 1\} \subset \Psi(\log 2)$ is not tight and $\Psi(\log 2)$ is not compact. ■

> **Remark :** Consequently, a full LDP will, in general, not hold, since it would require the entropy function to have compact level sets. In this paper, we prove that there is a *weak* LDP for *all* Markov chains. Obviously, a full LDP will hold only for some restricted classes of chains.

♦♦♦

2.B.2 Continuity properties of the entropy function

The proof of the lower bound of the LDP needs continuity properties of the entropy function. According to Definition 2.20, the graph Γ will be assumed, in this section, to correspond to an irreducible transition matrix P. Note that $H_d(A\|P)$ is finite only if $A \in \mathcal{M}_s(\Gamma)$.

Proposition 2.18 (continuity) *Let Γ' be a finite subgraph of Γ. Then $H_d(.\|P)$ is finite and continuous on $\mathcal{M}_s(\Gamma')$.*

Proof : The entropy function is a finite sum, for $(i,j) \in \Gamma'$, of finite continuous functions $A \mapsto a_{ij} \log(a_{ij}/a_i P_{ij})$, since $P_{ij} > 0$, $\forall i, j$. ■

Proposition 2.19 (exact l.s.c.) *Let $A \in \mathcal{M}_s(\Gamma)$. There exists a sequence of measures $A(n) \in \mathcal{M}_s(\Gamma)$, with finite support, converging to A, such that their entropy converges to $H_d(A\|P)$.*

$$\lim_{n\to\infty} A(n) = A \quad \textit{and} \quad \lim_{n\to\infty} H_d(A(n)\|P) = H_d(A\|P).$$

Proof : Assume $H_d(A\|P) < \infty$. By Proposition 2.22, there exists a decomposition into cycles $A = \sum_{i=1}^{\infty} u_i L(C_i)$, from which one can define the sequence

$$A(n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} u_i L(C_i) + \sum_{i=n+1}^{\infty} u_i L(C_1) \ \in \mathcal{M}_s(\Gamma) \stackrel{n\to\infty}{\longrightarrow} A. \tag{2.44}$$

In addition, let $K \stackrel{\text{def}}{=} \operatorname{Supp}(C_1)$. Since K is finite, there exists q_0 such that

$$K \subset \{1, \ldots, q_0\}^2.$$

Introduce the measure $\pi_{ij} \stackrel{\text{def}}{=} a_i P_{ij}$ and the ratios

$$\begin{cases} r_{ij} \stackrel{\text{def}}{=} a_{ij}/\pi_{ij}, \\ r_{ij}(n) \stackrel{\text{def}}{=} a_{ij}(n)/\pi_{ij}, \\ r_i(n) \stackrel{\text{def}}{=} a_i(n)/a_i. \end{cases}$$

By construction,

$$-e^{-1} \le r_{ij}(n) \log r_{ij}(n) \le \max\{0, r_{ij} \log r_{ij}\}, \quad \forall (i,j) \notin K, \tag{2.45}$$
$$-e^{-1} \le r_i(n) \log r_i(n) \le 0, \qquad \forall i > q_0. \tag{2.46}$$

These inequalities are derived from special forms of the decomposition (2.44). Indeed, writing for the sake of shortness $f(x) = x \log x$,

$$\begin{aligned} H_d(A(n)\|P) \;=\; & \sum_{(i,j)\in K} \pi_{ij} f\big(r_{ij}(n)\big) - \sum_{i\le q_0} a_i f\big(r_i(n)\big) \\ & + \sum_{(i,j)\in K^c} \pi_{ij} f\big(r_{ij}(n)\big) - \sum_{i> q_0} a_i f\big(r_i(n)\big) \end{aligned} \tag{2.47}$$

The first two sums in (2.47) are finite and converge, because f is continuous and there is finite number of terms. Similarly, using the bounds in (2.45)–(2.46), one sees the last two sums in (2.47) are also finite. Thus, Lebesgue's theorem implies that

$$H_d(A(n)\|P) \stackrel{n\to\infty}{\longrightarrow} H_d(A\|P).$$

When $H_d(A\|P) = \infty$, then a finite number of terms are sufficient to prove that $H_d(A(n)\|P)$ is larger than any fixed number, so that the limit is still true. The proof of Theorem 2.19 is terminated. ∎

2.C Cycles and balanced decomposition

The main goal of this section is to provide in Proposition 2.22 an effective decomposition of balanced measure, in order to analyze the continuity of the entropy function. To this end, convenient tools are graph theory [12], convex sets properties [123, 11, 96] and the bible [20].

A graph will be described as a subset Γ of $E \times E = E^2$, where E is countable. This appendix is independent of Markov chains, although all graphs considered hereafter are indeed associated with Markov transition matrices.

Definition 2.20 *The graph Γ of a Markov transition matrix P is the set of all possible transitions, i.e. $(i,j) \in \Gamma \iff P_{ij} > 0$.*

Clearly, Γ is strongly connected if, and only if, P is irreducible. A cycle is described as an n-tuple $(x_1, \dots, x_n)$, with all $(x_i, x_{i+1}) \in \Gamma$. It is minimal if, and only if, it passes through each vertex at most once, so that the number of minimal cycles is at most countable. These cycles are denoted by $(C_i)_{i \in \mathbb{N}}$. They will play an important role, essentially because any cycle has at least one minimal cycle. The empirical measure $L(C)$ of a cycle C belongs to $\mathcal{M}_s(\Gamma)$. The following proposition introduces some approximations that are needed for the method of change of measure.

Proposition 2.21 *Let Γ be a strongly connected graph, $A \in \mathcal{M}_s(\Gamma)$ with a finite support. There exists a finite subgraph $\Gamma' \subset \Gamma$ and balanced measures $A(n)$ converging to A, such that* $\operatorname{Supp} A(n)$ *is strongly connected and belongs to Γ'.*

Proof : Let $\{x_1, \dots, x_n\}$ be the support of A. Since Γ is strongly connected, there exists a Γ–path $T_{i,j}$ from x_i to x_j, hence $C = T_{1,2} \dots T_{n,1}$ is a cycle connecting all points of the support of A.
Taking $\Gamma' = \operatorname{Supp}(C) \cup \operatorname{Supp}(A)$ and $A(n) = n^{-1}L(C) + (1 - n^{-1})A$, the proposition is proved. ∎

We also need finite approximations. One can easily guess that the $L(C_i)$ generate $\mathcal{M}_s(\Gamma)$, but, when Γ is countable, there is also a countable number of minimal cycles, hence $\mathcal{M}_s(\Gamma)$ is neither finitely generated nor compact, so that a Krein–Millman like theorem must be proved for our special situation and this is the content of the next Proposition 2.22.
Convex sets will be considered in a topological vector space denoted by $\mathcal{M}_s^{ev}(\Gamma)$, the set of *finite signed* balanced measures on Γ. Endowed with the L_1-norm

$$\|A\| = \sum_{ij} |a_{ij}| < \infty,$$

$\mathcal{M}_s^{ev}(\Gamma)$ is even a Banach space. The cone $\mathcal{M}_s^{+}(\Gamma) \subset \mathcal{M}_s^{ev}(\Gamma)$ of *positive* balanced measures is the right object to study, since the extreme points of $\mathcal{M}_s(\Gamma)$ correspond to extreme generatrices of the cone, but we will strive to focus rather on simplicity than on full generality.

Proposition 2.22 (balanced decomposition) *The set of balanced measures having their support in* Γ *is the closed convex hull generated by the empirical measures of the minimal cycles of* Γ*, that is*

$$\mathcal{M}_s(\Gamma) = \overline{\text{co}}\left(\left\{L(C_i)\right\}_{i\in\mathbb{N}}\right). \tag{2.48}$$

Proof : Since $\mathcal{M}_s(\Gamma)$ is closed, $\overline{\text{co}}(\{C_i\}_{i\in\mathbb{N}}) \subset \mathcal{M}_s(\Gamma)$, because the closed convex hull is the closure of the convex hull.

Conversely, take $A \in \mathcal{M}_s(\Gamma)$, and consider $A - uL(C_1)$. This is a balanced measure, positive for $u = 0$. Introduce the sequences

$$\begin{cases} u_{i+1} = \max\left\{u : B_i - uL(C_i) \in \mathcal{M}_s^+(\Gamma)\right\}, \\ B_{i+1} = B_i - u_i L(C_i), \ \forall i \geq 1, \quad \text{with } B_1 = A. \end{cases} \tag{2.49}$$

This yields a decomposition with non-negative coefficients $u_1, u_2, \ldots$ and

$$B = A - \sum_i u_i L(C_i)$$

is positive, since $\mathcal{M}_s^+(\Gamma)$ is closed. Suppose now $B \neq 0$. Then there exists x_1, such that $b_{x_1} = \varepsilon > 0$. Since B is positive and balanced,

$$b_x = \sum_j b_{xj} = \sum_i b_{ix}, \ \forall x,$$

so that one can define by induction the sets

$$\mathcal{S}_1 \stackrel{\text{def}}{=} \{x_1\}, \quad \mathcal{S}_{n+1} \stackrel{\text{def}}{=} \{x_{n+1} : b_{x_n x_{n+1}} > 0, \quad \forall x_n \in \mathcal{S}_n\} \neq \emptyset,$$

where

$$\sum_{i\in\mathcal{S}_n} \sum_{j\in\mathcal{S}_{n+1}} b_{ij} \geq \varepsilon. \tag{2.50}$$

If a cycle could be extracted from $\mathcal{S}_1, \mathcal{S}_2, \ldots$, then B would possess a minimal cycle C_l, and u_l would not satisfy (2.49). On the other hand, if all $\mathcal{S}_n$ would be disjoint, then $\|B\|$ would be as large as the quantity obtained by summing (2.50) over all $n \geq 1$, which is infinite. Both conclusions are contradictory. Consequently,

$$B = 0, \quad A = \sum_i u_i L(C_i),$$

which implies as well that $\mathcal{M}_s(\Gamma) \subset \overline{\text{co}}(\{L(C_i)\}_{i\in\mathbb{N}})$. Proposition 2.22 is proved. ∎

Chapitre 3

Large Deviation Principle for Markov Chains in Continuous Time

Arnaud de La Fortelle
March 2000

Abstract

Let E be a denumerable state space, Y_t be an homogeneous Markov process on E with generator R. We introduce the *empirical generator* G_t of Y_t, and prove strong LDP local bounds for it. This allows to prove the weak LDP in a very general setting, for irreducible non-explosive Markov processes, not necessarily ergodic. Sanov's theorem is obtained by a contraction argument from the weak LDP for G_t.

In our opinion this is an improvement with respect to the existing literature, since LDP in the Markov case requires in general, either E to be finite, or strong uniformity conditions, which important classes of chains do not verify, e.g. bounded jump networks. Moreover the empirical generator together with the representation of the rate function as an entropy allow to prove nice properties (uniqueness, continuity, convexity). It also leads to applications in simulation (importance sampling) and in the evaluation of the rate function for sample path LDP in networks. Finally it seems that some technical problems can be reduced to convex programs which can be run with fast algorithms.

keywords Large deviation, weak LDP, Markov, countable, empirical measure, Sanov, entropy, information.

3.1 Introduction

Large deviations theory has been widely studied and the range of objects for which an LDP holds is nowadays very large (see [41, 40] and references therein). There are also several approaches [15, 27, 31, 46, 86, 87, 112, 127] by which one can demonstrate these principles, such as subadditivity, convex transforms or change of measures.
The goal of this study is to propose an LDP for a new object, *the empirical generator*, and to show that it is a natural object from which lots of interesting corollaries can be derived. The framework is constituted by continuous time Markov processes on a countable state space, which includes many applications, e.g. queueing networks.
In discrete time this would be a "level 2.5" LDP, i.e. Sanov's theorem for the pair empirical measure (see [61] for the definition of the level). But in discrete time (X_i, X_{i+1}) itself is a Markov chain so that the level 2.5 LDP can be derived from the level 2 LDP, the classical Sanov's theorem.
This is no longer true in continuous time since there is no such thing like $(X_t, X_{t+\mathrm{d}t})$. Actually, the continuous time entropy[1] H_c has been recently introduced (see [92]), a representation formula of Sanov's rate function has been proved using H_c (see [10]), but there is no LDP. Our results proves that H_c is really a rate function, and that the representation formula is just a contraction of the LDP. Moreover the validity is extended to a countable, instead of finite, state space.
A recent, thorough search showed that the LDP[2] was in fact proved in [76], before the two previously cited works [10, 92]. It was published in *Studia Scientiarum Mathematicarum Hungarica*, but surprisingly enough neither this very interesting paper nor his author are mentioned in the literature devoted to large deviations. But the idea of empirical generator as well as the information-theoretic interpretation of the rate function are already there. This can be explained by a wall between different mathematical fields: Hornik [74, 75, 76] was interested in statistics (hence the idea of measuring the generator), namely optimal tests, when Kesidis and Walrand [10], Baldi and Piccioni [92] (see also [22]) were rather focused on simulation (importance sampling); our point of view is probabilistic, since further results exhibit a simple relationship in queueing networks between the tails of stationary distributions and the entropy function. This multiplicity of separate fields shows in itself the broad interest of such results.

---

[1] We shall denote by an index c [resp. d] the entropy function related to the continuous [resp. discrete] time.

[2] The state space is assumed to be finite. The proof relies on Perron-Fröbenius theorem, so that our approach does differ significantly.

Using the empirical generator has theoretical as well as practical interest. By this mean, Sanov theorem can be derived by contraction for irreducible non-explosive Markov processes and therefore the rate function is easily shown to possess good properties (uniqueness, continuity, convexity), that are not proved in such a general framework. Indeed, in a countable state space, the full LDP does not necessarily hold, and this whole study only requires the weak LDP. It is also an important part of the work to show that, under suitable assumptions, almost all operations usually presented for the sole full LDP (such as contraction) can be performed on weak LDP. Moreover, we prove stronger bounds than LDP bounds, since we are able to prove that the entropy is the Ruelle-Lanford function for the empirical generator. The LDP is then an immediate corollary (see [97]) and this allows additionally to prove the uniqueness of the rate function, which is almost never done for weak LDP.

The practical interest is to show how to extend techniques requiring a change of measure, e.g. importance sampling (see [22]). Besides, adapted empirical generators can greatly simplify the evaluation of sample path LDP for networks, and even be a (practical) mean to decide between ergodicity and transience.

The organization of the paper is the following: the empirical generator G_t and the entropy function H_c are defined and analyzed in Section 3.2. H_c is proved to be the Ruelle-Lanford function of G_t in Section 3.3, which yields the LDP. Properties and operations on weak LDP are discussed in Section 3.4. Section 3.5 then derives Sanov's theorem and proves all the usual representations of the rate function to be equal.

In conclusion we wish to emphasize that this method, even though the main object is called a generator, is absolutely not limited to the Markov case. For example, this result can be extended very naturally to semi-Markov processes. The state space can also be generalized, but then more involved analysis is required.

3.2 Generators and entropy

The entropy function being in the heart of the LDP, it is useful to study its properties. On the other hand, this analysis is needed in order to demonstrate the large deviations assertion.

Because our purpose is to extend its result, we refer the reader to [34] for detailed properties of balanced measures and of the entropy function for *discrete time* Markov chains.

3.2.1 The empirical generator

Let E be a countable space and $(Y_t)_{t\geq 0}$ be a Markov process on E. Define the following notation

- T_n is the n-th time of jump (with $T_0 = 0$),
- $N_t \stackrel{\text{def}}{=} \sum_{n\geq 0} \mathbb{1}_{\{T_n \leq t\}}$ is the number of jumps at time t ($N_0 = 1$),
- $X_{n+1} \stackrel{\text{def}}{=} Y_{T_n}$ is the embedded chain ($Y_0 = X_1$),
- L_t is the empirical measure of Y_t and $\hat{L}_n$ the pair empirical measure of the embedded chain X_n,

$$L_t \stackrel{\text{def}}{=} \frac{1}{t}\int_0^t \delta_{Y_s}\,\mathrm{d}s, \tag{3.1}$$

$$\hat{L}_n \stackrel{\text{def}}{=} \frac{1}{n}\left(\sum_{i=1}^{n-1} \delta_{X_i X_{i+1}} + \delta_{X_n X_1}\right), \tag{3.2}$$

- $\mathcal{M}_+(E)$ [resp. $\mathcal{M}_1(E)$] is the set of finite positive [resp. probability] measures on E.

Now, the pair empirical measure $\hat{L}_n$ has the particular feature to be *balanced.*

Definition 3.1 (balanced measures) *The set $\mathcal{M}_s(E^2)$ of balanced measures is the subset of $\mathcal{M}_+(E^2)$, with both identical 1–dimensional projections*

$$A \in \mathcal{M}_s(E^2) \iff A(E,.) = A(.,E) \in \mathcal{M}_+(E). \tag{3.3}$$

Let Γ be a graph on E. $\mathcal{M}_s(\Gamma) \subset \mathcal{M}_s(E^2)$ is the set of balanced measures with support included in Γ.

For $A \in M_s(E^2)$, let us put

- $a_{xy} \stackrel{\text{def}}{=} A(\{(x,y)\})$ the 2–dimensional law;
- $a_x \stackrel{\text{def}}{=} A(\{x\}\times E)$, the unique 1–dimensional projection;
- $A_{xy} \stackrel{\text{def}}{=} A(\{(x,y)\}|\{x\}\times E)$, the conditional law, and hence $a_{xy} = a_x A_{xy}$.

Definition 3.2 (empirical generator) *Let Y_t be a Markov process on a countable space E. The empirical generator associated to Y_t is defined by*

$$G_t \stackrel{\text{def}}{=} \left(L_t, \frac{N_t}{t}\hat{L}_{N_t}\right).$$

$\mathcal{G}_s \stackrel{\text{def}}{=} \mathcal{M}_1(E) \times \mathcal{M}_s(E_*^2)$ denotes the set of empirical generators, where E_*^2 is the set of off-diagonal pairs of E. Let $G = (\pi, A)$ be an empirical generator; G is said irreducible if A is irreducible and the support of G is the support of A.
Let Γ be a subgraph of E_*^2. $\mathcal{G}_s(\Gamma) \subset \mathcal{G}_s(E_*^2)$ is the set of empirical generators with in Γ. For a generator G, we shall use the notation $\mathcal{G}_s(G)$ for $\mathcal{G}_s\big(\mathrm{Supp}(G)\big)$.

The following proposition explain why we still call the pair $G = (\pi, A)$ a *generator.*

Proposition 3.3 *Let E be a countable state space. There is a one-to-one mapping between the irreducible empirical generators (π, A) and the generators L_{xy} of irreducible ergodic Markov process on E such that $\sum_{x,y} \pi_x L_{xy}$ is finite; this correspondence is given by*

$$L_{xy} = \frac{a_{xy}}{\pi_x} \quad \text{if } x \neq y \qquad \text{and} \quad L_{xx} = -\frac{a_x}{\pi_x}. \tag{3.4}$$

Proof : In this correspondence, L_{xy} can be infinite if $\pi_x = 0$ when $a_x > 0$. There is formally no difficulty to handle this case which corresponds to an instantaneous state x, but we shall not treat it since an instantaneous transition almost never occurs.
Let $G = (\pi, A)$ be an irreducible empirical generator. It is associated a unique generator L_{xy} by (3.4). The transition probabilities for L are $-L_{xy}/L_{xx} = A_{xy}$ for $x \neq y$, hence L is irreducible. Since π is a stationary distribution for L, L is the generator of an irreducible ergodic Markov process.
Reciprocally consider an irreducible ergodic Markov process Y_t with generator L_{xy}, i.e. with intensities $-L_{xx}$ and transition probabilities A_{xy}. There exists a unique stationary distribution π for Y_t. If $\sum_{x,y} \pi_x L_{xy} < \infty$, $(\pi_x, \pi_x L_{xy})$ is an irreducible empirical generator, since $(\pi_x L_{xy})$ is a *finite* balanced measure. Obviously both correspondences are inverse of each other, hence the one-to-one correspondence. ∎

We shall sometimes denote by l the total mass $\sum_{xy} a_{xy}$ of A. Proposition 3.3 makes it easy to interpreted l, π_x, A_{xy} as the mean intensity of jumps, the stationary distribution and the routing. Therefore a_{xy} can be regarded as the mean number of jumps from x to y per unit time. Obviously these quantities exist if, and only if, the generator is ergodic.
When L is not irreducible, there might exist several stationary distributions which determine several empirical generators. However, if L and π are known, (π, A) is uniquely determined. The worst case is $l = 0$ (i.e. $A = 0$, there is

no jump); the mapping $(\pi, A) \to L_{xy}$ is ill-defined for all x such that $\pi_x = 0$, and $L_{xx} = 0$ for all x such that $\pi_x > 0$. Actually the routing is not well defined, but it does not matter as long as it concerns the entropy function. Actually, finding a convenient topology for $\mathcal{G}_s$ is enlightening for all these special cases.

Topological considerations : The topology for the discrete case is defined by the distance induced by the norm of the underlying Banach space of finite signed measures,

$$d(A, A') = \|A - A'\|_1 \stackrel{\text{def}}{=} \sum_{x,y \in E} |a_{xy} - a'_{xy}|.$$

which yields the same topology on $M_s(E^2)$ as the $\|.\|_\infty$ norm, even though the distances are not equivalent.

The underlying vector space is here the space of signed generators $\mathcal{G}_s^{\text{ev}}(E)$, when π, A or L_{xy} are finite, signed, and where A still verifies the balance equation (3.3). The generator L has in fact to be defined by a pair (L, π) where π is L-invariant, i.e. $\pi L = 0$. The sum and multiplication by a scalar is defined by $G = \alpha G' + \beta G''$,

$$\begin{cases} \pi = \alpha\pi' + \beta\pi'', \\ a_{xy} = \alpha a'_{xy} + \beta a''_{xy} \end{cases} \qquad \text{or} \qquad \begin{cases} \pi_x = \alpha\pi'_x + \beta\pi''_x, \\ L_{xy} = \frac{\alpha\pi'_x L'_{xy} + \beta\pi''_x L''_{xy}}{\alpha\pi'_x + \beta\pi''_x} \end{cases}$$

It is easily seen that the real variable is $\pi_x L_{xy}$. It is now natural to define the null element as $A = 0$ *and* $\pi = 0$ or equivalently $\pi = 0$ for the (L_{xy}, π_x) representation.

$\mathcal{G}_s^{\text{ev}}(E)$ is equipped with the norm

$$\|L\|_1 \stackrel{\text{def}}{=} \begin{cases} \sum_{x \in E} |\pi_x| + \sum_{x,y \in E} |\pi_x L_{xy}|, \\ \sum_{x \in E} |\pi_x| + \sum_{x,y \in E} |a_{xy}|, \end{cases}$$

and $\mathcal{G}_s(E)$ by the distance induced by this norm. It is now clear that the two definitions are equivalent, for the representation, the operations and the topology.

♦♦♦

3.2.2 Entropy

Definition 3.4 (discrete time entropy) *Let P be a transition matrix and $A \in \mathcal{M}_s(E^2)$. The discrete time relative entropy of A with respect to P is* *defined by*

$$H_d(A\|P) \stackrel{\text{def}}{=} \sum_{x,y\in E} a_{xy} \log\left(\frac{A_{xy}}{P_{xy}}\right), \tag{3.5}$$

with the usual convention

$$0\log(0/0) = 0, \quad 1/0 = +\infty, \quad 0\log 0 = 0.$$

We shall hereafter also denote by P the support of the transition matrix P_{xy}.

Proposition 3.5 *The discrete time relative entropy $H_d(A\|P)$ enjoys the following properties:*

- *H_d is positive, convex and lower semi-continuous with respect to A;*
- *H_d is null if, and only if, one can write $A = P$ as a Markov kernel, i.e. $A_{xy} = P_{xy}$, for all x such that $a_x > 0$ and for all $y \in E$;*
- *Let Γ be a finite subgraph of P. Then $H_d(.\|P)$ is finite and continuous on $\mathcal{M}_s(\Gamma)$;*
- *Let $A \in \mathcal{M}_s(P)$. There exists a sequence of measures $A(n) \in \mathcal{M}_s(P)$, irreducible, with finite support, converging to A, such that their entropy converges to $H_d(A\|P)$.*

$$\lim_{n\to\infty} A(n) = A \quad \text{and} \quad \lim_{n\to\infty} H_d(A(n)\|P) = H_d(A\|P).$$

Note that the closed level sets of H_d are not necessarily compact, hence a full LDP in discrete time does not hold in general (see [34]). This remark is still valid in continuous time.

Definition 3.6 (continuous time entropy) *Let R be a generator and $G = (\pi, A)$ an empirical generator. Denote by L_{xy} the corresponding generator (see (3.4)). The relative entropy of G with respect to R is defined by*

$$H_c(G\|R) \stackrel{\text{def}}{=} \sum_{x\neq y\in E} \pi_x I_p(L_{xy}\|R_{xy}), \tag{3.6}$$

$$\text{where } I_p(\nu\|\lambda) \stackrel{\text{def}}{=} \nu\log\frac{\nu}{\lambda} - \nu + \lambda,$$

with the usual convention of Definition 3.4.

Information theory interpretation The continuous time entropy function H_c has a twofold information theory significance, depending on how the Markov process is considered.

First note that $I_p(\nu\|\lambda)$ is the information gain by unit time when the intensity of a Poisson process is changed from λ to ν. It is linked with the rate function I_e for i.i.d. exponentially distributed random variables by

$$I_p(\nu\|\lambda) = \nu I_e(\nu^{-1}\|\lambda^{-1}) = \nu f(\lambda\nu^{-1}),$$

where $f(u) = u - 1 - \log u$ is the rate function when the mean is 1 and $I_e(.\|x)$ is the rate function when the mean is x. The interpretation is straightforward. The inter-arrival times are exponentially distributed with mean ν^{-1} instead of λ^{-1} so that the "pay-off" for each jump is $I_e(\nu^{-1}\|\lambda^{-1})$, multiplied by the number of jumps by unit time ν.

On one hand, the process of jump from x to y occurs as a Poisson process of intensity L_{xy} instead of R_{xy} for a fraction π_x of time, so that the mean information gain is deduced to be (3.6). On the other hand, the jumps at x are exponentially distributed with parameter L_x instead of R_x, so that the change of distribution is "rewarded" $I_e(L_x^{-1}\|R_x^{-1})$ each time, and there is a number of passage a_x per unit time. Moreover the change of routing P to A is rewarded $H_d(A\|P)$ per unit time. Thus it yields the other form

$$H_c(G\|R) = H_d(A\|P) + \sum_{x\in E} a_x I_e(L_x^{-1}\|R_x^{-1}). \tag{3.7}$$

The equality between (3.6) and (3.7) is easily checked. This information theory heuristic is exactly the sketch of an alternative proof using conditioning (as in [34]). Since we propose a stronger method, this proof is not detailed here.

Proposition 3.7 *The relative entropy $H_c(G\|R)$ is positive, lower semi-continuous and convex with respect to $G = (\pi, A)$. It is null if, and only if, $L_{xy} = R_{xy}$ for all $x, y \in E$.*

Proof : All those properties are immediately derived from the same properties of I_p, by (3.6). ∎

As in the discrete time, the proof of the LDP requires continuity properties of the entropy function. Hereafter R will be assumed to be an irreducible generator, and for the sake of simplicity $\mathcal{G}_s(R)$ will denote the empirical generators on the graph of R. Note that $H_c(G\|R)$ is finite only if $G \in \mathcal{G}_s(R)$.

Proposition 3.8 *Let Γ be a finite subgraph of* $\mathrm{Supp}(R)$. *Then $H_c(.\|R)$ is finite continuous on $\mathcal{G}_s(\Gamma)$. It has compact level sets.*

Proof : By (3.6), the entropy function is a finite sum, for $(x,y) \in \Gamma$, of continuous finite functions with compact level sets. ∎

Proposition 3.9 (exact l.s.c.) *Let $G \in \mathcal{G}_s(R)$. There exists a sequence of generators $G(n) \in \mathcal{G}_s(R)$, irreducible, with finite support, converging to G, such that their entropy converge to $H_c(G\|R)$.*

$$\lim_{n\to\infty} G(n) = G \quad \textit{and} \quad \lim_{n\to\infty} H_c(G(n)\|R) = H_c(G\|R).$$

Proof : By (3.7), this is an extension of the same property for the discrete entropy function H_d (Proposition 3.5); there exists $A(n)$ converging to A such that $A(n)$ is irreducible, with finite support and $H_d(A(n)\|P)$ converges to $H_d(A\|P)$. The proof is completed by taking $G(n) = (\pi(n), A(n))$ where $\pi(n)$ is an approximation of π that has the same support as $A(n)$. ∎

3.3 LDP for the empirical generator

Let E be a countable state space, and Y_t be a Markov process with generator R. We prove here that the entropy H_c is a Ruelle-Lanford function for the empirical generator G_t, which is a stronger result than a weak LDP. For the notion of Ruelle-Lanford functions and the related theorems, we refer the reader to Lewis and Pfister [97] (or see Section 3.4.2).

Theorem 3.10 *Let Y_t be a non-explosive Markov process on a countable state space E with generator R and initial distribution ν. The entropy function $H_c(.\|R)$ is a Ruelle-Lanford function for the empirical generator G_t.*

The proof relies on an exponential change of measure, from which strong upper and lower bounds are derived. Using the continuity properties of the entropy function, these bounds are extended to all empirical generators. Finally $H_c(.\|R)$ is deduced to be a Ruelle-Lanford function for the empirical generator G_t.

Change of measure Let $G = (\pi, A)$ be a generator and L_{xy} be the corresponding generator[3] by (3.4). In order to define properly h (see below), the support of G is restricted to $\mathrm{Supp}(R)$, i.e. $G \in \mathcal{G}_s(R)$,

- $h(x,y) \stackrel{\text{def}}{=} \begin{cases} \log(L_{xy}/R_{xy}) \in \mathbb{R} \cup \{-\infty\} & \text{if } x \in \mathrm{Supp}(G), \\ 0 & \text{otherwise,} \end{cases}$

[3]Note that the relation (3.4) allows to define L_{xy} even though Proposition 3.3 is valid only for irreducible G with finite support.

- $\Lambda(x) \stackrel{\text{def}}{=} \sum_{y\neq x} R_{xy}(e^{h(x,y)} - 1)$,
- and the process,

$$M_t \stackrel{\text{def}}{=} \exp\left\{\sum_{i=1}^{N_t-1} h(X_i, X_{i+1}) - \int_0^t \Lambda(Y_s)\,\mathrm{d}s\right\}. \tag{3.8}$$

To prove that M_t is a martingale, one first calculate the derivative in 0

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_x\left[M_t\right]\Big|_{t=0} = \sum_{y\neq x} R_{xy}(e^{h(x,y)} - 1) - \Lambda(x) = 0, \quad \forall x \in E. \tag{3.9}$$

Using the Markov property, this means that the derivative is null for all $t \geq 0$, so that $\mathbb{E}\left[M_t\right] = M_0 = 1$. Then, using again the Markov property, this proves that $\mathbb{E}\left[M_s|\mathcal{F}_t\right] = M_t$ for all $s \geq t \geq 0$, i.e. M_t is a martingale with respect to $\mathcal{F}_t$.
A new probability measure is defined by

$$\mathbb{P}^*\left[B\right] \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbb{1}_{\{B\}} M_t\right], \qquad \forall B \in \mathcal{F}_t.$$

It is again classical that under the probability $\mathbb{P}^*$, Y_t is a Markov process with intensity L_{xy},

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}^*_x[f(Y_t)]\Big|_{t=0} = \begin{cases} \sum_{y\neq x} L_{xy}(f(y) - f(x)), & \text{if } x \in \mathrm{Supp}(G), \\ \sum_{y\neq x} R_{xy}(f(y) - f(x)), & \text{if } x \notin \mathrm{Supp}(G). \end{cases}$$

We proved that Y_t is a Markov process with generator L_{xy} on $\mathrm{Supp}(G)$ which is absorbing under $\mathbb{P}^*$, and generator R_{xy} outside $\mathrm{Supp}(L)$. We shall denote by $\mathrm{Supp}(\mathbb{P}^*)$ the graph of all possible transitions under $\mathbb{P}^*$, i.e. the pairs (x, y) of $\mathrm{Supp}(G)$ when $x \in \mathrm{Supp}(G)$ and of $\mathrm{Supp}(R)$ when $x \notin \mathrm{Supp}(G)$.
To this change of measure is associated the linear mapping

$$\phi_{R,G}(\pi', A') \stackrel{\text{def}}{=} \sum_{x,y\in E} a'_{xy} h(x, y) - \sum_{x\in E} \pi'_x \Lambda(x), \tag{3.10}$$

where $\pi' \in \mathcal{M}_1(E)$ and $A' \in \mathcal{M}_+(\mathrm{Supp}(\mathbb{P}^*))$. If G has finite support, $h(x, y)$ and $\Lambda(x)$ are finite when $(x, y) \in \mathrm{Supp}(\mathbb{P}^*)$ so that $\phi_{R,G}$ is finite and continuous on $\mathcal{G}(E) \stackrel{\text{def}}{=} \mathcal{M}_1(E) \times \mathcal{M}_+(\mathrm{Supp}(\mathbb{P}^*))$ and a simple calculation shows that

$$\phi_{R,G}(G) = H_c(G\|R). \tag{3.11}$$

Strong bounds This change of measure is particularly well suited for the following additive version of the empirical generator G_t,

$$\widetilde{G}(t_1, t_2) \stackrel{\text{def}}{=} \left(\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \delta_{Y_u} \, \mathrm{d}u, \; \frac{1}{t_2 - t_1} \sum_{i=N_{t_1}}^{N_{t_2}-1} \delta_{X_i X_{i+1}} \right) \in \mathcal{G}(E),$$

since

$$\log M_t = t\phi_{R,G} \left(\widetilde{G}(0,t) \right). \tag{3.12}$$

The difference between $\widetilde{G}(0,t)$ and G_t [resp. $\mathcal{G}(E)$ and $\mathcal{G}_s(E)$] is that the first one is not necessarily balanced. Moreover, they are related by

$$G_t = \widetilde{G}(0,t) + \frac{1}{t}(0, \delta_{Y_t Y_0}). \tag{3.13}$$

For the purpose of Ruelle-Lanford function, it is necessary to obtain much stronger bounds than for proving the LDP. Here, the lower bound holds for compact sets (instead of open sets) and the upper bound holds for open sets (instead of compact sets). Now, it is easier to prove these bounds for $\widetilde{G}(0,t)$ than for G_t.

Lemma 3.11 *Let Y_t be a Markov process with irreducible generator R_{xy}, then*

$$\lim_{r \to 0} \sup_{K \subset B(G,r)} \liminf_{t \to \infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in K\right] \quad \geq \quad -H_c(G \| R), \tag{3.14}$$

$$\lim_{r \to 0} \limsup_{t \to \infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in B(G,r)\right] \quad \leq \quad -H_c(G \| R), \tag{3.15}$$

where K denotes compact sets. (3.14) holds for any irreducible generator $G \in \mathcal{G}_s(R)$ with finite support while (3.15) holds for any $G \in \mathcal{G}_s(R)$.

Proof : The technical proof that follows is nothing but the statement: *under $\mathbb{P}^*$, Y_{t_0} reaches* $\mathrm{Supp}(G)$ *with non-null probability. Then Y_t remains in* $\mathrm{Supp}(G)$ *because it is absorbing, and by ergodicity $\widetilde{G}(t_0, t)$ converges to G.* The problem is to exhibit compact sets for the lower bound.

Let $G \in \mathcal{G}_s(R)$ be an irreducible generator with finite support. Choose x_0 such that $\nu(x_0) > 0$ and $x_1 \in \mathrm{Supp}(G)$. From the irreducibility of R, there exists t_1 with $\mathbb{P}^*\left[Y_0 = x_0, \; Y_{t_1} = x_1\right] > 0$. Since the space of càdlàg functions

$\mathbb{C}([0,t_1],E)$ is Polish, there exists a compact set K_0 of trajectories on $[0,t_1]$ such that

$$K_0 \subset \{Y_0 = x_0\} \cap \{Y_{t_1} = x_1\}, \quad \text{and} \quad \mathbb{P}^*\left[\{Y_t,\ t \le t_1\} \in K_0\right] = C > 0. \tag{3.16}$$

$\widetilde{G}(0,t_1)$ is a continuous functional of the trajectory $\{Y_t, t \in [0,t_1]\}$, so that

$$K_1 = \left\{\widetilde{G}(0,t_1)(\omega),\ \forall \omega \in K_0\right\}$$

is compact. But $\mathrm{Supp}(G)$ is absorbing for $\mathbb{P}^*$, hence the set of states is now restricted to a finite set.
Define K_2 to be the closed ball of center G and radius $\varepsilon > 0$, in $\mathcal{G}(\mathrm{Supp}(G))$. It is a compact set since G has finite support and $l = \sum_{x,y} a_{xy}$ is bounded. G is irreducible, thus the Markov process is ergodic, once in $\mathrm{Supp}(G)$, hence $\widetilde{G}(0,t)$ tends to G and there exists t_2 such that

$$\mathbb{P}^*_{x_1}\left[\widetilde{G}(0,t) \in K_2\right] \ge \frac{1}{2}, \qquad \forall t \ge t_2. \tag{3.17}$$

Now, the two parts have to be joined. By the additivity of $\widetilde{G}$,

$$\widetilde{G}(0,t) = \frac{t_1}{t}\widetilde{G}(0,t_1) + \left(1 - \frac{t_1}{t}\right)\widetilde{G}(t_1,t).$$

Fix $\varepsilon \le 1$ and define the compact set

$$K \stackrel{\text{def}}{=} \bigcup_{0 \le u \le \varepsilon} \{uK_1 + (1-u)K_2\} \ \subset \mathcal{G}(E).$$

Note that K is included in a ball of center G and radius 2ε in $\mathcal{G}(E)$. Now, using the Markov property, by (3.16) and (3.17)

$$\mathbb{P}^*\left[\widetilde{G}_t \in K\right] \ge \mathbb{P}^*\left[\widetilde{G}(0,t_1) \in K_1\right] \mathbb{P}^*_{x_1}\left[\widetilde{G}(0,t-t_1) \in K_2\right] \ge \frac{C}{2},$$

for all $t \ge t_1/\varepsilon$ and $t - t_1 \ge t_2$.
Using the definition of $\mathbb{P}^*$ and by (3.12),

$$\begin{aligned} \frac{1}{t}\log \mathbb{P}\left[\widetilde{G}_t \in K\right] &= \frac{1}{t}\log \mathbb{E}^*\left[\mathbb{I}_{\{\widetilde{G}_t \in K\}} M_t^{-1}\right] \\ &\ge -\sup_{G' \in K} \phi_{R,G}(G') + \frac{1}{t}\log \mathbb{P}^*\left[\widetilde{G}_t \in K\right]. \end{aligned} \tag{3.18}$$

Since the probability for $\widetilde{G}_t$ to belongs to K is greater than $C/2 > 0$ for t large enough,

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[\widetilde{G}_t \in K\right] \geq -\sup_{G'\in K} \phi_{R,G}(G').$$

Now $\phi_{R,G}$ is continuous and its value at G is $H_c(G\|R)$. Moreover K is included in a small neighborhood of G, so that the strong lower bound holds,

$$\lim_{r\to 0} \sup_{K\subset B(G,r)} \liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[\widetilde{G}_t \in K\right] \geq -H_c(G\|R), \tag{3.19}$$

where K denotes compact sets.
To convert this result for G_t is easily done by (3.13). Note that the second term $t^{-1}(0, \delta_{Y_tY_0})$ belongs to a compact K_3 when Y_t lies in the finite support of G and $Y_0 = x_0$. In addition it is very small when t is large. Formally, define

$$K' \stackrel{\text{def}}{=} \bigcup_{0\leq u\leq\varepsilon} \{uK_1 + (1-u)K_2 + ut_1^{-1}K_3\} \subset \mathcal{G}(E).$$

Then

$$\widetilde{G}(0,t) \in K \implies G_t \in K'.$$

Moreover K' is included in a ball of center G and radius 3ε as soon as $t \geq t_1/\varepsilon$. Finally, by (3.19), the lower bound (3.14) is valid for irreducible $G \in \mathcal{G}_s(R)$ with finite support.

The upper bound requires a really nice property of the functional $\phi_{R,G}$. Actually, since the upper bound (3.15) must be valid for any $G_0 \in \mathcal{G}_s(R)$, the continuity of $\phi_{R,G}$ can not be used. Reversing the inequality of (3.18) yields, for any measurable set $O \subset \mathcal{G}(E)$,

$$\frac{1}{t} \log \mathbb{P}\left[\widetilde{G}(0,t) \in O\right] \leq -\inf_{G'\in O} \phi_{R,G}(G'), \tag{3.20}$$

where the probability has been bounded by 1. Moreover the empirical generator G that defines the change of measure and the corresponding functional $\phi_{R,G}$ can be chosen arbitrarily. By (3.13), the difference between G_t and $\widetilde{G}(0,t)$ is bounded by $1/t$ so that

$$G_t \in B(G_0,\varepsilon) \implies \widetilde{G}(0,t) \in B(G_0, 2\varepsilon), \qquad \forall t > 1/\varepsilon,$$

hence, by (3.20),

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in B(G_0,\varepsilon)\right] \leq -\inf_{G'\in B(G_0,2\varepsilon)} \phi_{R,G}(G'). \tag{3.21}$$

A deep and interesting property of $\phi_{R,G}$ has now to be used: $\phi_{R,G}$ is a subdifferential of H_c at G. It is indeed linear and one checks that

$$H_c(G'\|R) - H_c(G\|R) \geq \phi_{R,G}(G') - \phi_{R,G}(G) = \langle \phi_{R,G}, G' - G \rangle, \quad \forall G' \in \mathcal{G}_s(R).$$

Now, fix $\eta > 0$. By the lower semi-continuity of H_c, there exists $\varepsilon > 0$ such that the infimum of H_c over $B(G_0, 3\varepsilon)$ is greater than $H_c(G_0\|R) - \eta$. G is chosen so that it minimizes the entropy function over the closed ball $B(G_0, 2\varepsilon)$ (this minimum exists). Since H_c and $B(G_0, 2\varepsilon)$ are convex, necessarily

$$\langle \phi_{R,G}, G' - G \rangle \geq 0, \qquad \forall G' \in B(G_0, 2\varepsilon).$$

Finally, using the above relation, (3.11) and (3.21), the upper bound (3.15) is derived from

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in B(G_0, \varepsilon)\right] \leq -\phi_{R,G}(G) \leq -H_c(G_0\|R) + \eta.$$

The lemma is proved ■

Proof of Theorem 3.10: Ruelle-Lanford function In a first step, the lower bound (3.14) has to be extended to all empirical generators.
Let $G \in \mathcal{G}_s(E)$, not necessarily irreducible nor with finite support and $r > 0$. By Proposition 3.9, for all $\varepsilon > 0$, there exists an irreducible finite support generator $G' \in B(G, r)$ such that $|H_c(G'\|R) - H_c(G\|R)| < \varepsilon$. Since $B(G, r)$ is open, there exists $r' > 0$ such that $B(G', r')$ is included in $B(G, r)$, and, by (3.14) for all $\eta > 0$ there exists a compact set $K \subset B(G', r')$ such that

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in K\right] \geq -H_c(G'\|R) - \eta \geq -H_c(G\|R) - \eta - \varepsilon.$$

Since this is true for all $\varepsilon > 0$ and all $\eta > 0$, (3.14) is valid for G, hence the extension of Lemma 3.11.
Remind that the upper and lower rate functions $\overline{I}$ and $\underline{I}$ are defined by (see Definition 3.15),

$$\underline{I}(G) \stackrel{\text{def}}{=} -\lim_{r\to 0} \liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in B(G, r)\right], \tag{3.22}$$

$$\overline{I}(G) \stackrel{\text{def}}{=} -\lim_{r\to 0} \sup_{K\subset B(G,r)} \limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in K\right], \tag{3.23}$$

This extension of Lemma 3.11 yields

$$\overline{I}(G) \leq \underline{I}(G) \leq H_c(G\|R) \leq \overline{I}(G),$$

for all generators $G \in \mathcal{G}_s(R)$. This means $\underline{I}(G) = \overline{I}(G) = H_c(G\|R)$ for these generators. Both bounds are proved: H_c is a Ruelle-Lanford function for G_t.
■

Theorem 3.12 (LDP) *Let Y_t be a non-explosive Markov process on a countable state space E. The empirical generator at fixed time G_t verifies a weak LDP with rate function H_c. If E is finite, G_t verifies a full LDP with good rate function H_c.*

Proof : The first part of the theorem is [97, theorem 3.1] (or Theorem 3.16 here). For the second part, it is only necessary to extend the upper bound from compact to closed sets. The rate function is known to be good (Proposition 3.8) but anyway this is implied by the fullness of the LDP.
The upper bound is valid for closed sets F such that there exists an upper bound l_0 for $l = \sum_{x,y} a_{xy}$; they are compact since E is finite so that $\mathcal{M}_1(E)$ and $\mathcal{M}_s(E_*^2)$ are compact. The extension of this bound is obtained by considering the asymptotics of $\mathbb{P}\left[N_t/t \geq l_0\right]$.
In fact, for a Poisson process with intensity r, for all $l_0 > r$

$$\mathbb{P}\left[N_t/t > l_0\right] = e^{-rt} \sum_{n > l_0 t} \frac{(rt)^n}{n!} \leq e^{-rt} \frac{(rt)^{l_0 t}}{(l_0 t)!} \frac{l_0}{l_0 - r},$$

because $n! \geq n_0! n_0^{n-n_0}$ for $n \geq n_0$. This bound also holds for the Markov process Y_t with $\bar{r} \stackrel{\text{def}}{=} \sup\{R_x : x \in E\} < \infty$, so that

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[N_t/t > l_0\right] \leq -\bar{r} + l_0 - l_0 \log \frac{l_0}{\bar{r}} \stackrel{l_0\to\infty}{\longrightarrow} -\infty$$

Hence, writing $\Psi(l_0)$ the set of all generators (π, A) verifying $l \leq l_0$, for all $C > 0$, there exists $l_0 < \infty$ such that

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in \Psi^c(l_0)\right] \leq -C.$$

By the lower bound (3.14), this implies that H_c is greater than C on $\Psi^c(l_0)$. Now consider $C > \inf_F H_c(.\|R)$ and a corresponding l_0. F is contained in the union of $F \cap \Psi(l_0)$ and $\Psi^c(l_0)$, so that

$$\begin{aligned} &\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in F\right] && (3.24) \\ \leq\ & -\min\left\{\inf_{G \in F\cap\Psi(l_0)} H_c(G\|R), C\right\} = -\inf_{G\in F} H_c(G\|R) \end{aligned}$$

because the infimum of H_c is reached on $F \cap \Psi(l_0)$. The exponential tightness for G_t is proved, hence the extension of the weak LDP to a full LDP. ■

Explosive process : If the process Y_t can jump an infinite number of times in a finite duration, an LDP can not be derived this way, yet Lemma 3.11 is almost true. The stumbling block is the new beginning of the process after a time of explosion: it can create a new non-balanced state, and finally (3.13) no more holds. This means that $\widetilde{G}_t$ does not tend necessarily toward $\mathcal{G}_s(E)$. The rate function is not necessarily infinite outside $\mathcal{G}_s(E)$ and can be less than H_c on $\mathcal{G}_s(E)$. What can be proved with this method is an LDP for

$$\mathbb{E}\left[\mathbb{1}_{\{G_t\in .\}}\mathbb{1}_{\{Y_t \text{ does not explode before } t\}}\right].$$

Even for an explosive process, this substochastic process is well-defined (see Feller [65]) and uniquely characterized by its generator R.

♦♦♦

3.4 Operations on weak LDP

A full LDP is often required in the literature for important corollaries of LDP, such as Varadhan's integral lemma or the contraction principle. In fact, under some natural restrictions, the same theorems applies for weak LDP.

3.4.1 Definition and extension

Definition 3.13 (Weak LDP) *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E. It satisfies a weak LDP with rate function I if*

- *(W1) $I:\ \mathcal{M}_1(E) \mapsto \mathbb{R}_+$ is a lower semi-continuous function;*
- *(W2) for all open sets $O \subset \mathcal{M}_1(E)$,*

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(O) \geq -I(O); \tag{3.25}$$

- *(W3) for all compact sets $K \subset \mathcal{M}_1(E)$,*

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(K) \leq -I(K); \tag{3.26}$$

denoting by $I(F)$ the infimum of I over a set $F \subset \mathcal{M}_1(E)$.

To handle closed sets, we introduce an *exponentially tight sequence*:

Assumption (H1) *There exists a sequence $\{E_n \subset E,\ n \geq 1\}$ of closed sets such that, for all $n \geq 1$,*

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon\big(E_n^c\big) \leq -n. \tag{3.27}$$

Such a sequence is called an exponentially tight sequence (w.r.t. μ_ε).

Theorem 3.14 *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E satisfying a weak LDP with rate function I. If there exists an exponentially tight sequence E_n (see Assumption (H1)), then the upper bound (3.26) holds for all closed sets F that have compact intersection with all E_n.*

Proof : Let $F \subset \mathcal{M}_1(E)$ be a closed set with compact intersection with all E_n and $n > I(F)$. By (H1) there exists a closed set E_n such that (3.27) is satisfied. By the lower bound, this implies that $I(x) \geq C$ for all $x \in E_n^c$, therefore $I(F) = I(F \cap E_n)$.
Then divide F into $F \cap E_n$, which is compact by assumption and into $F \cap E_n^c$ which is included in E_n^c. Then,

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(F) \leq -\min\{I(F \cap E_n); C\} = -I(F),$$

where the first term of the min is due to the the upper bound (3.26) and the second to the exponential tightness in θ. Hence the upper bound. ∎

Note that the regularity requirement on $\{E_n,\ n \geq 1\}$ is light: E_n needs only to be closed. In the classical theory, in order to improve the bounds from compact to closed sets, the rate function I is required to have compact level sets and μ_ε to be exponentially tight. One easily checks that these assumptions means, in our framework, that $E_n = I^{-1}([0,n])$ in which case (3.27) is the usual exponential tightness assumption. However, there are other interesting cases where the exponential tightness does not hold but where this theorem has interesting consequences (see Theorem 3.25).

3.4.2 Ruelle-Lanford functions and uniqueness

A first step to the identification is the classical uniqueness, which is not usually given in the framework of weak LDP. The following results are contained in [97].

Definition 3.15 *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E. A Ruelle-Lanford function is uniquely defined by $\overline{I} = \underline{I}$ with*

$$\underline{I}(x) \stackrel{\text{def}}{=} -\lim_{r\to 0}\liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(\mathcal{V}_r), \tag{3.28}$$

$$\overline{I}(x) \stackrel{\text{def}}{=} -\lim_{r\to 0}\sup_{K\subset\mathcal{V}_r}\limsup_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(K), \tag{3.29}$$

where $\mathcal{V}_r$ is a decreasing (when $r \to 0$) base of open neighborhoods of $x \in E$ and K are compact sets. Both $\overline{I}$ and $\underline{I}$ are lower semi-continuous.

Theorem 3.16 (Existence) *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E. If $\underline{I} \le \overline{I}$ in Definition 3.15, then μ_ε verifies a weak LDP with rate function I, for any lower semi-continuous I satisfying $\underline{I} \le I \le \overline{I}$.*

Theorem 3.17 (Uniqueness) *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E. If it satisfies a weak LDP with rate function I, either there is a Ruelle-Lanford function which is the unique rate function I, or there is no uniqueness.*

Proof : A necessary and sufficient condition for an LDP to hold is

$$\limsup_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(K) \le \liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(O),$$

for all open sets O and all compact sets $K \subset O$, which ensures that $\overline{I} \ge \underline{I}$ in Definition 3.15.
Since it governs a LDP, I must satisfy $\underline{I} \le I \le \overline{I}$. Moreover, any lower semi-continuous function I verifying $\underline{I} \le I \le \overline{I}$ is a rate function (see Theorem 3.16).
Now, either I is a Ruelle-Lanford function and the uniqueness is proved, or there is a strict inequality $\overline{I} > \underline{I}$ and there are several rate functions. ■

3.4.3 Contraction

Hereafter, we prove a somewhat restricted contraction principle, that needs (H1) and a further assumption.

Assumption (H2) *Let $\{E_n,\ n \ge 1\}$ be an exponentially tight sequence and $f:\ E \to F$ be a continuous mapping on the Hausdorff topological space F. It is assumed that for all compact sets $K \subset E$ and all n*

$$f^{-1}(K) \cap E_n \quad \textit{is compact.}$$

Then f is said to have compact level sets w.r.t. $\{E_n,\ n \ge 1\}$.

Theorem 3.18 (Contraction) *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E which satisfies a weak LDP with rate function I. Let $f: E \to F$ be a continuous mapping on the Hausdorff topological space F. If Assumptions (H1) and (H2) are satisfied, then $\mu_\varepsilon \circ f^{-1}$ verifies a weak LDP with rate function J,*

$$J(y) \stackrel{\text{def}}{=} I\big(f^{-1}(y)\big) = \inf_{x \in f^{-1}(y)} I(x). \tag{3.30}$$

Proof : There are three points (W1)–(W3) to verify (see Definition 3.13). The lower bound (W2) is immediate, the upper bound (W3) is an application of Theorem 3.14, there is only the lower semi-continuity left (W1).
Let $y_m \in F$ be a sequence converging to y. We shall denote by K the compact set $\{y_1, y_2, \ldots, y\}$. It is possible to assume

$$\liminf_{m \to \infty} J(y_m) = C < \infty,$$

otherwise the lower semi-continuity is obvious. Then extract a sequence such that $J(y_m)$ converges to C. By (H1) there exists $E_n \subset E$ such that (3.27) is satisfied with $n > C$. Let $\varepsilon > 0$, there exists $x_m \in f^{-1}(y_m)$ such that $I(x_m) \leq J(y_m) + \varepsilon$. For m large enough, $J(y_m)$ is near C so that x_m belongs to $f^{-1}(K) \cap E_n$ which is compact by (H2). Thus it is possible to extract a sub-sequence x_m converging to x. Now f is continuous, so that $f(x) = y$ and by the lower semi-continuity of I,

$$\lim_{m \to \infty} J(y_m) + \varepsilon \geq \liminf_{m \to \infty} I(x_m) \geq I(x) \geq J(y).$$

Since this is true for all $\varepsilon > 0$, the lower semi-continuity is proved. ■

This idea to use an exponentially tight sequence $\{E_n,\ n \geq 1\}$ contains the two cases where this theorem has been proved; first the full LDP, where $E_n = I^{-1}([0, n])$ is compact. In this case (H1) is the usual exponential tightness assumption; and second the weak contraction principle as in [97] where $f^{-1}(K_1)$ is required to be compact for compact K_1, in which case $E_n = E$. In both cases the Assumption (H2) is satisfied but this method allows further results such as the contraction of the empirical generator to the empirical measure in the Markov case.
The following properties will be useful in the Markov case, but their interest are not restrained to this framework.

Lemma 3.19 *Suppose Assumptions (H1) and (H2) hold as in Theorem 3.18. For all $y \in F$ such that $J(y) < \infty$, the set of infima y such that $J(y) = I(f^{-1}(y))$ is a non-empty and compact set.*

Proof : Let $n > J(y)$ be finite. By (H1), there exists E_n such that (3.27) holds. The infimum $J(y)$ is equal to $I\big(f^{-1}(y) \cap E_n\big)$ (see the proof of Theorem 3.14). By (H2), $f^{-1}(y) \cap E_n$ is a compact set. Since I is lower semi-continuous, the infimum is reached on this compact. Now the set of infima is exactly $I^{-1}(J(y)) \cap f^{-1}(y)$ which is closed, and the proof is completed. ■

Proposition 3.20 *Suppose Assumptions (H1) and (H2) hold as in Theorem 3.18. If I is a Ruelle-Lanford function for μ_ε then J is a Ruelle-Lanford function for $\mu_\varepsilon \circ f^{-1}$.*

Proof : Let $\varepsilon > 0$. For each $x \in E$, let $\mathcal{V}_x$ be a small open neighborhood of x, such that $I(\mathcal{V}_x) \geq I(x) - \varepsilon$. Let $y \in F$ such that $J(y) < \infty$; otherwise the lower bound is trivial. By Lemma 3.19, $K = f^{-1}(y) \cap I^{-1}(J(y))$ is a non-empty compact set. Therefore it is possible to find a finite covering $\bigcup_{i \leq n} \mathcal{V}_{x_i}$ of K. By assumption,

$$I\left(\bigcup_{i\leq n} \mathcal{V}_{x_i}\right) \geq \min_{i \leq n} I(x_i) - \varepsilon \geq I(K) - \varepsilon.$$

Let $\mathcal{V}'_r$ be a decreasing base of open neighborhoods of y. By continuity of f, when r tends to 0, $f^{-1}(\mathcal{V}_r)$ converges to K. Thus there exists $r > 0$ such that $f^{-1}(\mathcal{V}'_r)$ is included in $\bigcup_{i \leq n} \mathcal{V}_{x_i}$. Then, defining $\underline{J}$ and $\overline{J}$ by (3.28)–(3.29),

$$\underline{J}(y) \geq \lim_{r \to 0} I\big(f^{-1}(\mathcal{V}'_r)\big) \geq I\left(\bigcup_{i=1}^{n} \mathcal{V}_{x_i}\right) \geq I(x) - \varepsilon \geq J(y) - \varepsilon.$$

Since this is true for all $\varepsilon > 0$, $\underline{J}(y) \geq J(y)$. Similarly,

$$\begin{aligned} \overline{J}(y) &\leq \lim_{r \to 0} \sup_{K \subset \mathcal{V}'_r} I\big(f^{-1}(K)\big) \\ &\leq \lim_{r \to 0} I\big(f^{-1}(\mathcal{V}'_r)\big) \leq I(x) + \varepsilon \leq J(y) + \varepsilon, \end{aligned}$$

using the lower semi-continuity of I to bound $I\big(f^{-1}(\mathcal{V}'_r)\big)$ for r small enough. This implies $\overline{J}(y) \leq J(y)$. But $\overline{J} \geq \underline{J}$ since $\mu_\varepsilon \circ f^{-1}$ verifies a weak LDP, so that $J = \overline{J} = \underline{J}$ and the proof is completed. ■

Proposition 3.21 (exact l.s.c.) *Suppose Assumptions (H1) and (H2) hold as in Theorem 3.18 and additionally assume that $F = f(E)$. If $\mathcal{B} \subset E$ is a dense subset such that I is the least lower semi-continuous function coinciding with I on $\mathcal{B}$, then $f(\mathcal{B})$ is a dense subset of F such that J is the least lower semi-continuous function coinciding with J on $f(\mathcal{B})$.*

Proof : The set $f(\mathcal{B})$ is dense in $f(E)$ by the continuity of f.
Let $y \in F$ such that $J(y) < \infty$. By Lemma 3.19 there exists $x \in E$ such that $y = f(x)$ and $I(x) = J(y)$. By assumption there exists a sequence x_n converging to x such that $I(x_n)$ converges to $I(x)$. Then

$$\limsup_{n\to\infty} J(f(x_n)) \leq \lim_{n\to\infty} I(x_n) = I(x) = J(y),$$

hence $J(f(x_n))$ converges to $J(y)$ by the lower semi-continuity of J. Now f is continuous, thus $f(x_n)$ converges to $y = f(x)$.
If $J(y) = \infty$, by the lower semi-continuity of J, the limit of $J(y_n)$ is infinite for all y_n converging to y. The proof is completed. ∎

Proposition 3.22 (convexity) *Suppose Assumptions (H1) and (H2) hold as in Theorem 3.18. If f is affine and I is convex then J is convex.*

Proof : The proof is immediate by using the properties of I and f. ∎

3.4.4 Varadhan's integral lemma

The general result that has to be extended is

Theorem 3.23 *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E which satisfies a weak LDP with rate function I. Let $f : E \to \mathbb{R}$ be a continuous function. Then, for all open sets O and compact sets K in E,*

$$\liminf_{\varepsilon\to 0} \varepsilon \log \int_O e^{f(x)/\varepsilon} \mu_\varepsilon(\,\mathrm{d}x) \geq -\inf_{x\in O}\{I(x) - f(x)\}, \tag{3.31}$$

$$\limsup_{\varepsilon\to 0} \varepsilon \log \int_K e^{f(x)/\varepsilon} \mu_\varepsilon(\,\mathrm{d}x) \leq -\inf_{x\in K}\{I(x) - f(x)\}. \tag{3.32}$$

Proof : Since f is finite and continuous, f is locally bounded and the proof is straightforward. ∎

To extend this result to the usual Varadhan's integral lemma, one needs severe restriction.

Theorem 3.24 *Let $\mu_\varepsilon \subset \mathcal{M}_1(E)$ be a family of probability measures on a Hausdorff topological space E which satisfies a weak LDP with rate function I. Let $f : E \to \mathbb{R}$ be a continuous function. Suppose Assumptions (H1) and (H2) holds as in Theorem 3.18 and that $f^{-1}([0,n])$ is an exponentially tight sequence w.r.t. $\mu_\varepsilon^f = e^{f/\varepsilon}\mu_\varepsilon$, then*

$$\lim_{\varepsilon\to 0} \varepsilon \log \int e^{f(x)/\varepsilon} \mu_\varepsilon(\,\mathrm{d}x) = -\inf_{x\in E}\{I(x) - f(x)\}. \tag{3.33}$$

Proof : By assumption, for all $n \geq 1$, there exists $[a, b] \subset \mathbb{R}$ such that $K = E_n \cap f^{-1}([a, b])$ is compact and

$$\limsup_{\varepsilon \to 0} \varepsilon \log \int_{K^c} e^{f(x)/\varepsilon} \mu_\varepsilon(\,\mathrm{d}x) \leq -n.$$

Applying (3.32), this means that

$$\limsup_{\varepsilon \to 0} \varepsilon \log \int e^{f(x)/\varepsilon} \mu_\varepsilon(\,\mathrm{d}x) \leq -\min\Big\{ \inf_{x \in K} \{I(x) - f(x)\}, n \Big\}. \tag{3.34}$$

If I is uniformly infinite, (3.34) means that the limit (3.33) is verified and infinite. Otherwise, let $n > \inf_{x \in E}\{I(x) - f(x)\}$. The lower bound (3.31) proves that the infimum is reached on K, so that (3.34) implies the upper bound

$$\limsup_{\varepsilon \to 0} \varepsilon \log \int e^{f(x)/\varepsilon} \mu_\varepsilon(\,\mathrm{d}x) \leq -\inf_{x \in E}\{I(x) - f(x)\}.$$

Combining with (3.31) for $O = E$, it follows that (3.33) holds. ■

3.5 One-dimensional Sanov's theorem

Note that all operations of Section 3.4 can be performed on the LDP for empirical generators G_t, for example the analog of Varadhan's integral lemma Theorem 3.23 or Theorem 3.24. But we shall further only explore the contraction so as to obtain the one-dimensional Sanov's theorem.

3.5.1 Contraction

Instantiating Theorem 3.18 for $f(G) = \pi$, one can checks in this case that the compactness assumption ($f^{-1}(K)$ compact for K compact) is *not* generally satisfied, so that the LDP for the one-dimensional empirical measure can not be directly derived. In the remark of page 154, we have shown that the real feature is to ensure the finiteness of the speed of the jumps. However, when the only measured parameter of the process is the empirical measure L_t, it is impossible to guarantee a finite number of jumps in a finite interval of time. There must therefore be a guarantee that R corresponds to a non-explosive process. Unfortunately it is not easy to give conditions for non-explosion and the lack of structure on E prevents us to treat easily systems with unbounded intensities, like the $M/M/\infty$ queue. A good way for ensuring both the uniqueness of the process and the contraction principle is to bound the intensities. Otherwise the sketch of the proof has to be adapted case by case.

Theorem 3.25 *Let Y_t be an irreducible Markov process with generator R and assume that the intensities R_i are bounded by $\bar{r}$. Then*

$$I_0(\pi\|R) \stackrel{\text{def}}{=} \inf_{G=(\pi,A)} H_c(G\|R), \tag{3.35}$$

is the Ruelle-Lanford function for the empirical measures L_t (defined by equation (3.1)). Therefore L_t satisfies a weak LDP with the unique, convex rate function I_0 that is the smallest lower semi-continuous function equal to $I_0(\pi\|R)$ on the finite support irreducible[4] *measures $\pi \in M_1(E)$.*

Proof : Hereafter f refers to the continuous projection $f(G) = \pi$. The heart of the proof is that $\theta(G) = l = \sum a_{xy}$ defines an exponentially tight sequence $E_n \stackrel{\text{def}}{=} \theta^{-1}([0, l_n])$ such that (H1) and (H2) are valid. (H1) has already been proved in Theorem 3.12 under the assumption of bounded intensities. The main difficulty is to prove (H2): if $K \subset \mathcal{M}_1(E)$ is compact and $l_n \in \mathbb{R}_+$, both $f^{-1}(K)$ and E_n are closed, but none is compact.
Since $K \subset \mathcal{M}_1(E)$ is compact, for all $\eta > 0$ there exists a compact $K_1 \subset E$ such that

$$\pi(K_1) = \sum_{x\in K_1} \pi_x \leq \eta, \qquad \forall \pi \in K.$$

Let $G = (\pi, A)$ be an element of $f^{-1}(K) \cap E_n$ and denote $a(F) = \sum_{x\in F} a_x$. Then

$$\begin{aligned} H_c(G\|R) &\geq \sum_{x\notin K_1} \pi_x I_p\Big(\frac{a_x}{\pi_x}\Big\| R_x\Big) \\ &\geq \pi(K_1^c) I_p\Big(\frac{a(K_1^c)}{\pi(K_1^c)}\Big\| \sum_{x\in K_1^c} \pi_x R_x\Big) \\ &= I_p\Big(a(K_1^c)\Big\| \pi(K_1^c) \sum_{x\in K_1^c} \pi_x R_x\Big). \end{aligned} \tag{3.36}$$

The second line is a consequence of the convexity of $I_p(.\|.)$ with respect to both variables and the last of its homogeneity. Let $\varepsilon > 0$. Fix $\eta \leq \varepsilon$ and the corresponding K_1 such that

$$I_p(\varepsilon\|\eta^2\bar{r}) > l_n \text{ and } \varepsilon > \eta^2\bar{r}.$$

The term $\pi(K_1^c)\sum_{x\in K_1^c} \pi_x R_x$ in (3.36) is less than $\eta^2\bar{r}$, so that, since $I_p(x\|y)$ is increasing w.r.t. x and decreasing w.r.t. t when $x > y$,

$$H_c(G\|R) \geq I_p(\varepsilon\|\eta^2\bar{r}) > l_n, \quad \forall \pi \in K,$$

[4] A measure $\pi \in \mathcal{M}_1(E)$ is said *irreducible* if its support is connected in the support (graph) of G.

for all G such that $a(K_1^c) > \varepsilon$. This means that, for all $\varepsilon > 0$, there exists a compact set $K_1 \subset E$ such that,

$$G \in f^{-1}(K) \cap E_n \implies \begin{cases} a(K_1^c) \leq \varepsilon, \\ \pi(K_1^c) \leq \varepsilon, \end{cases}$$

i.e. $f^{-1}(K) \cap E_n$ is tight. Since it is closed, it is compact. (H2) is proved.

Here comes the arsenal developed in Section 3.4. By Proposition 3.20, $I_0(.\|R)$ is a Ruelle-Lanford function for $L_t = f(G_t)$. By Theorem 3.16 and Theorem 3.17, L_t verifies a weak LDP with unique rate function $I_0(.\|R)$. Since f is linear and $H_c(.\|R)$ is convex, by Proposition 3.22, $I_0(.\|R)$ is also convex. By Proposition 3.9 and Proposition 3.21, the finite support irreducible measures are a dense subset of $\mathcal{M}_1(E)$ and $I_0(.\|R)$ is the least lower semi-continuous function equal to $I_0(\pi\|R)$ on the finite support irreducible measures $\pi \in \mathcal{M}_1(E)$.
Note that, by Lemma 3.19, $I_0(\pi\|R) = H_c(G\|R)$ for a non-empty compact set of generators G. ■

3.5.2 Identification of the 1-dimensional rate function

Exactly as in the discrete case (see [34]) there are several ways of representing the rate function in Sanov's theorem for the one-dimensional empirical measures

$$L_t \stackrel{\text{def}}{=} \frac{1}{t} \int_0^t \delta_{Y_u} \,\mathrm{d}u.$$

Actually, the formulations depend on the spaces of $E \mapsto \mathbb{R}$ functions usually considered: $C_b(E)$ (bounded continuous functions) and $U(E)$, the space of continuous bounded functions h with $\inf h > 0$. Note that in discrete spaces, all functions are continuous. The other expressions of the rate function are given in terms of transforms of the quantity

$$\Lambda(f) \stackrel{\text{def}}{=} \limsup_{t\to\infty} \frac{1}{t} \log \mathbb{E}\left[\exp \int_0^t f(Y_u)\,\mathrm{d}u\right],$$

and of the kernel K^f satisfying

$$K_t^f = P_t + \int_0^t P_{t-s}(fK_s^f)\,\mathrm{d}s,$$

i.e. formally the kernel with generator $L^f = L+f$. The Feynman-Kac formula

$$K_t^f\phi(x) = \mathbb{E}_x\left[\phi(Y_t)\exp\left\{\int_0^t f(Y_u)\,\mathrm{d}u\right\}\right]$$

connects both quantities.
An almost exhaustive list of the main rate functions is given below.

$$\begin{cases} I_0(\pi) = \inf\limits_{G=(\pi,A)} H_c(G\|R), \\ I_1(\pi) = \sup\limits_{h\in U(E)} -\sum\limits_{x\in E} \pi_x \dfrac{Rh(x)}{h(x)}, \\ I_2(\pi) = \sup\limits_{f\in C_b(E)} \left[\langle f,\pi\rangle - \Lambda(f)\right], \\ I_3(\pi) = \sup\limits_{f\in C_b(E)} \left[\langle f,\pi\rangle + \log \mathcal{R}(K_1^f)\right], \\ I_4(\pi) = \lim\limits_{\varepsilon\to 0} \dfrac{1}{\varepsilon} I_d(\pi\|P^\varepsilon), \end{cases}$$

where $\mathcal{R}$ denotes the convergence parameter ($-\log \mathcal{R}$ is sometimes called logarithmic spectral radius), $P^\varepsilon_{xy} = \mathbb{P}_x\left[Y_\varepsilon = y\right]$ the transition matrix at time ε for the process Y_t and I_d the rate function in the discrete time.
Inequalities under very general conditions between the four last expressions are proved in [41]. Actually there are equalities when the process is Feller continuous. However, these expressions are known to be rate functions only for restricted classes of Markov processes. Apart from the classical uniform condition, there is not as much work as in the discrete time for which the identification can be deduced from [43] (see also [27, 28, 42, 112, 115]).

Remark : The problem is to compare these expressions to I_0. It is not possible to use the uniqueness of I_0 as a rate function because the rate functions I_1–I_4 are not proved to be rate functions in a general enough framework. It is hereafter assumed that the intensities are bounded which is the condition of Theorem 3.25 but this restriction can be somewhat relaxed.

Let π be a finite support irreducible measure. Optimizing I_0 and I_1 yields

$$\sum_{y\neq x} \pi_y \frac{R_{yx}h(x)}{h(y)} = \sum_{y\neq x} \pi_x \frac{R_{xy}h(y)}{h(x)}, \tag{3.37}$$

$$\sum_{y\neq x} \pi_y L_{yx} = \sum_{y\neq x} \pi_x L_{xy}, \tag{3.38}$$

which are equivalent since $L_{xy} = R_{xy}h(y)/h(x)$ through Lagrange's multipliers; in fact the multipliers are $\log h(x)$ and correspond to the linear constraints (3.38) for L. Since H_c is strictly convex, since π is irreducible and since the constraints (3.38) are linear (so that the set where H_c is to be minimized is convex), there is a unique solution

L_{xy}, from which a unique solution for h is derived, up to a constant. Finally it is easily checked that $I_0 = I_1$, and since I_1–I_4 are equal, all rate functions are equal for the finite support irreducible measures.

Note that the definition of I_4 also works for the generators. Formally, if $P^\varepsilon = I + \varepsilon R + o(\varepsilon)$ and $A^\varepsilon = I + \varepsilon G + o(\varepsilon)$ for $G = (\pi, A)$ then π is invariant for all A^ε and

$$\begin{aligned}
H_d(A^\varepsilon \| P^\varepsilon) &= \sum_{x,y \in E} \pi_x A^\varepsilon_{xy} \log \frac{A^\varepsilon_{xy}}{P^\varepsilon_{xy}} \\
&= \sum_{x \in E} \pi_x \left(\sum_{y \neq x} \varepsilon L_{xy} \log \frac{L_{xy}}{R_{xy}} + \log \frac{1 - L_x}{1 - R_x} \right) + o(\varepsilon) \\
&= \varepsilon H_c(G \| R) + o(\varepsilon).
\end{aligned}$$

This calculation is valid for irreducible finite support generators. Indeed, for such a generator there exists a stochastic semi-group of transition corresponding to the generator, and A^ε can be defined similarly to P^ε. Eventually

$$H_c(G \| R) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} H_d(A^\varepsilon \| P^\varepsilon).$$

It is now necessary to prove that I_4 is not too large for the other measures. Let π be a measure and choose a generator G that has the same entropy (by Lemma 3.19). By Proposition 3.9 choose irreducible finite support $G(n)$ converging to G such that their entropy converges to $H_c(G\|R)$. $\pi(n)$ denotes the invariant measure of $G(n)$. As above, for such an irreducible finite support generator $G(n)$ it is possible to define $A^\varepsilon(n)$. This ensures that $\pi(n)$ is the invariant probability of $A^\varepsilon(n)$. Since $I_d(\pi\|P) = \inf_{A:\ A(.,E)=\pi} H_d(A\|P)$,

$$I_d(\pi(n) \| P^\varepsilon) \leq H_d(A^\varepsilon(n) \| P^\varepsilon) \leq \varepsilon H_c(G(n) \| R) + o(\varepsilon).$$

$H_c(G(n)\|R)$ converges to $H_c(G\|R)$ and $\pi(n)$ to π, so that

$$I_4(\pi) \leq \limsup_{n \to \infty} I_4(\pi(n)) \leq H_c(G \| R) = I_0(\pi \| R),$$

by the lower semi-continuity of I_4. This proves $I_4 \leq I_0$. Now, by the continuity property in Theorem 3.25, I_0 is known to be the least lower semi-continuous function such that $I = I_0$ for all finite support irreducible measures. But on these measures, $I_0 = I_1 = I_4$ so that finally $I_4 = I_0$ for all measures.

♦♦♦

Chapitre 4

Large deviations for polling systems

Franck Delcoigne & Arnaud de La Fortelle
March 2000

Abstract

In this paper, we prove a sample path large deviation principle for a rescaled process $n^{-1}Q_{nt}$, where Q_t represents the joint number of clients at time t in a polling system with N nodes, one server and Markovian routing. The goal of the paper is to identify the rate function. We introduce a so called empirical generator consisting of Q_t and of two empirical measures associated with S_t, the position of the server at time t. One of the main step is to derive large deviations bounds for a localized version of the empirical generator. The analysis relies on a suitable change of measure and on a representation of fluid limits for polling systems. Finally, the rate function is solution of a meaningful convex program.

keywords Large deviations, polling system, fluid limits, empirical generator, change of measure, contraction principle, entropy, convex program.

4.1 Introduction

Consider a polling system consisting of N nodes attended by a single server and denote by $\mathcal{S} \stackrel{\text{def}}{=} \{1, \ldots, N\}$ the set of nodes. At node i, arrivals of clients form a Poisson process with rate λ_i. Each customer at node i requires service, whose duration is exponentially distributed with parameter μ_i. Let $\rho_i \stackrel{\text{def}}{=} \lambda_i/\mu_i$, the intensity factor at node i. When the server arrives at a busy node, say i, it serves one customer and then moves to some node, chosen via some ergodic routing matrix $\mathbf{P} = (P_{ij})_{i,j\in\mathcal{S}}$ with invariant measure $\eta = (\eta_i)_{i\in\mathcal{S}}$. If it reaches an empty node, then it immediately switches to some other node, still chosen according to $\mathbf{P}$. The switch-over time to go from node i to node j, for $i,j \in \mathcal{S}$, is exponentially distributed with mean τ_{ij}.

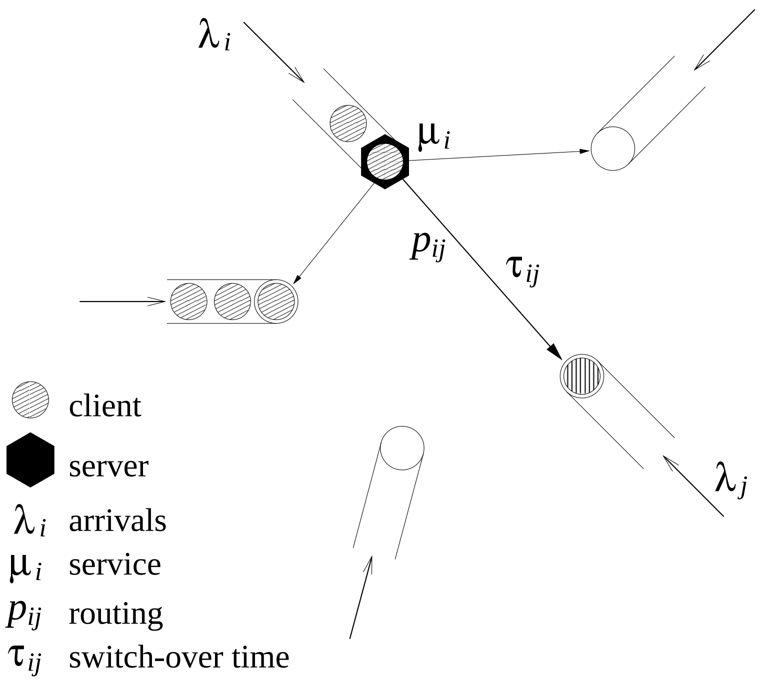

Figure 4.1: A polling system.

All stochastic input sequences (inter arrival times, services, switch-over times) are supposed to be mutually independent. When the join number of clients and the position of the server at time 0 are respectively given by $x = (x_1, \dots, x_N)$ and s, $Q(t,x,s) = \big(q_1(t,x,s), \dots, q_N(t,x,s)\big)$ and $S(t,x,s)$ represent the join number of clients at each node and the position of the server at time t. As a rule, we shall write $S(t,x,s) = i$ if the server is serving some customer at node i and $S(t,x,s) = ij$ if it is in transit between nodes i and j. Set $\mathcal{S}_0 \stackrel{\text{def}}{=} \mathcal{S} \cup \mathcal{S}^2$, the state space of the server. Then

$$X_{x,s} \stackrel{\text{def}}{=} \Big\{ \big(Q(t,x,s), S(t,x,s)\big),\ t \geq 0 \Big\}$$

is a Markov process with generator R such that

$$Rf(x,s) = \sum_{(y,s')\in\mathbb{Z}_+^N\times\mathcal{S}_0} q(x,s;y,s')\Big(f(y,s') - f(x,s)\Big), \qquad \forall (x,s)\in\mathbb{Z}_+^N\times\mathcal{S}_0,$$

where $f \in \mathcal{B}\big(\mathbb{Z}_+^N \times \mathcal{S}_0\big)$ and

$$q(x,s;y,s') = \begin{cases} \lambda_i, & \text{if } y = x + e_i, s' = s, \ \forall i \in \mathcal{S}, \\ \mu_i P_{ij}, & \text{if } x_i > 0, s = i, y = x - e_i, s' = ij, \ \forall i,j \in \mathcal{S}, \\ \dfrac{1}{\tau_{ji}}, & \text{if } x_i > 0, s = ji, y = x, s' = i, \ \forall i,j \in \mathcal{S}, \\ \dfrac{1}{\tau_{ji}} P_{il}, & \text{if } x_i = 0, s = ji, y = x, s' = il, \ \forall i,j,l \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases}$$

Whenever no confusion arises, the initial state (x,s) will be dropped. Let us recall that, if $\{x_n,\ n \geq 0\}$ is a sequence of $\mathbb{Z}_+^N$ such that $|x_n| \to \infty$, then every limiting point in distribution of the sequence of processes $|x_n|^{-1}X_{x_n,s}$ will be called a fluid limit. A precise definition of fluid limits for polling systems are given in Section 4.2.2.
Let us introduce now a definition and a notation which will be of constant use in the sequel:

Definition *For every $x = (x_1, \dots, x_N) \in \mathbb{R}_+^N$, denote by $\Lambda(x)$ the set of indices i such that $x_i > 0$. If Λ is a subset of $\mathcal{S}$, it is called face Λ:*

$$\{x \in \mathbb{R}_+^N \ : \quad x_i > 0, \ \forall i \in \Lambda, \ x_i = 0, \ \forall i \in \Lambda^c\}.$$

- A^c and $\mathbb{1}_{\{A\}}$ will denote respectively the complementary and the indicator function of a set A;
- $\mathcal{B}(E)$, $\mathcal{M}_+(E)$ and $\mathcal{M}_1(E)$, respectively, the sets of bounded functions, of positive measures and of probability measures on a space E;
- $D\left([0,T], \mathbb{R}^N\right)$, the space of right continuous functions $f : [0,T] \to \mathbb{R}^N$ with left limits, endowed with the Skorokhod metric denoted by d_d;
- $\mathcal{C}\left([0,T], \mathbb{R}^N\right)$, the space of continuous functions equipped with the metric of the uniform convergence denoted by d_c;

- $\mathcal{PL}\left([0,T],\mathbb{R}^N\right)$, $\mathcal{AC}\big([0,T],\mathbb{R}^N\big)$ and $\mathcal{L}^p\big([0,T],\mathbb{R}^N\big)$, respectively, the spaces of piecewise linear functions whose derivative has only finitely many discontinuities, of absolutely continuous functions and of functions f such that $|f|^p$ is integrable with respect to the Lebesgue measure.

A huge literature has been devoted to the study of polling systems because of their wide range of applicability. In [19, 52, 68], the necessary and sufficient conditions of ergodicity has been established for systems with one or several servers under a rich variety of service policies. However, the problem of determining the invariant measure for such systems is still open. Even, for limited policies, the mean waiting time can be computed only under symmetry assumptions [21]. The reader is referred to [130] for an overview about polling systems. In the present paper, a sample path large deviation principle or a sample path LDP for the rescaled process $n^{-1}Q_{nt}$ is established. This could be a preliminary step in order to obtain large deviations estimates for the stationary distribution. In view of future applications, some particular attention is devoted to the computation of the rate function governing the sample path LDP. All this program falls into the framework of LDP for Markov processes with discontinuous statistics i.e those for which the coefficients of their generator are not spatially continuous.

It seems that one of the first paper dealing with such processes is [59], where large deviations problems for Jackson networks were investigated using partial differential equations techniques. Quite recently, the LDP for a large class of Markov processes with discontinuous statistics has been proved in [56]. Roughly speaking, the authors of [56] express the logarithm of large deviation probabilities as the minimal cost of some stochastic optimal-control problem, and the limit of the optimal cost is shown to exist by means of a sub-additivity argument. However, the rate function is not explicit. Note that in [58], an explicit upper bound of large deviations involving Legendre transforms is proved. While for Jackson networks and some processor sharing models this bound is tight [4], in general the problem of the lower bound remains open. It is worth emphasizing that in our case, Q_t is not a Markov process so that the polling model does not satisfy the assumptions of [56]. Besides, the rate function can be explicitely described in terms of entropy functions.

Until now, the identification of the rate function has been carried out in some particular cases and usually for low dimensional systems. In [44], using the contraction principle, the exponential decay of the stationary distribution of the waiting time is computed for a two dimensional tandem networks taking advantage that it can be expressed simply as a continuous function of the input processes. It should be noted that in this setting, a sample path LDP for processes with independent increments over infinite intervals of time is needed [45]. General results were obtained in [54, 83] where the LDP has been established for random walks whose generator has a discontinuity along an hyperplane. These results are applied in [83] to compute the exponential decay of the stationary distribution of ergodic random walks in $\mathbb{Z}_+^2$. Nevertheless, in such examples, there are at most two boundaries with codimension one or two where discontinuity arise. Ultimately, the identification of the rate function governing the LDP for Jackson networks has been carried out in [4, 82].

Following [56], in order to get a sample path LDP for polling systems, the main step is to prove the forthcoming large deviations local bounds:

Theorem 4.1 (Local bounds) *Take $x \in \mathbb{R}_+^N$ and $D \in \mathbb{R}^N$ such that $D_i = 0$, $\forall i \in \Lambda(x)^c$. Then, for any τ satisfying $x_i + D_i\tau > 0$, $\forall i \in \Lambda(x)$,*

$$\begin{aligned} &\lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n}\inf_{|y-nx|<\epsilon n}\log \mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q(t,y)-nx-Dt|<\delta n\right] \quad (4.1)\\ = &\lim_{\delta\to 0}\lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\sup_{|y-nx|<\epsilon n}\log \mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q(t,y)-nx-Dt|<\delta n\right]\\ \stackrel{\text{def}}{=} &-\tau L(\Lambda(x), D). \end{aligned}$$

Note that for $D_i = 0$, $\forall i \in \Lambda(x)^c$, the conditions $x_i + D_i\tau > 0$, $\forall i \in \Lambda(x)$ is equivalent to $x + Dt$ lies in the face $\Lambda(x)$, for all $t \in [0, \tau]$.

In establishing Theorem 4.1, one must know in some sense how the different transition rates have to be modified in order that Q_t follows a given drift D. This means that rather than studying Q_t itself, we focus on the so called empirical generator

$$G_t \stackrel{\text{def}}{=} \left(\frac{N_t}{t}\hat{L}_{N_t}, L_t, \frac{Q_t - Q_0}{t}\right),$$

where

- N_t is the number of jumps of the server till t;
- $\{S_n,\ n \geq 0\}$ is the embedded process of the server just before it jumps. Note that it is not a Markov chain;
- $\hat{L}_n \stackrel{\text{def}}{=} \dfrac{1}{n}\sum_{i=0}^{n}\delta_{S_i} \in \mathcal{M}_1(\mathcal{S}_0)$ is the empirical measure of the process S_n;

- $L_t \stackrel{\text{def}}{=} \dfrac{1}{t}\displaystyle\int_0^t \delta_{S_u} du \in \mathcal{M}_1(\mathcal{S}_0)$ is the empirical measure of the process S_t.

The set of empirical generators is denoted by $\mathcal{G}$ (see Definition 4.6). On the long run, exactly like the pair empirical measure for Markov chains, the empirical generator tends to be *balanced*: this is (4.2), which means that the server exits a node i as often as it goes toward i. The set of balanced generators will be denoted by $\mathcal{G}_s$ (see Definition 4.9). When the transitions are restrained to lie in the support of $\mathbf{P}$ (i.e. $a_{ij} = 0$ if $P_{ij} = 0$), the set is denoted by $\mathcal{G}_s(\mathbf{P})$.

Definition *The set $\mathcal{G}_s^\Lambda$ of balanced generators is defined by the triples $(A, \pi, D) \in \mathcal{M}_+(\mathcal{S}_0) \times \mathcal{M}_1(\mathcal{S}_0) \times \mathbb{R}^\Lambda$ that verify*

$$\begin{aligned} \overline{a}_i \stackrel{\text{def}}{=} \sum_{j\in\mathcal{S}} a_{ji} &= \sum_{j\in\mathcal{S}} a_{ij}, && \forall i \in \mathcal{S}, \\ a_i &\le \overline{a}_i, && \forall i \in \Lambda^c, \\ a_i &= \overline{a}_i, && \forall i \in \Lambda, \\ a_i + D_i &\ge 0, && \forall i \in \Lambda. \end{aligned} \tag{4.2}$$

We denote by $\mathbb{R}^\Lambda$, for the sake of simplicity, the subspace of $D \in \mathbb{R}^N$ with $D_i = 0$ for $i \in \Lambda^c$.

The elements of a generator (A, π, D) can be seen as mean numbers per unit time of

a_i services at i ($\bar{a}_i$ passages through i);

a_{ij} passages from i to j; we shall define the matrix $\bar{A} \stackrel{\text{def}}{=} (a_{ij})$: the notation is consistent since $\bar{a}_i = \sum_j a_{ij}$ (see equation 4.2);

π_i time spent in service at i (π_{ij} time spent from i to j);

D_i length increase of queue i.

Note that, in general, for each generator (A, π, D) corresponds a polling system, described by its intensities $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})$. The correspondence is given by equations (4.24)–(4.27). Additionally, a usefull property of $\mathcal{G}_s^\Lambda$ is to be convex.
In Theorem 4.19, we prove large deviations bounds with a rate function $H(.\|R)$ defined in Definition 4.12, for a uniform version of G_t^Λ, where G_t^Λ is the empirical generator associated to a *localized* polling system X^Λ. The

transition mechanism describing the evolution of X^Λ is identical to X's except that the components indexed by Λ for X^Λ can be negative. In order to prove this result, we use a change of measure (chosen in a restricted class) which gives rise to a new polling system, for which the fluid limits can be completely characterized (see Theorems 4.4 and 4.5), and this is a key ingredient in the proof of the lower bound. This ensures that for D as in Theorem 4.1, after a suitable change of measure, X lies in any neighborhood of D. This way of reasoning is quite classical and was used among others in [82] for identifying the rate function governing the LDP for Jackson networks. However, owing to the presence of S_t, our approach has much more in common with the method used in [32] to prove the weak Sanov LDP for jump Markov processes in continuous time. Moreover, the function $H(.\|R)$ governing the large deviations bounds for G_t^Λ appears as an entropy and is then easily seen to possess good properties.

Definition *Let $R = (\lambda_i, \mu_i, \tau_{ij}, P_{ij})$ denotes the generator of the polling system, $G = (A, \pi, D) \in \mathcal{G}_s^\Lambda$ be a generator and $(\tilde\lambda_i, \tilde\mu_i, \tilde\tau_{ij}, \tilde P_{ij})$ its representation as a polling system. Using the notation $\bar A$ for the balanced matrix (a_{ij}), the relative entropy of G with respect to R is*

$$H(G\|R) \stackrel{\text{def}}{=} \sum_{i\in\mathcal{S}} \Big(I_p(\tilde\lambda_i\|\lambda_i) + \pi_i I_p(\tilde\mu_i\|\mu_i)\Big) + \sum_{i,j\in\mathcal{S}} \pi_{ij} I_p(\tilde\tau_{ij}^{-1}\|\tau_{ij}^{-1}) + H_d(\bar A\|P),$$

where $H_d(\bar A\|P) \stackrel{\text{def}}{=} \sum_{i,j\in\mathcal{S}} a_{ij}\log\left(\dfrac{a_{ij}}{\bar a_i P_{ij}}\right)$ *and* $I_p(\nu\|\lambda) \stackrel{\text{def}}{=} \nu\log\dfrac{\nu}{\lambda} - \nu + \lambda.$

This entropy has an easy interpretation in terms of information theory (see Section 4.3.3). Moreover it is positive, finite, continuous and strictly convex on $\mathcal{G}_s^\Lambda(\mathbf{P})$; it is infinite otherwise; it is null if, and only if, $G = R$ and it has compact level sets. Finally the local rate function $L(\Lambda, D)$ is better expressed as

Definition *The rate function $L(\Lambda, D)$ is defined by*

$$L(\Lambda, D) \stackrel{\text{def}}{=} \inf_{G\in f_\Lambda^{-1}(D)} H(G\|R), \quad \forall D \in \mathbb{R}^\Lambda, \tag{4.3}$$

where $f_\Lambda : \mathcal{G}_s^\Lambda \mapsto \mathbb{R}^\Lambda$ is the projection $f_\Lambda(G) = D$. For all $x \in \mathbb{R}_+^N$, $L(x, D)$ is defined to be $L(\Lambda(x), D)$.

At this stage, as an application of a kind of contraction principle, one can get Theorem 4.1. Since $L(\Lambda, D)$ is the rate function derived from $H(.\|R)$ by a contraction principle, many properties can be derived without much effort and it turns out that $L(\Lambda, D)$ is solution of a convex program: (4.3). In general, for Markov processes in $\mathbb{Z}_+^N$, the fluid limits cannot be characterized, even in the case of maximal spatial homogeneity. So, it seems that the sole traditional change of measure would be ineffective for the identification of the rate function in this general setting.

Now, the rate function $I_T(.)$ for the sample path LDP is expressed as

$$I_T(\varphi) = \begin{cases} \displaystyle\int_0^T L\left(\varphi(t), \dot{\varphi}(t)\right) dt, & \text{if } \varphi \text{ is absolutely continuous,} \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.4}$$

Remark 1 : $I_T(.)$ is defined by all the values $L(x, D)$, $x \in \mathbb{R}_+^N$ and $D \in \mathbb{R}^N$ such that $D_i = 0$, $\forall i \in \Lambda(x)^c$. Indeed, assume that for some t, $\varphi_i(t) = 0$ and $\dot{\varphi}_i(t)$ exists. Since $\varphi_i(t) \le \varphi_i(s)$ for all s, this implies $\dot{\varphi}_i(t) \le 0$. Then, necessarily $\dot{\varphi}_i(t) = 0$. Moreover, φ being absolutely continuous, $\dot{\varphi}_i(t)$ exists for almost all t.

♦♦♦

Remark 2 : In Theorem 4.2, it is proved that the rate function $I_T(.)$ has the expected form given in (4.4). It is worth noting that in [56], the rate function governing the LDP was defined as the lower semi-continuous regularization of a functional defined on the set of piece-wise linear functions. Nevertheless, it was conjectured that the rate function takes a form like in (4.4).

♦♦♦

Set

$$Q^n_{x,s} \stackrel{\text{def}}{=} \left\{ \frac{1}{n} Q(nt, [nx], s),\ t \ge 0 \right\},$$

and

$$\Phi_x(K) = \{\varphi \in D([0,T], \mathbb{R}_+^N) : I_T(\varphi) \le K,\ \varphi(0) = x\}.$$

Using the Markov property and the local bounds, one deduces easily the large deviations bounds for the probability that the process evolves around some piecewise linear function, which are expressed using $I_T(.)$ (see Proposition 4.23). Then using various properties of the rate function $I_T(.)$ and the exponential tightness of $\{Q^n_{x,s},\ n \ge 1\}$, one can get the main result of the paper: the sample path LDP.

Theorem 4.2 (Sample path LDP) *The sequence $\{Q^n_{x,s},\ n \ge 1\}$ satisfies a LDP in $D\big([0,T], \mathbb{R}_+^N\big)$ with good rate function $I_T(.)$: for every $T > 0$, $x \in \mathbb{R}_+^N$ and s,*

1. *If $C \subset \mathbb{R}_+^N$ is compact, so is $\bigcup_{x \in C} \Phi_x(K) \subset \mathcal{C}([0,T], \mathbb{R}_+^N)$.*

2. *for each closed set F of $D([0,T],\mathbb{R}_+^N)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[Q^n_{x,s} \in F\right] \leq -\inf\{I_T(\phi),\ \phi \in F, \phi(0) = x\};$$

3. *for each open set O of $D([0,T],\mathbb{R}_+^N)$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[Q^n_{x,s} \in O\right] \geq -\inf\{I_T(\phi),\ \phi \in O, \phi(0) = x\}.$$

The organization of the paper is the following one. In Section 4.2, the notion of localized polling system is discussed as well as fluid limits; the local bounds (4.1) are then restated in a more convenient way using bounds for such systems. In Section 4.3, the empirical generator of localized polling system is introduced as well as the entropy $H(.\|R)$ and the local rate function $L(x, D)$. The connections between empirical generator and fluid limits are discussed together with the properties of $H(.\|R)$ and $L(x, D)$. In Section 4.4, we derive large deviations bounds for the empirical generator of localized polling system. Then, using some kind of contraction principle, the local bounds (4.1) are proved. In Section 4.5, the sample path LDP (Theorem 4.2) is proved.

4.2 Localized polling systems and fluid limits

4.2.1 Localized polling systems

The choice of x, D and τ in Theorem 4.1 ensures that the bounds (4.1) will not depend on the transition mechanism when some components indexed by $\Lambda(x)$ are null. This allows one to restate (4.1) in terms of bounds for a local model without boundary conditions on the nodes $i \in \Lambda(x)$. More precisely, for each subset Λ of $\mathcal{S}$, let us define the process

$$X^\Lambda_{x,s} \stackrel{\text{def}}{=} \left\{\left(Q^\Lambda(t,x,s), S^\Lambda(t,x,s)\right),\ t \geq 0\right\}$$

where $Q^\Lambda(t,x,s) = \left(q_1^\Lambda(t,x,s), \ldots, q_N^\Lambda(t,x,s)\right)$ and $S^\Lambda(t,x,s)$ represent the joint number of clients at each node, the position of the server at time t and where (x,s) stands for the initial state. The transition mechanism defining the evolution of X^Λ is identical to X's except that there is no boundary condition on the nodes $i \in \Lambda$: the components $q_i^\Lambda(t,x,s)$ for $i \in \Lambda$ may be negative. X^Λ is then called a localized polling system. All the notation defined for polling systems in Section 4.1 are adapted in a straightforward way for localized polling systems. Note that if $Q^{\Lambda}_{\Lambda^c}(t,x,s)$ denotes $\{q_i^{\Lambda}(t,x,s),\ i \in \Lambda^c\}$, then

$$\left\{\left(Q^{\Lambda}_{\Lambda^c}(t,x,s), S^{\Lambda}(t,x,s)\right),\ t \geq 0\right\}$$

is a Markov process.

Definition *By an abuse of notation, one says that $X^{\Lambda}_{x,s}$ is ergodic if*

$$\left\{\left(Q^{\Lambda}_{\Lambda^c}(t,x,s), S^{\Lambda}(t,x,s)\right),\ t \geq 0\right\}$$

is ergodic.

Take τ satisfying $x_i + D_i\tau > 0$, $\forall i \in \Lambda(x)$. Then for ϵ sufficiently small and all y with $|y - nx| < \epsilon$

$$\begin{aligned} & \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] \\ = \ & \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q^{\Lambda(x)}(t,y) - nx - Dt| < \delta n\right]. \end{aligned}$$

Since $x_i = 0$, $\forall i \in \Lambda(x)^c$, the distribution of $X^{\Lambda}_{y,s}$ is invariant with respect to the shift nx. Then, one can restate the local bounds (4.1) in term of local bounds for $X^{\Lambda(x)}_{0,s}$.

Proposition 4.3 *Let $x \in \mathbb{R}^N_+$ and $D \in \mathbb{R}^N$ such that $D_i = 0$, $\forall i \in \Lambda(x)^c$. Then, for τ satisfying $x_i + D_i\tau > 0, \forall i \in \Lambda(x)$,*

$$\begin{aligned} \lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty} \frac{1}{n\tau} \inf_{|y-nx|<\epsilon} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] &= \\ \lim_{\delta\to 0}\liminf_{n\to\infty} \frac{1}{n\tau} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q^{\Lambda(x)}(t,0) - Dt| < \delta n\right], & \quad (4.5) \end{aligned}$$

and

$$\begin{aligned} \lim_{\delta\to 0}\lim_{\epsilon\to 0}\limsup_{n\to\infty} \frac{1}{n\tau} \sup_{|y-nx|<\epsilon} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] &= \\ \lim_{\delta\to 0}\limsup_{n\to\infty} \frac{1}{n\tau} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q^{\Lambda(x)}(t,0) - Dt| < \delta n\right], & \quad (4.6) \end{aligned}$$

the limits being independent of τ.

Proof : Although lengthy, this step is rather classical and mimics mainly the proof of [82, Proposition 4.1]. It is deferred to Appendix 4.A. ■

4.2.2 Fluid limits for localized polling systems

In this section, we recall some results of [19, 52, 68] about the convergence of the sequence $\{Q^n_{q,s}\}$ and the ergodicity of localized polling systems. These facts will be needed when proving the lower bound for the empirical generator. Assume that P is irreducible and denote by η the (unique) stationnary distribution of P, $\tau \stackrel{\text{def}}{=} \sum_{i,j\in\mathcal{S}} \eta_i P_{ij}\tau_{ij}$ and $\rho_k \stackrel{\text{def}}{=} \lambda_k/\mu_k$. Then,

Theorem 4.4 *X^Λ is ergodic if, and only if,*

$$\sum_{k\in\Lambda^c} \rho_k + \frac{\lambda_i}{\eta_i}\left(\tau + \sum_{k\in\Lambda}\frac{\eta_k}{\mu_k}\right) < 1, \quad \forall i \in \Lambda^c.$$

For any sequence $\{x_n,\ n\geq 1\}$ of $\mathbb{Z}_+^N$ with $|x_n|\to\infty$, it exists a subsequence $\{n_k,\ k\geq 1\}$ such that:

1. *in probability, $\displaystyle\lim_{k\to\infty}\frac{1}{|x_{n_k}|}Q^\Lambda(0,x_{n_k},s) = \overline{Q}^\Lambda(0)$;*

2. *uniformly over all compact sets, the following limits hold in probability:*

$$\lim_{k\to\infty}\left(\frac{N^\Lambda_{n_k t}}{n_k}\hat{L}^\Lambda_{N_{n_k t}}, tL^\Lambda_{n_k t}, \frac{1}{|x_{n_k}|}Q^\Lambda(n_k t, x_{n_k}, s)\right) = \left(\overline{A}^\Lambda_t, \overline{L}^\Lambda_t, \overline{Q}^\Lambda(t)\right),$$

where $\overline{A}^\Lambda_t \in \mathcal{M}_+(\mathcal{S}_0)$ and $t^{-1}\overline{L}^\Lambda_t \in \mathcal{M}_1(\mathcal{S}_0)$, for all $t>0$. Moreover, $\overline{A}^\Lambda, \overline{L}^\Lambda, \overline{Q}^\Lambda$ are Lipschitz and satisfy the following set of relations:

$$\overline{Q}^\Lambda_i(t) = \overline{Q}^\Lambda_i(0) + \lambda_i t - \overline{A}^\Lambda_t\{i\}, \quad \forall i\in\mathcal{S}, \tag{4.7}$$

$$\sum_{j\in\mathcal{S}}\overline{A}^\Lambda_t\{ij\} = \sum_{j\in\mathcal{S}}\overline{A}^\Lambda_t\{ji\}, \quad \forall i\in\mathcal{S}, \tag{4.8}$$

$$\overline{A}^\Lambda_t\{i\} = \sum_{j\in\mathcal{S}}\overline{A}^\Lambda_t\{ij\}, \quad \forall i\in\Lambda, \tag{4.9}$$

$$\overline{A}^\Lambda_t\{i\} \leq \sum_{j\in\mathcal{S}}\overline{A}^\Lambda_t\{ij\}, \quad \forall i\in\Lambda^c, \tag{4.10}$$

$$\overline{A}^\Lambda_t\{ij\} = P_{ij}\sum_{k\in\mathcal{S}}\overline{A}^\Lambda_t\{ik\}, \quad \forall i,j\in\mathcal{S}, \tag{4.11}$$

$$\overline{A}^\Lambda_t\{ij\} = \frac{1}{\tau_{ij}}\overline{L}^\Lambda_t\{ij\}, \quad \forall i,j\in\mathcal{S}, \tag{4.12}$$

$$\overline{A}^\Lambda_t\{i\} = \mu_i\overline{L}^\Lambda_t\{i\}, \quad \forall i\in\mathcal{S}. \tag{4.13}$$

Remark : Note that in some sense, $\overline{A}_t^\Lambda\{s\}$ and $\overline{L}_t^\Lambda\{s\}$ represent respectively the number of visits and the time spent in state s by the server till time t.

♦♦♦

Now, let us restate Theorem 4.4 for localized polling systems with initial condition $(0, s)$ since by Proposition 4.3 this will be sufficient in order to prove the local bounds of large deviation. Moreover, for our purpose, it will be sufficient to assume the stability condition (4.14). It is worth noting that the fluid limits obtained in this case are linear.

Theorem 4.5 *Assume the following stability condition:*

$$\sum_{k\in\Lambda^c} \rho_k + \frac{\lambda_i}{\eta_i}\left(\tau + \sum_{k\in\Lambda}\frac{\eta_k}{\mu_k}\right) \leq 1, \quad \forall i \in \Lambda^c. \tag{4.14}$$

Uniformly over all compact sets, the following limits hold in probability:

$$\lim_{n\to\infty}\left(\frac{N_{nt}^\Lambda}{n}\hat{L}_{N_{nt}}^\Lambda, tL_{nt}^\Lambda, \frac{1}{n}Q^\Lambda(nt, 0, s)\right) = \big(tA, t\pi, tD\big),$$

where $A \in \mathcal{M}_+(\mathcal{S}_0)$, $\pi \in \mathcal{M}_1(\mathcal{S}_0)$ and $D \in \mathbb{R}^N$. Moreover,

$$a_s = \begin{cases} \lambda_i, & \textit{if} \quad s = i \in \Lambda^c, \\ C\eta_i, & \textit{if} \quad s = i \in \Lambda, \\ C\eta_i P_{ij}, & \textit{if} \quad s = ij, \ \forall i, j \in \mathcal{S}, \end{cases} \tag{4.15}$$

$$\pi_s = \begin{cases} \dfrac{a_i}{\mu_i}, & \textit{if } s = i \in \mathcal{S}, \\ a_{ij}\tau_{ij}, & \textit{if } s = ij, \ \forall i, j \in \mathcal{S}. \end{cases} \tag{4.16}$$

$$D_i = \lambda_i - a_i, \quad \forall i \in \mathcal{S}, \tag{4.17}$$

$$\textit{where} \qquad C = \frac{1 - \sum_{k\in\Lambda^c}\rho_k}{\tau + \sum_{k\in\Lambda}\frac{\eta_k}{\mu_k}}.$$

Proof : Let us briefly sketch the proof. Put

$$\overline{W}_t^\Lambda = \sum_{k\in\Lambda^c}\frac{\overline{Q}_k^\Lambda(t)}{\mu_k}.$$

Note that $\overline{W}_t^\Lambda \geq 0$ and $\overline{W}_0^\Lambda = 0$. Moreover, $\overline{W}^\Lambda$ is Lipschitz and so absolutely continuous. Then, as in [68, Lemma 1] or [52, Lemma 3.1], using (4.14) one can prove that at all regular point the derivative of $\overline{W}_t$ is negative. Hence $\overline{W}_t = 0$, $\forall t$. This implies that $\overline{Q}_k^\Lambda(t) = 0$, $\forall k \in \Lambda^c$, $\forall t \geq 0$, and by (4.7)

$$\overline{A}_t^\Lambda\{i\} = \lambda_i t, \ \forall i \in \Lambda^c. \tag{4.18}$$

Moreover, (4.8) ensures that it exists C_t such that

$$\sum_{j\in\mathcal{S}} \overline{A}_t^\Lambda\{ij\} = C_t \eta_i, \ \forall i \in \mathcal{S}.$$

Hence, by (4.11),

$$\overline{A}_t^\Lambda\{ij\} = C_t \eta_i P_{ij}, \ \forall i, j \in \mathcal{S}. \tag{4.19}$$

Hence, using (4.18), (4.19) together with (4.9), (4.12) and (4.13), one has

$$t^{-1}\overline{L}_t^\Lambda\{s\} = \begin{cases} \rho_i, & \text{if} \quad s = i \in \Lambda^c, \\ \dfrac{C_t}{t}\dfrac{\eta_i}{\mu_i}, & \text{if} \quad s = i \in \Lambda, \\ \dfrac{C_t}{t}\eta_i P_{ij}\tau_{ij}, & \text{if} \quad s = ij, \ \forall i, j \in \mathcal{S}, \end{cases}$$

$t^{-1}\overline{L}_t^\Lambda$ being a probability for all t,

$$C_t = t\frac{1 - \displaystyle\sum_{k\in\Lambda^c} \rho_k}{\tau + \displaystyle\sum_{k\in\Lambda} \frac{\eta_k}{\mu_k}}.$$

Now, the other relations can be derived directly. ∎

4.3 Generators and entropy

The purpose of this section is to define a new representation of the classical generator, which is a kind of description for the long run behavior. We shall also call it generator. It fits very well with fluid limits and large deviations and is easily measured. Moreover it is a natural object inasmuch as operations and properties can be naturally defined for these objects. This section is a generalization for polling systems of the generators introduced in [32] for Markov processes.

The entropy is here a meaningful function which has the classical properties that are expected for such an object, e.g. positivity, continuity and convexity.

4.3.1 Empirical generators

The empirical generator is the central object of this paper, from which most of the properties are derived. It describes rather naturally the measured behavior of a chain.

Definition 4.6 *Let $X^\Lambda = \{Q^\Lambda(t,x,s), S^\Lambda(t,x,s)\}_{t\geq 0}$ be a localized polling system. Define S_n^Λ to be the embedded process of the server just before it jumps*[1]*, and,*

- *N_t^Λ, the number of jumps of the server till t;*
- *$\hat{L}_n^\Lambda \stackrel{\text{def}}{=} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \delta_{S_i^\Lambda} \in \mathcal{M}_1(\mathcal{S}_0)$, the empirical measure of the process S_n^Λ;*
- *$L_t^\Lambda \stackrel{\text{def}}{=} \dfrac{1}{t}\displaystyle\int_0^t \delta_{S_u^\Lambda} du \in \mathcal{M}_1(\mathcal{S}_0)$, the empirical measure of the process S_t^Λ;*
- *$G_t^\Lambda \stackrel{\text{def}}{=} \left(\dfrac{N_t^\Lambda}{t}\hat{L}^\Lambda_{N_t^\Lambda}, L^\Lambda_{N_t^\Lambda}, \dfrac{Q_t^\Lambda - Q_0^\Lambda}{t}\right)$, the empirical generator associated to the polling system X^Λ.*

We shall denote by $\mathcal{G} \stackrel{\text{def}}{=} \mathcal{M}_+(\mathcal{S}_0) \times \mathcal{M}_1(\mathcal{S}_0) \times \mathbb{R}^N$ the space of empirical generators. Let $\Gamma \subset \mathcal{S}^2$ be a subgraph on the nodes $\mathcal{S}$, then the space of empirical generators with transitions in Γ is denoted by $\mathcal{G}(\Gamma) \stackrel{\text{def}}{=} \mathcal{M}_+(\mathcal{S} \cup \Gamma) \times \mathcal{M}_1(\mathcal{S}\cup\Gamma) \times \mathbb{R}^N$.

For the sake of brevity we shall drop the Λ for the measures N_t^Λ, $\hat{L}_n^\Lambda$, L_t^Λ and G_t^Λ.

4.3.2 Generators

We shall need the notion of balanced measure in order to describe the transitions of the polling system. It is a well-known object that appears in LDP for discrete time Markov chains [34].

Definition 4.7 (balanced measures) *The set $\mathcal{M}_s(E^2)$ of balanced measures is the set of finite positive measures on E^2, with both identical 1–dimensional projections*

$$A \in \mathcal{M}_s(E^2) \iff A(E,.) = A(.,E) \in \mathcal{M}_+(E). \tag{4.20}$$

Let Γ be a graph on E. $\mathcal{M}_s(\Gamma) \subset \mathcal{M}_s(E^2)$ is the set of balanced measures with support included in Γ.

[1]Note that S_n^Λ is not a Markov chain, because the state of the queues is missing.

For $A \in \mathcal{M}_s(E^2)$, let us put

- $a_{ij} \stackrel{\text{def}}{=} A(\{(i,j)\})$ the 2–dimensional law;
- $a_i \stackrel{\text{def}}{=} A(\{i\} \times E)$, the unique 1–dimensional projection;
- $A_{ij} \stackrel{\text{def}}{=} A(\{(i,j)\} | \{i\} \times E)$, the conditional law, and hence $a_{ij} = a_i A_{ij}$.

Note that we shall use the identity $A_{ij} = B_{ij}$ for $A, B \in \mathcal{M}_s(E^2)$. This does not mean equality $A = B$ for the two-dimensional law, but only for the conditional law.
Now, the analysis of polling systems has some specific features that must be reflected in the mathematical structure. The balance equation (4.20) just states that, in a Markov chain, a process enters a state as often as it exits it. In a polling system, the server goes to a queue as many times as it goes from that queue, but it stops and serves less times.

Definition 4.8 *For $A \in \mathcal{M}_+(\mathcal{S}_0)$, let $\overline{a}_i = \sum_{j\in\mathcal{S}} a_{ij}$ for all i in $\mathcal{S}$. Denote $\mathcal{M}_s(\mathcal{S}_0)$ the set of elements $A \in \mathcal{M}_+(\mathcal{S}_0)$ satisfying*

$$\sum_{j\in\mathcal{S}} a_{ji} \;=\; \sum_{j\in\mathcal{S}} a_{ij} \stackrel{\text{def}}{=} \overline{a}_i, \quad \forall i \in \mathcal{S}, \tag{4.21}$$

$$a_i \;\le\; \overline{a}_i, \qquad \forall i \in \mathcal{S}. \tag{4.22}$$

For all subset Λ of $\mathcal{S}$, denote $\mathcal{M}_s^\Lambda(\mathcal{S}_0)$ the set of elements $A \in \mathcal{M}_+(\mathcal{S}_0)$ satisfying (4.21) for all $i \in \mathcal{S}$, (4.22) for all $i \in \Lambda^c$, and the equality in (4.22) for all $i \in \Lambda$.

The equality in (4.22) means that, when the server arrives at a saturated queue, then it serves a client. Therefore $\mathcal{M}_s(\mathcal{S}_0) = \mathcal{M}_s^\emptyset(\mathcal{S}_0)$. Better, when the constraints (4.22) become equalities, i.e. when Λ increases, the set $\mathcal{M}_s^\Lambda(\mathcal{S}_0)$ decreases:

$$\Lambda \subset \Lambda' \implies \mathcal{M}_s^{\Lambda'}(\mathcal{S}_0) \subset \mathcal{M}_s^\Lambda(\mathcal{S}_0). \tag{4.23}$$

Note that, by (4.21), the matrix $\overline{a}_{ij} \stackrel{\text{def}}{=} a_{ij}$ associated to A is a balanced measure on $\mathcal{S}^2$. Its conditional law is $\overline{A}_{ij} \stackrel{\text{def}}{=} a_{ij}/\overline{a}_i$.

Definition 4.9 (generators) *For all subset Λ of $\mathcal{S}$, the set $\mathcal{G}_s^\Lambda$ of generators is defined by*

$$\mathcal{G}_s^\Lambda \stackrel{\text{def}}{=} \left\{(A,\pi,D) \in \mathcal{M}_s^\Lambda(\mathcal{S}_0) \times \mathcal{M}_1(\mathcal{S}_0) \times \mathbb{R}^\Lambda \ :\ a_i + D_i \ge 0,\ \forall i \in \Lambda\right\},$$

where we denote by $\mathbb{R}^\Lambda$, for the sake of simplicity, the subspace of $D \in \mathbb{R}^N$ with $D_i = 0$ for $i \in \Lambda^c$.
The characteristics of a generator such as irreducibility or the support refer to the routing part A of the generator. Let Γ be a subgraph of $\mathcal{S}^2$, then $\mathcal{G}_s^\Lambda(\Gamma) \subset \mathcal{G}_s^\Lambda$ is the set of generators with support included in Γ.
We shall denote by l the overall mean intensity of server jumps, $l = \sum_{s \in \mathcal{S}_0} a_s$.

Note that $\mathcal{G}_s^\Lambda$ does not decrease with Λ as $\mathcal{M}_s^\Lambda$. Indeed the drift part $\mathbb{R}^\Lambda$ increases with Λ.

Proposition 4.10 *There is a one-to-one mapping between the irreducible generators of $\mathcal{G}_s^\Lambda$ and the stable irreducible localized polling systems. (A, π, D) corresponds to the fluid limit of a localized polling system with parameters $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})$, i.e. equations (4.15)–(4.17), when $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})$ are derived from (A, π, D) by equations (4.24)–(4.27).*

Proof : Let $G = (A, \pi, D) \in \mathcal{G}_s^\Lambda$ be an irreducible generator. It is associated a unique localized polling system $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})_{i,j \in \mathcal{S}}$ by the relations

$$\begin{aligned} \tilde{\lambda}_i &= a_i + D_i, && (4.24)\\ \tilde{\mu}_i &= \frac{a_i}{\pi_i}, && (4.25)\\ \tilde{\tau}_{ij} &= \frac{\pi_{ij}}{a_{ij}}, && (4.26)\\ \tilde{P}_{ij} &= \overline{A}_{ij} = \frac{a_{ij}}{\bar{a}_i}. && (4.27) \end{aligned}$$

It must be now proved that $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})_{i,j \in \mathcal{S}}$ verifies the stability condition (4.14). By (4.21) and (4.27), $\bar{a}_i$ is an invariant measure for $\tilde{P}$. For G is irreducible, so is $\tilde{P}$, thus $\bar{a}_i$ is the unique invariant measure, up to a multiplicative constant. The invariant probability of $\tilde{P}$ is then $\tilde{\eta}_i = C\bar{a}_i$. Since $G \in \mathcal{G}_s^\Lambda$, the condition $a_i \leq \bar{a}_i$ for $i \in \Lambda^c$ can be written

$$\sum_{k \in \Lambda^c} \pi_k + \frac{a_i}{\bar{a}_i} \left(\sum_{k,l \in \mathcal{S}} \pi_{kl} + \sum_{k \in \Lambda} \pi_k \right) \leq 1, \qquad \forall i \in \Lambda^c. \tag{4.28}$$

Since $D_i = 0$ for $i \in \Lambda^c$, and $a_i = \bar{a}_i$ for $i \in \Lambda$, (4.28) is equivalent to (4.14). Therefore $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})_{i,j \in \mathcal{S}}$ is a stable localized polling system.
Reciprocally, let $(\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})_{i,j \in \mathcal{S}}$ be a stable irreducible localized polling system. The relations (4.15)–(4.17) define a unique triple (A, π, D) such that A verifies the balance relation (4.21) and that $a_i = \bar{a}_i$ for $i \in \Lambda$, by (4.8) and (4.9). The stability condition $a_i \leq \bar{a}_i$ for $i \in \Lambda^c$ is immediately derived from

the fluid limit equation (4.10). Therefore $A \in \mathcal{M}_s^\Lambda$. Since $(\tilde\lambda_i, \tilde\mu_i, \tilde\tau_{ij}, \tilde P_{ij})_{i,j\in\mathcal{S}}$ is stable, $D_i = 0$ for $i \in \Lambda^c$, hence $(A, \pi, D) \in \mathcal{G}_s^\Lambda$.

We have constructed two mappings between $\mathcal{G}_s^\Lambda$ and $\mathcal{P}^\Lambda$, mainly verifying that the stability conditions are the same. We have now to prove that these mappings are inverse of each other. Fix (A, π, D), then makes the correspondence

$$(A, \pi, D) \to (\tilde\lambda_i, \tilde\mu_i, \tilde\tau_{ij}, \tilde P_{ij}) \to (\tilde A, \tilde\pi, \tilde D).$$

Remind that we proved just above that $\tilde\eta_i = C\bar a_i$; now, by (4.15), $\tilde a_i = \tilde C \tilde\eta_i$ for $i \in \Lambda$. Calculations similar to (4.28) yield that $C\tilde C = 1$, then it is straight to prove $\tilde A = A$, $\tilde\pi = \pi$ and $\tilde D = D$. The bijection is proved. ■

The triple (A, π, D) of a generator is easily divided into a server part (A, π) and the drift part D. The server part a_s is interpreted as the mean number of passages through s per unit time, the stationary distribution π and the routing $\overline{A}$.

When G is not irreducible, there might exist several stationary distributions which determine several triples. However, if $(\tilde\lambda_i, \tilde\mu_i, \tilde\tau_{ij}, \tilde P_{ij})$ and one stationary distribution η for the routing $\tilde P$ are known, (A, π, D) is uniquely determined. When $\overline{a}_i = 0$, the mapping is ill-defined but it does not lead to undetermined forms if a stationary distribution π_i is defined. Actually the routing is not well defined, but it does not matter as long as it concerns the entropy function.

Generators and empirical generators : Note that the set of empirical generators $\mathcal{G}$ is not included in any $\mathcal{G}_s^\Lambda$ since the measure $\hat L_{N_t}$ is *not necessarily balanced.* Indeed there is generally two states where it is not balanced: if X begins in i and if it is in $j \neq i$ at time t, the server exits i one more time than it enters and the balance equation (4.21) is not satisfied for i. This is also true for the end point j, but for none of the other states. Finally this effects is bounded by $2/t$ for $t^{-1}N_t\hat L_{N_t}$ so that it vanishes when the time tends to infinity.

For the purpose of large deviations, bounds should be obtained for open and closed sets in $\mathcal{G}$. Since the distance between $\mathcal{G}$ and $\mathcal{G}_s^\Lambda$ tends to 0, the entropy has only to be defined on $\mathcal{G}_s^\Lambda$ and is infinite otherwise. Therefore the study of large deviations has only to be done locally around $\mathcal{G}_s^\Lambda$.

♦♦♦

4.3.3 Entropy

Definition 4.11 *The entropy in discrete time for a balanced measure A with respect to a stochastic matrix* $\mathbf{P}$ *is defined by*

$$H_d(A\|P) \stackrel{\text{def}}{=} \sum_{i,j\in E} a_{ij} \log\left(\frac{A_{ij}}{P_{ij}}\right), \tag{4.29}$$

with the usual convention $0\log(0/0) = 0$, $1/0 = +\infty$ *and* $0 \log 0 = 0$. *The continuous-time relative entropy of Poisson processes of intensities* ν *and* λ *is defined by*

$$I_p(\nu\|\lambda) \stackrel{\text{def}}{=} \nu \log\frac{\nu}{\lambda} - \nu + \lambda \tag{4.30}$$

We recall the properties of H_d and I_p, when E is finite. We shall also denote by $\mathbf{P}$ the graph of $\mathbf{P}$, i.e. all transitions such that $P_{ij} > 0$.

- $H_d(.\|P)$ is positive, it is null if, and only if $A_{ij} = P_{ij}$;
- $H_d(.\|P)$ is 1-homogeneous and convex; if P is irreducible, $H_d(.\|P)$ is strictly convex, i.e. $H(A\|P) = H(B\|P)$ if, and only if, $A_{ij} = B_{ij}$;
- $H_d(.\|P)$ is finite and continuous on $\mathcal{M}_s(\mathbf{P})$, infinite otherwise; it has compact level sets.
- $I_p(\nu\|\lambda)$ is positive, 1-homogeneous, continuous and strictly convex with respect to the pair (ν, λ) whenever $\lambda > 0$;
- $I_p(\nu\|\lambda)$ is null if, and only if, $\nu = \lambda$; for all λ, it has compact level sets in ν.

Definition 4.12 (relative entropy) *Let* $R = (\lambda_i, \mu_i, \tau_{ij}, P_{ij})$ *denotes the parameters of the polling system,* $G = (A, \pi, D) \in \mathcal{G}_s^\Lambda$ *be a generator and* $(\tilde\lambda_i, \tilde\mu_i, \tilde\tau_{ij}, \tilde P_{ij})$ *its representation as a polling system. The relative entropy of* G *with respect to* R *is*[2]

$$\begin{aligned} H(G\|R) \stackrel{\text{def}}{=} & \sum_{i\in\mathcal{S}} \left(I_p(\tilde\lambda_i\|\lambda_i) + \pi_i I_p(\tilde\mu_i\|\mu_i)\right) \\ & + \sum_{i,j\in\mathcal{S}} \pi_{ij} I_p(\tilde\tau_{ij}^{-1}\|\tau_{ij}^{-1}) + H_d(\overline{A}\|P). \end{aligned} \tag{4.31}$$

[2]Note that the relative entropy $H(G\|R)$ is independent of Λ.

This entropy has an easy interpretation in terms of information theory. The relative entropy can be defined as the *mean information gain*. $H(.\|R)$ is decomposed in (4.31) as the sum of the information gain for the arrivals $I_p(\tilde{\lambda}_i\|\lambda_i)$, the information gain for the service time $I_p(\tilde{\mu}_i\|\mu_i)$ multiplied by the time π_i spent at each queue, similarly the information gain for the transit time $\pi_{ij}I_p(\tilde{\tau}_{ij}^{-1}\|\tau_{ij}^{-1})$ and the information gain for the routing $H_d(\overline{A}\|P)$. For this last term, it can be decomposed ($H_d(A\|P)$ is 1-homogeneous w.r.t. A) into $\overline{a}$ which is the mean number of routing decisions by unit time, and $H_d(\overline{A}/\overline{a}\|P)$ which is the information gain for each routing decision. This natural interpretation leads to natural properties.

Proposition 4.13 *The relative entropy $H(.\|R)$ is positive, finite, continuous and strictly convex on $\mathcal{G}_s^\Lambda(\mathbf{P})$; it is infinite otherwise; it is null if, and only if, $G = R$. It has compact level sets.*

Proof : All those properties are immediately derived from the same properties of H_d and I_p, under the condition that μ_i, λ_i and τ_{ij} are strictly positive and finite, and that $\mathbf{P}$ is irreducible, conditions that are naturally satisfied here. ■

This is an easy but useful lemma for the LDP.

Lemma 4.14 *The set of irreducible generators is a dense subset of $\mathcal{G}_s^\Lambda(\mathbf{P})$.*

Proof : Let $G \in \mathcal{G}_s^\Lambda(\mathbf{P})$. For $\varepsilon > 0$, $(1-\varepsilon)G + \varepsilon R$ is irreducible because R is irreducible. It converges to G when ε tends to 0, hence the lemma. ■

4.3.4 Properties of the local functional $L(x, D)$

Definition 4.15 *The rate function $L(\Lambda, D)$ is defined by*

$$L(\Lambda, D) \stackrel{\text{def}}{=} \inf_{G \in f_\Lambda^{-1}(D)} H(G\|R), \quad \forall D \in \mathbb{R}^\Lambda, \tag{4.32}$$

where $f_\Lambda : \mathcal{G}_s^\Lambda \mapsto \mathbb{R}^\Lambda$ is the projection $f_\Lambda(G) = D$. For all $x \in \mathbb{R}_+^N$, $L(x, D)$ is defined to be $L(\Lambda(x), D)$.

Note that $L(\Lambda, D)$ is a rate function derived by the contraction of $H(.\|R)$, which is a good rate function (see Proposition 4.13). Even though the corresponding LDP are not proved, since f_Λ is linear, L inherits lots of good properties of H (see [32]).

Proposition 4.16 *The rate function $L(\Lambda, D)$ is positive, finite, continuous and strictly convex with respect to $D \in \mathbb{R}^\Lambda$; it is null if, and only if, D is the drift of the localized polling system; it has compact level sets. Moreover, the infimum $G \in \mathcal{G}_s^\Lambda$ such that $H(G\|R) = L(\Lambda, D)$ is reached at a unique point $G(D) \in \mathcal{G}_s^\Lambda(\mathbf{P})$, and $G(D)$ is a continuous function of D.*

Proof : Since f_Λ is linear and $H(.\|R)$ is strictly convex, by Proposition 3.22, $L(\Lambda, .)$ is convex. The compactness of the level sets is immediately derived from the compactness of the level sets of $H(.\|R)$.
Since $H(.\|R)$ is strictly convex, since its level sets are compact and since $L(\Lambda, D)$ is the solution of a convex program, the infimum $G(D)$ is reached at a unique point, which is obviously in $\mathcal{G}_s^\Lambda(\mathbf{P})$ where H is finite.
Let $D_n \stackrel{n\to\infty}{\longrightarrow} D$. Since $H(.\|R)$ has compact level sets and since it is continuous on $\mathcal{G}_s^\Lambda(\mathbf{P})$, it is easily seen that the sequence $\{G(D_n),\ n \geq 1\}$ is relatively compact. Let $\{G(D_{n_k}),\ k \geq 1\}$ be a subsequence such that $G(D_{n_k}) \stackrel{k\to\infty}{\longrightarrow} G$. $H(.\|R)$ being continuous,

$$L(\Lambda, D_{n_k}) = H(G(D_{n_k})\|R) \stackrel{k\to\infty}{\longrightarrow} H(G\|R).$$

By definition of L, $H(G\|R) \geq L(\Lambda, D)$. Now $G(D) = (A, \pi, D)$, and consider $G_{n_k} = (A, \pi, D_{n_k})$. Since H is continuous and since G_{n_k} tends to $G(D)$,

$$L(\Lambda, D) = H(G(D)\|R) = \lim_{k\to\infty} H(G_{n_k}\|R) \geq \lim_{k\to\infty} H(G(D_{n_k})\|R) = H(G\|R).$$

Then, $H(G\|R) = H(G(D)\|R)$, and by the uniqueness of the infimum, $G = G(D)$. So, $G(D_n) \stackrel{n\to\infty}{\longrightarrow} G(D)$. Finally $G(D)$ is continuous and so is $L(\Lambda, D)$. ∎

Remark : We shall extend Definition 4.15 for all $D \in \mathbb{R}^N$ by extending the validity of the equation (4.32). First, extend the set $\mathcal{G}_s^\Lambda$ of generators to

$$\overline{\mathcal{G}}_s^\Lambda \stackrel{\text{def}}{=} \left\{(A, \pi, D) \in \mathcal{M}_s^\Lambda(\mathcal{S}_0) \times \mathcal{M}_1(\mathcal{S}_0) \times \mathbb{R}^N \ :\ a_i + D_i \geq 0,\ \forall i\right\}.$$

This actually means that all localized generators are considered, not only the stable ones. The projection $\bar{f}_\Lambda :\ \overline{\mathcal{G}}_s^\Lambda \mapsto \mathbb{R}^N$ is the analog of f_Λ, and now

$$L(\Lambda, D) \stackrel{\text{def}}{=} \inf_{G \in \bar{f}_\Lambda^{-1}(D)} H(G\|R)$$

is defined for all $D \in \mathbb{R}^N$. Of course the definitions coincide for D belonging to $\mathbb{R}^\Lambda$.

One checks this extension preserves all properties proved here, particularly the lower semi-continuity and the convexity, which is convenient for the proof of the sample path LDP. Moreover, as underlined in remark 1 p. 172, these values are almost never used, so that finally this extension does not modify the rate function I_T.

♦♦♦

Lemma 4.17 *$L(\Lambda, D)$ is increasing w.r.t. Λ. $L(x, D)$ is jointly lower semi-continuous w.r.t. x and D.*

Proof : By Definition 4.15, $H(L\|R)$ does not depend of Λ and the domain $f_\Lambda^{-1}(D)$ where the infimum is taken decreases w.r.t. Λ; therefore $L(\Lambda, D)$ is increasing w.r.t. Λ. Actually $(A, \pi, D) \in f_\Lambda^{-1}(D)$ if, and only if $\pi \in \mathcal{M}_1(\mathcal{S}_0)$ and $A \in \mathcal{M}_s^\Lambda(\mathcal{S}_0)$ with $a_i + D_i \geq 0$ if $i \in \Lambda$, so that, by (4.23), it decreases w.r.t. Λ.
Let (x_n, D_n) tends to (x, D). Since all strictly positive coordinates of x have to be strictly positive for n large enough,

$$\exists n_0, \quad \Lambda(x) \subset \Lambda(x_n), \qquad \forall n \geq n_0.$$

Therefore, using both the continuity w.r.t. D and the growth of $L(., D)$,

$$\begin{aligned} &\liminf_{n\to\infty} \big(L(x_n, D_n) - L(x, D)\big) \\ \geq\ & \lim_{n\to\infty} \inf_{\Lambda \supset \Lambda(x)} \big(L(\Lambda, D_n) - L(\Lambda, D)\big) + \liminf_{n\to\infty} \big(L(x_n, D) - L(x, D)\big) \geq 0. \end{aligned}$$

Hence the lower semi-continuity with respect to both variables x and D. ■

Lemma 4.18 *There exists $M \in \mathbb{R}$ such that,*

$$L(x, D) \geq \frac{1}{2}\|D\| \log \|D\|, \qquad \forall x \in \mathbb{R}_+^N,\ \forall \|D\| \geq M. \tag{4.33}$$

Proof : Assume that $|D_j| = \|D\| \stackrel{\text{def}}{=} \max_i |D_i|$. From the definitions of $L(x, D)$ (Definition 4.15), of $H(G\|R)$ (Definition 4.12) and from the positivity of I_p and H_d,

$$L(x, D) \geq \inf_{L \in f_\Lambda^{-1}(D)} \big(I_p(a_j + D_j\|\lambda_j) + I_p(a_j\|\pi_j\mu_j)\big).$$

But, $I_P(x\|y)$ is increasing in x and decreasing in y for $x \geq y$ and necessarily $a_j \geq -D_j$, so that,

$$L(x, D) \geq \min\Big\{I_p(|D_j|\|\lambda_j)\ ,\ I_p(|D_j|\|\mu_j)\Big\}, \quad \forall |D_j| \geq \max\{\lambda_j, \mu_j\}.$$

Now $I_P(x\|y) \sim x \log x$ when x tends to infinity, so that there exists $M(y)$ such that

$$I_p(x\|y) \geq \frac{1}{2} x \log x, \quad \forall x \geq M(y).$$

Therefore (4.33) holds with

$$M = \max_{i \in \mathcal{S}} \left\{ M(\lambda_i), M(\mu_i), \lambda_i, \mu_i \right\}.$$

The lemma is proved. ∎

4.4 LDP for the empirical generator

It is to be stressed that the following Theorem 4.19 relies on the one-to-one mapping between empirical generators and polling systems expressed in Proposition 4.10, which itself depends on the linear description of the fluid limit when the polling system starts from an empty state (see Theorem 4.5). Thus, we shall only consider throughout this section that the queues are initially empty[3] ($Q^\Lambda(0) = 0$). Now, by Proposition 4.3, the problem of large deviations has been reduced precisely to this case.

We shall prove in this section a LDP for the combination of the empirical generator *and*

$$\left\{ \sup_{s \in [0,t]} |Q^\Lambda(s) - sD| < t\delta \right\},$$

defined for any $D \in \mathbb{R}^\Lambda$. We shall also define the projection $f_\Lambda : \mathcal{G}_s^\Lambda \mapsto \mathbb{R}^\Lambda$ by $f_\Lambda(G) = D$, and the ball $B(G, r)$ of center G and radius r. Note that $f_\Lambda(B(G, r)) = B(D, r)$, the second ball being in $\mathbb{R}^N$, since the distance is a max.

Theorem 4.19 (Generator's local bounds) *Let Λ be a face and $G = (\pi, A, D) \in \mathcal{G}_s^\Lambda$. Then,*

$$\begin{aligned} & \lim_{\delta \to 0} \liminf_{t \to \infty} \frac{1}{t} \log \mathbb{P} \left[G_t \in B(G, \delta),\ \sup_{s \in [0,t]} |Q^\Lambda(s) - sD| < t\delta \right] && (4.34) \\ = \ & \lim_{\delta \to 0} \limsup_{t \to \infty} \frac{1}{t} \log \mathbb{P} \left[G_t \in B(G, \delta),\ \sup_{s \in [0,t]} |Q^\Lambda(s) - sD| < t\delta \right] && (4.35) \\ = \ & -H(G\|R). \end{aligned}$$

[3] For this reason, the initial condition $(0, s)$ in $Q^\Lambda(t, 0, s)$ will be dropped.

The proof will be done in four steps. First, a change of measure, so as to treat only neighborhoods of a generator G; second the lower and upper local LDP bounds for particular generators; third a continuity argument, in order to extend the bounds to all generators; fourth the exponential tightness of the measures.

4.4.1 Exponential change of measure

Fix $G \in \mathcal{G}_s^\Lambda$. In this section the bijection of Proposition 4.10 is used to describe G indifferently as $G = (A, \pi, D)$ or as $G = (\tilde{\lambda}_i, \tilde{\mu}_i, \tilde{\tau}_{ij}, \tilde{P}_{ij})$. In the next section, fluid limit results will be used. Since these results are restricted to irreducible polling systems, the generator G will henceforth be assumed to be irreducible. Secondly, in order for the mapping h_2 to be properly defined (see below), $\tilde{P}_{ij}$ is assumed to be null when P_{ij} is, i.e. $G \in \mathcal{G}_s^\Lambda(\mathbf{P})$.
The first step is to amount the problem to neighborhoods of G. For this purpose we introduce a change of measure. Consider a localized polling system X^Λ with generator $R = (\lambda_i, \mu_i, \tau_{ij}, P_{ij})$. Then define

- the vector $\alpha \in \mathbb{R}^N$ by $\alpha_i \stackrel{\text{def}}{=} \log \frac{\tilde{\lambda}_i}{\lambda_i}$, for all $i \in \mathcal{S}$;
- the mapping $h_1 : \mathcal{S}_0 \mapsto \mathbb{R}$ by $\begin{cases} h_1(i) \stackrel{\text{def}}{=} \alpha_i + \log \frac{\tilde{\mu}_i}{\mu_i}, \\ h_1(ij) \stackrel{\text{def}}{=} \log \frac{\tau_{ij}}{\tilde{\tau}_{ij}}, \end{cases}$ for all $i, j \in \mathcal{S}$;
- the mapping $h_2 : \mathcal{S}_0 \mapsto \mathbb{R}$ by $\begin{cases} h_2(ij) \stackrel{\text{def}}{=} \log \frac{\tilde{P}_{ij}}{P_{ij}}, \\ h_2(i) \stackrel{\text{def}}{=} 0, \end{cases}$ for all $i, j \in \mathcal{S}$;
- the compensator $K : \mathcal{S}_0 \mapsto \mathbb{R}$ by

$$\begin{cases} K(i) \stackrel{\text{def}}{=} \displaystyle\sum_{k \in \mathcal{S}} (\tilde{\lambda}_k - \lambda_k) + \tilde{\mu}_i - \mu_i, \\ K(ij) \stackrel{\text{def}}{=} \displaystyle\sum_{k \in \mathcal{S}} (\tilde{\lambda}_k - \lambda_k) + \tilde{\tau}_{ij}^{-1} - \tau_{ij}^{-1}, \end{cases} \quad \forall i, j \in \mathcal{S};$$

- and the process M_t^Λ by

$$\exp \left\{ < \alpha, Q^\Lambda(t, x, s) - x > + \sum_{i=0}^{N_t^\Lambda - 1} h_1(S_i^\Lambda) + h_2(S_{i+1}^\Lambda) - \int_0^t K(S_v^\Lambda) \, \mathrm{d}v \right\}.$$

Since K has been exactly defined so that[4]

$$K(s)=\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\exp\left\{<\alpha,Q^\Lambda(t,x,s)-x>+\sum_{i=0}^{N_t^\Lambda-1}h_1(S_i^\Lambda)+h_2(S_{i+1}^\Lambda)\right\}\right]_{t=0},$$

it is easily checked that the derivative of $\mathbb{E}\left[M_t^\Lambda\right]$ at $t=0$ is null. Then using the Markov property, one can get that the derivative is null for all $t\geq 0$, so that $\mathbb{E}\left[M_t^\Lambda\right]=1$. Using again the Markov property, this proves that $\mathbb{E}\left[M_t^\Lambda|\mathcal{F}_s\right]=M_s^\Lambda$, for all $t\geq s\geq 0$, hence $\{M_t^\Lambda,\ t\geq 0\}$ is a martingale w.r.t. the natural filtration $\mathcal{F}_t$.
Then define a new probability measure by

$$\tilde{\mathbb{P}}\left[B\right]\stackrel{\text{def}}{=}\mathbb{E}\left[\mathbb{I}_{\{B\}}M_t^\Lambda\right],\quad \forall B\in\mathcal{F}_t.$$

It is a matter of routine to show that under $\tilde{\mathbb{P}}$, X^Λ is again a Markov process. Moreover if $q^\Lambda(x,s;y,s')$ denotes the intensity of jump from (x,s) to (y,s') of X^Λ under $\mathbb{P}$, then the generator X^Λ under $\tilde{\mathbb{P}}$ is given by

$$\sum_{y,s'}q^\Lambda(x,s;y,s')e^{<\alpha,y-x>+\mathbb{I}_{\{s\neq s'\}}(h_1(s)+h_2(s'))}\big(f(y,s')-f(x,s)\big).$$

Hence under $\tilde{\mathbb{P}}$, X^Λ depicts the evolution of a localized polling system without boundary conditions on the nodes belonging to Λ. The arrival rate at node i is $\lambda_i e^{\alpha_i}=\tilde{\lambda}_i$ whereas the service rate is equal to $\mu_i e^{-\alpha_i+h_1(i)}=\tilde{\mu}_i$. The intensity of switch-over time and the probability of routing between i and j are respectively given by $\tau_{ij}^{-1}e^{h_1(ij)}=\tilde{\tau}_{ij}^{-1}$ and $P_{ij}e^{h_2(ij)}=\tilde{P}_{ij}$.
This is the change of measure we wanted.

4.4.2 Upper and lower bounds

The ball $B(G,\delta)$ is precisely defined by

$$B(G,\delta)\stackrel{\text{def}}{=}\Big\{G'\in\mathcal{G}(\mathbf{P}):\ \|G'-G\|_\infty<\delta\Big\}.$$

Define the continuous mapping $\phi_{R,G}:\mathcal{G}(\mathbf{P})\mapsto\mathbb{R}$ by

$$\phi_{R,G}(G')\stackrel{\text{def}}{=}<\alpha,D'>+\sum_{s\in\mathcal{S}_0}a'_s h_1(s)+\sum_{i,j\in\mathcal{S}}a'_{ij}h_2(ij)-\sum_{s\in\mathcal{S}_0}\pi'_s K(s).\qquad(4.36)$$

[4] Note that the derivative is independent of Λ and x, so that they are dropped.

Simple manipulations using (4.24)–(4.27) yield $\phi_{R,G}(G) = H(G\|R)$. Note that $\phi_{R,G}$ is finite everywhere as soon as $H(G\|R)$ is finite, since then there are no infinite terms in (4.36). The change of measure has been done so that

$$\log M_t^\Lambda = t\phi_{R,G}(G_t).$$

For the sake of brevity, we shall denote by $E_t(G,\delta)$ the event

$$E_t(G,\delta) \stackrel{\text{def}}{=} \left\{ G_t \in B(G,\delta),\ \sup_{s\in[0,t]} |Q^\Lambda(s) - sD| < t\delta \right\}.$$

Applying the change of measure and the previous relation yields

$$\begin{aligned} \frac{1}{t}\log \mathbb{P}\left[E_t(G,\delta)\right] &= \frac{1}{t}\log \tilde{\mathbb{E}}\left[\left(M_t^\Lambda\right)^{-1} \mathbb{1}_{E_t(G,\delta)}\right] \\ \geq\ & -\sup_{G'\in B(G,\delta)} \phi_{R,G}(G') + \frac{1}{t}\log \tilde{\mathbb{P}}\left[E_t(G,\delta)\right]. \end{aligned} \tag{4.37}$$

Now, it has been shown through fluid limit that, under $\tilde{\mathbb{P}}$, G_t converges to G in probability and $Q^\Lambda(s,0)$ converges to sD uniformly for $s \in [0,t]$, so that the probability of the event $E_t(G,\delta)$ converges to 1. Moreover, by the continuity of $\phi_{R,G}$, the first term of (4.37) tends to $\phi_{R,G}(G) = H(G\|R)$, hence the local lower bound,

$$\lim_{\delta\to 0}\liminf_{t\to\infty} \frac{1}{t}\log \mathbb{P}\left[E_t(G,\delta)\right] \geq -H(G\|R). \tag{4.38}$$

Then, reversing the inequality in (4.37) yields

$$\frac{1}{t}\log \mathbb{P}\left[E_t(G,\delta)\right] \leq -\inf_{G'\in B(G,\delta)} \phi_{R,G}(G') + \frac{1}{t}\log \tilde{\mathbb{P}}\left[E_t(G,\delta)\right],$$

which turns to the local upper bound by the same argument,

$$\lim_{\delta\to 0}\limsup_{t\to\infty} \frac{1}{t}\log \mathbb{P}\left[\tilde{\mathbb{P}}\left[E_t(G,\delta)\right]\right] \leq -H(G\|R). \tag{4.39}$$

Now the irreducibility assumption will be relaxed. First note that the upper bound (4.39) is obtained by bounding the probability $\tilde{\mathbb{P}}[E_t(G,\delta)]$ by 1, thus it is valid for reducible generators in $\mathcal{G}_s^\Lambda(\mathbf{P})$.

By Lemma 4.14, the set of irreducible generators is dense in $\mathcal{G}_s^\Lambda(\mathbf{P})$. For any $G \in \mathcal{G}_s^\Lambda(\mathbf{P})$ and any $\delta > 0$, it is possible to find an irreducible $G' \in \mathcal{G}_s^\Lambda(\mathbf{P})$ such that the distance with G is less than δ. Hence there exists $\delta' > 0$ such that $B(G',\delta') \subset B(G,\delta)$. Therefore

$$\liminf_{t\to\infty} \frac{1}{t}\log \mathbb{P}\left[E_t(G,\delta)\right] \geq \liminf_{t\to\infty} \frac{1}{t}\log \mathbb{P}\left[E_t(G',\delta')\right] \geq -H(G'\|R).$$

Since this is true for any G' close to G, by continuity of H on $\mathcal{G}_s^\Lambda(\mathbf{P})$, (4.38) is also true on $\mathcal{G}_s^\Lambda(\mathbf{P})$ for reducible generator.
If $G \notin \mathcal{G}_s^\Lambda(\mathbf{P})$, by Proposition 4.13, $H(G\|P) = \infty$. Obviously (4.38) is valid. When G does not belong to $\mathcal{G}_s^\Lambda(\mathbf{P})$ which is closed, there exists $\delta > 0$ such that $B(G, \delta)$ is entirely outside $\mathcal{G}_s^\Lambda(\mathbf{P})$. The probability for such an event is null. Hence the bound (4.39) is true for all generators.

4.4.3 Exponential tightness

We shall prove here that the measures are exponentially tight, that is, for all $C \in \mathbb{R}$, there exists a compact set $K \subset \mathcal{G}$ such that

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in K^c\right] \leq -C. \tag{4.40}$$

For a Poisson process with intensity bounded by r, the measured intensity N_t/t is bounded by,

$$\mathbb{P}\left[N_t/t > l_0\right] = e^{-rt} \sum_{n > l_0 t} \frac{(rt)^n}{n!} \leq e^{-rt} \frac{(rt)^{l_0 t}}{(l_0 t)!} \frac{l_0}{l_0 - r}.$$

Using Stirling's expansion, the logarithm of the right part is shown to be about $-tl_0 \log l_0$ when t and l_0 are large, which yields the exponential tightness.
Now, the intensity of jumps for the server and for the arrivals is bounded by

$$r \stackrel{\text{def}}{=} \max_{i,j\in\mathcal{S}} \left\{ \lambda_i, \mu_i, \frac{1}{\tau_{ij}} \right\}.$$

Therefore, taking l_0 big enough, $K \stackrel{\text{def}}{=} \{(A, \pi, D) : l \leq l_0\}$, where $l = \sum_{s\in\mathcal{S}_0} a_s$, is a compact set verifying the bound (4.40).

The exponential tightness allows classically to extend a weak LDP to a full LDP. We did not prove an LDP here, but the exponential tightness is required for the kind of contraction that follows.

4.4.4 Contraction

Since $f_\Lambda(G) = D$ is a continuous mapping from $\mathcal{G}_s^\Lambda$ into $\mathbb{R}^\Lambda$, the contraction principle applied to the local bounds (4.34)–(4.35) suggests the following contraction.

Theorem 4.20 (Local bounds) *Let Λ be a face and $D \in \mathbb{R}^\Lambda$. Then,*

$$\lim_{\delta\to 0}\liminf_{t\to\infty}\frac{1}{t}\log \mathbb{P}\left[\sup_{s\in[0,t]}|Q^\Lambda(s)-sD|<t\delta\right] \tag{4.41}$$

$$= \lim_{\delta\to 0}\limsup_{t\to\infty}\frac{1}{t}\log \mathbb{P}\left[\sup_{s\in[0,t]}|Q^\Lambda(s)-sD|<t\delta\right] = -L(\Lambda,D), \tag{4.42}$$

where $L(\Lambda,.)$ is the good rate function

$$L(\Lambda,D)\stackrel{\text{def}}{=}\inf_{G\in f_\Lambda^{-1}(D)} H(G\|R), \quad \forall D\in\mathbb{R}^\Lambda. \tag{4.43}$$

Proof : The sketch of the proof is exactly the same as a proof of the contraction principle, except that it is restricted around $\mathcal{G}_s^\Lambda$. First note that,

$$f_\Lambda^{-1}(B(D,\delta)) = \bigcup_{G\in f_\Lambda^{-1}(D)} B(G,\delta). \tag{4.44}$$

The lower bound is checked with a sequence $G(n)$ such that $H(G(n)\|R)$ converges to $L(\Lambda,D)$.
For the upper bound, consider the right part of (4.44). From this set, it can be extracted by the exponential tightness a compact set $K\subset\mathcal{G}$, so that

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathbb{P}\left[G_t\in K, \sup_{s\in[0,t]}|Q^\Lambda(s)-sD|<t\delta\right]$$
$$= \limsup_{t\to\infty}\frac{1}{t}\log\mathbb{P}\left[\sup_{s\in[0,t]}|Q^\Lambda(s)-sD|<t\delta\right].$$

Fix $\varepsilon>0$. By (4.35), for all $G\in\mathcal{G}_s^\Lambda$, there exists $\delta>0$ such that

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathbb{P}\left[G_t\in B(G,\delta), \sup_{s\in[0,t]}|Q^\Lambda(s)-sD|<t\delta\right]\le -H(G\|R)+\varepsilon.$$

Since $K\cap f_\Lambda^{-1}(D)$ is compact, it can be covered by a finite number of balls, as in (4.44). Moreover there exists δ_0 such that

$$K\cap f_\Lambda^{-1}(B(D,\delta_0))\subset\bigcup_{i=1}^n B(G(i),\delta(i)) \quad \text{with } f_\Lambda(G_i)=D.$$

But $B(D, \delta)$ is decreasing with δ so that the previous equation is valid for $\delta \le \delta_0$. Finally

$$\begin{aligned}&\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{P}\left[G_t \in K, \sup_{s\in[0,t]} |Q^\Lambda(s) - sD| < t\delta\right] \\ \le \quad & -\min_i H(G(i)\|R) + \varepsilon \le -L(\Lambda, D) + \varepsilon.\end{aligned}$$

This is valid for all $\varepsilon > 0$ and $\delta < \delta_0$, hence the upper bound. Since (4.41) is less than (4.42), they are both equal to $-L(\Lambda, D)$. ■

4.5 The sample path LDP

First, we verify that the rate function $I_T(.)$ possesses the usual properties. For, recall that

$$\Phi_x(K) = \{\varphi \in D([0,T], \mathbb{R}_+^N) : I_T(\varphi) \le K, \varphi(0) = x\}.$$

Then

Proposition 4.21 *(i). Assume $I_T(\varphi) \le K$ for some K. Then, for all $\epsilon > 0$, there exists $\delta > 0$ independent of φ such that for any collection of non overlapping intervals $[t_j, t_{j+1}]$ in $[0,T]$ with $\sum_j t_{j+1} - t_j = \delta$,*

$$\sum_j |\varphi(t_{j+1}) - \varphi(t_j)| \le \epsilon.$$

(ii). $I_T(.)$ is lower semi-continuous in $\Big(D\big([0,T], \mathbb{R}_+^N\big), d_d\Big)$.

(iii). If $C \subset \mathbb{R}_+^N$ is compact, so is $\bigcup_{x\in C} \Phi_x(K) \subset \mathcal{C}([0,T], \mathbb{R}_+^N)$.

(iv). Let $\varphi \in \mathcal{AC}\big([0,T], \mathbb{R}_+^N\big)$ with $I_T(\varphi) < \infty$. Then, for all $\epsilon > 0$, it exists $\varphi_\epsilon \in \mathcal{PL}\big([0,T], \mathbb{R}_+^N\big)$ such that:

(a) $d_c(\varphi_\epsilon, \varphi) \le \epsilon$,

(b) $I_T(\varphi_\epsilon) \le I_T(\varphi) + \epsilon$.

Proof : As in [128, Lemma 5.18], one can prove (i) using Lemma 4.18.

By (i), in order to prove the lower semi-continuity of $I_T(.)$, it is sufficient to consider sequences of absolutely continuous functions. Since on $\mathcal{C}\big([0,T], \mathbb{R}_+^N\big)$, the metrics d_c and d_d are equivalent, one can use d_c. Now, using Lemma 4.18, since $L(x, D)$ is lower semi-continuous in (x, D) and convex with respect to D (see Lemma 4.17 and Proposition 4.16), [85, Theorem 3 of Section 9.1.4] yields (i).

(iii) is a consequence of (i) and (ii) (see [128, Proposition 5.46]).

For (iv), similar results were proved in [57, Lemma 7.5.4] in the case of a random walk with discontinuity along an hyperplane. Nevertheless, a different proof is provided in Section 4.C. ■

Remark : (ii) and (iv) imply that for $\varphi \in \mathcal{AC}([0,T], \mathbb{R}_+^N)$ with $I_T(\varphi) < \infty$, it exists a sequence $\{\varphi_n,\ n \geq 1\}$ with $\varphi_n \in \mathcal{PL}([0,T], \mathbb{R}_+^N)$ for all n and satisfying

$$\lim_{n\to\infty} d_c(\varphi_n, \varphi) = 0 \text{ and } \lim_{n\to\infty} I_T(\varphi_n) = I_T(\varphi).$$

♦♦♦

Using irreducibility properties of X, one gets from Theorem 4.1 large deviations bounds for the probability that the process drifts along some linear path:

Proposition 4.22 (Linear bounds) *Let $x \in \mathbb{R}_+^N$ and $D \in \mathbb{R}^N$ satisfying $x + D\tau \in \mathbb{R}_+^N$. Then*

$$\begin{aligned}
&\lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n}\inf_{|y-nx|<\epsilon n}\log \mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q(t,y) - nx - nDt| < \delta n\right]\\
=\ &\lim_{\delta\to 0}\lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\sup_{|y-nx|<\epsilon n}\log \mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q(t,y) - nx - nDt| < \delta n\right]\\
=\ &-\tau L\big(\Lambda(D)\cup\Lambda(x), D\big).
\end{aligned}$$

In Theorem 4.1, $x + Dt$ is supposed to lie in the face $\Lambda(x)$ for all $t \in [0, \tau]$. In Proposition 4.22, it is assumed that $x + Dt$ evolves in the face $\Lambda(D) \cup \Lambda(x)$ only for $t \in]0, \tau[$. In fact, the boundary effects possibly occuring for t in the neighborhood of 0 or τ can be ignored. The derivation of Proposition 4.22 from Theorem 4.1 was done in [56]. A simple proof of this purely technical fact is provided in appendix 4.A. Using the Markov property and Proposition 4.22, one deduces easily the following large deviations bounds for the probability that the process evolves around some piecewise linear function:

Proposition 4.23 (Piecewise linear bounds) *Let* $\varphi \in \mathcal{PL}\left([0,T], \mathbb{R}_+^N\right)$ *and set* $x \stackrel{\text{def}}{=} \varphi(0)$. *Then*

$$
\begin{aligned}
-I_T(\varphi) &= \lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n}\inf_{|y-nx|<\epsilon n}\log \mathbb{P}\left[\sup_{t\in[0,nT]}|Q(t,y)-n\varphi(t)|<\delta n\right] \\
&= \lim_{\delta\to 0}\lim_{\epsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\sup_{|y-nx|<\epsilon n}\log \mathbb{P}\left[\sup_{t\in[0,nT]}|Q(t,y)-n\varphi(t)|<\delta n\right].
\end{aligned}
$$

In order to prove Theorem 4.2, the piecewise linear interpolation of $Q^n_{x,s}$ denoted $Z^n_{x,s}$ is needed. For, n being fixed, put

$$t_j^n = j\frac{T}{n},\ j=0,1,\dots,n.$$

Then $Z^n_{x,s}$ is defined by

$$Z^n_{x,s}(t) = \left(\frac{n}{T}t - j\right)Q^n_{x,s}(t^n_{j+1}) + \left((j+1)-\frac{n}{T}t\right)Q^n_{x,s}(t^n_j),\ \forall t\in[t^n_j, t^n_{j+1}].$$

The following property is needed in proving the upper bound of Theorem 4.2.

Proposition 4.24 1. *For all* $\delta>0$, *uniformly in* $x\in\mathbb{R}_+^N$,

$$\limsup_{n\to\infty}\frac{1}{n}\log \mathbb{P}\left[\sup_{t\in[0,nT]}|Z^n_{x,s}(t)-Q^n_{x,s}(t)|\geq\delta\right] = -\infty.$$

2. *Fix* C *a compact set of* $\mathbb{R}_+^N$. *For each* $M<\infty$, *there is a compact* $\mathcal{K}$ *of* $\mathcal{C}\left([0,T],\mathbb{R}_+^N\right)$ *such that*

$$\limsup_{n\to\infty}\frac{1}{n}\log \mathbb{P}\left[Z^n_{x,s}\in\mathcal{K}\right] \leq -M,\ \forall x\in C.$$

Proof : This is a straight adaptation of [58, Lemmas 2.4 and 2.5]. ∎

Proof of Theorem 4.2 Although technical, the extension of Proposition 4.23 to Theorem 4.2 is rather standard using Propositions 4.21 and 4.24. The lower bound is obtained using Proposition 4.23 and the remark p. 193. The upper bound is derived via Propositions 4.21 (iii), 4.23 and 4.24. The reader is referred to [56, Section 5] for details. ∎

4.6 Conclusion and comments

As it emerges, in proving a LDP for $Q(t)/t$, it can be fruitful to work at a higher level. In fact, rather than studying Q_t itself, we focused on the empirical generator which allows one to know how the different transition rates have to be modified in order that Q_t stays near a given drift D. Then using the explicit expression of fluid limits for polling systems, one can select in this case the *unique* change of measure under which the system follows a prescribed empirical generator. Finally, using a contraction principle, the local rate function can be identified as the solution of an optimization problem, in our case a convex program. As it was pointed out, in general for Markov processes in $\mathbb{Z}_+^N$, the fluid limits cannot be characterized. So it seems hard to identify the rate function for arbitrary queueing systems using directly the present methodology. Nevertheless, in our opinion, it could be applied successfully to other networks. For instance, let us briefly discuss the well known example of Jackson networks.

Consider an open network consisting of N nodes and denote the set of nodes by $\mathcal{S} \stackrel{\text{def}}{=} \{1, \ldots, N\}$. At node i, arrivals of clients form a Poisson process with rate λ_i. Each customer at node i requires service, whose duration is exponentially distributed with parameter μ_i. When a client has been served at node i, then it goes to node j with probability P_{ij} or exits the system with probability P_{i0}. $Q_t \stackrel{\text{def}}{=} \big(q_1(t), \ldots, q_N(t)\big)$ is a Markov process where $q_i(t)$ represents the number of clients at node i at time t. For $i \in \mathcal{S}$, $j \in \mathcal{S} \cup \{0\}$, let

- $N_n\{i, j\}$, represents among the n first transitions, the number of customers served at node i and going to node j.

In this case, the empirical generator is the process G_t defined by

$$G_t \stackrel{\text{def}}{=} \left(\frac{N_t}{t} \hat{L}_{N_t}, L_t, \frac{Q_t - Q_0}{t} \right),$$

where

- N_t is the number of transitions till t;
- $\hat{L}_n \stackrel{\text{def}}{=} \dfrac{1}{n} N_n \in \mathcal{M}_+\Big(\mathcal{S} \times \big(\mathcal{S} \cup \{0\}\big)\Big)$ is a subprobability. It represents in some sense the empirical measure of the embedded Markov chain except that it does not take into account the arrivals of clients in the network;

- $L_t = \big(L_1(t), \dots, L_N(t)\big)$ where

$$L_i(t) \stackrel{\text{def}}{=} \frac{1}{t}\int_0^t \delta_{q_i(u)\neq 0}\, du.$$

  $L_i(t)$ is the time average the server at node i is not idle.

Like for polling systems, in order to identify $L(\Lambda, D)$ with $D_i = 0$, $\forall i \in \Lambda^c$, one considers localized Jackson networks X^Λ. The transition mechanism describing the evolution of X^Λ is identical to X's except that the components indexed by Λ can be negative. Now, let us introduce the set of localized empirical generators for X^Λ.

Definition 4.25 (Localized empirical generators) *The set $\mathcal{G}^\Lambda$ of localized empirical generators is defined by the triples $(A, \pi, D) \in \mathcal{M}_+(\mathcal{S}_0) \times [0,1]^N \times \mathbb{R}^\Lambda$ that verify*

$$\begin{aligned} \pi_i &= 1, \forall i \in \Lambda, \\ D_i - \sum_{j\in\mathcal{S}} a_{ji} + a_i &\geq 0, \forall i \in \mathcal{S}, \end{aligned}$$

where a_i stands for $\sum_{j\in\mathcal{S}\cup\{0\}} a_{ij}$. We denote by $\mathbb{R}^\Lambda$, for the sake of simplicity, the subspace of $D \in \mathbb{R}^N$ with $D_i = 0$ for $i \in \Lambda^c$.

Remark : Note that $\mathcal{G}^\Lambda$ is a convex set.

♦♦♦

To each generator $(A, \pi, D) \in \mathcal{G}^\Lambda$ corresponds a localized Jackson network, described by its intensities $(\tilde\lambda_i, \tilde\mu_i, \tilde P_{ij})$. The correspondence is given by

$$\begin{aligned} \tilde\lambda_i &= D_i - \sum_{j\in\mathcal{S}} a_{ji} + a_i, \\ \tilde\mu_i &= \frac{a_i}{\pi_i}, \\ \tilde P_{ij} &\stackrel{\text{def}}{=} \frac{a_{ij}}{a_i}. \end{aligned}$$

Definition 4.26 (relative entropy) *Let $R = (\lambda_i, \mu_i, P_{ij})$ denotes the generator of the Jackson network, $G = (A, \pi, D) \in \mathcal{G}_s^\Lambda$ be a localized empirical generator and $(\tilde\lambda_i, \tilde\mu_i, \tilde P_{ij})$ its representation as a localized Jackson network. The relative entropy of G with respect to R is*[5]

$$H(G\|R) \stackrel{\text{def}}{=} \sum_{i\in\mathcal{S}} \Big(I_p(\tilde\lambda_i \| \lambda_i) + \pi_i I_p(\tilde\mu_i \tilde P_{ij} \| \mu_i P_{ij}) \Big),$$

[5] Note that the relative entropy $H(G\|R)$ is independent of Λ.

where $I_p(\nu \| \lambda) \stackrel{\text{def}}{=} \nu \log \frac{\nu}{\lambda} - \nu + \lambda$.

This entropy has an easy interpretation in terms of information theory. $H(.\|R)$ is decomposed as the sum of the information gains, first $I_p(\tilde{\lambda}_i \| \lambda_i)$ for the arrivals, second $I_p(\tilde{\mu}_i \tilde{P}_{ij} \| \mu_i P_{ij})$ for the service times and the routing, multiplied by the time π_i that the server at node i is not idle.

Definition 4.27 *The rate function* $L(\Lambda, D)$ *is defined by*

$$L(\Lambda, D) \stackrel{\text{def}}{=} \inf_{G \in f_\Lambda^{-1}(D)} H(G \| R), \quad \forall D \in \mathbb{R}^\Lambda,$$

where $f_\Lambda : \mathcal{G}^\Lambda \mapsto \mathbb{R}^\Lambda$ *is the projection* $f_\Lambda(G) = D$. *For all* $x \in \mathbb{R}_+^N$, $L(x, D)$ *is defined to be* $L(\Lambda(x), D)$.

Like for polling systems, $L(\Lambda, D)$ appears naturally as the solution of a convex program. This was proved in [82] after suitable change of variable. In our case, the main point will be to establish the Theorem 4.19. This can be done using the preceding discussion, a suitable change of measure and the representation of fluid limits for Jackson networks. The reader is reffered to [82, Sections 5 and 6] for the last two points.

Appendix

4.A From local bounds to piecewise linear functions

In this section, we shall use strong properties of irreducibility of the Markov process X. Denote

$$p = \min_{s,s',y,y'} \frac{q(y,s;y',s')}{\sum_{y'',s''} q(y,s;y'',s'')} > 0.$$

Let $x \in \mathbb{Z}_+^N$. $x+e_i$ can be reached with probability larger than p in one step from x. $\mathbf{P}$ being irreducible, $x-e_i$ can be reached with probability larger than p^J in less than J steps from x. Moreover $J \leq 3N$. Then

Lemma 4.28 *For all $x, y \in \mathbb{Z}_+^N$, there exists some path in $\mathbb{Z}_+^N$,*

$$i_1 = x, \ldots, i_k = y$$

with length less than $J|x-y|$, i.e $k \leq J|x-y|$, and the embedded Markov chain associated to X follows this path with probability larger than $p^{J|x-y|}$. Since the jumps of X are bounded by one, the path can be chosen so that

$$|i_j - x| \leq |x-y| + J, \ \forall j = 1, \ldots, k.$$

The following notation will be used in this section:

$$q_* = \min_{s,y} \sum_{y'',s''} q(y,s;y'',s'') \text{ and } q^* = \max_{s,y} \sum_{y'',s''} q(y,s;y'',s'').$$

Lemma 4.29 *Let $0 \leq t_0 \leq t_1$. If $|y_0 - y_1| \leq C$, then for all x with $0 \leq x \leq \dfrac{t_1 - t_0}{J|y_1 - y_0|}$,*

$$\begin{aligned} &\mathbb{P}\left[Q(t_0) = y_0, Q(t_1) = y_1, \sup_{t\in[t_0,t_1]} |Q(t) - y_0| \leq C + J\right] \\ \geq \ & p^{J|y_1-y_0|}\left(1 - e^{-q_* x}\right)^{J|y_1-y_0|} e^{-q^*(t_1-t_0)}. \end{aligned}$$

Proof : It is sufficient to exhibit some particular event which has the desirable properties. By Lemma 4.28, there exists some path in $\mathbb{Z}_+^N$,

$$i_1 = y_0, \ldots, i_k = y_1$$

with length less than $J|y_0 - y_1|$ and the embedded Markov chain associated to X follows this path with probability larger than $p^{J|y_0-y_1|}$. Moreover

$$|i_j - y_0| \le |y_1 - y_0| + J \le C + J, \ \forall j = 1, \dots, k.$$

Now, the process is forced to stay in each state i_j a time less than x. This is done with probability larger than $\left(1 - e^{-q_* x}\right)^{J|y_1-y_0|}$ since there are at most $J|y_0 - y_1|$ jumps. When the process has reached y_1, it is forced to stay in y_1 a time larger than $t_1 - t_0$. This is done with probability larger than $e^{-q^*(t_1-t_0)}$. Hence, the probability of such event for X is larger than $p^{J|y_1-y_0|}\left(1 - e^{-q_* x}\right)^{J|y_1-y_0|}e^{-q^*(t_1-t_0)}$. This achieves the proof. ■

Using the preceding lemma, one can deduce some bound on the probability that X stays in some tube centered around some affine function.

Lemma 4.30 *Let $0 \le t_0 \le t_1$, $x_0 \in \mathbb{R}_+^N$ and $D \in \mathbb{R}^N$ satisfying $x_0 + D(t - t_0) \in \mathbb{R}_+^N$, $\forall t$ with $t_0 \le t \le t_1$. Denote φ, the function defined by $\varphi(t) = x_0 + D(t - t_0)$. Set $x_1 \stackrel{\text{def}}{=} \varphi(t_1)$. Assume*

$$|x_0 - y_0| \wedge |x_1 - y_1| < |D|(t_1 - t_0) < \frac{C}{3}.$$

Then

$$\begin{aligned} &\mathbb{P}\left[Q(t_0) = y_0, Q(t_1) = y_1, \sup_{t\in[t_0,t_1]} |Q(t) - \varphi(t)| \le 2C + J\right] \\ \ge \ & p^{3|D|J(t_1-t_0)}\left(1 - e^{-\frac{q_*}{3|D|J}}\right)^{3|D|J(t_1-t_0)} e^{-q^*(t_1-t_0)} \stackrel{\text{def}}{=} \psi(D)^{t_1-t_0}. \end{aligned}$$

Proof : We have

$$|y_1 - y_0| < 3|D|(t_1 - t_0) < C. \tag{4.45}$$

If $0 \le x \le \dfrac{t_1 - t_0}{J|y_1 - y_0|}$, Lemma 4.29 and the triangular inequality imply

$$\begin{aligned} &\mathbb{P}\left[Q(t_0) = y_0, Q(t_1) = y_1, \sup_{t\in[t_0,t_1]} |Q(t) - \varphi(t)| \le 2C + J\right] \\ \ge \ & p^{J|y_1-y_0|}\left(1 - e^{-q_* x}\right)^{J|y_1-y_0|} e^{-q^*(t_1-t_0)}, \\ \ge \ & p^{3|D|J(t_1-t_0)}\left(1 - e^{-q_* x}\right)^{3|D|J(t_1-t_0)} e^{-q^*(t_1-t_0)}, \end{aligned}$$

where the last inequality follows by (4.45) and since $0 < p < 1$ and $0 < 1 - e^{-q_* x} < 1$. The proof is concluded remarking that (4.45) implies $\dfrac{1}{3|D|J} \le \dfrac{t_1 - t_0}{J|y_1 - y_0|}$. ■

Proof of Proposition 4.22 If $x+Dt \in \Lambda(D)$, $\forall t \in [0,\tau]$, then the result is nothing than Theorem 4.1. Nevertheless, it is known only that $x+Dt \in \Lambda(D)$ in the open interval $]0,\tau[$. In order to circumvent this difficulty, fix Δ positive with $2\Delta < \tau$. In $[n\Delta, n(\tau-\Delta)]$, we will apply Theorem 4.1 and then let Δ tends to 0. More precisely, if $\delta' > \delta$, then

$$\mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] \leq$$

$$\mathbb{P}\left[\sup_{t\in[0,n\Delta]} |Q(t,y) - nx - Dt| < \delta n, \sup_{t\in[n\Delta,n(\tau-\Delta)]} |Q(t,y) - nx - Dt| < \delta' n\right].$$

Letting n tends to infinity and then successively ϵ, δ, δ' to 0, Theorem 4.1 yield

$$\begin{aligned} &\lim_{\delta\to 0}\lim_{\epsilon\to 0}\limsup_{n\to\infty} \frac{1}{n} \sup_{|y-nx|<\epsilon n} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] \\ \leq\ & -(\tau - 2\Delta) L\big(\Lambda(D), D\big). \end{aligned}$$

Letting Δ tends to 0, one obtains

$$\begin{aligned} &\lim_{\delta\to 0}\lim_{\epsilon\to 0}\limsup_{n\to\infty} \frac{1}{n} \sup_{|y-nx|<\epsilon n} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] \\ \leq\ & -\tau L\big(\Lambda(D), D\big). \end{aligned} \tag{4.46}$$

Let us prove

$$\begin{aligned} &\lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty} \frac{1}{n} \inf_{|y-nx|<\epsilon n} \log \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] \\ \geq\ & -\tau L\big(\Lambda(D), D\big). \end{aligned} \tag{4.47}$$

In $[n\Delta, n(\tau-\Delta)]$, we will apply Theorem 4.1 and in $[0,n\Delta] \cup [n(\tau-\Delta), n\tau]$, Lemma 4.30. Put $2C_n + J = \delta n$. Moreover, taking into account the order in which the limits are taken, assume without loss of generality that $J \leq \frac{1}{2}\delta n$. In order to use Lemma 4.30, $[0, n\Delta]$ will be divided in K_δ intervals,

$$\left[\frac{(i-1)\Delta n}{K_\delta}, \frac{i\Delta n}{K_\delta}\right], \quad i = 1, \dots, K_\delta, \quad \text{with } K_\delta \stackrel{\text{def}}{=} \left\lfloor \frac{12\Delta|D|}{\delta} \right\rfloor + 1.$$

We apply Lemma 4.30 on each interval with $2C_n + J$ and $n\Delta/K_\delta$ replacing respectively $2C + J$ and $t_1 - t_0$. Note that

$$|D|\frac{n\Delta}{K_\delta} \leq \frac{\delta n}{12} \leq \frac{C_n}{3}.$$

Now, each process $\{Q(nt, y, s),\ t \geq 0\}$ is forced to be at time $i\Delta n/K_\delta$ at some point $x_i^{(N)}$, $i = 0, \ldots, K_\delta$. $x_i^{(N)}$ is chosen in such a way that

$$\begin{aligned} x_0^{(N)} &= y, \\ \left| x_i^{(N)} - x - D\frac{i\Delta n}{K_\delta} \right| &\leq N,\ i = 1, \ldots, K_\delta. \end{aligned}$$

Taking into account the order in which the limits are taken, assume

$$\epsilon n \leq |D| \frac{n\Delta}{K_\delta} \leq \frac{\delta n}{12} \leq \frac{C_n}{3}.$$

Then, Lemma 4.30 and the Markov property give

$$\mathbb{P}\left[\sup_{t \in [0, n\tau]} |Q(t, y) - nx - Dt| < \delta n \right] \geq$$
$$\psi(D)^{\Delta n} \mathbb{P}\left[|Q(\Delta n) - x + D\Delta n| < N, \sup_{t \in [n\Delta, n\tau]} |Q(t, y) - nx - Dt| < \delta n \right].$$

Let $\delta' < \delta$ satisfies

$$\delta' n \leq |D| \frac{n\Delta}{K_\delta} \leq \frac{\delta n}{12} \leq \frac{C_n}{3}.$$

Again Lemma 4.30 and the Markov property give

$$\mathbb{P}\left[\sup_{t \in [0, n\tau]} |Q(t, y) - nx - Dt| < \delta n \right] \geq$$
$$\psi(D)^{2\Delta n} \mathbb{P}\left[|Q(\Delta n) - x + D\Delta n| < N, \sup_{t \in [n\Delta, n(\tau - \Delta)]} |Q(t, y) - nx - Dt| < \delta' n \right].$$

Letting n tends to infinity and then successively ϵ, δ', δ to 0, one gets using Theorem 4.1

$$\begin{aligned} &\lim_{\delta \to 0} \lim_{\epsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \inf_{|y - nx| < \epsilon n} \log \mathbb{P}\left[\sup_{t \in [0, n\tau]} |Q(t, y) - nx - Dt| < \delta n \right] \\ \geq\ & 2\Delta \log \psi(D) - L(\Lambda(D), D)(\tau - 2\Delta). \end{aligned}$$

Let Δ tends to 0 in order to obtain (4.47). Then (4.46) and (4.47) give the result. Proposition 4.22 is proved. ■

Proof of Proposition 4.23 This can be derived in a straight way using Proposition 4.22 and the Markov property. ■

4.B Proof of Proposition 4.3

Let us prove equation (4.5). Take τ satisfying $x_i + D_i\tau > 0$, $\forall i \in \Lambda(x)$. Then for ϵ sufficiently small and all y with $|y - nx| < \epsilon$

$$\begin{aligned} & \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q(t,y) - nx - Dt| < \delta n\right] \\ = \; & \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q^{\Lambda(x)}(t,y) - nx - Dt| < \delta n\right]. \end{aligned}$$

Then, note that

$$\begin{aligned} & \lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n\tau}\inf_{|y-nx|<n\epsilon}\log\mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q(t,y)-nx-Dt|<\delta n\right] \\ = \; & \lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n\tau}\inf_{|y-nx|<n\epsilon}\log\mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q^{\Lambda(x)}(t,y)-nx-Dt|<\delta n\right], \\ = \; & \lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n\tau}\inf_{|y|<n\epsilon}\log\mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q^{\Lambda(x)}(t,y)-Dt|<\delta n\right], \end{aligned}$$

where the last relation holds because the distribution of $X^{\Lambda}_{y,s}$ is invariant with respect to the shift nx since $x_i = 0$, $\forall i \in \Lambda(x)^c$. Hence, it is sufficient to prove that

$$\begin{aligned} & \lim_{\delta\to 0}\lim_{\epsilon\to 0}\liminf_{n\to\infty}\frac{1}{n\tau}\inf_{|y|<n\epsilon}\log\mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q^{\Lambda(x)}(t,y)-Dt|<\delta n\right] \\ & \geq \lim_{\delta\to 0}\liminf_{n\to\infty}\frac{1}{n\tau}\log\mathbb{P}\left[\sup_{t\in[0,n\tau]}|Q^{\Lambda(x)}(t,0)-Dt|<\delta n\right]. \end{aligned}$$

Taking into account the order in which the limits are taken, assume without loss of generality that $\epsilon n + J \leq \frac{1}{2}\delta n$. Hence, by Lemma 4.29

$$\begin{aligned} & \inf_{|y|<n\epsilon}\mathbb{P}\left[Q^{\Lambda}(\epsilon n, y, s) = 0, \sup_{t\in[0,\epsilon n]}|Q^{\Lambda}(t,y,s)| \leq \frac{n\delta}{2}\right] \\ \geq \; & p^{J\epsilon n}\left(1 - e^{-\frac{q_*}{J}}\right)^{J\epsilon n} e^{-q^*\epsilon n} \stackrel{\text{def}}{=} \tilde{\gamma}^{\epsilon n}. \end{aligned} \tag{4.48}$$

Finally, if $|n\epsilon D| < n\delta/2$, this being in no way a restriction, (4.48) and the Markov property yield

$$\begin{aligned}
&\inf_{|y|<n\epsilon} \mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q^\Lambda(t,y) - Dt| < \delta n\right] \\
&\quad \geq \inf_{|y|<n\epsilon} \mathbb{P}\left[Q^\Lambda(\epsilon n, y, s) = 0, \sup_{t\in[0,\epsilon n]} |Q^\Lambda(t,y,s)| \leq \frac{n\delta}{2}\right] \\
&\qquad\qquad\qquad \times \mathbb{P}\left[\sup_{t\in[0,n(\tau-\epsilon)]} |Q^\Lambda(t,0,s) - Dt - D\epsilon n| \leq n\delta\right], \\
&\quad \geq \tilde{\gamma}^{\epsilon n}\mathbb{P}\left[\sup_{t\in[0,n\tau]} |Q^\Lambda(t,0,s) - Dt| \leq \frac{n\delta}{2}\right].
\end{aligned}$$

Taking the different limits, one can get (4.5). The proof of (4.6) is similar and is therefore omitted. ■

4.C Proof of Proposition 4.21 (iv)

First, one can assume that $|\dot{\varphi}(t)| \leq K$, a.e, for some K. Indeed,

Lemma 4.31 *Let $\varphi \in \mathcal{AC}\big([0,T],\mathbb{R}_+^N\big)$ with $I_T(\varphi) < \infty$. Then, for all $\epsilon > 0$, there exist $K_\epsilon > 0$, $\varphi_\epsilon \in \mathcal{AC}\big([0,T],\mathbb{R}_+^N\big)$ such that:*

(i). $|\dot{\varphi}_\epsilon(t)| \leq K_\epsilon$, *a.e,*

(ii). $d_c(\varphi_\epsilon,\varphi) \leq \epsilon$,

(iii). $I_T(\varphi_\epsilon) \leq I_T(\varphi) + \epsilon$.

Proof : The demonstration is achieved by introducing a time change as in [57, Lemma 6.5.3]. The detailed proof is omitted. ■

In order to prove the assertion (iv) of Proposition 4.21, it is sufficient to construct a sequence $\psi^{(n)}$ in $\mathcal{PL}\big([0,T],\mathbb{R}_+^N\big)$ with the following properties:

(i). $\lim_{n\to\infty} d_c(\psi^{(n)},\varphi) = 0$,

(ii). $\displaystyle\lim_{n\to\infty} \int_0^T \mathbb{1}_{\{\Lambda(\psi^{(n)}(t))\neq\Lambda(\varphi(t))\}} dt = 0$,

(iii). $|\dot{\psi}^{(n)}(t)| \leq K$, a.e.

(iv). $\lim_{n\to\infty} \dot{\psi}^{(n)}(t) = \dot{\varphi}(t)$, a.e.

Indeed, if these assumptions are fulfilled, then

$$
\begin{aligned}
|I_T(\psi^{(n)}) - I_T(\varphi)| \;\; &\le \;\; \sup_{\Lambda} \int_0^T |L(\Lambda, \dot{\psi}^{(n)}(t)) - L(\Lambda, \dot{\varphi}(t))| dt \\
&+ \;\; 2 \sup_{x \le K} \sup_{\Lambda} \big(L(\Lambda, x) \big) \int_0^T \mathbb{1}_{\{\Lambda(\psi^{(n)}(t)) \ne \Lambda(\varphi(t))\}} dt.
\end{aligned}
$$

Since $L(\Lambda, x)$ is continuous with respect to x, the first term tends to 0 using properties (iii), (iv) and dominated convergence theorem. The second one tends to 0 by (ii).

In a first step, one constructs a sequence $\varphi^{(n)} \in \mathcal{PL}\big([0,T], \mathbb{R}_+^N\big)$ satisfying (i), (ii), (iii). As it will emerge, one can prove that $\dot{\varphi}^{(n)}$ tends to $\dot{\varphi}$ only for the weak topology in $\mathcal{L}^1\big([0,T], \mathbb{R}^N\big)$. In a second step, the use of Mazur's theorem, see [53, corollary 14] page 422, allows one to construct from $\{\varphi^{(n)}, n \ge 1\}$ a sequence $\{\psi^{(n)}, n \ge 1\}$ satisfying (i) to (iv). More precisely:

Lemma 4.32 *There exists $\{\varphi^{(n)}, n \ge 1\}$ with $\varphi^{(n)} \in \mathcal{PL}\big([0,T], \mathbb{R}_+^N\big)$ satisfying:*

(i). $\lim_{n\to\infty} d_c(\varphi^{(n)}, \varphi) = 0$.

(ii). The sequence $\{\varphi^{(n)}, n \ge 1\}$ satisfies

$$
\lim_{n\to\infty}\!\uparrow \big\{t \in [0,T] : \varphi_k^{(n)}(t) > 0\big\} = \big\{t : \varphi_k(t) > 0\big\}, \quad \forall k \in \mathcal{S}.
$$

As a consequence

$$
\begin{aligned}
&\lim_{n\to\infty}\!\downarrow \big\{t \in [0,T] : \Lambda(\varphi(t)) \ne \Lambda(\varphi^{(n)}(t))\big\} \\
= \;\; &\lim_{n\to\infty}\!\downarrow \bigcup_{k \in \mathcal{S}} \Big(\big\{t : \varphi_k(t) > 0\big\} / \big\{t : \varphi_k^{(n)}(t) > 0\big\} \Big)
\end{aligned}
$$

and

$$
\lim_{n\to\infty} \int_0^T \mathbb{1}_{\{\Lambda(\varphi(t)) \ne \Lambda(\varphi^{(n)}(t))\}} dt = 0.
$$

(iii). $|\dot{\varphi}_k^{(n)}(t)| \le K$, *a.e,* $\forall k \in \mathcal{S}$.

(iv). $\dot{\varphi}^{(n)}$ *tends to* $\dot{\varphi}$ *for the weak topology in* $\mathcal{L}^1\big([0,T], \mathbb{R}^N\big)$.

Proof : $\{\varphi^{(n)}, n \geq 1\}$ is constructed component by component. φ being continuous, $O_k \stackrel{\text{def}}{=} \{t : \varphi_k(t) > 0\}$ is open and can be written as a countable union of disjoint open intervals,

$$O_k = \sum_{i \geq 0} \Big( ]a_k(i), b_k(i)[\Big),$$

with $\varphi_k\big(a_k(i)\big) = \varphi_k\big(b_k(i)\big) = 0$. In the following, $O_k^{(n)}$ will stand for the partial sum $\sum_{i=0}^{n} \big(]a_k(i), b_k(i)[\big)$. Denote

$$\eta_k = \int_0^T \mathbb{I}_{\{O_k\}}(t)dt \quad \text{and} \quad \eta_k^{(n)} = \int_0^T \mathbb{I}_{\{O_k^{(n)}\}}(t)dt.$$

In order to define $\varphi^{(n)}$ in $O_k^{(n)}$, take some grid $\{t_k^{(n)}(i)\}_i$ in $O_k^{(n)}$. Then in $O_k^{(n)}$, $\varphi^{(n)}$ is defined as the interpolation of φ with respect to the grid $\{t_k^{(n)}(i)\}_i$. Moreover, owing to the continuity of φ, it is possible to choose the grid in such a way that

$$|\varphi_k^{(n)}(t) - \varphi_k(t)| \leq \frac{1}{n}, \quad \forall t \in O_k(n).$$

For $t \notin O_k^{(n)}$, put $\varphi_k^{(n)}(t) = 0$. Besides, since $\varphi_k\big(a_k(i)\big) = 0$ and $|\dot{\varphi}_k(t)| \leq K$ a.e, for $t \in O_k/O_k^{(n)}$,

$$\begin{aligned}
&= |\varphi_k^{(n)}(t) - \varphi_k(t)| \\
&= |\varphi_k(t)| \leq \sum_{i > n} \mathbb{I}_{\{t \in]a_k(i), b_k(i)[\}} \Big(\varphi_k\big(a_k(i)\big) + \int_{a_k(i)}^t \dot{\varphi}_k(u)du \Big), \\
&\leq K \int_0^T \mathbb{I}_{\{O_k/O_k^{(n)}\}}(u)du = K(\eta_k - \eta_k^{(n)}),
\end{aligned}$$

where $\eta_k - \eta_k^{(n)}$ goes to 0 as n tends to infinity. Hence, the preceding discussion implies (i).

On a other hand, it is clear by construction that $O_k^{(n)}$ is increasing with n to O_k. Hence (ii) follows. Since $\varphi_k^{(n)}$ is the interpolation of φ_k and $|\dot{\varphi}_k|$ is bounded by K a.e, there is (iii).

Since $|\dot{\varphi}^{(n)}(t)| \leq K$ a.e, for all n and $L(\Lambda, D)$ is continuous with respect to D, $I_T(\varphi^{(n)})$ is uniformly bounded. Using Lemma 4.18, it is clear that

$$\lim_{r \to \infty} \sup_n \int_0^T \mathbb{I}_{\{|\dot{\varphi}^{(n)}(t)| \geq r\}} |\dot{\varphi}^{(n)}(t)| dt = 0.$$

Hence, quite classically, one obtains: $\forall \epsilon > 0, \exists \delta > 0$ such that

$$\int_0^T \mathbb{I}_{\{A\}}(t)dt \leq \delta \implies \int_0^T \mathbb{I}_{\{A\}}(t)|\dot{\varphi}^{(n)}(t)|dt < \epsilon, \quad \forall n.$$

Then, Dunford-Pettis theorem (see [53, corollary 11 p. 294]) implies that $\{\dot{\varphi}^{(n)}, n \geq 1\}$ is relatively compact for the weak topology in $\mathcal{L}^1\big([0,T], \mathbb{R}^N\big)$. Let f be a limiting point of $\{\dot{\varphi}^{(n)}, n \geq 1\}$. First,

$$\lim_{n\to\infty} \varphi^{(n)}(t_2) - \varphi^{(n)}(t_1) = \lim_{n\to\infty} \int_{t_1}^{t_2} \dot{\varphi}^{(n)}(s)ds = \int_{t_1}^{t_2} f(s)ds, \quad \forall t_1, t_2.$$

Since $d_c(\varphi^{(n)}, \varphi)$ tends to 0,

$$\varphi(t_2) = \varphi(t_1) + \int_{t_1}^{t_2} f(s)ds, \quad \forall t_1, t_2.$$

Radon Nikodym's theorem ensures that $f(t) = \dot{\varphi}(t)$ a.e. So $\{\dot{\varphi}^{(n)}, n \geq 1\}$ converges to $\dot{\varphi}$ for the weak topology in $\mathcal{L}^1\big([0,T], \mathbb{R}^N\big)$. ■

Proof of Proposition 4.21 (iv) : Now, let us construct $\{\psi^{(n)}, n \geq 1\}$ from $\{\varphi^{(n)}, n \geq 1\}$. Since $\{\dot{\varphi}^{(n)}, n \geq 1\}$ tends to $\dot{\varphi}$ for the weak topology in $\mathcal{L}^1\big([0,T], \mathbb{R}^N\big)$, Mazur's theorem implies that it exists a sequence $\{f^{(n)}, n \geq 1\}$, each $f^{(n)}$ being a convex combination of the $\dot{\varphi}^{(n)}$ such that $\{f^{(n)}, n \geq 1\}$ converges strongly to $\dot{\varphi}$ in $\mathcal{L}^1\big([0,T], \mathbb{R}^N\big)$. Denote

$$f^{(n)} = \sum_{k=1}^{s(n)} \alpha_{nk} \dot{\varphi}^{(m_{nk})} \quad \text{with} \quad \sum_{k=1}^{s(n)} \alpha_{nk} = 1, \alpha_{nk} \geq 0.$$

Set

$$\psi^{(n)} = \sum_{k=1}^{s(n)} \alpha_{nk} \varphi^{(m_{nk})}, \quad \forall n.$$

$\psi^{(n)} \in \mathcal{PL}\big([0,T], \mathbb{R}_+^N\big)$ and clearly $\dot{\psi}^{(n)} = f^{(n)}$. Hence, possibly extracting some subsequence, one can suppose that $\dot{\psi}^{(n)}$ converges to $\dot{\varphi}$ a.e. Without loss of generality, one can assume that $\lim_{n\to\infty} \min_k m_{nk} = \infty$. Then $d_c(\psi^{(n)}, \varphi)$ tends to 0 with n. Moreover, it is clear that $|\dot{\psi}^{(n)}(t)| \leq K$, a.e. Finally, if one denotes $m_n^* = \max_k m_{nk}$, using (ii) of Lemma 4.32, one obtains

$$\big\{t \in [0,T] : \Lambda(\psi^{(n)}(t)) \neq \Lambda(\varphi(t))\big\} = \big\{t \in [0,T] : \Lambda(\varphi^{(m_n^*)}(t)) \neq \Lambda(\varphi(t))\big\}.$$

Hence, again by (ii) of Lemma 4.32,

$$\lim_{n\to\infty} \int_0^T \mathbb{I}_{\{\Lambda(\psi^{(n)}(t)) \neq \Lambda(\varphi(t))\}} dt = 0.$$

The construction of $\{\psi^{(n)}, n \geq 1\}$ is achieved. ■

Bibliographie

[1] Albert, A. Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Ann. Math. Statist. 33* (1962), 727–753.

[2] Anantharam, V. How large delays build up in a $GI/G/1$ queue. *Queueing Systems Theory Appl. 5*, 4 (1989), 345–367.

[3] Ash, R. *Information theory.* Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1965. Interscience Tracts in Pure and Applied Mathematics, No. 19.

[4] Atar, R., and Dupuis, P. Large deviations and queueing networks: methods for rate function identification. *Stochastic Process. Appl. 84*, 2 (1999), 255–296.

[5] Azencott, R. Grandes déviations et applications. In *Eighth Saint Flour Probability Summer School—1978 (Saint Flour, 1978).* Springer, Berlin, 1980, pp. 1–176.

[6] Azencott, R., and Ruget, G. Mélanges d'équations différentielles et grands écarts à la loi des grands nombres. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 38*, 1 (1977), 1–54.

[7] Bahadur, R. R. On the asymptotic efficiency of tests and estimates. *Sankhyā 22* (1960), 229–252.

[8] Bahadur, R. R., and Ranga Rao, R. On deviations of the sample mean. *Ann. Math. Statist. 31* (1960), 1015–1027.

[9] Bahadur, R. R., and Zabell, S. L. Large deviations of the sample mean in general vector spaces. *Ann. Probab. 7*, 4 (1979), 587–621.

[10] Baldi, P., and Piccioni, M. A representation formula for the large deviation rate function for the empirical law of a continuous time Markov chain. *Statist. Probab. Lett. 41*, 2 (1999), 107–115.

[11] Barbu, V., and Precupanu, T. *Convexity and optimization in Banach spaces*, second ed., vol. 10 of *Mathematics and its Applications (East European Series).* D. Reidel Publishing Co., Dordrecht, 1986.

[12] Berge, C. *Graphs*, vol. 6 of *North-Holland Mathematical Library.* North-Holland Publishing Co., Amsterdam, 1985. Second revised edition of part 1 of the 1973 English version.

[13] Blackwell, D., and Kendall, D. The martin boundary of Pólya's urn scheme, and an application to stochastic population growth. *J. Appl. Probability 1* (1964), 284–296.

[14] Bolthausen, E. Laplace approximations for sums of independent random vectors. *Probab. Theory Relat. Fields 72*, 2 (1986), 305–318.

[15] Bolthausen, E. Markov process large deviations in τ-topology. *Stochastic Process. Appl. 25*, 1 (1987), 95–108.

[16] Bolthausen, E., and Schmock, U. On the maximum entropy principle for uniformly ergodic Markov chains. *Stochastic Process. Appl. 33*, 1 (1989), 1–27.

[17] Borovkov, A. A., and Korshunov, D. A. Probabilities of large deviations for one-dimensional Markov chains. I. Stationary distributions. *Teor. Veroyatnost. i Primenen. 41*, 1 (1996), 3–30. Translation in *Theory Probab. Appl,* 41 (1996), 1 1–24.

[18] Borovkov, A. A., and Mogulskii, A. A. Large deviations for stationary Markov chains in a quarter plane. In *Probability theory and mathematical statistics (Tokyo, 1995).* World Sci. Publishing, River Edge, NJ, 1996, pp. 12–19.

[19] Borovkov, A. A., and Schassberger, R. Ergodicity of a polling network. *Stochastic Process. Appl. 50*, 2 (1994), 253–262.

[20] Bourbaki, N. *Éléments de mathématique. Fasc. XV. Livre V: Espaces vectoriels topologiques. Chapitre I: Espaces vectoriels topologiques sur un corps valué. Chapitre II: Ensembles convexes et espaces localement convexes.* Hermann, Paris, 1966. Actualités Scientifiques et Industrielles, No. 1189. Deuxième édition revue et corrigée.

[21] Boxma, O., and Weststrate, J. Waiting time in polling systems with markovian server routing. In *Messung, Modellierung und Bewertung von Rechensysteme* (1989), Springer-Verlag, pp. 89–104.

[22] Bucklew, J. A., Ney, P., and Sadowsky, J. S. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Probab. 27*, 1 (1990), 44–59.

[23] Cattiaux, P., and Léonard, C. Minimization of the Kullback information for some Markov processes. In *Séminaire de Probabilités, XXX*, vol. 1626 of *Lecture Notes in Math.* Springer, Berlin, 1996, pp. 288–311.

[24] Comets, F., and Zeitouni, O. Information estimates and Markov random fields. *Markov Process. Related Fields 5*, 3 (1999), 269–291.

[25] Csiszár, I., Cover, T. M., and Choi, B. S. Conditional limit theorems under Markov conditioning. *IEEE Trans. Inform. Theory 33*, 6 (1987), 788–801.

[26] Csiszár, I., and Körner, J. *Coding theorems for discrete memoryless systems.* Probability and Mathematical Statistics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1981.

[27] de Acosta, A. Upper bounds for large deviations of dependent random vectors. *Z. Wahrsch. Verw. Gebiete 69*, 4 (1985), 551–565.

[28] de Acosta, A. Large deviations for empirical measures of Markov chains. *J. Theoret. Probab. 3*, 3 (1990), 395–431.

[29] de Acosta, A. Moderate deviations for empirical measures of Markov chains: lower bounds. *Ann. Probab. 25*, 1 (1997), 259–284.

[30] de Acosta, A., and Chen, X. Moderate deviations for empirical measures of Markov chains: upper bounds. *J. Theoret. Probab. 11*, 4 (1998), 1075–1110.

[31] de Acosta, A., and Ney, P. Large deviation lower bounds for arbitrary additive functionals of a Markov chain. *Ann. Probab. 26*, 4 (1998), 1660–1682.

[32] de La Fortelle, A. Large deviation principle for Markov chains in continuous time. Tech. Rep. 3877, INRIA, 2000. To appear in *Problems of Information Transmission.*

[33] de La Fortelle, A., and Fayolle, G. Large deviation principle for a Markov chain with a countable state space. Tech. Rep. 3503, INRIA, 1998.

[34] DE LA FORTELLE, A., AND FAYOLLE, G. Large deviation principle for Markov chains in discrete time. Tech. Rep. 3791, INRIA, 1999.

[35] DELCOIGNE, F., AND DE LA FORTELLE, A. Large deviations for polling systems. Tech. Rep. 3892, INRIA, 2000.

[36] DELLACHERIE, C., AND MEYER, P.-A. *Probabilités et potentiel. Chapitres IX à XI*, revised ed. Hermann, Paris, 1983. Théorie discrète du potentiel. [Discrete potential theory], Actualités Scientifiques et Industrielles [Current Scientific and Industrial Topics], 1410.

[37] DEMBO, A., AND KUELBS, J. Refined Gibbs conditioning principle for certain infinite-dimensional statistics. *Studia Sci. Math. Hungar. 34*, 1-3 (1998), 107–126.

[38] DEMBO, A., AND ZEITOUNI, O. *Large deviations techniques and applications.* Jones and Bartlett Publishers, Boston, MA, 1993.

[39] DEMBO, A., AND ZEITOUNI, O. Refinements of the Gibbs conditioning principle. *Probab. Theory Related Fields 104*, 1 (1996), 1–14.

[40] DEMBO, A., AND ZEITOUNI, O. *Large deviations techniques and applications*, second ed. Springer-Verlag, New York, 1998.

[41] DEUSCHEL, J.-D., AND STROOCK, D. W. *Large deviations*, vol. 137 of *Pure and Applied Mathematics.* Academic Press Inc., Boston, MA, 1989.

[42] DINWOODIE, I. H. Identifying a large deviation rate function. *Ann. Probab. 21*, 1 (1993), 216–231.

[43] DINWOODIE, I. H., AND NEY, P. Occupation measures for Markov chains. *J. Theoret. Probab. 8*, 3 (1995), 679–691.

[44] DOBRUSHIN, R. L., AND PECHERSKY, E. A. Large deviations for tandem queueing systems. *J. Appl. Math. Stochastic Anal. 7*, 3 (1994), 301–330.

[45] DOBRUSHIN, R. L., AND PECHERSKY, E. A. Large deviations for random processes with independent increments on infinite intervals. In *Probability theory and mathematical statistics (St. Petersburg, 1993).* Gordon and Breach, Amsterdam, 1996, pp. 41–74.

[46] Donsker, M. D., and Varadhan, S. R. S. On some problems of large deviations for Markov processes. In *Proceedings of the 40th Session of the International Statistical Institute (Warsaw, 1975), Vol. 1. Invited papers* (1975), vol. 46, pp. 409–416, 417–418 (1976). With discussion.

[47] Donsker, M. D., and Varadhan, S. R. S. Asymptotic evaluation of certain Markov process expectations for large time. I, II, III. *Comm. Pure Appl. Math. 28* (1975, 1976), 1–47; **28** 279–301; **29** 389–461.

[48] Donsker, M. D., and Varadhan, S. R. S. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Comm. Pure Appl. Math. 36*, 2 (1983), 183–212.

[49] Doob, J. L. Discrete potential theory and boundaries. *J. Math. Mech. 8* (1959), 433–458; erratum 993.

[50] Doob, J. L. *Classical potential theory and its probabilistic counterpart*, vol. 262 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science].* Springer-Verlag, New York, 1984.

[51] Doob, J. L., Snell, J. L., and Williamson, R. E. Application of boundary theory to sums of independent random variables. In *Contributions to probability and statistics.* Stanford Univ. Press, Stanford, Calif., 1960, pp. 182–197.

[52] Down, D. On the stability of polling models with multiple servers. *J. Appl. Probab. 35*, 4 (1998), 925–935.

[53] Dunford, N., and Schwartz, J. T. *Linear Operators. I. General Theory.* Interscience Publishers, Inc., New York, 1958. With the assistance of W. G. Bade and R. G. Bartle. Pure and Applied Mathematics, Vol. 7.

[54] Dupuis, P., and Ellis, R. S. Large deviations for Markov processes with discontinuous statistics. II. Random walks. *Probab. Theory Related Fields 91*, 2 (1992), 153–194.

[55] Dupuis, P., and Ellis, R. S. Large deviation analysis of queueing systems. In *Stochastic networks.* Springer, New York, 1995, pp. 347–365.

[56] Dupuis, P., and Ellis, R. S. The large deviation principle for a general class of queueing systems. I. *Trans. Amer. Math. Soc. 347*, 8 (1995), 2689–2751.

[57] Dupuis, P., and Ellis, R. S. *A weak convergence approach to the theory of large deviations.* John Wiley & Sons Inc., New York, 1997. A Wiley-Interscience Publication.

[58] Dupuis, P., Ellis, R. S., and Weiss, A. Large deviations for Markov processes with discontinuous statistics. I. General upper bounds. *Ann. Probab. 19*, 3 (1991), 1280–1297.

[59] Dupuis, P., Ishii, H., and Soner, H. M. A viscosity solution approach to the asymptotic analysis of queueing systems. *Ann. Probab. 18*, 1 (1990), 226–255.

[60] Ellis, R. S. Large deviations for a general class of random vectors. *Ann. Probab. 12*, 1 (1984), 1–12.

[61] Ellis, R. S. *Entropy, large deviations, and statistical mechanics*, vol. 271 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Science].* Springer-Verlag, New York, 1985.

[62] Ellis, R. S. Large deviations for the empirical measure of a Markov chain with an application to the multivariate empirical measure. *Ann. Probab. 16*, 4 (1988), 1496–1508.

[63] Fayolle, G., Iasnogorodski, R., and Malyshev, V. *Random walks in the quarter-plane.* Springer-Verlag, Berlin, 1999. Algebraic methods, boundary value problems and applications.

[64] Fayolle, G., Malyshev, V. A., and Men′shikov, M. V. *Topics in the constructive theory of countable Markov chains.* Cambridge University Press, Cambridge, 1995.

[65] Feller, W. On the integro-differential equations of purely discontinuous Markoff processes. *Trans. Amer. Math. Soc. 48* (1940), 488–515.

[66] Feller, W. *An introduction to probability theory and its applications. Vol. II.*, second ed. John Wiley & Sons Inc., New York, 1971.

[67] Freĭdlin, M. I., and Wentzell, A. D. *Random perturbations of dynamical systems.* Springer-Verlag, New York, 1984. Translated from the Russian by Joseph Szücs.

[68] Fricker, C., and Jaibi, M. Stability of multi-server polling models. Tech. Rep. 3347, INRIA, 1998.

[69] Gallager, R. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[70] Gärtner, J. On large deviations from an invariant measure. *Teor. Verojatnost. i Primenen. 22*, 1 (1977), 27–42.

[71] Genon-Catalot, V., and Picard, D. *Éléments de statistique asymptotique.* Springer-Verlag, Paris, 1993.

[72] Gīhman, Ĭ. Ī., and Skorohod, A. V. *The theory of stochastic processes. II.* Springer-Verlag, New York, 1975. Translated from the Russian by Samuel Kotz, Die Grundlehren der Mathematischen Wissenschaften, Band 218.

[73] Hamming, R. W. *Coding and information theory.* Prentice-Hall Inc., Englewood Cliffs, N.J., 1980.

[74] Hornik, K. Asymptotically optimal tests of hypotheses on the generator of a finite Markov process. *J. Statist. Plann. Inference 23*, 3 (1989), 345–352.

[75] Hornik, K. Exponential rate optimal estimation of the generator of a finite Markov process. *Statist. Decisions 8*, 2 (1990), 101–113.

[76] Hornik, K. On some large deviation probabilities for finite Markov processes. *Studia Sci. Math. Hungar. 26*, 2-3 (1991), 355–362.

[77] Hunt, G. A. Markoff processes and potentials. *Proc. Nat. Acad. Sci. U.S.A. 42* (1956), 414–418.

[78] Hunt, G. A. Markoff processes and potentials. I, II. *Illinois J. Math. 1* (1957), 44–93, 316–369.

[79] Hunt, G. A. Markoff processes and potentials. III. *Illinois J. Math. 2* (1958), 151–213.

[80] Hunt, G. A. Markoff chains and Martin boundaries. *Illinois J. Math. 4* (1960), 313–340.

[81] Ibragimov, I. A., and Linnik, Y. V. *Independent and stationary sequences of random variables.* Wolters-Noordhoff Publishing, Groningen, 1971. With a supplementary chapter by I. A. Ibragimov and V. V. Petrov, Translation from the Russian edited by J. F. C. Kingman.

[82] Ignatiouk-Robert, I. Large deviations of Jackson networks. Tech. Rep. 14/99, Université de Cergy-Pontoise, 1999. To appear in *Annals of App. Probab.*

[83] Ignatyuk, I. A., Malyshev, V. A., and Shcherbakov, V. V. The influence of boundaries in problems on large deviations. *Uspekhi Mat. Nauk 49*, 2(296) (1994), 43–102.

[84] Iltis, M. Sharp asymptotics of large deviations in $\mathbb{R}^d$. *J. Theoret. Probab. 8*, 3 (1995), 501–522.

[85] Ioffe, A. D., and Tihomirov, V. M. *Theory of extremal problems.* North-Holland Publishing Co., Amsterdam, 1979. Translated from the Russian by Karol Makowski.

[86] Iscoe, I., Ney, P., and Nummelin, E. Large deviations of uniformly recurrent Markov additive processes. *Adv. in Appl. Math. 6*, 4 (1985), 373–412.

[87] Jain, N. C. Large deviation lower bounds for additive functionals of Markov processes. *Ann. Probab. 18*, 3 (1990), 1071–1098.

[88] Kaĭmanovich, V. A., and Vershik, A. M. Random walks on discrete groups: boundary and entropy. *Ann. Probab. 11*, 3 (1983), 457–490.

[89] Kakutani, S. Two-dimensional Brownian motion and harmonic functions. *Proc. Imp. Acad. Tokyo 20* (1944), 706–714.

[90] Kakutani, S. Markoff process and the Dirichlet problem. *Proc. Japan Acad. 21* (1945), 227–233 (1949).

[91] Kemeny, J. G., Snell, J. L., and Knapp, A. W. *Denumerable Markov chains*, second ed. Springer-Verlag, New York, 1976. With a chapter on Markov random fields, by Griffeath, D., *Graduate Texts in Mathematics* 40.

[92] Kesidis, G., and Walrand, J. Relative entropy between Markov transition rate matrices. *IEEE Trans. Inform. Theory 39*, 3 (1993), 1056–1057.

[93] Kullback, S., and Leibler, R. A. On information and sufficiency. *Ann. Math. Statistics 22* (1951), 79–86.

[94] Kurkova, I. A., and Malyshev, V. A. Martin boundary and elliptic curves. *Markov Process. Related Fields 4*, 2 (1998), 203–272.

[95] Lanford, O. Entropy and equilibrium states in classical statistical mechanics. *Lecture Notes in Phys. 20* (1973), 1–113.

[96] Lay, S. R. *Convex sets and their applications.* Robert E. Krieger Publishing Co. Inc., Malabar, FL, 1992. Revised reprint of the 1982 original.

[97] Lewis, J., and Pfister, C. Thermodynamic probability theory: some aspects of large deviations. *Uspekhi Mat. Nauk 50*, 2(302) (1995), 47–88. Translation in *Russian Math. Surveys* 50 (1995), No 2, 279–317.

[98] Lindqvist, B. On the loss of information incurred by lumping states of a Markov chain. *Scand. J. Statist. 5*, 2 (1978), 92–98.

[99] Linnik, J. V. An information-theoretic proof of the central limit theorem with Lindeberg conditions. *Teor. Verojatnost. i Primenen 4* (1959), 311–321.

[100] Liptser, R. S. Large deviations for empirical measures of Markov processes (discrete time, the noncompact case). *Teor. Veroyatnost. i Primenen. 41*, 1 (1996), 65–88.

[101] Majewski, K. Large deviations of the steady-state distribution of reflected processes with applications to queueing systems. *Queueing Systems Theory Appl. 29*, 2-4 (1998), 351–381.

[102] Martin, R. S. Minimal positive harmonic functions. *Trans. Amer. Math. Soc. 49* (1941), 137–172.

[103] Meda, A., and Ney, P. The Gibbs conditioning principle for Markov chains. In *Perplexing problems in probability.* Birkhäuser Boston, Boston, MA, 1999, pp. 385–398.

[104] Miclo, L. Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de Markov finies. In *Séminaire de Probabilités, XXXI*, vol. 1655 of *Lecture Notes in Math.* Springer, Berlin, 1997, pp. 136–167.

[105] Milgram, M. Barycentric entropy of Markov systems. *Cybernet. Systems 12*, 1-2 (1981), 141–178.

[106] Miller, H. D. A convexivity property in the theory of random variables defined on a finite Markov chain. *Ann. Math. Statist. 32* (1961), 1260–1270.

[107] Miller, H. D. A convexivity property in the theory of random variables defined on a finite Markov chain. *Ann. Math. Statist. 32* (1961), 1260–1270.

[108] Molčanov, S. A. Martin boundary for the direct product of Markov chains. *Teor. Verojatnost. i Primenen 12* (1967), 353–358.

[109] Molčanov, S. A. Martin boundary of the direct product of Markov processes. *Uspehi Mat. Nauk 22*, 2 (134) (1967), 125–126.

[110] Molčanov, S. A. The Martin boundary of a direct product of Markov processes. *Sibirsk. Mat. Ž. 11* (1970), 370–380.

[111] Ney, P. Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$. *Ann. Probab. 11*, 1 (1983), 158–167.

[112] Ney, P., and Nummelin, E. Markov additive processes: large deviations for the continuous time case. In *Probability theory and mathematical statistics, Vol. II (Vilnius, 1985).* VNU Sci. Press, Utrecht, 1987, pp. 377–389.

[113] Ney, P., and Spitzer, F. The Martin boundary for random walk. *Trans. Amer. Math. Soc. 121* (1966), 116–132.

[114] Northshield, S. On the spectrum and Martin boundary of homogeneous spaces. *Statist. Probab. Lett. 22*, 4 (1995), 275–279.

[115] Nummelin, E. *General irreducible Markov chains and nonnegative operators.* Cambridge University Press, Cambridge, 1984.

[116] O'Brien, G. L., and Vervaat, W. Capacities, large deviations and loglog laws. In *Stable processes and related topics (Ithaca, NY, 1990).* Birkhäuser Boston, Boston, MA, 1991, pp. 43–83.

[117] Orey, S. Large deviations in ergodic theory. In *Seminar on stochastic processes, 1984 (Evanston, Ill., 1984).* Birkhäuser Boston, Boston, Mass., 1986, pp. 195–249.

[118] Overbeck, L. Martin boundaries of some branching processes. *Ann. Inst. H. Poincaré Probab. Statist. 30*, 2 (1994), 181–195.

[119] Picardello, M., and Woess, W. Examples of stable Martin boundaries of Markov chains. In *Potential theory (Nagoya, 1990)*. de Gruyter, Berlin, 1992, pp. 261–270.

[120] Picardello, M. A., and Woess, W. Martin boundaries of Cartesian products of Markov chains. *Nagoya Math. J. 128* (1992), 153–169.

[121] Rényi, A. *Calcul des probabilités. Avec un appendice sur la théorie de l'information*. Dunod, Paris, 1966. Traduit de l'allemand par C. Bloch. Collection Universitaire de Mathématiques, No. 21.

[122] Revuz, D. *Markov chains*. North-Holland Publishing Co., Amsterdam, 1975. North-Holland Mathematical Library, Vol. 11.

[123] Rockafellar, R. T. *Convex analysis*. Princeton University Press, Princeton, N.J., 1970. Princeton Mathematical Series, No. 28.

[124] Rudin, W. *Real and complex analysis*, third ed. McGraw-Hill Book Co., New York, 1987.

[125] Ruelle, D. Correlation functionals. *J. Mathematical Phys. 6* (1965), 201–220.

[126] Ruelle, D. A variational formulation of equilibrium statistical mechanics and the Gibbs phase rule. *Comm. Math. Phys. 5* (1967), 324–329.

[127] Sanov, I. N. On the probability of large deviations of random variables. In *Select. Transl. Math. Statist. and Probability, Vol. 1*. Inst. Math. Statist. and Amer. Math. Soc., Providence, R.I., 1961, pp. 213–244.

[128] Shwartz, A., and Weiss, A. *Large deviations for performance analysis*. Stochastic Modeling Series. Chapman & Hall, London, 1995. Queues, communications, and computing, With an appendix by Robert J. Vanderbei.

[129] Stroock, D. W. *An introduction to the theory of large deviations*. Springer-Verlag, New York, 1984.

[130] Takagi, H. Queuing analysis of polling models. *ACM Comput. Surveys 20*, 1 (1988), 5–28.

[131] Takahashi, Y. Entropy functional (free energy) for dynamical systems and their random perturbations. In *Stochastic analysis (Katata/Kyoto, 1982)*. North-Holland, Amsterdam, 1984, pp. 437–467.

[132] Veršik, A. M., and Kaĭmanovič, V. A. Random walks on groups: boundary, entropy, uniform distribution. *Dokl. Akad. Nauk SSSR 249*, 1 (1979), 15–18.

[133] Wu, L. M. Grandes déviations pour les processus de Markov essentiellement irréductibles. I. Temps discret. *C. R. Acad. Sci. Paris Sér. I Math. 312*, 8 (1991), 609–614.

Résumé

La théorie des grandes déviations a pour objet l'étude asymptotique d'*événements rares.* Cette thèse développe la technique dite de *changement de mesure* et ses relations avec l'*entropie.*

Après une présentation des techniques de base, on expose différents aspects du changement de mesure, quelques propriétés de l'entropie et les résultats attenants (PGD faibles...) pour les variables i.i.d. et les chaînes de Markov (en temps discret ou continu). On montre comment certains objets ou concepts *naturels* apparaissent — générateurs empiriques, fonctions de Ruelle-Lanford, etc. On s'aperçoit alors que l'analyse peut être poussée dans plusieurs directions : théorie du PGD faible, frontière de Martin (étude de comportements non-stationnaires), calcul de fonctions « quasi-harmoniques » dans certains réseaux (en vue d'estimer les queues de la distribution stationnaire), développement asymptotique d'ordre plus élevé (et calcul de variance). Ces techniques sont généralisables aux processus semi-markoviens. L'entropie est un dénominateur commun à tous ces problèmes.

La seconde partie de ce document présente les résultats achevés et en cours de publication : PGD pour les chaînes de Markov et pour une classe de réseaux à polling.

Mots-clefs Grandes déviations, PGD faible, entropie, changement de mesure, générateur empirique, chaînes de Markov, réseaux de files d'attente, frontière de Martin.

Abstract

This thesis focuses on the *change of measure* and its links with the *entropy* in the context of large deviations. We first present basic large deviations techniques settling the framework; the change of measure is set out together with properties of the entropy and related results (weak LDP ...) for i.i.d. variables and Markov processes. Then some notions somehow naturally appear: Ruelle-Lanford functions, empirical generators, etc. The analysis can be pursued further into several directions: weak LDP theory, Martin boundary, calculus of "quasi-harmonic" functions in networks (related to the tails of stationary distributions), asymptotic calculations and extension of these techniques to semi-Markov processes. A unity of structure in these problems is exhibited by means of the entropy function.
The second part of the thesis contains articles to appear: LDP for Markov chains in discrete and continuous time; LDP for some polling networks.

Keywords Large deviations, weak LDP, entropy, change of measure, empirical generator, Markov chains, queueing networks, Martin boundary.