HAL
open science

Réseaux de réactions : de l'analyse probabiliste à la réfutation

Vincent Picard

To cite this version:

Vincent Picard. Réseaux de réactions : de l'analyse probabiliste à la réfutation. Bio-informatique [q-bio.QM]. Université de Rennes, 2015. Français. NNT : 2015REN1S093 . tel-01246180v2

HAL Id: tel-01246180

https://theses.hal.science/tel-01246180v2

Submitted on 18 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Vincent PICARD

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Systèmes Aléatoires
ISTIC - UFR Informatique - Électronique

Réseaux de réactions : de l'analyse probabiliste à la réfutation

Thèse soutenue à Rennes le 16 décembre 2015

devant le jury composé de :

Frédérique BASSINO
Professeure à l'Université Paris 13 / *Rapporteure*

Vincent DANOS
Directeur de recherche au CNRS / *Rapporteur*

Paolo BALLARINI
Maître de conférence à l'École Centrale de Paris / *Examinateur*

Madalena CHAVES
Chargée de recherche à INRIA / *Examinatrice*

Olivier RIDOUX
Professeur à l'Université de Rennes 1 / *Examinateur*

Anne SIEGEL
Directrice de recherche au CNRS / *Directrice de thèse*

Jérémie BOURDON
Professeur à l'Université de Nantes / *Co-directeur de thèse*

稲妻に
悟らぬ人の
尊さよ

Inazuma ni
Satoranu hito no
Toutosa yo

Comme il est admirable
Celui qui ne pense pas "la vie est éphémère"
En voyant un éclair

松尾 芭蕉　Matsuo Bashō (1644-1694)

Remerciements

Cette thèse n'aurait jamais vu le jour sans le soutien indéfectible de mes deux directeurs de thèse Anne et Jérémie. Merci Anne, chercheuse impressionnante à la ténacité légendaire mais aussi directrice de thèse très humaine et compréhensive. Merci Jérémie, grand chercheur enthousiaste et talentueux avec qui j'ai pu partager mon amour pour les fonctions génératrices. Merci à tous les deux de m'avoir guidé pendant toutes ces années et de m'avoir donné la chance d'exprimer ma créativité. Anne, Jérémie, merci, si j'ai acquis une certitude pendant la thèse c'est que je n'aurais jamais trouvé de meilleurs directeurs de thèse que vous.

Je remercie tous les membres du jury de me faire l'honneur d'évaluer cette thèse et je suis en particulier très reconnaissant à Frédérique Bassino et Vincent Danos d'avoir bien voulu accepter les rôles de rapporteurs du manuscrit.

Je suis infiniment reconnaissant à ma famille qui, malgré ses origines modestes, se consacre sans faillir à l'instruction de ses enfants. Elle m'a continuellement encouragé dès le plus jeune âge à développer mon goût pour les sciences et à poursuivre de longues études. Cette thèse qui marque la fin de ma vie d'étudiant est aussi un cadeau de remerciement à toute ma famille.

Je remercie Marie-Noëlle Georgeault, Isobelle Kelly et Marie Le Roïc d'avoir grandement simplifié ma vie administrative de doctorant. Un grand merci à Isobelle et Marie d'avoir organisé la soutenance à mi-parcours ainsi que la soutenance finale.

J'ai eu de nombreux professeurs dans ma vie qui ont su me transmettre leur savoir, mais je souhaite en remercier quelques uns en particulier qui m'ont marqué profondément. Merci à Éric Hakenholz, véritable amoureux des mathématiques, qui a su me faire partager sa passion dès le collège : la quadrature du cercle, les coniques, le grand théorème de Fermat ! Éric, c'est vraiment grâce à toi que je me suis plongé dans les mathématiques ! Merci à Sylvie Guetienne qui a su aiguiser mon esprit avec de nombreuses énigmes mathématiques et qui m'a remis nombre de ses anciens polys de fac. Merci à Étienne et Françoise Trabbia pour leur enseignement des mathématiques, rigoureux et de qualité. Merci surtout de m'avoir encouragé et préparé avec succès à intégrer les classes préparatoires du lycée Louis-le-Grand, situé alors à des milliers de kilomètres de mon foyer. Un grand merci à Alain Pommellet, Laurent Chéno et Sylvie Dancre qui ont toujours cru au potentiel de leurs tau-

pins et taupines, cela m'a permis d'intégrer une É.N.S. Enfin, merci au professeur Michel Pierre de nous avoir donné de si beaux cours d'analyse : des mathématiques belles, simples, profondes et puissantes.

Lors de mes études à l'É.N.S. j'ai fait la rencontre de nombreux chercheurs qui m'ont mené petit à petit sur la voie du doctorat. Je souhaite les remercier sincèrement : François Coste, Dominique Lavenier, Marie-France Sagot, et encore une fois Anne et Jérémie. Un grand merci à Damien Éveillard pour le grand intérêt qu'il porte à ma recherche. Merci également à Cédric Lhoussaine et Adrien Richard de m'avoir invité à présenter ma recherche dans leur séminaire d'équipe. Je remercie également Julien Clément, Blaise Genest, Cédric Lhoussaine, Damien Eveillard, d'avoir bien voulu être jury de la soutenance à mi-parcours de la thèse. Cette thèse a aussi été l'opportunité de beaucoup voyager. J'ai tout d'abord voyagé dans la belle Bretagne et je remercie les biologistes de Roscoff, Robert Bellé, Patrick Cormier, Julia Morales et Odile Mulner-Lorillon de m'avoir parlé de biologie et d'oursins avec passion. Ensuite merci à Anne, Jérémie et à Alejandro Maass de m'avoir permis de présenter mes travaux de thèse au Chili. J'ai aussi eu la chance unique de séjourner plusieurs mois au Japon, un pays dont je suis tombé amoureux. Je ne peux que remercier du fond du cœur les personnes qui ont rendu cela possible. Encore merci à mes directeurs de thèse mais aussi à Charlotte Truchet. Je remercie chaleureusement Philippe Codognet de m'avoir accueilli au sein de son équipe à l'université de Tokyo. Merci à mon *sempai* Jean-François Baffier et à Thomas Silverston pour leur grande sympathie. Un merci tout particulier à Ryuko Nakamura.

La thèse étant un exercice long et difficile j'ai aussi traversé des périodes pénibles. À ces moments, j'ai toujours pu trouver du réconfort auprès de mes amis que je remercie et sans qui cette thèse ne serait pas parvenue à son terme. Merci mille fois à mes amis de longue date : Jenny, Aurore et Antoine. Merci à tous les amis qui m'ont entouré pendant la thèse : Gaëlle, Sylvain qui m'a entre autres aidé à mettre en forme ce manuscrit, Guillaume compagnon de table et de voyage, Mathilde pour nos petits-déjeuners, Clovis, Valentin, Alexan qui m'a rendu la vue, Anaïs et ses ondes positives, Geoffroy, Siva, sa discrétion légendaire, son amour des chiens, Renaud compagnon de rire et de bureau, Cyril, Clau, Coline, Aymeric, Charles, Julie, Nico, Guillaume C., Victo et Jean. Merci aussi à Jérémy. Je remercie également Catherine Bellannée pour son grand soutien. J'ai aussi été bien encouragé par mon club d'échecs, merci à Jonathan Demanghon, Ludovic Lejarre, Denis Monroy et Maryse Le Guen. Merci aussi au talentueux Alexandre Astier qui sait me faire rire et réfléchir même aux heures les plus sombres.

Merci sincèrement à tous !

Table des matières

Remerciements 1

Table des matières 3

Index des notations 7

Introduction 9

I Préliminaires, état de l'art 17

1 Modèles dynamiques et stationnaires différentiels 19
- 1.1 Réseaux de réactions . . . 20
  - 1.1.1 Définition . . . 20
  - 1.1.2 Représentations . . . 22
  - 1.1.3 Petits exemples . . . 23
  - 1.1.4 Exemple de la synthèse protéique cap-dépendante de la cellule œuf chez l'oursin . . . 24
- 1.2 Dynamiques par équations différentielles . . . 26
  - 1.2.1 Définition . . . 27
  - 1.2.2 Lois de flux . . . 27
  - 1.2.3 Limites : lois, paramètres et passages à l'échelle . . . 31
  - 1.2.4 Exemple de la synthèse protéique chez l'oursin . . . 35
- 1.3 Analyse stationnaire des flux et réfutation . . . 36
  - 1.3.1 Cône d'équilibre des flux stationnaires et utilisation de données de pentes 36
  - 1.3.2 Méthodes par contraintes reposant sur l'équilibre des flux . . . 38
  - 1.3.3 Réfutation d'un modèle 1 voie dans le modèle oursin . . . 39
- 1.4 Les limites du déterminisme . . . 42

2 Modèles dynamiques probabilistes 45
- 2.1 Modélisation dynamique Markovienne . . . 47
- 2.2 De l'équation maîtresse aux moments . . . 49
- 2.3 Comparaison avec la dynamique différentielle . . . 50
- 2.4 Résolution de l'équation maîtresse . . . 55
  - 2.4.1 Résolution exacte . . . 56

2.4.2 Approximations des moments 57
2.4.2.1 Méthodes des moments clos 58
2.4.2.2 L'approximation de bruit linéaire 58
2.4.2.3 Conclusion 59
2.4.3 Méthode de Monte-Carlo 60
2.5 Conclusion 63

II Approximation de Bernoulli du régime stationnaire en dynamique stochastique et applications 65

3 Approximation de Bernoulli du régime stationnaire 67
3.1 Discrétisation de la dynamique stochastique 68
3.1.1 Définition 68
3.1.2 Validité 69
3.1.3 Illustration 70
3.2 Dynamique de Bernoulli 72
3.2.1 Définition 74
3.2.2 Expression analytique des espérances et variances 74
3.2.3 Théorème central limite pour la dynamique de Bernoulli 76
3.2.4 Interprétation en tant que marches aléatoires 77
3.3 Analyse stationnaire et validité de l'approximation 78
3.3.1 Probabilités de réactions stationnaires 79
3.3.2 Comparaison des espérances 80
3.3.3 Comparaison des matrices de covariances 84
3.4 Conclusion 86

4 Applications à la validation de modèles 89
4.1 Méthodes par contraintes pour les moments d'ordre 1 et 2 90
4.1.1 Table des contraintes 90
4.1.2 Effet d'un bruit blanc expérimental 92
4.1.3 Exemples 92
4.1.3.1 Illustration de l'intérêt des contraintes de moments d'ordre 2 . 93
4.1.3.2 Exemple d'un réseau métabolique jouet 95
4.2 Ellipsoïdes de confiance 97
4.2.1 Définition 97
4.2.2 Étude des cas dégénérés de la loi limite 100
4.2.2.1 P-invariants 101
4.2.2.2 Cas du processus de comptage 101
4.2.2.3 Caractérisation 102
4.2.3 Application à la réfutation de modèles 104
4.3 Contraintes par l'exploitation d'un rapport de taux de production 105
4.3.1 Convergence du rapport de taux de production 106
4.3.2 Nouveau tableau de contraintes 109
4.4 Conclusion 110

5 Vérification de propriétés asymptotiques sur les réseaux stationnaires 113
5.1 Syntaxe et sémantique 114
5.1.1 Syntaxe 114
5.1.2 Sémantique 115
5.2 Satisfaisabilité et validité 118
5.3 Exemple 121
5.4 Conclusion 122

Conclusion 123
Contexte 123
Résultats 124
Perspectives 127

Bibliographie 137

Table des figures 139

Index des notations

	Matrices et espaces vectoriels
$M_{n,m}(X)$	Ensemble des matrices de taille $n \times m$ à coefficients dans X
$\mathbb{I}_n$	Matrice identité de taille $n \times n$
A^t	Transposée de la matrice A
$\operatorname{diag} \vec{x}$	Matrice diagonale dont les coefficients diagonaux sont les composantes du vecteur $\vec{x}$ dans le même ordre
$\mathscr{O}_n(\mathbb{R})$	Ensemble des matrices réelles orthogonales de taille $n \times n$
$\ker A$	Noyau de la matrice A *i.e.* ensemble des vecteurs $\vec{x}$ tels que $A\vec{x} = \vec{0}$
$\operatorname{vect} X$	Sous-espace vectoriel engendré par l'ensemble X
	Réseaux de réactions
n	Nombre d'espèces
m	Nombre de réactions
$\alpha = (\alpha_{i,j})_{1 \le i \le n, 1 \le j \le m}$	Matrice de consommation des espèces par les réactions
$\beta = (\alpha_{i,j})_{1 \le i \le n, 1 \le j \le m}$	Matrice de production des espèces par les réactions
$S = \beta - \alpha$	Matrice de stœchiométrie du réseau de réactions
$(\nu_j)_{1 \le j \le m}$	Colonnes de la matrice S
	Théorie des probabilités
$\mathbb{P}(A)$ ou $\mathbb{P}[A]$	Probabilité de l'événement A
$\mathbb{P}(A\|B)$	Probabilité de l'événement A sachant l'événement B
$\mathbb{E}\, X$	Espérance de la variables aléatoire X
$\mathbb{E}(X\|Y)$	Espérance conditionnelle de la variable (resp. vecteur) aléatoire X sachant la variable (resp. vecteur) aléatoire Y
$\operatorname{Var} X$	Variance de la variable aléatoire X
$\mathbb{C}\operatorname{ov} \vec{X}$	Matrice de variance-covariance du vecteur aléatoire $\vec{X}$
$\mathscr{N}(\vec{\mu}, C)$	Loi normale multivariée de centre $\vec{\mu}$ et de matrice de variance-covariance C
$\mathscr{E}(\lambda)$	Loi exponentielle de paramètre λ
$\xrightarrow{\mathscr{L}}$	Convergence en loi

	Dynamique différentielle
$(\vec{x}(t))_{t\in\mathbb{R}^+}$	Dynamique markovienne à temps continu
k_j	Constante de réaction cinétique de la réaction R_j
f_j	Flux de la réaction R_j
	Dynamiques stochastiques
c_j	Constante de réaction stochastique de la réaction R_j
$(\vec{x}(t))_{t\in\mathbb{R}^+}$	Dynamique markovienne à temps continu
$(\vec{y}(k))_{k\in\mathbb{N}}$	Dynamique markovienne à temps discret issue de $(\vec{x}(t))_{t\in\mathbb{R}^+}$
$(\vec{z}(k))_{k\in\mathbb{N}}$	Dynamique de Bernoulli
$\Sigma, \Sigma(\vec{x})$	Ensemble des états, ensemble des états accessibles depuis l'état $\vec{x}$
$h_j(\vec{x})$ $(1 \leq j \leq m)$	Propension de la réaction R_j dans l'état $\vec{x}$
$p_j(\vec{x})$ $(1 \leq j \leq m)$	Probabilité de la réaction R_j dans l'état $\vec{x}$
$\vec{\mathfrak{p}}$	Loi stationnaire de $(\vec{p}(\vec{y}(k)))_{k\in\mathbb{N}}$

Introduction

Informatique et biologie des systèmes

Si la *science informatique* se définit essentiellement en tant que science du *calcul* elle est aussi en grande partie une science des *modèles*. Là où le mathématicien s'applique davantage à démontrer des résultats dans un modèle donné, à bâtir des théories reposant sur une fondation d'axiomes donnés, l'informaticien, lui, préfère jouer avec un grand nombre de modèles, changer les règles du jeu, comprendre ce que l'on peut faire dans un cas et pas dans l'autre et savoir comment exploiter au mieux ces règles. L'informatique repose en effet sur quatre grands piliers théoriques inter-dépendants : l'algorithmique, la calculabilité et la complexité, les langages formels et la logique. L'algorithmique s'attache à comprendre comment réaliser une tâche le plus efficacement possible étant donné un modèle de calcul, c'est-à-dire un jeu d'instructions élémentaires. La théorie de la calculabilité et de la complexité explore les fondements de la notion de calcul en se reposant sur des *modèles de calcul* élémentaires et en particulier celui de la *machine de Turing*. La théorie des langages formels traite des *modèles d'expression* (automates finis, automates à pile, grammaires, *etc*) qui déterminent quels textes, quelles suites de symboles, peuvent être exprimés et reconnus automatiquement. Enfin, en logique l'étude de *la théorie des modèles* permet de comprendre quelles théories et structures mathématiques peuvent s'exprimer par un langage logique pouvant être manipulé par une machine. Il n'est donc pas étonnant de constater l'importance croissante de l'informatique dans les domaines de la modélisation, en physique, en chimie, en biologie, en astronomie.

En particulier, l'informatique participe de manière cruciale à la *biologie des systèmes*, un champ inter-disciplinaire qui se consacre à la modélisation de systèmes biologiques complexes. La spécificité de cette discipline est de s'intéresser à la modélisation de mécanismes biologiques globaux (par exemple, le cycle circadien, les mécanismes d'un cancer, *etc*) dont le fonctionnement est la résultante d'*interactions complexes* entre de nombreux acteurs, à des échelles spatiales et temporelles très hétérogènes : gènes, ARN messagers, régulation par micro-ARN, complexes de protéines, métabolites, tissus, organes, transmission de signaux, *etc*. La biologie des systèmes s'oppose donc à la conception réductionniste qui consiste à comprendre le tout en étudiant séparément chacune des parties. Au contraire, elle repose sur le principe du *tout valant plus que la somme des parties*, ici le projet ambitieux du chercheur est de parvenir à concevoir, comprendre et analyser des modèles

construits à partir de ces différents acteurs hétérogènes pour étudier la fonction biologique complexe qui résulte de leurs interactions. Kitano [K+01] a identifié quatre objectifs de la biologie des systèmes.

1. **Concevoir la structure** des modèles, c'est-à-dire déterminer les acteurs et leurs interactions.
2. Déterminer le **comportement dynamique** c'est-à-dire obtenir à partir de la structure l'évolution temporelle des quantités d'acteurs dans le système
3. **Identifier** les acteurs clefs du système, c'est-à-dire comprendre quels acteurs vont être responsables d'un certain comportement. Cette analyse proche de la théorie du contrôle a comme exemple d'application l'identification de cibles thérapeutiques pour la conception de médicaments.
4. **Synthétiser** des systèmes, ce que l'on appelle la *biologie de synthèse*, c'est-à-dire être capable de construire *ab initio* des systèmes biologiques ayant une fonction biologique souhaitée.

L'informatique participe de manière importante à l'effort de recherche dans chacun de ces domaines. Tout d'abord elle dispose d'un large éventail d'objets théoriques qui fournissent des langages de description de ces modèles d'agents en interactions : les graphes, les réseaux d'automates, les réseaux de Petri. De plus, elle dispose d'une vaste littérature théorique permettant d'étudier les comportements dynamiques de ces modèles.

Dynamiques des réseaux de réactions

Cette thèse concerne essentiellement les deux premiers objectifs. Notre objet d'étude sera le *réseau de réactions* qui est un formalisme très général, équivalent en informatique à celui du *réseau de Petri*, et qui fournit un langage général pour définir la structure (premier objectif) des modèles en biologie des systèmes. Il a aussi l'avantage d'être un langage commun entre les diverses communautés, biologistes, mathématiciens, informaticiens, chimistes qui travaillent en biologie de systèmes. L'analyse quantitative (deuxième objectif) de ces systèmes de réactions couplées est un centre d'intérêt majeur en biologie des systèmes. Cette thèse s'intéresse à l'interaction entre ces deux objectifs c'est-à-dire que l'on va non seulement s'intéresser à l'obtention de la dynamique du réseau à partir de sa structure, ce qui a déjà été grandement traité dans la littérature, mais surtout on cherche aussi à comprendre comment des informations dynamiques, à savoir des mesures quantitatives expérimentales, permettent d'obtenir des informations sur la structure.

Essentiellement deux cadres de modélisations de leur dynamique ont été introduits [Hel08] : les équations différentielles ordinaires à l'échelle de la population de cellules et les chaînes de Markov à l'échelle de la cellule individuelle.

Dans le premier cas, les *équations différentielles ordinaires* (EDO) fournissent des trajectoires déterministes décrivant le comportement moyen d'une population. L'évolution temporelle des quantités d'espèces $\vec{x}$ est alors déterminée par un système d'équations différentielles ordinaires couplées de type $\frac{d\vec{x}}{dt} = S\vec{f}(\vec{x})$ où S est la matrice de stœchiométrie du

système et $\vec{f}$ est un vecteur de *flux* dépendant des quantités de matière courantes. Généralement, la fonction $\vec{f}$ est donnée par la loi d'action de masse bien que d'autres types de lois (Michaelis-Menten, Droop, ...) peuvent aussi être utilisées. Lorsque les quantités d'espèces initiales sont connues ainsi que les lois dynamiques et leurs paramètres, ces équations différentielles déterminent la trajectoire du système qui peut être obtenue approximativement à l'aide d'algorithmes d'analyse numérique.

Lorsque le réseau de réactions est trop grand ou que les données expérimentales sont insuffisantes, il est difficile de déterminer les paramètres de ces équations différentielles. Une analyse alternative est de considérer l'*état stationnaire* du système, où les concentrations des réactifs sont supposées être constantes suite à un équilibre de leurs productions et de leurs consommations (formellement, l'équation $S\vec{f} = 0$ est vérifiée). À partir d'informations stœchiométriques, les *approches par contraintes* consistent à trouver une valeur appropriée des flux $\vec{f}$ à l'état stationnaire tels que $S\vec{f} = \vec{0}$, plus quelques contraintes biologiques additionnelles. Cette approche nommée *flux balance analysis* (FBA) [PRP04, OTP10] est intensément présente dans la littérature de biologie des systèmes.

Systèmes de contraintes et réfutation

Parmi les nombreuses applications de ces méthodes basées sur les contraintes de flux à l'équilibre, on peut les appliquer à des problèmes de *validation de modèles*, c'est-à-dire, déterminer si un réseau de réactions est consistant avec certaines observations expérimentales ou non. Pour cela, on tente de réfuter un réseau de réactions en aboutissant à un système de contraintes inconsistant. Par exemple, considérons le réseau de réactions suivant

$$(\text{modèle } 1) \left\{ \begin{array}{lcl} R_1: & A & \rightarrow & B+C \\ R_2: & B & \rightarrow & A+2D \end{array} \right. \tag{1}$$

et supposons qu'on arrive à déterminer expérimentalement que les réactifs (et donc les flux) sont à l'équilibre et que les taux moyens de production de C et D sont égaux. D'un point de vue contraintes, les flux à l'équilibre doivent alors vérifier les équations

$$(\text{modèle } 1) \left\{ \begin{array}{ll} -f_1 + f_2 = 0 & (\text{équilibre de A}) \\ f_1 - f_2 = 0 & (\text{équilibre de B}) \\ f_1 = 2f_2 & (\text{productions de } C \text{ et } D \text{ comparables}) \end{array} \right. \tag{2}$$

qui n'admettent pas de solution. Ainsi, si on suppose que nos mesures expérimentales sont justes il faut rejeter le modèle de réactions proposé. Cependant, il existe des cas où les contraintes de flux à l'équilibre ne sont pas suffisantes pour aboutir à une réfutation. Considérons par exemple les deux réseaux de réactions suivants

$$(\text{modèle } 1) \left\{ \begin{array}{lcl} R_1: & A & \rightarrow & B+C \\ R_2: & B & \rightarrow & A+2D \end{array} \right. \qquad (\text{modèle } 2) \left\{ \begin{array}{lcl} R_1: & A & \rightarrow & B+C+2D \\ R_2: & B & \rightarrow & A \end{array} \right. \tag{3}$$

mais en supposant cette fois ci qu'on observe une production moyenne de D double par rapport à celle de C. Une approche reposant sur les flux indique que dans les deux modèles, tous les flux $\vec{f} = (f_1, f_2)$ tels que $f_1 = f_2$ satisfont l'équilibre des réactifs A et B et vérifient que le taux d'accumulation de C vaut f_1 tandis que celui de D vaut $2f_1$. Ainsi, les deux systèmes sont cohérents vis à vis des mesures moyennes obtenues. Toutefois, les régimes stationnaires de ces deux réseaux de réactions *peuvent* être distingués l'un de l'autre. Intuitivement, dans le modèle 1, les quantités de C et D doivent être anti-corrélées alors qu'elles devraient être corrélées dans le système 2. Pour formaliser cette intuition, nous devons travailler à l'échelle de l'individu où les fluctuations stochastiques existent. Cela peut être observé sur la figure 1 sur laquelle des trajectoires individuelles des deux réseaux ont été générées à l'aide d'un algorithme probabiliste défini et étudié dans le chapitre 2. Il apparaît que la moyenne et la variance des deux réseaux sont comparables, ce qui ne permet pas de les distinguer. Par contre, les systèmes se distinguent clairement par leurs covariances (ligne noire) entre C et D (avec une corrélation de -1 approximativement pour le modèle 1 et 1 pour le modèle 2).

Encouragés par cet exemple, une motivation majeure de cette thèse est d'*introduire un cadre de méthodes par contraintes* qui généralise celui de l'équilibre des flux au cas stochastique ce qui permettra de réfuter et donc discriminer des modèles en tenant compte des informations de variabilité mesurées par la variance et la covariance, c'est-à-dire, les *moments d'ordre 2*. Cette approche est fortement motivée par le développement récent des méthodes expérimentales en biologie *single-cell* [HZ11] (fluoresence, imagerie microscopique, spectrométrie de masse) qui fournira à l'avenir de plus en plus de données à l'échelle de l'individu.

Comment prendre en compte les (co)variances ?

Afin d'aboutir à de telles contraintes issues de mesures expérimentales sur les moments d'ordre deux, il est naturel de considérer la seconde approche dynamique pour les réseaux de réactions qui permet de raisonner à l'échelle de l'individu. En effet, à l'aide de la modélisation par *chaînes de Markov*, on peut générer de nombreuses trajectoires aléatoires dont on sait déterminer non seulement la moyenne mais aussi les (co)variances. Ces modèles probabilistes fournissent donc une description stochastique des trajectoires d'un individu dont on peut étudier les caractéristiques aléatoires. D'un point de vue probabiliste, la loi des trajectoires de ces modèles (dont dépendent les moments) est déterminée par *l'équation maîtresse chimique* (CME).

Malheureusement dans la plupart des cas, la CME ne peut être résolue, même numériquement, en raison d'un nombre d'équations différentielles égal à la taille de l'espace d'états accessibles de ce système. Pour cette raison, D. T. Gillespie a popularisé une méthode de Monte-Carlo qui prend la forme d'un algorithme [Gil76, Gil77] de génération de trajectoires aléatoires. La loi des trajectoires obtenues est alors la solution de la CME. Cet algorithme nommé *algorithme de Gillespie* ou encore l'*algorithme de simulation stochastique* (SSA)

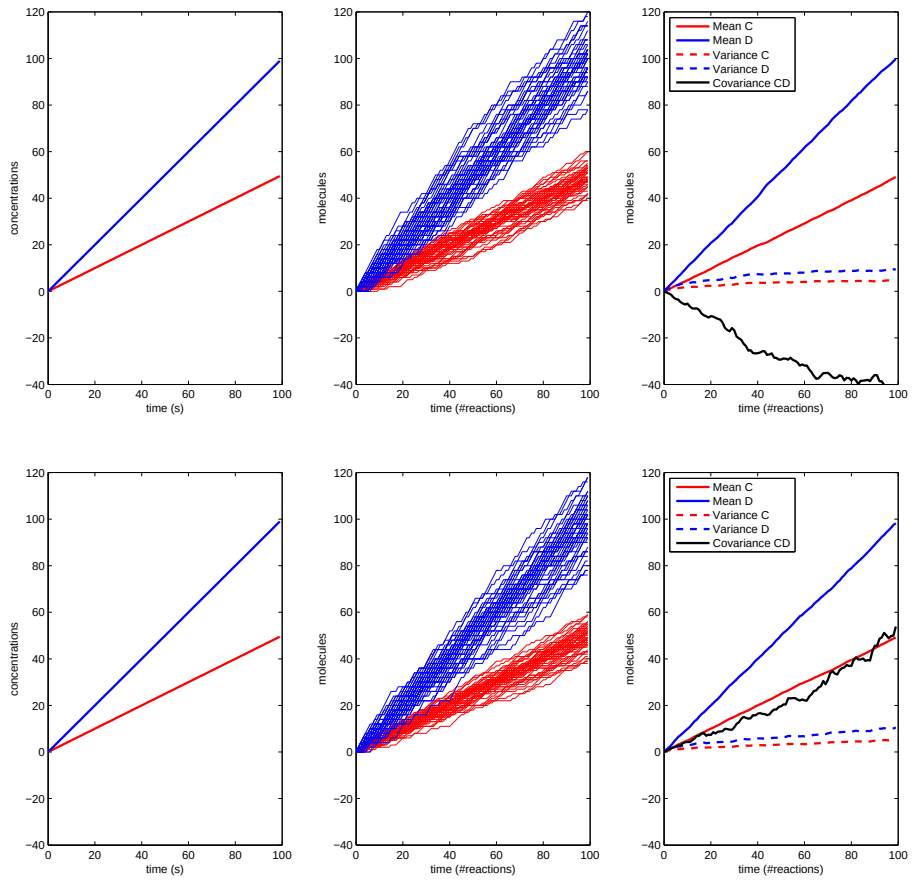


FIGURE 1 – **Réseaux distinguables par leurs moments d'ordre 2** Dynamiques différentielles et stochastiques pour le modèle 1 (première ligne) et le modèle 2 (seconde ligne) décrits dans Eq. (3). Les courbes rouges (resp. bleues) représentent les quantités de C (resp. D). La première colonne représente la solution de l'équation différentielle issue de la loi d'action de masse. La seconde colonne représente 50 générations de trajectoires aléatoires obtenues à l'aide d'un algorithme de simulation stochastique. Les paramètres cinétiques et stochastiques ont été fixés à 1 et il y a $1000A$ et $1000B$ initialement. La troisième colonne représente les estimations des espérances, des variances et des covariances croisées entre C et D. Ces estimations sont obtenues à partir des 50 trajectoires de la seconde colonne. On constate que seule la trajectoire des covariances de C et D permettent de distinguer ces réseaux.

est devenu le pilier fondateur des modélisations dynamiques stochastiques en biologie des systèmes [Wil12]. De multiples applications biologiques ont été présentées [MA97, ARM98] et des améliorations de l'algorithme ont été proposées [GB00]. Ainsi, l'algorithme de Gillespie et ses variantes peuvent être vu comme le pendant probabiliste des algorithmes de résolution numérique des EDOs en modélisation dynamique différentielle. Une autre approche

courante en modélisation stochastique sont les méthodes des moments clos (CMM) dont en particulier l'approximation de bruit linéaire (LNA), également nommée Ω-expansion. Ces méthodes introduisent un système d'équations différentielles pour la trajectoire des moments (jusqu'à un certain ordre fixé) qui lorsqu'il est résolu permet d'en obtenir une approximation. Dans le cas de l'approximation de bruit linéaire, la méthode consiste à décomposer la trajectoire en une partie déterministe obtenue à l'aide des équations de la loi d'action de masse et une partie aléatoire prenant la forme d'un bruit gaussien dont les paramètres à chaque instant t peuvent être calculés par résolution d'un système d'équations différentielles.

Cependant, comme nous le verrons dans le chapitre 2, l'utilisation du SSA ou des méthodes de moments clos conduisent à des problèmes similaires à l'utilisation des EDOs : une complète connaissance du système est nécessaire en ce qui concerne le type de lois cinétiques ainsi que leurs paramètres c'est-à-dire les constantes cinétiques de réaction. Dans tous les cas ces paramètres sont difficiles à déterminer en particulier quand la taille des réseaux de réactions est grande. De plus il est également difficile de définir un analogue du vecteur de flux stationnaire qu'on puisse facilement mettre en relation analytique avec les trajectoires des moments obtenues.

L'approche d'équilibre des flux (FBA) contourne les difficultés de la dynamique différentielle grâce à une hypothèse de stationnarité, dans cette thèse nous suivrons donc cette même direction mais dans un cadre stochastique. Nous utilisons la notion de stationnarité pour approximer le comportement stochastique des chaînes de Markov quand le système est proche de l'équilibre. Dans cette approximation qu'on nommera *dynamique de Bernoulli* et qui sera étudiée dans le chapitre 3, les probabilités de tirages $\vec{p}$ des réactions sont des constantes ce qui conduit à des trajectoires de type *marches aléatoires*, entièrement déterminées par la matrice de stœchiométrie et par un vecteur de *probabilités de réactions* $\vec{p}$. On montre qu'il est possible d'analyser cette dynamique et établir des relations analytiques entre les probabilités de réactions et la trajectoire des moments. Par exemple on obtient dans le chapitre 3 un théorème central limite décrivant entièrement le comportement asymptotique de cette dynamique en fonction de $\vec{p}$ et de la stœchiométrie S.

Résultat 0.1.

$$\frac{1}{\sqrt{k}}\left(\vec{z}(k)-\left(\vec{z}(0)+kS\vec{p}\right)\right)\xrightarrow[k\to+\infty]{\mathscr{L}}\mathcal{N}\left(\vec{0},W(S,\vec{p})\right), \tag{4}$$

où la matrice de variance-covariance réduite vaut $W(S,\vec{p})=S\left(\mathrm{diag}(\vec{p})-\vec{p}\vec{p}^{\,t}\right)S^t$.

Le vecteur de probabilités de réactions est donc un bon candidat pour être l'analogue stochastique du vecteur de flux stationnaire. Après avoir introduit et étudié la dynamique de Bernoulli nous présentons plusieurs applications dont les systèmes de contraintes capables de prendre en compte les (co)variances (chapitre 4). Nous présentons également une formalisation logique (chapitre 5) dans laquelle des propriétés générales sur les moments d'ordre 1 ou 2 peuvent être formellement exprimées et vérifiées à l'aide des systèmes de contraintes introduits précédemment.

Plan de la thèse

Chapitre 1 – Modèles dynamiques et stationnaires différentiels Le premier chapitre dresse l'état de l'art de la sémantique déterministe ou différentielle des réseaux de réactions. Nous définissons précisément la notion de réseau de réactions et nous l'illustrons à l'aide d'exemples fictifs et réels. Nous montrons comment on peut obtenir des systèmes d'équations différentielles pour décrire la dynamique de ces réseaux mais que ces systèmes reposent sur la connaissance de lois physiques d'évolution pour les réactions ainsi que sur la détermination de nombreux paramètres cinétiques. Ces paramètres étant difficiles à inférer, on présente l'analyse stationnaire, dont le *flux balance analysis*, qui permet d'étudier des réseaux de plus grande taille. En particulier, on montrera comment on peut mettre en relation les flux stationnaires et les pentes moyennes et en déduire des systèmes de contraintes. Nous verrons comment on peut se servir de ces systèmes de contraintes pour obtenir des réfutations de réseaux avec un exemple biologique concret chez l'oursin. Malheureusement, cette approche ne peut être utilisée pour prendre en compte des informations de (co)variances obtenues à partir de plusieurs trajectoires. Par exemple, on ne peut pas distinguer les deux modèles présentés dans l'introduction de cette thèse. Un objectif majeur de cette thèse est d'adapter cette méthode pour pouvoir prendre en compte les moments d'ordres supérieurs des trajectoires.

Chapitre 2 – Modèles dynamiques probabilistes Puisque l'on souhaite prendre en compte des informations de moments, il est naturel de s'intéresser à la dynamique stochastique des réseaux de réactions. Dans ce chapitre, présentant l'état de l'art en modélisation dynamique stochastique, on montre comment elle permet d'obtenir une dynamique des moments. Nous introduisons l'équation chimique maîtresse qui gouverne la loi des trajectoires et *a fortiori* celle des moments mais qui ne peut être résolue en pratique. Nous présentons alors les méthodes de résolution approximatives à savoir les méthodes de moments clos et la simulation de Monte-Carlo (algorithme de Gillespie). Ces méthodes ne permettent pas de s'affranchir d'une connaissance parfaite des lois et paramètres cinétiques du réseau, qui sont difficiles à déterminer lorsque la taille des réseaux et grande. De plus, ces méthodes ne permettent pas directement d'obtenir une analogie directe avec les méthodes de flux stationnaires. En particulier elles n'exhibent pas d'analogue au vecteur de flux stationnaires qu'on puisse relier analytiquement à la trajectoire des espérances et des (co)variances. Cela ne peut être obtenu qu'en considérant des hypothèses de stationnarité introduites dans le chapitre suivant.

Chapitre 3 – Approximation de Bernoulli en régime stationnaire Dans ce chapitre, on introduit la dynamique de Bernoulli en tant qu'approximation stationnaire de la dynamique probabiliste classique. On montre dans un premier temps qu'il est possible d'analyser cette dynamique, c'est-à-dire qu'on détermine des expressions analytiques pour les moments de ses trajectoires. On détermine aussi son comportement asymptotique à l'aide d'un théorème central limite dont on détermine entièrement les paramètres. De plus, cette dyna-

mique est paramétrée par un vecteur de probabilités de réactions $\vec{p}$ qui joue l'analogue du vecteur de flux stationnaires en dynamique différentielle. Nous concluons l'étude de cette dynamique en étudiant la qualité de l'approximation proposée. En particulier on montre que cette approximation est valide dans le cadre de la limite thermodynamique.

Chapitre 4 – Applications à la validation de modèles Dans ce chapitre applicatif, on montre comment utiliser les résultats d'approximation du Chapitre 3. Nous montrons comment obtenir des systèmes de contraintes analogues aux contraintes du FBA mais qui sont obtenues à l'aide de mesures sur les espérances, variances et (co)variances d'un ensemble de trajectoires expérimentales données. Ces contraintes portent sur le vecteur $\vec{p}$ de probabilités de réactions qui est l'équivalent du vecteur de flux dans les méthodes d'analyse de l'équilibre des flux. Nous illustrons ces contraintes sur quelques exemples et en particulier on résout l'exemple des systèmes non distinguables par les espérances mais distinguables par leurs covariances présenté dans cette introduction. Un résultat secondaire de ce chapitre est la possibilité de définir des ellipsoïdes de confiance pour les trajectoires du réseau. Ces ellipsoïdes peuvent être utilisées pour confronter des trajectoires données avec un réseau de réactions donné. Nous donnons un exemple de réfutation reposant sur ces ellipsoïdes. Ces ellipsoïdes nous permettent également de démontrer un théorème de convergence sur le rapport des taux de production de deux espèces, dans le cadre stochastique. Ce théorème permet d'établir d'autres contraintes sur $\vec{p}$ en exploitant un rapport de taux productions mesuré sur une seule trajectoire uniquement.

Chapitre 5 – Vérification de propriétés asymptotiques sur les réseaux stationnaires Nous introduisons une formalisation logique des résultats précédents. La logique permet d'exprimer de façon plus « naturelle » des connaissances sur les moments. Par exemple la question « est-ce qu'une variance petite sur une sortie du réseau entraîne une covariance négative sur deux autres sorties ? » peut s'exprimer dans ce langage. Nous montrons comment la validité de ces formules logiques peut être déterminée à l'aide de plusieurs systèmes de contraintes linéaires, quadratiques. Nous étudions également la complexité théorique de cette question.

Je vous souhaite une très bonne lecture.

Première partie

Préliminaires, état de l'art

Chapitre 1

Modèles dynamiques et stationnaires différentiels

La *modélisation dynamique* a pour but de décrire l'évolution temporelle des quantités de matière présentes dans un système biologique. L'obtention d'un modèle quantitatif prédictif est un outil précieux pour le biologiste. Elle lui permet de tester des hypothèses biologiques en explorant *in silico* le comportement du système étudié par exemple lorsque des quantités d'espèces chimiques sont modifiées ou lorsque certaines réactions chimiques sont ajoutées ou supprimées. Il économise non seulement des ressources mais il peut également effectuer des observations inaccessibles par l'expérience. Réussir à construire un bon modèle dynamique prédictif peut donc être un atout précieux pour le biologiste.

L'objectif premier de ce chapitre préliminaire est d'introduire les concepts fondamentaux des modèles dynamiques reposant sur les équations différentielles à travers d'exemples réels et fictifs. Nous expliquerons comment obtenir de tels modèles mais nous montrerons aussi pourquoi leur mise en œuvre peut être difficile même pour des systèmes de petite taille en raison principalement des problèmes liés à l'apprentissage des paramètres. La compréhension de ces éléments permet de mieux appréhender l'introduction des *approches par contraintes* qui sont au cœur de cette thèse. Dans le cas des méthodes différentielles, l'ajout d'une hypothèse de *stationnarité* permet d'aboutir à des systèmes de contraintes de flux à l'équilibre. Ce système de contraintes forme un nouveau modèle qui décrit le comportement stationnaire c'est-à-dire l'équilibre du système. Il permet de répondre à des questions concernant les flux de matières à l'équilibre tout en contournant les problèmes délicats relatifs à la détermination des lois cinétiques et des paramètres corrects. Ce chapitre introduira également ces approches par contraintes en se concentrant sur leur application dans le contexte de la *réfutation de modèle* qui constitue également un thème important de cette thèse. La question est alors de déterminer si le modèle de réactions proposé est compatible avec certaines observations expérimentales.

Un dernier objectif de ce chapitre est de souligner les carences des méthodes différentielles non seulement en ce qui concerne les hypothèses qui les fondent caractérisées par

des quantités de matière de natures continues et déterministes mais aussi du point de vue de leurs applications. Ces manques invitent à repenser entièrement les questions abordées en les plaçant dans le contexte d'une modélisation probabiliste, plus réaliste mais aussi plus riche en termes applicatifs. Cela fait l'objet du chapitre suivant.

1.1 Réseaux de réactions

Il convient dès à présent de définir plus précisément l'objet d'étude de cette thèse. Les méthodes présentées s'appliquent à une large classe de modèles biologiques ou écologiques dont le point commun est qu'ils peuvent se décrire en tant qu'ensemble de *réactions* agissant sur un ensemble d'*espèces*. Ainsi, les réseaux de réactions peuvent être vus comme un *langage* permettant d'exprimer des modèles. Ce formalisme est grandement répandu en modélisation de systèmes biologiques ou écologiques et une présentation détaillée peut-être trouvée dans [Wil06].

1.1.1 Définition

Une *espèce* est une entité du système qui peut-être quantifiée et dont la quantité est susceptible d'évoluer au cours du temps. Une *réaction* est une action élémentaire du système qui consiste en la consommation de quantités entières d'un nombre fini d'espèces, dites *réactifs*, et la production de quantités entières d'un nombre fini d'espèces, dites *produits*. On la notera de manière classique sous la forme

$$\alpha_1 X_1 + \cdots + \alpha_n X_n \to \beta_1 X_1 + \cdots + \beta_n X_n$$

où $X_1, \ldots, X_n$ sont les noms des espèces impliquées dans la réaction, $(\alpha_i)_{i=1,\ldots,n} \in \mathbb{N}^n$ sont les quantités consommées et $(\beta_i)_{i=1,\ldots,n} \in \mathbb{N}^n$ sont les quantités produites. On considère donc d'un point de vue formel que toutes les réactions sont *irréversibles*, toutefois cela ne limite pas le pouvoir d'expressivité puisque toute réaction *réversible* peut se décomposer en deux réaction irréversibles.

Bien que la dénomination et la notation font allusion aux réactions chimiques impliquant des espèces chimiques, espèces et réactions doivent être pensées comme des *abstractions*. Par exemple, dans certains modèles de régulation génétique une espèce peut représenter un gène et sa quantité le niveau d'activation de ce gène sans qu'on puisse à proprement parler de *quantités de molécules de gènes*. D'autres exemples décrivant des processus de vie et de mort dans une population ou des systèmes écologiques de proies-prédateurs peuvent s'écrire en tant qu'ensembles de réactions bien qu'on ne traite pas dans ces modèles de quantités de molécules. Le pouvoir d'abstraction donne également la possibilité dans des modèles moléculaires cellulaires de nommer différemment une même molécule selon le compartiment moléculaire dans lequel elle réside.

Un point important est qu'une réaction définit une unité d'action du système c'est-à-dire que le modélisateur possède une liberté dans le choix de l'échelle d'étude du système. Par exemple, la mécanique de la respiration cellulaire aérobie nécessite un ensemble de réactions métaboliques mais peut être simplifié par une seule réaction

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O.$$

Le modélisateur peut alors décider de ne pas intégrer dans son modèle toute la complexité du mécanisme de respiration mais de le modéliser par une seule réaction. Une fois encore, on constate l'intérêt de considérer les réactions comme des abstractions.

On peut maintenant définir notre objet d'étude de base, le réseau de réaction, qui consiste en une liste finie de m réactions de la forme

$$\alpha_{1,j}X_1 + \cdots + \alpha_{n,j}X_n \rightarrow \beta_{1,j}X_1 + \cdots + \beta_{n,j}X_n$$

pour $j \in \{1, \ldots, m\}$, en voici une définition formelle.

Définition 1.1 (Réseau de réaction)**.** Un *réseau de réaction* est un quadruplet (n, m, α, β) où

— $n \in \mathbb{N}$ est le nombre d'espèces du réseau qu'on notera $X_1, \ldots, X_n$,

— $m \in \mathbb{N}$ est le nombre de réactions du réseau qu'on notera $R_1, \ldots, R_m$,

— $\alpha \in M_{n,m}(\mathbb{N})$ est une matrice d'entiers de taille $n \times m$ appelée *matrice de consommation* dont les coefficients $\alpha_{i,j}$ représentent la quantité d'espèce X_i consommée par la réaction R_j,

— $\beta \in M_{n,m}(\mathbb{N})$ est une matrice d'entiers de taille $n \times m$ appelée *matrice de production* dont les coefficients $\beta_{i,j}$ représentent la quantité d'espèce X_i produite par la réaction R_j.

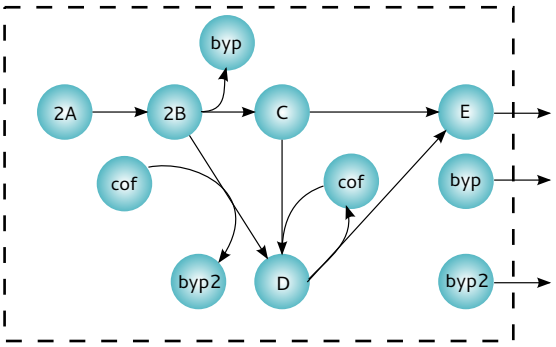
On dit que l'espèce X_i est un $réactif$ (resp. $produit$) de la réaction R_j lorsque $\alpha_{ij} > 0$ (resp. $\beta_{ij} > 0$). Une espèce peut donc être à la fois un réactif et un produit. On appellera $entrée$ tout réactif qui n'est produit par aucune des réactions et $sortie$ tout produit qui n'est réactif d'aucune réaction.

Une propriété importante d'un réseau de réaction est sa matrice de stœchiométrie.

Définition 1.2 (Matrice de stœchiométrie)**.** La *matrice de stœchiométrie* ou *matrice stœchiométrique* d'un réseau de réaction (n, m, α, β) est la matrice $S = \beta - \alpha$ de taille $n \times m$.

Les colonnes de la matrice S représentent les bilans de matière de chaque réaction. Tout au long de cette thèse, je prends garde à distinguer le réseau de réaction de sa matrice de stœchiométrie car la donnée de la matrice S ne suffit pas à définir le réseau. Par exemple les deux réseaux $\{2X \rightarrow X\}$ et $\{X \rightarrow \varnothing\}$ ont même matrice de stœchiométrie mais sont différents. La différence peut sembler anecdotique mais elle ne l'est pas car, comme expliqué plus loin, ces deux réseaux n'ont pas le même comportement dynamique.

La figure 1.1 donne un exemple de réseau de réactions issu d'un réseau métabolique de [PPW+03] accompagné de sa matrice de stœchiométrie. Nous en donnons aussi une représentation schématique souvent utilisée par les biologistes et modélisateurs pour se figurer la topologie du réseau. La matrice de stœchiométrie de ce réseau est donnée dans la figure 1.1.



$$
\begin{aligned}
2A &\rightarrow 2B \\
2B &\rightarrow \text{byp} + C \\
2B + \text{cof} &\rightarrow D + \text{byp2} \\
D &\rightarrow E + \text{cof} \\
C + \text{cof} &\rightarrow D \\
C &\rightarrow E
\end{aligned}
$$

$$
\begin{array}{c}
\text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{cof} \\ \text{byp} \\ \text{byp2}
\end{array}
\begin{pmatrix}
-2 & 0 & 0 & 0 & 0 & 0 \\
2 & -2 & -2 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & -1 & -1 \\
0 & 0 & 1 & -1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & -1 & 1 & -1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}
$$

FIGURE 1.1 – **Premier exemple de réseau de réactions.** Un réseau métabolique ([PPW+03]), sa modélisation en tant que réseau de réactions et la matrice de stœchiométrie associée. Dans la représentation graphique, les sorties sont mises en valeurs à droite parfois en les dédoublant

1.1.2 Représentations

Il existe deux représentations importantes des réseaux de réactions. La première est importante d'un point de vue fondamental car elle définit le réseau en tant que processus de

calcul. La seconde tout aussi importante est sa représentation sous forme de graphe qui permet sa représentation graphique, importante en biologie afin de pouvoir raisonner à l'aide de diagrammes.

Représentation par réseau de Petri D'un point de vue informatique théorique, les réseaux de réactions correspondent à la notion de *réseau de Petri* [Mur89]. Les réseaux de Petri ont été introduit dans le domaine de la vérification formelle en tant que modèle de calcul moins puissant que la machine de Turing, et donc plus facile à étudier, mais qui permet tout de même d'aborder les problématiques liées au parallélisme et au partage de ressources. Un réseau de Petri est constitué d'un ensemble de *places* pouvant accueillir un nombre fini de *jetons*. Les quantités de *jetons* sont susceptibles d'évoluer par l'action de *transitions* qui consomment des jetons dans certaines places et en produisent dans d'autre place.

Définition 1.3 (Réseau de Petri)**.** Un *réseau de Petri* est un quadruplet $(P, T, Pre, Post)$ où

— $P = \{p_1, \ldots, p_n\}$ est un ensemble fini de *places*,

— $T = \{t_1, \ldots, t_n\}$ est un ensemble fini de *transitions*,

— Pre est une matrice entière de taille $n \times m$ où $Pre_{i,j}$ indique le nombre de jetons de la place p_i consommés par la transition t_j,

— $Post$ est une matrice entière de taille $n \times m$ où $Post_{i,j}$ indique le nombre de jetons produits dans la place p_i par la transition t_j.

L'équivalence avec les réseaux de réactions, où les places correspondent aux espèces et les transitions aux réactions, est immédiate. Les deux terminologies sont équivalentes et l'utilisation de l'une plutôt que l'autre dépend essentiellement de la communauté concernée. Dans cette thèse je parle de réseaux de réactions qui sont plus familiers aux biologistes mais le discours aurait pu porter de manière équivalente sur les modèles utilisant les réseaux de Petri.

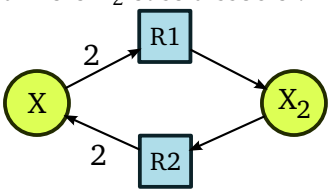
Représentation par graphe biparti Il est commode de pouvoir visualiser graphiquement un réseau de réaction (n, m, α, β). On utilise alors généralement une représentation en tant que graphe orienté biparti $G = (V, E)$ où les sommets sont les espèces et les réactions $V = \{X_1, \ldots, X_n\} \sqcup \{R_1, \ldots, R_m\}$. L'arc $X_i \to R_j$ (resp. $R_j \to X_i$) appartient à E si est seulement si X_i est un réactif (resp. produit) de R_j. On remarquera que c'est également la représentation graphique utilisée pour représenter les réseaux de Petri. En biologie des systèmes, cette représentation graphique des réseaux métaboliques est souvent nommée *carte métabolique*. Le lecteur trouvera des exemples de représentations graphiques dans la section suivante.

1.1.3 Petits exemples

À titre d'illustration, je présente deux exemples de réseaux de réactions. Le comportement dynamique de ses réseaux n'est pour l'instant pas défini, leur dynamique différentielle sera définie plus loin.

Exemple 1.1 (Dimérisation)**.** Un réseau peut tout d'abord représenter un ensemble de réactions chimiques. L'exemple suivant est celui de la dimérisation [Wil12] dans lequel deux molécules X identiques peuvent s'associer pour former un dimère X_2 et se dissocier.

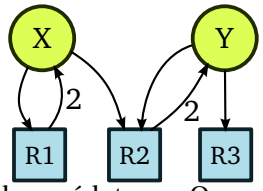
$$2X \to X_2$$
$$X_2 \to 2X$$



Toutefois comme je l'ai mentionné un réseau de réaction ne modélise pas nécessairement un ensemble de réactions chimiques. L'exemple qui suit est un cas particulièrement intéressant.

Exemple 1.2 (Réactions de Lotka-Volterra)**.** Lotka étudia en 1910 [Lot10] le système de réactions chimiques suivant et explora ses propriétés dynamiques oscillatoires remarquables. Il étendit ensuite la portée de ses résultats en considérant que le système pouvait être un bon modèle pour décrire les oscillations de populations de plantes et d'herbivores. De manière indépendante, Volterra proposa en 1927 [Vol27] le même modèle pour décrire aussi l'évolution de populations de proies et de prédateurs. On l'appelle aujourd'hui système de Lotka-Volterra.

$$X \to 2X \qquad \text{(reproduction)}$$
$$X + Y \to 2X \qquad \text{(prédation)}$$
$$Y \to \varnothing \qquad \text{(famine)}$$



Ici, l'espèce X représente les proies et l'espèce Y représente les prédateurs. On considère que les proies ont accès une quantité constante de nourriture contrairement aux prédateurs qui doivent se nourrir de proies. La reproduction des proies est donc constamment exponentielle (réaction 1) contrairement à celle des prédateurs qui est pondérée par celle des proies (réaction 2). Si un prédateur ne trouve pas de proie il finira par mourir (réaction 3).

1.1.4 Exemple de la synthèse protéique cap-dépendante de la cellule œuf chez l'oursin

Je présente maintenant un exemple un peu plus conséquent qui servira de fil rouge pour illustrer les concepts introduits dans ce chapitre : la modélisation dynamique par équations différentielles mais aussi l'analyse de l'équilibre des flux. On verra en particulier que même sur cet exemple concret de taille très réduite on rencontre immédiatement de nombreux problèmes en particulier en ce qui concerne la détermination des lois cinétiques et l'identification des paramètres.

Le choix de cet exemple est motivé par une étroite collaboration de mon équipe de recherche avec les biologistes de la station de recherche marine de Roscoff (France). Cette

collaboration a entraîné d'intéressantes discussions dans lesquelles nous avons pu nous confronter directement aux difficultés de la modélisation dynamique par équations différentielles.

La synthèse des protéines chez les procaryotes et les eucaryotes se décompose de manière très simplifiée en deux étapes :

— la *transcription* au cours de laquelle un gène est lu et transcrit en ARN messager (ARNm)

— et la *traduction* au cours de laquelle un ARN messager est traduit en protéine.

La cellule œuf de l'oursin contient initialement une réserve d'ARN messagers prêts à être traduits au moment de la fécondation. Les biologistes de la station de Roscoff s'intéressent donc au système qui permet cette traduction et plus spécifiquement ils s'intéressent au mécanisme de la *synthèse protéique cap-dépendante* [BPSC10]. En effet, la plupart des ARN messagers présents dans la cellule œuf possèdent une coiffe qui les oblige à être traités par un mécanisme spécifique de synthèse. Un modèle simplifié [LRML+14] de ce mécanisme de synthèse est décrit par le réseau de réaction suivant.

Exemple 1.3 (Synthèse protéique cap-dépendante chez l'oursin)**.**

$R1 : 4E + BP \rightarrow 4EBP$	(capture de 4E)
$R2 : 4EBP \rightarrow 4E + BP$	(libération de 4E)
$R3 : 4E + 4G \rightarrow 4E4G$	(association de 4E et 4G)
$R4 : 4E4G \rightarrow 4E + 4G$	(dissociation de 4E4G)
$R5 : 4E4G \rightarrow 4E4G + Prot$	(synthèse protéique)
$R6 : 4E4G \rightarrow 4E4G + BP$	(synthèse de BP)
$R7 : BP \rightarrow \varnothing$	(dégradation de BP)

Encore une fois, il s'agit d'un modèle très simple cachant une réalité beaucoup plus complexe [BPSC10]. Dans cette version simplifiée, la synthèse cap-dépendante repose sur le complexe moléculaire 4E4G qui recrute les ARN messagers dotés d'une coiffe pour les traduire. Avant la fécondation, la présence de ce complexe est faible car l'un de ses constituants, 4E, est fortement capturé par une protéine nommée BP pour *binding protein*.

Question biologique Au moment de la fécondation de l'œuf, les biologistes observent une diminution de la quantité totale de BP (sous sa forme libre et associée à 4E) associée à une augmentation de la quantité de 4E4G. Cela conduit à l'initiation de la synthèse protéique. Le réseau de réactions proposé avait pour but de répondre à une question essentielle : comment comprendre l'augmentation de la quantité du complexe 4E4G ? Pour cela deux hypothèses biologiques étaient envisagées :

— la première hypothèse était une *augmentation de la libération de 4E* (intensification de la réaction 2) qui conduit à une importante augmentation de 4E libre disponible et également une augmentation de la quantité de BP prête à être dégradée,

— la seconde hypothèse était une *augmentation de la dégradation de BP sous sa forme libre* (intensification de la réaction 7) conduisant par un *effet de pompe* à une diminution de la capture de 4E et donc à une quantité plus faible de 4EBP et une plus grande disponibilité de 4E.

Pour déterminer laquelle de ces deux hypothèses est valide nous avons maintenant besoin de comprendre ce que signifie l'*intensification d'une réaction* et comment obtenir l'*évolution des quantités de matières* à partir du réseau de réactions afin de pouvoir la comparer aux données expérimentales disponibles. Nous arrivons donc naturellement à la question de la dynamique des réseaux de réactions et nous reviendrons ultérieurement à cet exemple.

1.2 Dynamiques par équations différentielles

Nous allons maintenant définir la dynamique d'un réseau de réactions c'est-à-dire comment obtenir l'évolution temporelle des quantités de chaque espèce. En réalité nous devrions plutôt affirmer que nous allons définir *une* dynamique possible car il existe dans la littérature de nombreuses propositions. Ce chapitre traite des méthodes reposant sur les équations différentielles mais dans le chapitre suivant nous présenterons des méthodes reposant sur les chaînes de Markov. Il existe bien d'autres possibilités plus qualitatives que quantitatives dont nous reparlerons plus loin mais que l'on peut citer tout de suite : des dynamiques booléennes avec les réseaux booléens (*Boolean networks*) [Kau69, Kau93] et leur version stochastique les réseaux booléens probabilistes (*probabilistic Boolean networks*) [SDZ02], les réseaux de Thomas [Tho73], les frappes de processus [PMR12], ...

À l'heure actuelle, le paradigme le plus largement répandu chez les biologistes mais aussi chez les modélisateurs est celui des équations différentielles ordinaires dans lequel les quantités de matières sont représentées de *manière continue* sous forme de *concentrations* et *déterminées* par un système autonome d'équations différentielles ordinaires couplées. Cette approche a de nombreux avantages. C'est celle habituellement utilisée dans le domaine de la chimie, elle paraît donc à première vue adaptée à la description de phénomènes chimiques qui se passent au sein de la cellule. Un point non négligeable est qu'elle est conceptuellement simple : les équations différentielles font partie du langage commun à toutes les disciplines scientifiques et elles permettent donc une collaboration plus facile entre mathématiciens, informaticiens, physiciens, chimistes et biologistes. Elle a aussi le mérite de conduire à des simulations informatiques faciles grâce à aux algorithmes de résolution numérique qui peuvent aujourd'hui être réalisées rapidement sur n'importe quel ordinateur de bureau. Toutefois cette approche pose aussi de nombreuses limites autant sur ses hypothèses de départ (quantités continues, hypothèses cinétiques) que sur sa mise en pratique des cas concrets. Après avoir introduit les équations différentielles traditionnellement associées aux réseaux de réactions je détaillerai certaines de ces limites qui motivent les approches par contraintes et les approches stochastiques traitées dans le chapitre suivant.

1.2.1 Définition

Les méthodes dynamiques différentielles consistent à représenter la vitesse d'évolution des quantités de matière à l'aide d'un système d'équations différentielles ordinaires couplées et autonome de la forme générale

$$\frac{d\vec{x}(t)}{dt} = F(\vec{x}(t)) \qquad (t \in \mathbb{R}). \tag{1.1}$$

Le vecteur $\vec{x}(t)$ est un vecteur colonne à n composantes *réelles positives* représentant les *concentrations* de chaque espèce au temps t. On l'appellera également *trajectoire du système*. L'ensemble des valeurs susceptibles d'être prises par $\vec{x}(t)$ est l'octant positif $\mathbb{R}_+^n$, on dira que cet ensemble est l'*espace des états*. La fonction F s'appelle la *loi* d'évolution du système. Le système est dit *autonome* car F ne dépend pas du temps, autrement dit les vitesses des quantités de matière ne dépendent que de l'état courant du système et les lois physiques sont indépendantes du temps. En général F vérifie les hypothèses du théorème de Cauchy-Lipschitz portant sur l'existence et l'unicité de la solution à conditions initiales fixées. On aboutit donc à des trajectoires *continues* et *déterministes*.

1.2.2 Lois de flux

On souhaite maintenant que les solutions correspondent à l'évolution du réseau de réactions considéré, or, pour le moment nous n'avons établi aucune relation entre le réseau et la loi d'évolution du système. Pour cela nous allons introduire la notion de *flux* qui définit intuitivement la notion de *taux d'activité* d'une réaction.

Définition 1.4 (Lois de flux)**.** Soit un réseau de réactions (n, m, α, β) de matrice de stœchiométrie S. On dira que la dynamique différentielle suit *une loi de flux* si elle peut s'écrire

$$\frac{d\vec{x}(t)}{dt} = F(\vec{x}(t)) = S\vec{f}(\vec{x}(t)) \qquad (t \in \mathbb{R}) \tag{1.2}$$

où $\vec{f}$ est une fonction $\vec{f} : \mathbb{R}_+^n \to \mathbb{R}_+^m$ représentant *les flux* de chaque réaction dans un état donné. Le flux f_j de la réaction R_j ne dépendra que des concentrations des réactifs de la réaction R_j.

Le jeu consiste maintenant à définir des fonctions de flux *correctes* vis à vis du système vivant étudié. Traditionnellement, on utilise la *loi d'action de masse* mais il en existe beaucoup d'autres (des exemples fréquents sont donnés dans la table 1.1). Puisque chacune définit sa propre dynamique on rencontre la première difficulté des méthodes dynamiques qui présupposent que l'on connait les lois des flux de chaque réaction.

La loi la plus communément utilisée est la *loi d'action de masse*.

Définition 1.5 (Loi d'action de masse)**.** La *loi d'action de masse* est la dynamique différentielle définie par les flux

$$f_j(\vec{x}) = k_j \cdot \prod_{i=1}^{n} x_i^{\alpha_i} \tag{1.3}$$

où k_j est un paramètre cinétique nommé *constante cinétique* de la réaction R_j.

Cette loi tire son origine de l'étude de systèmes chimiques en équilibre thermodynamiques. Elle signifie que le flux d'une réaction est proportionnel à $x_i^{\alpha_i}$ pour chaque réactif X_i. Ce choix peut se comprendre lorsqu'on considère que le nombre de façons de choisir α_i molécules X_i parmi les N_i molécules X_i disponibles, c'est-à-dire $\binom{N_i}{\alpha_i}$, est asymptotiquement équivalent à $N_i^{\alpha_i}/\alpha_i! = \Omega^{\alpha_i} x_i^{\alpha_i}/\alpha_i!$ quand $N_i \to +\infty$. Si on admet que le flux d'une réaction est proportionnel au nombre de combinaisons de molécules de réactifs disponibles, on obtient alors l'équation (1.3) où les facteurs $\Omega^{\alpha_i}/\alpha_i!$ sont intégrés dans la constante cinétique.

Loi	**Type de réaction**	**Fonction de flux**	**Paramètres**	**Utilisation**
Action de masse [Hel08]	$\sum_{i=1}^{n} \alpha_i X_i \rightarrow \dots$	$f_j(\vec{x}) = k_j \cdot \prod_{i=1}^{n} x_i^{\alpha_i}$	constante de réaction k_j	Réactions chimiques simples
Michaelis-Menten [MM13]	$S \rightarrow P$	$f(\vec{x}) = \frac{V_{\max} x_S}{K_m + x_S}$	$V_{\max}, K_m$	Réactions enzimatiques
Hill (activation) [Alo06]	$X \rightarrow Y$	$f(x) = \frac{\beta x^n}{K^n + x^n}$	K, n, β	Activation de l'expression d'un gène
Hill (répression) [Alo06]	$X \rightarrow Y$	$f(x) = \frac{\beta}{1+(x/K)^n}$	K, n, β	Représsion de l'expression d'un gène

TABLE 1.1 – Exemples courants de lois cinétiques

Dimension physique de la constante cinétique La définition de la loi d'action de masse implique que la dimension physique de la constante cinétique dépend du type de réaction considérée. En effet l'unité du flux dans le système international est le $molL^{-1}s^{-1}$. À partir de l'équation (1.3), on en déduit que l'unité de k_j est le $mol^{1-\sum_{i=1}^{n}\alpha_i}L^{\sum_{i=1}^{n}\alpha_i-1}s^{-1}$. Par exemple, la constante cinétique de la réaction $X+Y \to XY$ s'exprime en $mol^{-1}Ls^{-1}$.

À titre d'illustration, nous allons définir dynamique de lois d'action de masse pour les deux exemples simples suivants. Ils nous serviront également de points de comparaison lorsque nous étudierons les méthodes dynamiques probabilistes dans le chapitre suivant.

Exemple 1.4. On considère un réseau à deux réactions

$$\begin{aligned} R_1 &: X \to Y \quad (k_1) \\ R_2 &: Y \to X \quad (k_2) \end{aligned}$$

avec comme concentrations initiales $(x(0), y(0)) = (1,0)$. La loi d'action de masse donne le système différentiel linéaire

$$\begin{aligned} \frac{dx}{dt} &= k_2 y - k_1 x \\ \frac{dy}{dt} &= k_1 x - k_2 y \end{aligned}$$

qui peut se réécrire sous forme matricielle

$$\frac{d\vec{x}}{dt} = A(k_1, k_2)\vec{x}$$

où $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$ est le vecteur des concentrations. La solution de ce système peut d'obtenir à l'aide de l'exponentielle de matrice $\vec{x}(t) = \exp(tA(k_1,k_2))\vec{x}(0)$ et on obtient après calcul de cette exponentielle

$$\begin{aligned} x(t) &= \frac{k_2}{k_1+k_2} + \frac{k_1}{k_1+k_2}\exp(-t(k_1+k_2)) \\ y(t) &= \frac{k_1}{k_1+k_2} - \frac{k_1}{k_1+k_2}\exp(-t(k_1+k_2)). \end{aligned}$$

L'exemple précédent a conduit à des équations différentielles linéaires pour la loi d'action de masse. La résolution de l'équation homogène associée se résume alors à un calcul d'exponentielle de matrices. Ces réseaux qu'on qualifiera de *linéaires* ont d'autres bonnes propriétés que nous exploiterons dans le chapitre suivant.

Définition 1.6 (Ordre d'une réaction et réseaux linéaires)**.** Dans un système de réactions (n, m, α, β), l'*ordre de la réaction* R_j est l'entier $\sum_{i=1}^{n}\alpha_{i,j}$. Le réseau est dit *linéaire* lorsque toutes ses réactions sont d'ordre 1 au plus.

Voici maintenant un exemple non linéaire.

Exemple 1.5. Considérons un système où il n'y a que des molécules X qui se désintègrent lorsqu'elles se rencontrent :

$$2X \to \varnothing \quad (k).$$

On suppose qu'il y a initialement une concentration $\vec{x}(0) = 4$, la loi d'action de masse nous donne comme dynamiques

$$\frac{dx}{dt} = -2kx^2. \tag{1.4}$$

Cette équation différentielle s'intègre facilement et on montre que l'unique solution satisfaisant les conditions initiales est

$$x(t) = \frac{4}{1+8kt}. \tag{1.5}$$

La quantité de molécules X décroit donc selon une fonction inverse d'après la loi d'action de masse.

Comme mentionné ci-dessus, la dynamique différentielle de ces deux exemples sera intéressante à comparer à leur dynamique stochastique.

1.2.3 Limites : lois, paramètres et passages à l'échelle

Les approches par équations différentielles sont pratiques car elles utilisent un formalisme simple et facile à comprendre. Comme montré dans les exemples précédents, on peut dans certains cas obtenir une solution analytique pour les trajectoires. Dans les autres cas il est en général possible d'utiliser des algorithmes classiques de résolution numérique (Runge-Kutta, ...) pour produire des simulations correctes. Dans le contexte de cette thèse qui est celui de la comparaison d'un modèle, le réseau de réactions, et de données, les séries temporelles expérimentales, il suffirait donc de comparer les séries temporelles obtenues par simulation numérique et celles obtenues par l'expérience pour conclure. Toutefois certaines limites, concernant la détermination des lois de flux et de leurs paramètres, méritent d'être soulignées. Elles nous conduiront dans un premier temps à considérer des approches par contraintes qui sont au cœur de cette thèse. Nous verrons par la suite les limites liées au déterminisme qui nous amèneront dans le chapitre suivant à considérer les approches stochastiques.

Détermination des lois Une limite très importante de l'approche présentée est que le modélisateur doit d'une part déterminer quelles lois cinétiques il associe à chaque réaction et d'autre part déterminer les valeurs des paramètres de ces lois. La loi d'action de masse est adaptée aux réactions élémentaires d'un système chimique homogènes à l'équilibre thermodynamique. Cependant, l'organisation du vivant est bien plus complexe et ne peut se résumer en une suite de *tubes à essais*. Un autre point est que les réactions ne représentent pas nécessairement des réactions chimiques élémentaires entre molécules dans un milieu homogène. On peut par exemple considérer l'exemple suivant de la régulation d'un gène [Wil06].

Exemple 1.6 (Expression régulée d'un gène)**.**

$$\begin{aligned} g + TF &\rightarrow g - TF \\ g - TF &\rightarrow g + TF \\ g - TF + ARNP &\rightarrow g - TF + ARNP + ARNm \end{aligned}$$

Dans cet exemple, g désigne la région promotrice située en amont d'un gène sur lequel peut se fixer un facteur de transcription TF de manière réversible. Lorsque ce le facteur de transcription est en place, l'ARN polymérase ($ARNP$) peut alors transcrire le gène en ARN messager ($ARNm$). Ici, on a donc affaire à un ou des sites d'une seule macromolécule (l'ADN) (qui n'est donc pas répartie de façon homogène) sur lesquels peuvent se fixer des facteurs de transcription. On constate souvent l'apparition de phénomènes de saturations autour du site de fixation qui conduisent, lorsqu'on cherche à établir des dynamiques à considérer d'autres lois de flux telles que la $loi\ de\ Hill$ aussi appelée $sigmoïde$ [Alo06]. Cette fonction prend en compte la possible saturation de la région promotrice. Dans les cas plus complexes, il peut même y avoir plusieurs facteurs de transcription qui doivent s'assembler dans un certain ordre [Wil06]. Un autre cas où la loi d'action de masse n'est pas adaptée est celui où la réaction considérée est en fait une réaction bilan, synthétisant un mécanisme complexe de plusieurs réactions élémentaires *réelles* non modélisées. Un exemple connu est celui des réactions enzymatiques.

Exemple 1.7 (Réactions enzymatiques)**.** Une enzyme E peut se fixer à un substrat S pour former un produit P. Cela correspond à un réseau de trois réactions

$$S + E \leftrightarrow SE \rightarrow P + E. \tag{1.6}$$

En admettant que ces trois réactions admettent une dynamique d'action de masse et sous certaines hypothèses sur leurs vitesses relatives, on résume souvent ce réseau à une seule réaction

$$S \rightarrow P \tag{1.7}$$

mais en utilisant l'équation différentielle de Michaelis-Menten [MM13]

$$\frac{dp}{dt} = \frac{V_{\max} \cdot s}{s + K_M}. \tag{1.8}$$

Lorsque $s \rightarrow +\infty$, $dp/dt \rightarrow V_{max}$ de telle sorte que la quantité limitée d'enzymes est prise en compte sous forme de paramètres. On a ainsi simplifié le réseau en passant de trois à une réaction et de trois à deux espèces mais la conséquence est que la loi d'évolution est plus complexe avec deux paramètres au lieu d'un.

La détermination des lois est donc un choix de modélisation qui est en général déterminé par connaissance générale du type de mécanismes impliqués. Le tableau 1.1 donne quelques exemples de lois très utilisées en biologie des systèmes ainsi que leurs paramètres.

Identification des paramètres Nous aboutissons alors à un problème majeur : même lorsque des lois de flux correctes ont été choisies, il est nécessaire de trouver les valeurs adéquates des paramètres cinétiques. Cette détermination n'est pas aisée lorsque peu d'informations cinétiques (mesures de trajectoires) sont accessibles par l'expérience. Il faut très souvent se contenter d'ordres de grandeurs, en considérant que deux réactions ont la même fonction de flux, de mesures effectuées *in vitro* ou de mesures obtenues pour des réactions similaires chez d'autres organismes (comme par exemple dans [AFLB+12] ou [LRML+14]). On pourra trouver une présentation et une comparaison des méthodes d'inférence de paramètres existantes dans [MMB03]. En général, on formalise cette inférence comme un problème d'optimisation non linéaire qui consiste à déterminer les paramètres qui minimisent une certaine fonction (la distance entre les valeurs prévues par le modèle et les données expérimentales). De plus il arrive que les paramètres de certains systèmes différentiels ne puissent pas être déterminés. Le processus de naissance et de décès, aussi appelé processus de vie et de mort, en est un exemple [Wil06].

Exemple 1.8 (Processus de naissance et de décès)**.**

$$X \to 2X \qquad (\lambda) \qquad \text{(naissance)}$$
$$X \to \varnothing \qquad (\mu) \qquad \text{(mort)}$$

Les équations de la loi d'action de masse s'écrit

$$\frac{dx}{dt} = \lambda x - \mu x. \tag{1.9}$$

La solution analytique de cette équation différentielle est

$$x(t) = x_0 \exp((\lambda - \mu)t). \tag{1.10}$$

Imaginons que l'on puisse mesurer expérimentalement $x(t)$ au cours d'un intervalle de temps $[a, b]$. On pourra alors calculer $\ln x(t)$ qui sera une droite de pente exactement égale $\lambda - \mu$. Toutefois on ne pourra jamais en tirer d'information sur la valeur du taux de natalité λ ni sur celle du taux de mortalité μ. En effet, deux jeux de paramètres ayant la même différence $\lambda - \mu$ auront la même solution analytique, ils sont donc indistinguables. On dit que les paramètres λ et μ ne sont pas *identifiables*. Pour cette raison il est important d'être très critique lorsqu'on considère les paramètres fournis par une méthode d'inférence, puisqu'il peut exister plusieurs ensemble de valeurs de paramètres correspondant aux données à disposition (une infinité dans notre exemple). Par exemple, les méthodes de détermination de paramètres faisant intervenir le hasard (par exemple en utilisant la méthode CMA-ES [HO01]) pourront produire des valeurs de paramètres différents à chaque exécution en ce qui concerne les paramètres qui ne sont pas identifiables.

Passage à l'échelle Vu toutes ces difficultés pour construire un modèle différentiel, on se rend compte qu'on ne va pas pouvoir utiliser ces méthodes pour étudier des réseaux de plus grande taille. Ainsi, l'étude de [MMB03] montre que l'inférence de seulement 36 paramètres

nécessite plusieurs heures sur un ordinateur de bureau en utilisant les meilleurs méthodes. Les grands réseaux métaboliques [EP00] ou les réseaux de signalisation [PAK13] pouvant être constitués de plusieurs centaines de réactions, on est alors confronté à un problème de *passage à l'échelle* des méthodes différentielles. Ce problème mobilise aujourd'hui grandement la communauté de la biologie des systèmes qui s'oriente dans deux grandes directions complémentaires.

La première est de *réduire la taille des réseaux* c'est-à-dire considérer des réseaux de plus petite taille mais qui conservent certaines propriétés du réseau étudié. Ces méthodes exploitent en général les différentes échelles de temps des réactions mises en jeu [RGZL08, CSRS+04, WZJC04, MLN15] et reposent parfois sur des concepts mathématiques complexes telles que la tropicalisation [SFR14]. Parfois la taille trop importante du réseau est due à une forte combinatoire moléculaire qui conduit à un nombre trop grand d'espèces. C'est par exemple le cas lors des réactions de polymérisation ou lorsque une molécule peut être phosphorylée sur plusieurs sites. En effet le formalisme trop pauvre des réseaux de réactions oblige alors dans le premier cas à introduire une espèce pour chaque longueur de polymère (une espèce pour chaque chaîne de n monomères) et dans le second cas, une espèce pour chaque sous-ensemble de sites phosphorylés. Pour traiter cette explosion combinatoire des espèces, des *langages basés règles* [DFF+07, FDK+09] tels que *Kappa* (κ) [DL04] ont été proposés en s'inspirant de la théorie des langages de programmation concurrente. Ces langages permettent un plus haut niveau d'expression des phénomènes biologiques tout en permettant d'appliquer des méthodes d'interprétation abstraite.

La seconde direction est de réduire les questions que l'on se pose sur le système. Par exemple, on peut seulement s'intéresser aux présences ou absences d'espèces chimiques plutôt qu'à leur quantités exactes et on se retrouve alors à étudier des dynamiques abstraites Booléennes. Il existe une vaste littérature qui se propose de considérer des dynamiques abstraites des réseaux et d'étudier à quelles questions elle permettent de répondre : réseaux Booléens déterministes [Kau69, Kau93] et probabilistes [SDZ02], réseaux de Thomas [Tho73], processus de frappe [PMR11, PMR12, FPI+12]. Ces méthodes forment ce que l'on appelle *modélisation qualitative* qui permet d'aborder certains problèmes (vérification de propriétés logiques [CFS06], possibilité de production d'une espèce, étude des réseaux de signalisation [PAK13], étude de la multi-stationnarité et des oscillations [Sno98, RC07], etc) mais dont le problème est qu'elle perd l'aspect *quantitatif* de la modélisation. Une autre grande approche est de considérer une hypothèse supplémentaire de *stationnarité*, c'est-à-dire qu'on se propose d'étudier la dynamique du réseau lorsqu'elle est dans un *régime d'équilibre*. Cette hypothèse est l'origine d'un vaste corpus de méthodes d'analyse par contraintes des flux d'équilibre. Nous présentons ces approches dans la section suivante. Les travaux présentés dans cette thèse se situent dans cette catégorie, nous présentons également une approche par contraintes où les contraintes sont obtenues à l'aide de mesures statistiques sur une population de trajectoires.

1.2.4 Exemple de la synthèse protéique chez l'oursin

Explication de la synthèse protéique Revenons à l'exemple du modèle de la synthèse protéique cap-dépendante de la cellule œuf de l'oursin. Ce modèle est une très bonne illustration des difficultés rencontrées, décrites dans [LRML+14], lorsqu'on souhaite développer des modèles différentiels. Le but de ce modèle était de tester deux hypothèses pour expliquer les courbes de synthèses protéiques après fécondation à disposition :

1. la déstabilisation du complexe 4E-BP (intensification de la R2) et
2. l'augmentation de la dégradation de BP (intensification de la R7).

Ces deux hypothèses ont été interprétées en tant qu'augmentation des constantes cinétiques associées à ces deux réactions, augmentation d'un facteur restant à déterminer. Les facteurs corrects doivent permettre d'obtenir les courbes de synthèses protéiques observées par des méthodes de *western blot*. Le problème est qu'il est nécessaire d'identifier en premier lieu la valeur des constantes initiales. Pour cela, on a utilisé trois méthodes :

— la reprise de valeurs chez une espèce voisine,

— des mesures de certaines constantes par un dispositif physique de *résonance plasmique de surface*,

— l'exploitation de mesures de concentrations à l'équilibre avant la fécondation.

Malgré la combinaison de ces trois méthodes aucune constantes multiplicatives n'a permis d'expliquer correctement la synthèse protéique. Il a alors fallu considérer une modification progressive de ces constantes avec un facteur multiplicatif changeant de valeur sur un intervalle de temps $[0, T]$. Autrement dit, on a considéré dans cet étude un système d'équations différentielles non autonome. Grâce à l'ajout de ce paramètre, les auteurs de l'étude ont pu mettre en évidence que la synthèse protéique observée n'est possible que lorsqu'on combine simultanément les deux hypothèses (déstabilisation du complexe 4EBP et de la dégradation de BP). Ce résultat est intéressant biologiquement mais on note que la phase d'apprentissage des paramètres cinétiques est difficile même pour ce modèle de taille réduite.

Cas de la cyclineB, passage à l'échelle Une autre question des biologistes au sujet de cette synthèse protéique est de comprendre le cas particulier de la protéine de cyclineB. En effet, l'expérience montre que sa synthèse se comporte différemment :

1. le début de sa synthèse est retardée par rapport aux autres protéines,
2. le taux de synthèse semble être environ double par rapport à la moyenne des autres protéines.

Pour comprendre cette exception, les biologistes ont imaginé un réseau de réactions plus complexe en intégrant un second mécanisme de synthèse de la cyclineB. Cette extension oblige alors à distinguer les types d'ARN messagers ce qui conduit à un réseau bien plus grand comportant 26 réactions et 21 espèces représenté en figure 1.2. Vu les difficultés rencontrées pour inférer les paramètres du petit modèle, il n'a pas été possible de procéder à une étude similaire pour tenter de valider ou d'invalider ce nouveau modèle plus complexe.

Il est nécessaire d'aller vers d'autres types de méthodes, par exemple les méthodes d'analyse stationnaire décrites dans la suite de ce chapitre.

1.3 Analyse stationnaire des flux et réfutation

Nous présentons dans cette section les méthodes d'analyse des flux stationnaires à l'aide d'approches par contraintes. Nous commençons par définir le cône d'équilibre qui est l'espace des vecteurs de flux permettant un équilibre des réactifs. Nous établissons aussi le lien entre les flux stationnaires et la valeur des pentes moyennes. Enfin après avoir présenté quelques cas d'utilisations classiques de ces contraintes, nous les appliquons dans le cadre de la réfutation d'un réseau avec un exemple chez l'oursin.

1.3.1 Cône d'équilibre des flux stationnaires et utilisation de données de pentes

L'analyse de l'équilibre des flux [PRP04, OTP10] constitue une famille d'approches de modélisation reposant sur l'étude des solutions *stationnaires* des équations différentielles de flux introduites précédemment :

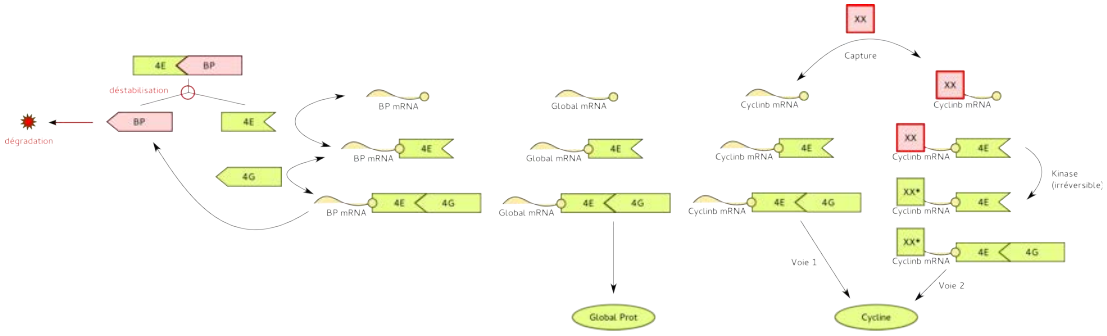
$$\frac{d\vec{x}(t)}{dt} = F(\vec{x}(t)) = S\vec{f}(\vec{x}(t)) \qquad (t \in \mathbb{R}).$$

Ici, la stationnarité signifie que les quantités de réactifs sont constantes c'est à dire $dx_i(t)/dt = 0$ pour tous les réactifs X_i. On permet toutefois aux produits de s'accumuler. Puisque la fonction de flux ne dépend pas des produits, elle est également constante $df(\vec{x})/dt = 0$ ce qui permet de considérer plus simplement sa valeur constante qu'on appelle *vecteur des flux stationnaires* $\vec{f}$. Pour vérifier la condition $dx_i(t)/dt = 0$ pour chaque réactif, le vecteur des flux stationnaire doit nécessairement appartenir à un cône d'équilibre dont voici la définition formelle.

Définition 1.7 (Cône d'équilibre des flux)**.** Soit S^* la matrice de stœchiométrie réduite issue de S en ne gardant que les lignes correspondant aux réactifs. On appelle *cône d'équilibre des flux* l'ensemble des vecteurs de flux suivant

$$\mathscr{C} = \left\{\vec{f} \geq \vec{0} \ / \ S^*\vec{f} = \vec{0}\right\}. \tag{1.11}$$

La contrainte $\vec{f} \geq \vec{0}$ correspond à notre choix de ne considérer que les réactions irréversibles. Ce n'est toutefois pas un facteur limitant, on peut toujours considérer le cas des réactions réversibles en introduisant deux réactions pour chacun des sens de la réaction. On remarque que $S^*\vec{f} = \vec{0}$ correspond à un système de contraintes linéaires pour $\vec{f}$ c'est pourquoi on parle d'*approches par contraintes*. En ce qui concerne les produits, on remarque que si $\vec{f}$ est constant, alors l'équation différentielle de la loi de flux impose que les pentes des produits sont constantes. Notons $\vec{\rho} = S\vec{f} = \frac{d\vec{x}}{dt}$ le vecteur des pentes de toutes les espèces, on a donc nécessairement $\rho_i = 0$ alors pour tous les réactifs. Cette dernière relation permet d'*établir un lien* entre le vecteur de flux $\vec{f}$ et les pentes $\vec{\rho}$ des sorties.



$$R1 : 4E + BP -> 4EBP$$
$$R2 = R1rev : 4EBP -> 4E + BP$$
$$R3 : BP -> \varnothing$$
$$R4 : 4E + BPRNA -> BPRBA4E$$
$$R5 = R4rev : BPRNA4E -> 4E + BPRNA$$
$$R6 : 4E + GlobalRNA -> GlobalRNA4E$$
$$R7 = R6rev : GlobalRNA4E -> 4E + GlobalRNA$$
$$R8 : 4E + CyclinbRNA -> CyclinbRNA4E$$
$$R9 = R8rev : CyclinbRNA4E -> 4E + CyclinbRNA$$
$$R10 : BPRNA4E + 4G -> BPRNA4E4G$$
$$R11 = R10rev : BPRNA4E4G -> BPRNA4E + 4G$$
$$R12 : GlobalRNA4E + 4G -> GlobalRNA4E4G$$
$$R13 = R12rev : GlobalRNA4E4G -> GlobalRNA4E + 4G$$
$$R14 : CyclinbRNA4E + 4G -> CyclinbRNA4E4G$$
$$R15 = R14rev : CyclinbRNA4E4G -> CyclinbRNA4E + 4G$$
$$R16 : CyclinbRNA + XX -> XXCyclinbRNA$$
$$R17 = R16rev : XXCyclinbRNA -> CyclinbRNA + XX$$
$$R18 : XXCyclinbRNA + 4E -> XXCyclinbRNA4E$$
$$R19 = R18rev : XXCyclinbRNA4E -> XXCyclinbRNA + 4E$$
$$R20 : XXCyclinbRNA4E -> XXCyclinbRNA4Eactive$$
$$R21 : XXCyclinbRNA4Eactive + 4G -> XXCyclinbRNA4E4G$$
$$R22 = R21rev : XXCyclinbRNA4E4G -> XXCyclinbRNA4Eactive + 4G$$
$$R23 : BPRNA4E4G -> BP + BPRNA + 4E + 4G$$
$$R24 : GlobalRNA4E4G -> GlobalProt + GlobalRNA + 4E + 4G$$
$$R25 : CyclinbRNA4E4G -> Cyclinb1 + CyclinbRNA + 4E + 4G$$
$$R26 : XXCyclinbRNA4E4G -> Cyclinb2 + XX + CyclinbRNA + 4E + 4G$$

FIGURE 1.2 – **Modèle de synthèse protéique à deux voies chez l'oursin.** Le but de ce modèle est d'introduire le cas particulier de la synthèse de la cyclineB qui serait synthétisée par deux voies, c'est-à-dire deux mécanismes distincts. Le mécanisme de la seconde voie est inspiré de celui d'une espèce marine voisine.

Une remarque importante dans cette démarche d'étude des cas stationnaires est que le vecteur de flux $\vec{f}$ est désormais une constante, les méthodes reposant sur ce type de contraintes ont l'important avantage de ne nécessiter aucune information sur les lois dynamiques ni sur les paramètres cinétiques associés. En quelque sorte le vecteur $\vec{f}$ est un résumé de toutes ces informations cinétiques et on cherche à déterminer directement $\vec{f}$ plutôt que des lois et leurs paramètres. L'approche est donc essentiellement stœchiométrique.

1.3.2 Méthodes par contraintes reposant sur l'équilibre des flux

Ainsi, l'hypothèse de stationnarité associée à des mesures de pentes sur certaines espèces produites par le réseau permet de définir un espace de solutions possibles pour le vecteur des flux stationnaires $\vec{f}$. Il reste à déterminer comment on exploite cet ensemble de valeurs et il existe dans la littérature une vaste gamme de méthodes reposant sur ce système de contraintes de flux à l'équilibre. Dans cette thèse on parlera de méthodes reposant sur l'équilibre des flux. La figure 1.3 donne de nombreux exemples de méthodes reposant sur l'équilibre des flux. La méthode la plus répandue est sans conteste le *flux balance analysis* (FBA) qui consiste à déterminer un flux $\vec{f}^*$ optimal qui maximise une certaine fonction biologique linéaire ϕ (typiquement la production de biomasse). D'un point de vue informatique, on se ramène alors à un problème de programmation linéaire

$$\begin{aligned} \text{Maximiser } & \phi(\vec{f}) \\ \text{Sous les contraintes } & S^* \vec{f} = \vec{0} \\ & \vec{f} \geq \vec{0}. \end{aligned}$$

Une autre approche appelée *flux variability analysis* (FVA) consiste à déterminer les bornes possibles pour chaque flux f_j parmi l'espace des flux permettant d'atteindre une certaine valeur de biomasse B. Cela peut également être obtenu à l'aide d'un programme linéaire

$$\begin{aligned} \text{Maximiser/minimiser } & f_j \\ \text{Sous les contraintes } & S^* \vec{f} = \vec{0} \\ & \phi(\vec{f}) = B \\ & \vec{f} \geq \vec{0}. \end{aligned}$$

Ces deux approches correspondent alors d'un point de vue informatique à la résolution de *programmes linéaires* pour laquelle on a développé des algorithmes très efficaces [DOW+55].

Le FBA est extrêmement utilisé en biologie des systèmes mais il a un défaut majeur : il repose sur l'hypothèse que la *nature optimise* la fonction ϕ. Il est donc difficile de donner un sens au vecteur solution $\vec{f}^*$. On préfèrera, et c'est l'esprit de cette thèse, considérer toutes les solutions possibles et en déterminer les conséquences. Le FVA est un bon exemple de ce genre d'approche qui ne repose pas sur une hypothèse d'optimisation de la nature : les bornes obtenues sont sûres compte-tenu de l'hypothèse de stationnarité et des contraintes choisies.

On peut également utiliser ces contraintes pour aboutir à des réfutations de réseau. En effet, si l'ensemble des solutions possibles déterminées par les contraintes est vide alors on peut en conclure soit qu'il n'y a pas de solutions stationnaires, soit que les données mesurées sont fausses, soit que le réseau de réactions n'est pas le bon. Cette approche est comparable au FVA car on détermine qu'il n'y a aucune valeur possible pour chacun des flux.

1.3.3 Réfutation d'un modèle 1 voie dans le modèle oursin

Comme illustration on se propose de réfuter dans l'exemple du modèle oursin, la théorie de la non existence d'un second mécanisme de synthèse de la CyclineB. On utilise l'approche stœchiométrique de l'étude par contraintes des *flux stationnaires* (FBA), ce qui permet de s'affranchir de connaissances sur les constantes cinétiques et des concentrations des métabolites. On caractérise l'état de fonctionnement stationnaire du système par son vecteur de flux $\vec{f}$ ayant autant de composantes que de réactions dans le système. On va ensuite utiliser une approche par contrainte pour intégrer nos hypothèses biologiques. En effet, nous avons montré qu'il est possible de relier des mesures de pentes stationnaires et les flux par l'équation $\vec{\rho} = S\vec{f} = \frac{d\vec{x}}{dt}$.

— **Le rapport de pentes**. Dans notre cas, on sait donner une estimation du rapport de pentes CyclineB/GlobalProt, ce qui correspond à la contrainte

$$\gamma_{inf} \leq \frac{f_{25} + f_{26}}{f_{24}} \leq \gamma_{sup} \tag{1.12}$$

qui peut s'écrire de manière équivalente avec des inéquations linéaires

$$\gamma_{inf} f_{24} \leq f_{25} + f_{26} \leq \gamma_{sup} f_{24}. \tag{1.13}$$

Les mesures expérimentales obtenues montrent que ce rapport se situe dans l'intervalle $[\gamma_{inf}, \gamma_{sup}] = [1,8; 2]$.

— **Les contraintes d'équilibre des réactifs**. En effet on a vu que pour avoir un régime stationnaire les réactifs doivent être en quantités constantes, c'est-à-dire de pente nulle. On obtient alors une contrainte d'équilibre pour chacun des réactifs. Par exemple, dans ce modèle l'équilibre de BP se traduit par

$$f_1 + f_3 = f_2 + f_{23}, \tag{1.14}$$

autrement dit, il y a autant de désintégrations de BP et de liaison avec 4E, que de synthèse de BP et de libération du complexe 4EBP.

— **Divers**. Dans le cas présent on se servira aussi de l'hypothèse

$$f_{24} = f_{25}$$

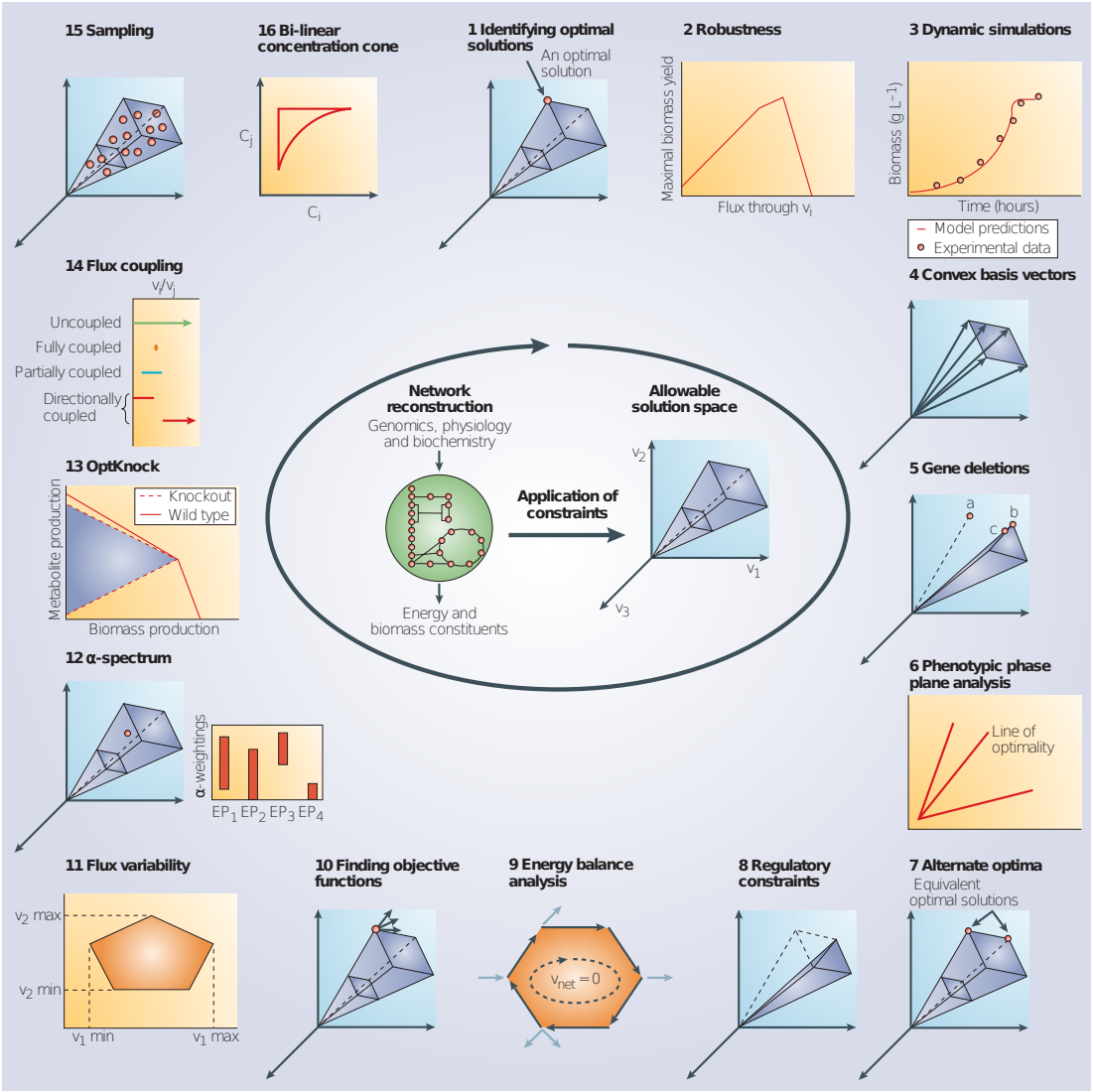


FIGURE 1.3 – **Quelques méthodes d'analyse reposant sur un système de contraintes des flux à l'équilibre.** (image extraite de [PRP04])

c'est-à-dire que la synthèse de la CyclineB par la voie $normale$ n'est pas plus importante que la moyenne. Cela peut s'obtenir en mesurant des quantités d'ARNm de cycline comparables à la moyenne et grâce aux équations de flux

$$f_{24} = k_{24}[GlobalRNA4E4G],$$
$$f_{25} = k_{25}[CyclinbRNA4E4G]$$

avec $k_{24} = k_{25}$ puisque les deux réactions correspondent au même mécanisme de transcription.

On obtient alors un grand système d'équations et d'inéquations linéaires. Toute solution de ce système correspond à un régime stationnaire (solution à pentes constantes) valide pour le système sous nos hypothèses et nos mesures expérimentales. S'il n'y a pas de solution c'est que le modèle ou une de nos hypothèse est fausse. Á partir de ces contraintes, nous nous sommes intéressés aux valeurs de (f_{25}, f_{26}) dans l'espace solution qui correspondent aux flux de productions de la cyclineB par les voies 1 et 2 respectivement. Ces valeurs peuvent s'obtenir par exemple en fixant par contrainte f_{25} et en minimisant/maximisant la fonction objectif $\vec{f} \mapsto f_{26}$. Pour calculer plus simplement ces bornes, nous avons omis les contraintes d'équilibre des réactifs ce qui donne une sur-approximation de l'espace des solutions. On obtient alors la figure 1.4. Sur la figure on observe alors que pour tout $\vec{f}$ solution,

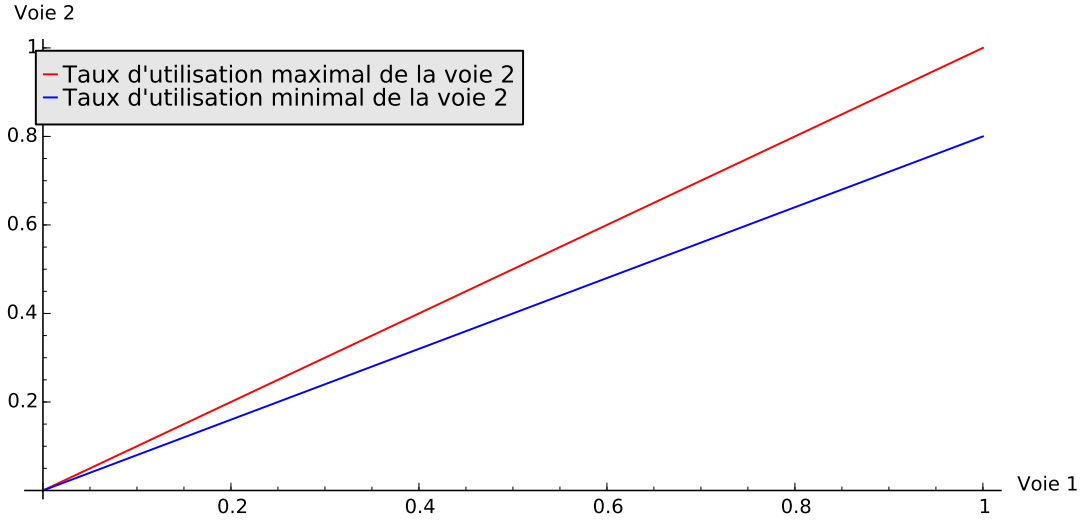


FIGURE 1.4 – **Comparaison des taux d'utilisation des deux voies dans le modèle de synthèse protéique chez l'oursin.** Ce graphique représente des bornes pour les valeurs possibles de f_{26} en fonction de celles de f_{25} et d'après les hypothèses que nous avons formulées.

$f_{25} > 0$ implique que $f_{26} > 0$. Cela peut être aussi montré par le calcul sur l'ensemble des contraintes. Une conséquence est qu'**on peut réfuter tout modèle à 1 voie**. En effet, considérer un modèle à une seule voie de production de cyclineB revient à ajouter la contrainte

$f_{26} = 0$ inhibant la seconde voie. Or, d'après notre observation $f_{26} = 0$ implique $f_{25} = 0$, autrement dit il n'y aurait pas de solutions stationnaires permettant de production de cyclineB. Pourtant on observe bien la production de cyclineB, ainsi le modèle à 1 seule voie peut être réfuté. D'un point de vue contraintes, cela signifie qu'il n'y a pas de solution à notre ensemble de contraintes si on ajoute les contraintes $f_{26} = 0$ (hypothèse 1 voie) et $f_{25} > 0$ (déduite de l'observation de synthèse de la cyclineB).

Résultat 1.1. *Dans le modèle de synthèse protéique chez l'oursin avec modélisation de la cyclineB (figure 1.2), si on suppose que le rapport de pentes* $CyclineB/GlobalProt$ *est compris dans l'intervalle* $[\gamma_{inf}, \gamma_{sup}] = [1,8;2]$ *et que l'on a des quantités comparables d'ARNm* $[GlobalRNA4E4G] \simeq [CyclinbRNA4E4G]$*, alors la voie 1 ne peut être active que si la voie 2 est active. Autrement dit, les deux voies sont nécessaires pour expliquer la mesure de pente double pour la cyclineB.*

1.4 Les limites du déterminisme

Dans ce chapitre, nous avons présenté les réseaux de réactions ainsi que leur modélisation dynamique par équations différentielles. Nous avons mis en évidence les difficultés rencontrées lorsqu'on souhaite utiliser en pratique de tels modèles, principalement en raison du problème de l'inférence des paramètres cinétiques. Nous avons illustré ces difficultés sur l'exemple concret de la modélisation de la synthèse protéique cap-dépendante dans la cellule œuf de l'oursin. Ainsi, les approches différentielles souffrent d'un problème de passage à l'échelle. Un des moyens d'étudier de plus grands réseaux est de considérer les approches stationnaires, dont la plus connue est le *flux balance analysis*, et qui permettent tout de même d'étudier les réseaux de réactions en régime stationnaire sans connaissances des lois et constantes cinétiques, en exploitant uniquement l'information stœchiométrique du réseau. En particulier nous avons montré comment on peut obtenir des contraintes sur les flux stationnaires en exploitant des mesures de pentes moyennes de production des espèces. Lorsque les systèmes de contraintes obtenus n'admettent pas de solutions alors on peut considérer (sous réserve d'existence du régime stationnaire) que le réseau peut être réfuté. Dans cette thèse on considérera donc que le problème du passage à l'échelle est résolu dans ce cas grâce aux méthodes par contraintes.

Toutefois, une limite importante de la sémantique différentielle sont les caractères continus et déterministes de la dynamique. L'avantage de l'utilisation des concentrations continues est qu'elles résument en une seule valeur un grand nombre d'informations (quantités de molécules, positions, vitesses, *etc*). L'inconvénient est que la dynamique par action de masse présuppose une homogénéité du système, son équilibre thermodynamique et des populations suffisamment grandes. En effet, puisque une seule valeur des concentrations à un instant t modélise un grand nombre d'états réels possibles, il y a plusieurs concentrations possibles à l'instant $t + \Delta t$ et donc fondamentalement cette modélisation ne devrait pas être déterministe. Les équations différentielles doivent donc être comprises comme un modèle traitant de *valeurs moyennes* et qui est justifié quand les quantités de matières sont

suffisantes. Dans les cas de faibles concentrations ou de compartiments de petits volumes il n'est plus satisfaisant. Ce cas n'est pas rare en biologie puisque des études [Wil09, ARM98, GZRI+06, MA97] montrent l'importance des comportements stochastiques en biologie.

Pour résumer, les modèles déterministes différentiels nous confrontent aux problèmes suivants :

1. le caractère déterministe et continu des solutions ne permet pas de rendre compte de la nature discrète et stochastique du vivant,
2. le caractère déterministe ne permet de ne considérer que l'échelle de la population et rend impossible l'exploitation de multiples trajectoires individuelles autrement que par la moyennisation de ces trajectoires.
3. la difficulté de déterminer les lois cinétiques associées aux flux des réactions ainsi que leurs paramètres cinétiques, ce qui rend difficile le passage à l'échelle.

Nous avons montré qu'il est possible de contourner cette dernière difficulté en ne considérant que la dynamique stationnaire des réseaux ce qui conduit à l'éventail des méthodes par contraintes du vecteur de flux stationnaire $\vec{f}$. Toutefois, ces méthodes ne permettent pas de contourner les deux premières difficultés. En particulier, seules des données de pentes $\vec{\rho}$ permettent de contraindre $\vec{f}$, or ces pentes étant déterministes on est obligé de considérer, lorsqu'on possède plusieurs trajectoires expérimentales, la valeur moyenne des pentes. En particulier l'exemple présenté dans l'introduction de la thèse, qui est de tenter de réfuter l'un des modèles suivants

$$(mod\grave{e}le\,1)\left\{\begin{array}{lcl} R_1: & A \rightarrow & B+C \\ R_2: & B \rightarrow & A+2D \end{array}\right. \qquad (mod\grave{e}le\,2)\left\{\begin{array}{lcl} R_1: & A \rightarrow & B+C+2D \\ R_2: & B \rightarrow & A \end{array}\right.$$

à l'aide de mesures de plusieurs trajectoires, ne peut être résolu. En effet, on a montré que ces deux réseaux ont un comportement moyen identique. Seule la prise en compte de moments d'ordre supérieurs (variance, covariances, *etc*) permet de les différencier.

Le projet de cette thèse est donc d'adapter les méthodes par contraintes existantes en dynamiques différentielles à un cadre stochastique dans lequel il est possible de former des contraintes à partir d'informations sur les moyennes, les variances, les covariances d'un ensemble de trajectoires. En conséquence de quoi, il est maintenant naturel de considérer la sémantique stochastique des réseaux de réactions qui permettent d'obtenir des distributions de trajectoires plutôt qu'une dynamique à solution unique.

Chapitre 2

Modèles dynamiques probabilistes

L'objectif du chapitre précédent était d'introduire la sémantique déterministe différentielle de la dynamique d'un réseau de réactions. Nous avons constaté que cette sémantique nécessitait une bonne connaissance des lois de flux du réseau et des paramètres cinétiques. Par conséquent le passage à l'échelle, c'est-à-dire obtenir un bon modèle dynamique pour des réseaux de plusieurs dizaines de réactions devient impossible. Nous avons également présenté les méthodes de contraintes de flux à l'équilibre qui permettent de s'affranchir de la connaissance des lois dynamiques et de leurs paramètres. Comme exemple d'application de ces méthodes par contraintes, nous avons montré comment on peut réfuter des réseaux en fonction de données sur les pentes moyennes des sorties, lorsque les données mesurées conduisent à un système de contraintes incompatibles.

Toutefois, nous avons conclu en soulignant quelques lacunes des approches différentielles. Tout d'abord, elles sont *déterministes* et ne permettent pas par conséquent de traiter la variabilité des individus, elles ne peuvent être appliquées que pour modéliser le comportement moyen d'une population. Ainsi lorsqu'on est capable de mesurer le comportement dynamique de plusieurs individus, on ne peut exploiter toute l'information mesurée car on est forcé de ne considérer que leurs moyennes. D'un point de vue théorique c'est une lacune importante car les biologistes et les modélisateurs [MA97, ARM98, GZRI+06, Wil06] ont mis en évidence l'importance de la stochasticité dans le vivant et la nécessité de pouvoir travailler sur les comportements individuels. D'un point de vue plus pratique, lorsqu'on souhaite vérifier la cohérence entre un réseau de réactions et des données temporelles on souhaiterait pouvoir exploiter les variances et les covariances. Nous avons en effet proposé un exemple dans lequel deux réseaux distincts ne présentaient pas de différence du point de vue de leurs dynamiques moyennes mais qu'on pouvait distinguer par leurs variances et leurs covariances.

Ainsi, il est naturel de s'intéresser à la *sémantique stochastique* des réseaux de réactions qui permet de modéliser des trajectoires aléatoires, leurs moyennes, leurs variances, leurs co-variances, *etc.* Cette sémantique a pour principe de ne plus déterminer une unique trajectoire pour l'évolution temporelle du réseau mais de déterminer à la place la *distribution*

ou *loi de probabilité* d'un ensemble de trajectoires possibles. L'objectif de ce chapitre est de présenter la littérature existante en biologie des systèmes autour des sémantiques stochastiques. Traditionnellement, celle-ci est définie par une *chaîne de Markov* qui détermine la loi de probabilité des trajectoires. Nous montrerons que la connaissance de la loi des trajectoires, permet d'obtenir des expressions pour les moments de la trajectoire (espérances, variances, co-variances, *etc*). Pour vérifier la cohérence d'un réseau de réactions avec un ensemble de trajectoires expérimentales, on peut alors en théorie comparer les estimations des moments de ces trajectoires avec l'expression théorique des moments de la chaîne de Markov.

Toutefois, en pratique, on se heurte à des difficultés encore plus grandes que pour la sémantique différentielle. Premièrement, tout comme en sémantique différentielle, la chaîne de Markov décrivant la sémantique stochastique se définit à l'aide de paramètres cinétiques stochastiques qui sont difficiles à déterminer, en particulier lorsque les réseaux sont de grande taille. Deuxièmement, le nombre d'états de la chaîne de Markov étudiée correspond aux nombres d'états accessibles par le réseau, c'est-à-dire la combinatoire tous les vecteurs de quantités de matières (soit le nombre exact de molécules de chaque espèce) qu'on peut atteindre à partir de l'état initial. Ce nombre est donc potentiellement *exponentiel* en fonction du nombre d'espèces présentes dans le réseau. En conséquence le calcul de la loi des trajectoires s'avère impossible en pratique.

Dans ce chapitre, on présente donc également les méthodes utilisées dans le domaine pour estimer la loi des trajectoires. La méthode qui constitue le fondement des méthodes stochastiques en biologie des systèmes est l'utilisation d'une approche de Monte-Carlo appelée *algorithme de Gillespie* ou encore *algorithme de simulation stochastique* (SSA). Cette méthode permet de générer des trajectoires aléatoires pour le réseau de réactions dont la distribution est correcte vis à vis de la sémantique stochastique. En d'autres termes, il s'agit de simuler la chaîne de Markov décrivant la dynamique stochastique. Puisque nous nous intéressons à l'exploitation des moments d'ordre 2, il est judicieux de présenter également d'autres approches classiques permettant d'obtenir une approximation des moments : *la méthode des moments clos* (CMM) et en particulier *l'approximation de bruit linéaire* (LNA). Dans toutes ces approches d'approximation des moments, de l'algorithme de Gillespie au LNA, une connaissance des lois et des paramètres cinétiques demeure nécessaire.

Ainsi, l'objectif cherché dans cette thèse est de proposer des pistes de développement de méthodes de contraintes, indépendantes des lois et paramètres cinétiques, analogues aux méthodes d'équilibre de flux en sémantique déterministe, mais permettant d'exploiter des informations sur les moments d'ordre supérieur des trajectoires. S'il existe une vaste littérature autour des méthodes d'équilibre des flux (FBA), les approches par contraintes en sémantique stochastique, comme celles présentées dans cette thèse, sont une thématique intéressante à explorer. Malheureusement, nous verrons que les méthodes classiques décrites dans ce chapitre, en plus des difficultés mentionnées ci-dessus, ne permettent pas d'identifier rapidement un vecteur, analogue au vecteur de flux à l'équilibre, qu'on puisse lier aux moments de la loi des trajectoires. Un tel vecteur ne s'obtient qu'en étudiant le cas stationnaire qui doit être correctement défini en sémantique stochastique. Ce travail sera

effectué dans le chapitre 3 et les applications en termes de méthodes de contraintes seront développées dans les chapitres 4 et 5.

2.1 Modélisation dynamique Markovienne

La sémantique stochastique [Gil07, Wil06, Wil09] des réseaux de réactions se définit en termes de *processus stochastique* à valeurs dans un espace discret. Dans ce cas, le vecteur $\vec{x}(t) \quad (t \in \mathbb{R}^+)$ est désormais un *vecteur aléatoire* prenant ses valeurs dans $\mathbb{N}^m$. Dans la littérature, les premiers travaux en modélisation dynamique stochastique remontent au moins aux années 1940 [Del40, Bar58, Bar59, McQ67] mais c'est surtout Gillespie [Gil76, Gil77, Gil00, Gil01, GP03, Gil07] qui a contribué à diffuser la modélisation dynamique stochastique grâce à son algorithme éponyme. L'*hypothèse fondamentale* de la sémantique stochastique est décrite par Gillespie [Gil76] en ces termes : la probabilité "moyenne" qu'une combinaison de molécules correspondant aux réactifs d'une réaction R_j réagisse dans le petit intervalle de temps $[t, t+\delta t]$ vaut $c_j \cdot \delta t + o_{\delta t \to 0}(\delta t)$. Ici c_j est une constante cinétique synthétisant divers paramètres physiques des molécules (leurs volumes, températures, *etc*). L'article de Gillespie [Gil76] décrit également les hypothèses physiques sous-jacentes à l'hypothèse fondamentale : une répartition spatiale homogène de la matière et un équilibre thermodynamique (mais non chimique). À partir de cette hypothèse fondamentale et d'hypothèses d'indépendance non explicitées dans son article, Gillespie aboutit à une description probabiliste de la dynamique des trajectoires.

D'un point de vue plus formel, l'hypothèse dynamique introduite est de considérer que les trajectoires du réseaux sont les réalisations d'une chaîne de Markov. Dans ce chapitre, on choisira d'introduire la sémantique stochastique dans le formalisme plus mathématique des chaînes de Markov à temps continu, ce qui est parfaitement équivalent à la présentation faite par Gillespie [AK11]. Ainsi, contrairement à la sémantique déterministe, on considèrera maintenant que $\vec{x}(t)$ est une *chaîne de Markov à temps continu* [Wil06], c'est-à-dire que le comportement du futur dépend de son état présent mais pas de son passé. Plus formellement, si $\Sigma \subset \mathbb{N}^m$ désigne l'espace des états, alors la chaîne de Markov $\vec{x}(t)$ vérifie la propriété de Markov :

$$\mathbb{P}\left[\vec{x}(t_{n+1}) = \vec{y} \mid \vec{x}(t_n) = \vec{x}, \vec{x}(t_{n-1}) = \vec{x}_{n-1}, \ldots, \vec{x}(t_0) = \vec{x}_0\right] = \mathbb{P}\left[\vec{x}(t_{n+1}) = \vec{y} \mid \vec{x}(t_n) = \vec{x}\right], \quad (2.1)$$

pour toutes suites d'instants $0 \leq t_0 < \cdots < t_{n-1} < t_n < t_{n+1}$ et pour tous états $\vec{x}_0, \ldots, \vec{x}_{n-1}, \vec{x}, \vec{y}$. On supposera également que la chaîne de Markov est *homogène* ce qui signifie que sa loi d'évolution ne dépend pas de l'instant considéré

$$\mathbb{P}\left[\vec{x}(t_1) = \vec{y} \mid \vec{x}(t_0) = \vec{x}\right] = \mathbb{P}\left[\vec{x}(t_1 - t_0) = \vec{y} \mid \vec{x}(0) = \vec{x}\right]. \quad (2.2)$$

Par rapport à la présentation de Gillespie, cela correspond au fait que les constantes c_j ne dépendent pas du temps. Cela est l'équivalent du caractère autonome des systèmes différentiels en sémantique déterministe.

Les chaînes de Markov à temps continu sont déterminées par leurs *taux de transitions* qui quantifient la *propension* de la chaîne de Markov à transiter d'un état à l'autre. Nous donnons ici une définition de la sémantique stochastique, qui reprend essentiellement la définition des chaînes de Markov à temps continu, mais adaptée aux changements d'états qui correspondent au déclenchement d'une réaction du réseau [AK11].

Définition 2.1 (Sémantique stochastique)**.** La sémantique stochastique d'un réseau de réactions (n, m, α, β) de matrice de stœchiométrie $S = \beta - \alpha = (\vec{v}_j)_{1 \leq j \leq m}$ est la chaîne de Markov homogène à temps continu $(\vec{x}(t))_{t \in \mathbb{R}^+}$ définie par les lois de transitions

$$\forall t \geq 0, \forall \vec{x} \in \mathbb{N}^m, \forall 1 \leq j \leq m, \ \mathbb{P}\left[\vec{x}(t+\tau) = \vec{x} \mid \vec{x}(t) = \vec{x}\right] = 1 - h_j(\vec{x})\tau + o_{\tau \to 0}(\tau),$$
$$\forall t \geq 0, \forall \vec{x} \in \mathbb{N}^m, \forall 1 \leq j \leq m, \ \mathbb{P}\left[\vec{x}(t+\tau) = \vec{x} + \vec{v}_j \mid \vec{x}(t) = \vec{x}\right] = h_j(\vec{x})\tau + o_{\tau \to 0}(\tau),$$

lorsque $\vec{x} + \vec{v}_j \in \mathbb{N}^m$. Les fonctions $(h_j)_{1 \leq j \leq m}$ sont appelées *propensions* des réactions $(R_j)_{1 \leq j \leq m}$

La définition indique donc que les probabilités conditionnelles de déclenchement des réactions sur un intervalle de temps infinitésimal $[t, t + dt]$ sont proportionnelles à dt et de constantes de proportionnalité $(h_j(\vec{x}))_{1 \leq j \leq m}$. Une propriété importante des chaînes de Markov à temps continu est qu'on peut déterminer les probabilités de transitions d'un état à l'autre, qui correspondent aux probabilités de transitions de la chaîne de Markov incluse (*embedded Markov chain*) [KT75]. Cela se traduit dans notre cas par la possibilité de calculer la loi de probabilité de la prochaine réaction.

Proposition 2.1. *Sachant que $\vec{x}(t) = \vec{x}$ la probabilité que la prochaine réaction soit R_j est $p_j(\vec{x}) = \frac{h_j(\vec{x})}{h_0(\vec{x})}$ où $h_0(\vec{x}) = \sum_{j=1}^{m} h_j(\vec{x})$. On appellera $\vec{p}(\vec{x}) = (p_j(\vec{x}))$ les* probabilités de réactions *et $h_0(\vec{x})$* l'activité *du système dans l'état $\vec{x}$.*

Cette proposition-définition sera utile tout au long de cette thèse où les probabilités de réactions ont une place centrale.

Lois cinétiques En comparaison avec la sémantique déterministe, les fonctions de propension h_j correspondent à la loi cinétique de la réaction correspondante. Afin d'obtenir la correspondance avec la sémantique stochastique classique telle que définie par Gillespie [Gil76], on doit choisir des propensions de type *loi d'action de masse* :

$$h_j(\vec{x}) = c_j \prod_{i=1}^{n} \binom{x_i}{\alpha_{i,j}}, \tag{2.3}$$

où c_j est la *constante cinétique stochastique* (mentionnée en introduction de cette section) de la réaction et $\binom{n}{k}$ les coefficients binomiaux de Newton. Les coefficients binomiaux servent à dénombrer le nombre total de *combinaisons d'instances d'espèces* pouvant réagir selon la réaction considérée. Ce nombre augmente donc avec les quantités de matières et lorsque ces dernières sont importantes, peut être approximé (à une constante multiplicative près) au produit des quantités de matières, d'où l'analogie avec la loi d'action de masse. Une

remarque importante est que cette formule fait intervenir les consommations des réactions $(\alpha_{i,j})$ montrant que la dynamique stochastique tout comme la dynamique différentielle, ne peut être définie uniquement par la stœchiométrie du réseau. Par exemple, $2X \to X$ et $X \to \varnothing$ ont même stœchiométrie mais des fonctions de propension différentes.

2.2 De l'équation maîtresse aux moments

Dans la section précédente nous avons introduit la sémantique stochastique dans laquelle la dynamique des réseaux de réactions est une chaîne de Markov à temps continu. Le but de cette démarche est d'obtenir un modèle dynamique qui puisse rendre compte de la variabilité des données et capable de décrire l'évolution des *moments de la trajectoire*, c'est-à-dire les espérances, les variances, co-variances, *etc.* En effet, on souhaite pouvoir confronter des mesures expérimentales de moments, obtenues à partir d'un ensemble de trajectoires expérimentales, à un réseau de réactions, pour parvenir éventuellement à le réfuter.

Commençons par donner une définition formelle des moments de la trajectoire.

Définition 2.2 (Moments)**.** Soit $(\vec{x}(t)) \in \mathbb{N}^n$ $(t \in \mathbb{R}^+)$ la dynamique stochastique d'un réseau de réactions, soit $\bar{m} = (m_1, \ldots, m_n)$ un multi-indice, le moment $\mu_{\bar{m}}$ d'indice $\bar{m}$ de la trajectoire $\vec{x}(t)$ est défini par

$$\mu_{\bar{m}} = \mathbb{E} \prod_{i=1}^{n} x_i^{m_i}. \tag{2.4}$$

L'*ordre* d'un moment d'indice $\bar{m}$ est $\sum_{i=1}^{n} m_i$.

Ainsi la trajectoire possède n moments d'ordre 1 qui sont les espérances $\mathbb{E}\, x_i$. Les moments d'ordre 2 correspondent aux valeurs $\mathbb{E}\, x_i x_j$ dont on peut en distinguer deux types :

— les *variances* $\mathbb{E}\, x_i^2$ qui décrivent la dispersion des trajectoires autour de leur trajectoire moyenne et

— les *co-variances* $\mathbb{E}\, x_i x_j$ $(i \neq j)$ qui décrivent les corrélations inter-espèces.

Si on prend compte de la symétrie $\mathbb{E}\, x_i x_j = \mathbb{E}\, x_j x_i$ il y a donc $n(n+1)/2$ moments d'ordre 2 qu'on représente souvent sous forme d'une matrice $(\mathbb{E}\, x_i x_j)_{1 \leq i,j,n}$ appelée *matrice de variance-covariance* ou plus simplement matrice de covariance. Il est important de bien remarquer que les moments ne sont pas des valeurs aléatoires.

À partir de l'hypothèse d'une dynamique markovienne, on souhaite pouvoir déterminer la valeur de ces moments. Cela est possible si on connait *la loi des trajectoires* c'est-à-dire

$$P(\vec{x}, t) = \mathbb{P}\left[\vec{x}(t) = \vec{x}\right], \tag{2.5}$$

pour instant t et état $\vec{x}$. Cette loi contient en effet toute l'information probabiliste sur la trajectoire. En particulier, par définition de l'espérance [KT75] on a la proposition suivante.

Proposition 2.2. *Le moment d'indice $\bar{m}$ est déduit de la loi de $\vec{x}(t)$ par la relation*

$$\mu_{\bar{m}}(t) = \sum_{\vec{x}\in\mathbb{N}^n} P(\vec{x},t) \prod_{i=1}^{n} x_i^{m_i}. \tag{2.6}$$

Comme cas particuliers, on obtient les formules suivantes pour les moments d'ordre 1 et 2 :

$$\mathbb{E}\,\vec{x}(t) = \sum_{\vec{x}\in\mathbb{N}^m} \vec{x}P(\vec{x},t), \tag{2.7}$$

$$\mathbb{E}\,x_i(t)^2 = \sum_{\vec{x}\in\mathbb{N}^m} x_i^2 P(\vec{x},t), \tag{2.8}$$

$$\mathbb{E}\,x_i(t)x_j(t) = \sum_{\vec{x}\in\mathbb{N}^m} x_i x_j P(\vec{x},t). \tag{2.9}$$

La théorie des chaînes de Markov à temps continu montre que cette loi est entièrement déterminée par l'*équation maîtresse* [KT75, Wil06] qu'on nomme plus spécifiquement *équation maîtresse chimique* (CME) dans le domaine de la modélisation stochastique des réseaux de réactions. L'équation maîtresse fournit un système d'équations différentielles ordinaires couplées mais contrairement aux méthodes différentielles, ce système régit l'évolution de la loi de probabilité de $\vec{x}(t)$ (qui est désormais un processus stochastique) et non $\vec{x}$ lui-même.

Proposition 2.3 (Équation maîtresse)**.** *La loi de probabilité $P(\vec{x},t)$ de la chaîne de Markov $(\vec{x}(t))$ $(t\in\mathbb{R})$ vérifie l'équation*

$$\frac{\partial P(\vec{x},t)}{\partial t} = \sum_{j=1}^{m} h_j(\vec{x}-\vec{v}_j)P(\vec{x}-\vec{v}_j,t) - h_j(\vec{x})P(\vec{x},t), \tag{2.10}$$

où $S = (\vec{v}_j)_{1\le j\le n}$ est la matrice de stœchiométrie du réseau étudié.

Dans le cas où l'ensemble des états accessibles est fini, on peut remarquer que l'équation maîtresse constitue un système d'équations différentielles ordinaires couplées où il y a une fonction inconnue $P(\vec{x},\bullet)$ pour chaque valeur possible de $\vec{x}$. On trouvera plus loin dans ce chapitre l'exemple 2.2 où l'on résout l'équation maîtresse et où l'on en déduit une expression des moments de la trajectoire.

2.3 Comparaison avec la dynamique différentielle

Lorsque l'on a introduit la sémantique différentielle, on l'a présentée comme une modélisation du comportement moyen d'une population d'individus régie par un même réseau de réactions. Puisque nous venons de démontrer que la sémantique stochastique permet de déterminer l'espérance de la trajectoire $\mathbb{E}\,\vec{x}(t)$ il est naturel de comparer cette espérance avec la solution des équations de lois d'action de masse. Nous verrons dans cette section

qu'**en général il n'y a pas égalité**. Cette différence, qui a déjà été étudiée dans la littérature [Kur72, Wil06] souligne certaines inexactitudes de la sémantique différentielle par lois d'action de masse.

Afin de comparer les deux sémantiques, on se propose donc d'étudier la valeur de $d\mathbb{E}\vec{x}/dt$ dans le cadre de la sémantique stochastique. Ce calcul peut se trouver dans [Wil06].

Proposition 2.4. *L'espérance du processus stochastique décrit par l'équation maîtresse vérifie la relation suivante*

$$\frac{d\mathbb{E}\vec{x}(t)}{dt} = \sum_{j=1}^{m} \mathbb{E}\, h_j(\vec{x}(t))\vec{v}_j. \tag{2.11}$$

Démonstration. En multipliant l'équation maîtresse pour un état fixé $\vec{u}$ par l'état lui-même $\vec{u}$, on obtient

$$\frac{\partial P(\vec{u},t)\vec{u}}{\partial t} = \sum_{j=1}^{m} h_j(\vec{u}-\vec{v}_j)P(\vec{u}-\vec{v}_j,t)\vec{u} - h_j(\vec{u})P(\vec{u},t)\vec{u}. \tag{2.12}$$

On somme toutes ces équations pour tous les états $\vec{u}$ atteignables et on obtient

$$\begin{aligned}
\frac{\partial \sum_{\vec{u}} P(\vec{u},t)\vec{u}}{\partial t} &= \sum_{\vec{u}}\sum_{j=1}^{m} h_j(\vec{u}-\vec{v}_j)P(\vec{u}-\vec{v}_j,t)\vec{u} - h_j(\vec{u})P(\vec{u},t)\vec{u} && (2.13)\\
&= \sum_{\vec{u}}\sum_{j=1}^{m} h_j(\vec{u}-\vec{v}_j)P(\vec{u}-\vec{v}_j,t)(\vec{u}-\vec{v}_j) + h_j(\vec{u}-\vec{v}_j)P(\vec{u}-\vec{v}_j,t)\vec{v}_j && (2.14)\\
&- h_j(\vec{u})P(\vec{u},t)\vec{u} && (2.15)\\
&= \sum_{j=1}^{m} \mathbb{E}\, h_j(\vec{x}(t))\vec{x}(t) + \mathbb{E}\, h_j(\vec{x}(t))\vec{v}_j - \mathbb{E}\, h_j(\vec{x}(t))\vec{x}(t), && (2.16)
\end{aligned}$$

c'est-à-dire

$$\frac{d\mathbb{E}\vec{x}(t)}{dt} = \sum_{j=1}^{m} \mathbb{E}\, h_j(\vec{x}(t))\vec{v}_j. \tag{2.17}$$

□

Cette proposition amène à deux remarques. Tout d'abord il est important de préciser que cette relation n'est pas (encore) une équation différentielle pour $\mathbb{E}\vec{x}$ c'est-à-dire qu'elle n'est pas de la forme $d\mathbb{E}\vec{x}/dt = F(\mathbb{E}\vec{x})$. Cependant, la relation a une forme proche de l'évolution de la concentration $\vec{\chi} = \vec{x}/\Omega$ (où Ω est le volume) par la loi d'action de masse

$$\frac{d\vec{\chi}}{dt} = \sum_{j=1}^{m} f_j(\vec{\chi})\vec{v}_j. \tag{2.18}$$

On est donc naturellement amené à comparer les vitesses de $\mathbb{E}\vec{x}(t)$ et $\Omega\vec{\chi}(t)$, c'est-à-dire les

quantités $h_j(\vec{x})$ et $\Omega f_j(\vec{\chi})$ qui peuvent être vues comme des polynômes en les variables x_i

$$\Omega f_j(\vec{\chi}) = \Omega k_j \prod_{i=1}^{n} \chi_i^{\alpha_{i,j}} = \frac{k_j}{\Omega^{\sum_{i=1}^{n} \alpha_{i,j}-1}} \cdot \prod_{i=1}^{n} x_i^{\alpha_{i,j}}$$
$$h_j(\vec{x}) = c_j \prod_{i=1}^{n} \binom{x_i}{\alpha_{i,j}}.$$

Lorsque les quantités de matières sont grandes *i.e.* $x_i \to \infty$, les coefficients du monôme dominant $\prod_{i=1}^{n} x_i^{\alpha_{i,j}}$ dans chacun de ces polynômes sont respectivement $\frac{k_j}{\Omega^{\sum_{i=1}^{n} \alpha_{i,j}-1}}$ et $\frac{c_j}{\prod_{i=1}^{n} \alpha_{i,j}!}$. Ainsi, lorsque les quantités de matière sont grandes ($x_i \to +\infty$) les deux polynômes sont asymptotiquement équivalents à condition de poser la règle de conversion

$$\frac{k_j}{\Omega^{\sum_{i=1}^{n} \alpha_{i,j}-1}} = \frac{c_j}{\prod_{i=1}^{n} \alpha_{i,j}!} \tag{2.19}$$

entre la constante cinétique k_j de la réaction R_j et sa constante stochastique de réaction c_j. Dans la suite, lorsqu'on souhaitera comparer les dynamiques stochastiques et la loi d'action de masse on adotera cette **règle conventionnelle de conversion** [Wil06] entre les paramètres cinétiques. On précisera par la suite dans quels cas les deux sémantiques sont effectivement comparables en termes de moyennes.

Réaction	Conversion
$\alpha_1 X_1 + \cdots + \alpha_n X_n \to \cdots$	$k/\Omega^{\sum_{i=1}^{n} \alpha_i - 1} = c/\prod_{i=1}^{n} \alpha_i!$
$\varnothing \to \cdots$	$k\Omega = c$
$X \to \cdots$	$k = c$
$X + Y \to \cdots$	$k/\Omega = c$
$2X \to \cdots$	$k/\Omega = c/2$

TABLE 2.1 – Règle conventionnelle de conversion entre constante cinétique k de la loi d'action de masse et la constante stochastique c de réaction

Limite thermodynamique et cas des systèmes de réactions linéaires On a utilisé dans la règle de conversion, le fait que les quantités de matières sont grandes $x_i \to +\infty$ pour ne conserver que la partie non négligeable dans l'expression des propensions. Cela est lié au premier cadre de validité de la loi d'action de masse vis-à-vis de la sémantique stochastique, c'est-à-dire $\vec{\mathbb{E}}x(t) = \vec{\chi}(t)$, qui est celui de la *limite thermodynamique*. Le passage à la limite thermodynamique consiste à augmenter à l'infini le volume du système tout en maintenant les concentrations constantes. Dans ce cadre, on a montré [Kur72] qu'il y avait bien égalité $\vec{\mathbb{E}}x(t) = \vec{\chi}(t)$. L'autre cas de validité détaillé dans [Wil06] est celui des réseaux de réactions linéaires. On dit qu'une réaction est *linéaire* si elle met en jeu au plus un réactif, c'est-à-dire qu'elle est d'ordre 0 ou 1. Les réseaux de réactions linéaires conduisent à des systèmes dynamiques plus simples. Par exemple, la loi d'action de masse pour les réseaux linéaires donne un système d'équations différentielles linéaires. Dans le cas linéaire, on a de plus

un lien exact entre espérance de $\vec{x}(t)$ décrit par l'équation maîtresse et les concentrations correspondantes $\vec{\chi}(t)$ définies par la loi d'action de masse.

Proposition 2.5. *Soit un système de réactions linéaires. Soit $\vec{x}(t)$ le processus stochastique décrit par l'équation maîtresse pour des quantités de molécules initiales $\vec{x}_0$. Soit $\vec{\chi}(t)$ l'évolution des concentrations décrite par la loi d'action de masse pour des concentrations initiales $\vec{\chi}_0 = \vec{x}_0/\Omega$, alors à tout instant $\mathbb{E}\,\vec{x}(t) = \Omega\vec{\chi}(t)$.*

Démonstration. La propension d'une réaction d'ordre 0 est une constante tandis que celle d'une réaction d'ordre 1 est linéaire, on peut donc toujours écrire $\mathbb{E}\,h_j(\vec{x}(t)) = h_j(\mathbb{E}\,\vec{x}(t))$. De plus pour les réactions d'ordre 1 on a $h_j = f_j$ en utilisant la règle de conversion des constantes cinétiques. Ainsi, d'après la proposition 2.4 l'espérance de $\vec{x}/\Omega$ vérifie l'équation différentielle de la loi d'action de masse

$$\frac{d\,\mathbb{E}\,\vec{x}(t)/\Omega}{dt} = \sum_{j=1}^{m} f_j(\mathbb{E}\,\vec{x}(t)/\Omega)\vec{v}_j. \tag{2.20}$$

□

Cette démonstration nous apprend aussi que la loi d'action de masse ne produit pas rigoureusement la trajectoire moyenne d'un système de réaction dès le moment où l'on a des réactions d'ordre 2 ou plus. En effet en général $\sum_{j=1}^{m} \mathbb{E}\,h_j(\vec{x}(t))\vec{v}_j \neq \sum_{j=1}^{m} h_j(\mathbb{E}\,\vec{x}(t))$ lorsque les propensions ne sont pas linéaires. Ce constat est intéressant lorsqu'on pense que pour beaucoup de biologistes et de modélisateurs, les méthodes différentielles et en particulier la loi d'action de masse correspondent au $vrai$ comportement du système. En réalité, ce n'est qu'un modèle qui a son domaine de validité.

Exemple 2.1. Pour illustrer cette différence, considérons la résolution stochastique de l'exemple 1.5 d'un réseau **non linéaire** déjà étudié dans le cadre différentiel.

$$2X \to \varnothing \quad (c),$$

avec pour conditions initiales $x(0) = 4$ molécules de X. Le système ne peut être que dans 3 états : $x = 4$, $x = 2$ ou $x = 0$. On notera pour $i \in \{1,2,3\}$, $p_i(t) = P(2i,t)$ la probabilité d'être dans l'état $2i$ à l'instant t. L'équation maîtresse nous donne le système différentiel

$$p_2' = -6cp_2 \tag{2.21}$$

$$p_1' = 6cp_2 - cp_1 \tag{2.22}$$

$$p_0' = cp_1, \tag{2.23}$$

avec comme conditions initiales $(p_2, p_1, p_0) = (1,0,0)$. Ce système triangulaire peut être résolu en traitant chaque ligne successivement et en utilisant la méthode de la variation de la constante pour traiter les seconds membres. On obtient alors l'unique solution

$$p_2(t) = e^{-6ct} \tag{2.24}$$

$$p_1(t) = \frac{6}{5}(1 - e^{-5ct})e^{-ct} \tag{2.25}$$

$$p_0(t) = \frac{1}{5}e^{-6ct} - \frac{6}{5}e^{-ct} + 1. \tag{2.26}$$

L'espérance se calcule alors facilement

$$\mathbb{E}\,x(t) = 2p_1(t) + 4p_2(t) = \frac{12}{5}e^{-ct} + \frac{8}{5}e^{-6ct}. \tag{2.27}$$

On constate ainsi une différence qualitative entre l'équation maîtresse qui prédit une décroissance exponentielle et la loi d'action de masse qui prédit une décroissance inverse.

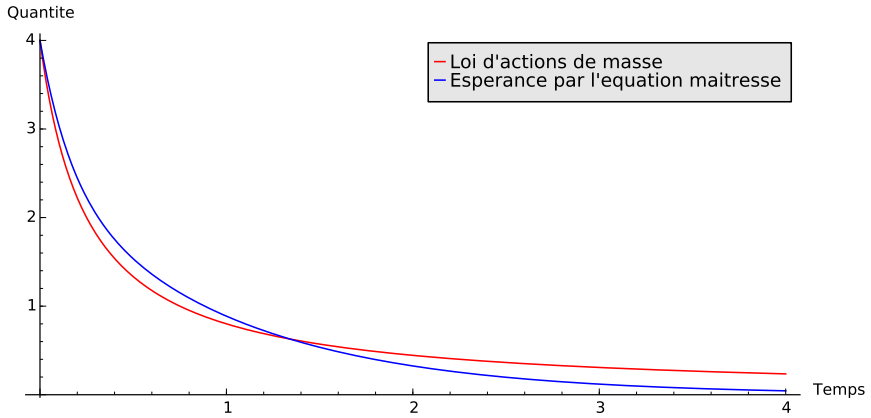


FIGURE 2.1 – **Comparaison entre loi d'action de masse et équation maîtresse pour l'exemple des désintégrations par paires.** En rouge, la loi d'action de masse prédit une décroissance inverse. En bleu, l'espérance issue de l'équation maîtresse prédit une décroissance exponentielle. Il est donc faux en général d'énoncer que la loi d'action de masse décrit les valeurs moyennes des quantités de matière.

Cet exemple est instructif car il permet de constater à quel point la dynamique stochastique est plus riche que la loi d'action de masse. En effet, considérons la variable aléatoire $M = \inf\{\tau / x(\tau) = 0\}$ c'est-à-dire le temps d'extinction de ce système (désintégration de toutes les molécules). Comme on ne peut quitter l'état $x = 0$, il est facile de constater par double inclusion que les événements $M \leq t$ et $x(t) = 0$ sont égaux, on peut donc obtenir la fonction de répartition de M

$$F_M(t) = p_0(t) = \frac{1}{5}e^{-6ct} - \frac{6}{5}e^{-ct} + 1, \tag{2.28}$$

ainsi que sa densité

$$f_M(t) = F'_M(t) = \frac{6c}{5}(e^{-ct} - e^{-6ct}). \tag{2.29}$$

On peut alors en déduire tous les moments de M et en particulier l'espérance de vie du système

$$\mathbb{E}\,M = \int_0^{+\infty} t f_M(t)\,dt = \frac{7}{6c}. \tag{2.30}$$

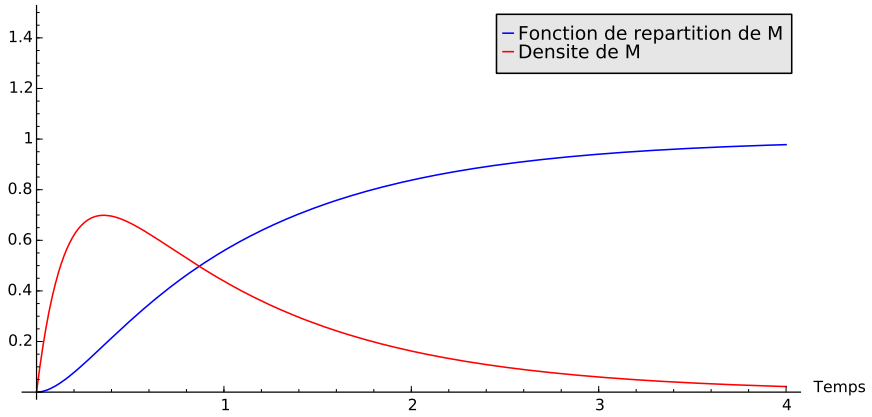


FIGURE 2.2 – **Loi de probabilité du temps de vie d'un système de molécules se désintégrant par paires.** Cette information ne peut-être obtenue par la lois d'action de masse.

Toutes ces informations, pourtant très pertinentes lorsqu'on s'intéresse à un tel système, **ne sont pas accessibles par la loi d'action de masse**. En effet, la trajectoire obtenue n'atteint jamais 0 et on ne peut décider que de façon arbitraire une valeur seuil signifiant que le système est mort. Il est ainsi impossible de parler de l'espérance de vie moyenne du système en sémantique déterministe.

En conclusion de cette comparaison, nous avons vu qu'en général il n'y a pas égalité entre la solution des équations d'action de masse et la courbe d'espérance prévue par l'équation maîtresse. Toutefois, cette comparaison est valable dans deux cas : d'une part le cas de la limite thermodynamique c'est-à-dire quand les quantités d'espèces sont importantes et d'autre part le cas des réseaux de réactions linéaires. Par ailleurs, a montré jusqu'à présent que la sémantique stochastique, reposant sur une modélisation plus fidèle à la physique, donnait des résultats qualitativement plus riches que la sémantique différentielle. Toutefois, nous allons maintenant voir que les complications surviennent lorsqu'on tente de résoudre l'équation maîtresse.

2.4 Résolution de l'équation maîtresse

Puisque résoudre l'équation maîtresse permet d'obtenir la loi de probabilité des trajectoires et donc leurs moments, il est naturel dans cette thèse de traiter la résolution de cette équation. Malheureusement, nous montrons dans cette section que sa résolution est impossible en pratique. Nous présentons étudions en premier lieu la résolution exacte de l'équation puis on étudiera les moyens de contourner cette résolution.

2.4.1 Résolution exacte

On s'intéresse à la résolution de l'équation maîtresse dans le cas simple où le nombre d'états est fini. Dans ce cas, l'équation maîtresse correspond à un système d'équations différentielles ordinaires, une pour chaque état $\vec{x}$, à coefficients constants. En théorie, on sait facilement résoudre les systèmes différentiels linéaires à coefficients constants. Concrètement, l'équation peut se réécrire sous la forme

$$\frac{d\vec{f}}{dt} = A\vec{f} \tag{2.31}$$

où $\vec{f}(t) = (P(\vec{x}_i, t))_i$ est le vecteur des fonctions inconnues et A est une matrice carrée constante dont la taille est le nombre d'états atteignables du système. La résolution de l'équation maîtresse se ramène alors au calcul de l'exponentielle de matrice $\vec{f}(t) = \exp(tA)\vec{f}(0)$. Toutefois, en pratique ce système est difficile à résoudre puisqu'il admet autant d'inconnues que d'états atteignables, ce qui doit être mis en regard des approches différentielles où le nombre d'inconnues est le nombre d'espèces chimiques. Ainsi, la matrice A introduite est en pratique de taille trop grande pour qu'on puisse facilement en calculer l'exponentielle. Pour cette raison, on admet que la résolution exacte de l'équation maîtresse n'est pas possible en pratique [Wil06] et on ne peut envisager résoudre l'équation maîtresse que dans des cas simples décrits dans [McQ67].

À titre d'exemple de résolution dans un cas simple on considère de nouveau l'exemple 1.4 dans le cadre stochastique.

Exemple 2.2. Soit l'exemple extrême d'un système à deux réactions

$$\begin{aligned} R_1 &: X \to Y \quad (c_1) \\ R_2 &: Y \to X \quad (c_2) \end{aligned}$$

avec comme quantités initiales une seule molécule de X. Le système n'a donc que deux états possibles, soit $(x, y) = (1, 0)$ qui est l'état initial, soit $(x, y) = (0, 1)$. Notons $f_X(t) = P((1, 0), t)$ la probabilité au temps t d'être dans l'état $(1, 0)$ et $f_Y(t) = P((0, 1), t)$ celle d'être dans l'état $(0, 1)$. L'équation maîtresse donne

$$\begin{aligned} \frac{df_X}{dt} &= c_2 f_Y - c_1 f_X \\ \frac{df_Y}{dt} &= c_1 f_X - c_2 f_Y \end{aligned}$$

qui est un système d'équations différentielles ordinaires couplées linéaire qui est dans le cas présent le même système que celui de la loi d'action de masse. Rappelons que ce n'est en général pas le cas puisque ce système possède autant d'équations que d'états atteignables et non d'espèces chimiques. En appliquant la même méthode de résolution à l'aide de l'exponentielle de matrice on obtient donc

$$\begin{aligned} f_X(t) &= \frac{c_2}{c_1 + c_2} + \frac{c_1}{c_1 + c_2} \exp(-t(c_1 + c_2)) \\ f_Y(t) &= \frac{c_1}{c_1 + c_2} - \frac{c_1}{c_1 + c_2} \exp(-t(c_1 + c_2)). \end{aligned}$$

Nous venons de résoudre l'équation maîtresse pour ce système, l'exemple étant suffisamment simple pour qu'on puisse y parvenir, c'est-à-dire qu'on obtient une expression analytique de la loi de probabilité $P(\vec{x}, t)$ du système à chaque instant t. À partir de cette information complète, on peut obtenir tous les moments du système et en particulier les valeurs moyennes. Dans ce cas particulier, elles se calculent d'autant plus facilement que $x(t)$ et $y(t)$ suivent des lois de Bernoulli. On obtient alors, en utilisant les formules des espérances et variances des lois de Bernoulli, les valeurs moyennes

$$\begin{aligned}\mathbb{E}\, x(t) &= \frac{c_2}{c_1+c_2} + \frac{c_1}{c_1+c_2}\exp(-t(c_1+c_2))\\ \mathbb{E}\, y(t) &= \frac{c_1}{c_1+c_2} - \frac{c_1}{c_1+c_2}\exp(-t(c_1+c_2)),\end{aligned}$$

mais aussi les variances

$$\begin{aligned}\operatorname{Var} x(t) &= \frac{c_1c_2}{(c_1+c_2)^2} + \frac{c_1^2-c_1c_2}{(c_1+c_2)^2}\exp(-t(c_1+c_2)) - \frac{c_1^2}{(c_1+c_2)^2}\exp(-2t(c_1+c_2))\\ \operatorname{Var} y(t) &= \frac{c_1c_2}{(c_1+c_2)^2} + \frac{c_1^2-c_1c_2}{(c_1+c_2)^2}\exp(-t(c_1+c_2)) - \frac{c_1^2}{(c_1+c_2)^2}\exp(-2t(c_1+c_2)).\end{aligned}$$

En conclusion cet exemple montre que dans certains cas de réseaux simples, on peut en résolvant l'équation maîtresse obtenir des formules analytiques pour décrire les moments d'ordre 1 et 2 (et plus) en fonction des paramètres cinétiques et de l'état initial. Si on est capable de mesurer expérimentalement ces moments alors on obtient des informations sur les paramètres cinétiques. Par exemple si on mesure asymptotiquement des bornes pour la variance de $\vec{y}(t)$: $A \leq \operatorname{Var}(+\infty) \leq B$ alors on est capable de savoir que $\frac{c_1c_2}{(c_1+c_2)^2} \in [A, B]$. Une telle résolution permet donc éventuellement d'aboutir à des réfutations en fonction des moments d'ordre 1 et 2 par une approche de contrainte des paramètres cinétiques. Malheureusement, il faut que la résolution au départ soit possible en premier lieu ce qui n'est envisageable que dans un nombre limité de cas [McQ67]. De plus, l'expression des moments obtenus sera dépendante des lois cinétiques supposées et de la topologie du réseau ce qui n'est pas le cas de l'approche par contraintes des flux en différentiel dans lequel l'expression des pentes moyennes était toujours une fonction linéaire du flux à l'équilibre, entièrement déterminée par la stœchiométrie du réseau. En conséquence, la classe algébrique des contraintes obtenues sur les paramètres cinétiques ne peut être facilement identifiée.

2.4.2 Approximations des moments

À terme on souhaite exploiter des informations sur les moments d'ordre 1 et 2. Nous avons vu qu'il est théoriquement possible de déterminer ces moments dans la dynamique stochastique en résolvant l'équation maîtresse. Cependant cette résolution est impossible en pratique sauf pour des exemples très simples. Il est donc naturel de s'intéresser aux techniques de résolutions approximatives. Une possibilité est de tenter d'approximer directement les moments, on présentera deux méthodes classiques de la littérature : la méthode des moments clos (CMM) [Hes08] et l'approximation de bruit linéaire (LNA) [VK92].

2.4.2.1 Méthodes des moments clos

La première méthode des moments clos a été introduite en 1957 par Whittle [Whi57] dans le cadre général de l'étude des moments d'une chaîne de Markov. Nous avons déjà défini (Définition 2.2) le moment d'indice $\bar{m}$ ainsi que son ordre. On note $\vec{\mu}_k$ le vecteur de tous les moments (dans un ordre arbitrairement fixé) d'ordre au plus k. Il est possible de montrer [SH06] à partir de l'équation maîtresse, qu'on peut obtenir une équation différentielle de la forme

$$\frac{d\vec{\mu}_k}{dt} = A\vec{\mu}_k + B\vec{\gamma} \tag{2.32}$$

où A et B sont des matrices et $\vec{\gamma}$ est un vecteur ne contenant que des moments d'ordre strictement supérieur à k. On dit alors que ce système d'équations différentielles est $ouvert$ et il ne peut être résolu car il contient plus d'inconnues que d'équations. Une méthode des moments clos consiste à résoudre le système

$$\frac{d\vec{\mu}_k}{dt} = A\vec{\mu}_k + B\vec{\varphi}(\vec{\mu}_k), \tag{2.33}$$

où $\vec{\varphi}$ est une approximation des moments d'ordre supérieurs strictement à k décrits précédemment en fonction des moments d'ordre k. On a donc $clos$ l'équation (2.32) et la résolution de ce nouveau système nous fourni une approximation des moments d'ordre k. Il y a donc autant de méthodes des moments clos que de manière de définir l'approximation $\vec{\varphi}$. On pourra se référer à [Hes08] pour une comparaison des différentes méthodes d'approximation existantes (*derivative matching*, troncature des cumulants, quasi-déterminisme, *etc*). Dans la pratique, les méthodes des moments clos ne sont donc pas générales et difficiles à appliquer car il faut produire des développements analytiques compliqués (et parfois même des approximations) pour obtenir l'équation ouverte puis utiliser une bonne approximation des moments supérieurs de cette équation ouverte. Dans la littérature, on trouve surtout le traitement assez récent du modèle logistique de croissance de population [Nås03b, Nås03a] qui malgré sa simplicité entraîne déjà de nombreuses complications ou encore la résolution du cas des systèmes linéaires [GLO05] (qui sont directement sous forme close).

Dans le cadre de la question qui nous intéresse, la méthode des moments clos n'est donc pas assez générale, puisque chaque système doit bénéficier d'un traitement analytique particulier. Il ne règle pas le problème de la nécessaire connaissance des lois et paramètres cinétiques. Enfin il ne fait pas clairement apparaître un vecteur analogue aux flux stationnaires $\vec{f}$ qui pourraient être contraints à partir de données sur les moments. L'approximation de bruit linéaire n'apporte pas de progrès dans ce domaine car elle peut être vue comme un cas particulier de la méthode des moments clos à l'ordre 2.

2.4.2.2 L'approximation de bruit linéaire

L'approximation de bruit linéaire, également appelée Ω-expansion, peut être considérée comme une forme de méthode des moments clos à l'ordre 2 puisqu'elle consiste aussi à

clore l'équation des moments

$$\frac{d\vec{\mu}_2}{dt} = A\vec{\mu}_2 + B\vec{\gamma} \tag{2.34}$$

en utilisant cette fois une approximation des moments supérieurs qui dépend d'une autre fonction $\vec{\phi}$, elle même déterminée par une équation différentielle propre :

$$\frac{d\vec{\phi}}{dt} = \psi(\vec{\phi}), \tag{2.35}$$

$$\frac{d\vec{\mu}_2}{dt} = A\vec{\mu}_2 + B\vec{\varphi}(\vec{\phi}, \vec{\mu}_2). \tag{2.36}$$

L'approximation de bruit linéaire utilise pour $\vec{\phi}$ la solution des équations de loi d'action de masse, c'est-à-dire qu'on suppose dans cette approximation que l'influence des fluctuations sur l'espérance est négligeable (tout comme dans la limite thermodynamique). L'équation $\frac{d\vec{\phi}}{dt} = \psi(\vec{\phi})$ représente donc ici l'équation d'action de masse. L'équation close est ensuite calculée en estimant les fluctuations $\chi(t)$ autour de la solution $\vec{\phi}$ à la limite thermodynamique. On pose alors $\vec{x}(t) = \Omega\vec{\phi}(t) + \Omega^{1/2}\vec{\chi}(t)$ et on peut obtenir (après des approximations complexes) une expression close pour les moments des fluctuations $\vec{\chi}(t)$ à $o(\Omega^{-1})$ près lorsque $\Omega \to +\infty$. Cette équation close permet ensuite d'obtenir des équations closes pour les moments d'ordre 1 et 2. Bien que le LNA ait été décrit dans des cas particuliers [VK92] il n'existe pas dans la littérature selon [EE03] de présentation générale de la méthode au cas multivarié. Les détails techniques ont donc été décrits dans les suppléments de [EE03] et pourra trouver une présentation générale et une implémentation logicielle sous forme de boîte à outils `Matlab` dans [Hes08]. Ainsi, le LNA fournit une approximation gaussienne des trajectoires où l'espérance est la solution de la loi d'action de masse $\vec{\phi}$ et la matrice variance-covariance est déterminée par un système d'équations différentielles non homogène faisant intervenir $\vec{\phi}$. La qualité de l'approximation obtenue par le LNA fait l'objet de recherches récentes [WGSP12].

2.4.2.3 Conclusion

Les méthodes d'approximations des moments de la loi de probabilité des trajectoires en sémantique stochastique, c'est-à-dire d'estimation des moments de la loi solution de l'équation maîtresse, reposent sur l'obtention d'un système d'équations différentielles clos décrivant la dynamique des moments jusqu'à un certain ordre fixé. La résolution de ce système permet alors d'avoir une expression en fonction du temps de chacun de ces moments. L'avantage de ces méthodes est qu'elles évitent l'explosion du nombre d'équations (liée à l'explosion du nombre des états accessibles) en conduisant à des systèmes différentiels où le nombre d'équations est réduit aux nombres de moments auxquels on s'intéresse. Ces méthodes de moments clos sont donc actuellement utilisées lorsqu'on souhaite déterminer facilement les moments des trajectoires, par exemple dans le cadre de la vérification probabiliste de réseaux de réactions [CKL15].

Une difficulté importante est qu'il est difficile de choisir *a priori* quelle méthode d'approximation doit être utilisée pour obtenir les équations différentielles closes des moments.

Dans [Hes08] on constate que la qualité des approximations des moments obtenues diffèrent selon les cas. Les recherches récentes (par exemple [BHP+15]) s'intéressent au problème de la détermination de la meilleure approximation pour résoudre un problème donné.

Ces méthodes s'appliquent difficilement dans le cadre de cette thèse, qui est la confrontation d'un réseau avec des données expérimentales. En effet, elles nécessitent de connaître les valeurs des paramètres cinétiques afin d'obtenir des moments approximés qu'on puisse comparer aux estimations expérimentales des moments. Or l'inférence des paramètres est un problème difficile.

Une méthode par contraintes dans laquelle des mesures expérimentales sur les moments fournissent des contraintes sur les paramètres cinétiques est envisageable mais on se heurte alors aux mêmes difficultés que pour la résolution exacte de l'équation maîtresse. En effet, l'expression des moments obtenue n'est pas générale, elle dépend à la fois de la topologie du réseau mais aussi de la méthode d'approximation utilisée pour obtenir le système différentiel clos des moments. Il est donc difficile d'exhiber un vecteur analogue au vecteur de flux stationnaire qu'on pourrait relier analytiquement aux moments asymptotiques et ceci d'une manière générale et indépendante des lois et paramètres cinétiques.

2.4.3 Méthode de Monte-Carlo

La sémantique stochastique que nous avons introduite propose de modéliser la dynamique comme une chaîne de Markov $\vec{x}(t)$ dont les changements d'états interviennent à des instants $0 = t_0 < t_1 < \dots$ qui correspondent chacun au déclenchement d'une réaction. Une trajectoire de la chaîne de Markov est alors entièrement déterminée par

— les durées inter-réactions $\tau_k = t_{k+1} - t_k$ pour $k \in \mathbb{N}$ et

— les numéros $(\mu_k)_{k\in\mathbb{N}^*}$ des réactions déclenchées aux instants (t_k) correspondants.

Ici les (τ_k) et (μ_k) sont des variables aléatoires qui permettent de déterminer la trajectoire grâce à l'équation générale

$$\forall k \in \mathbb{N}, \quad \vec{x}(t_{k+1}) = \vec{x}(t_k) + S\vec{e}_{\mu_{k+1}}, \tag{2.37}$$

où les $\vec{e}_i = (0, \cdots, 0, 1^{(i)}, 0, \cdots, 0)^\top$ sont les vecteurs de la base canonique de $\mathbb{R}^n$. Nous avons vu qu'il est difficile de déterminer la loi des trajectoires de la chaîne de Markov. Toutefois il est possible de générer des trajectoires aléatoires car on connait les lois de τ_k et μ_{k+1} conditionnellement à l'état courant de la chaîne de Markov. Ainsi, l'algorithme de simulation stochastique (SSA) [Gil76, Gil77, Gil07] est une méthode de Monte-Carlo qui se propose de générer aléatoirement les couples (μ_{k+1}, τ_k) en se reposant sur les quantités d'espèces disponibles $\vec{x}_{t_k}$ à chaque instant. Il existe différentes variations équivalentes de l'algorithme de Gillespie et nous présentons ici la variante de la *méthode directe* car elle correspond à la simulation classique des chaînes de Markov à temps continu [Wil06]. C'est également la première proposition de l'algorithme SSA proposée dans l'article de Gillespie [Gil76]. Cette simulation est décrite par Gillespie à partir de *l'hypothèse fondamentale* de la dynamique stochastique décrite en début de chapitre. En effet, cette hypothèse montre que chaque

combinaison de molécules met un certain temps aléatoire pour réagir, appelé *temps de réaction*, qui est distribué selon la loi exponentielle de paramètre c_j, la constante stochastique associée à la réaction considérée. On exploite alors les bonnes propriétés de la loi exponentielle (proposition 2.6 ci-dessous) pour obtenir l'algorithme de simulation stochastique d'un réseau de réactions, version *méthode directe*.

Proposition 2.6. *Supposons que le* temps de réaction *de chaque* combinaison possible *de molécules (i.e. d'instances) de réactifs soit exponentiellement distribué. Notons alors c_j le paramètre de cette loi exponentielle lorsque la combinaison de réactifs correspond à la réaction R_j. De plus supposons qu'il y a indépendance des événements de réactions de ces combinaisons d'espèces, alors*

1. *τ_k est exponentiellement distribué de paramètre $h_0(\vec{x}(t_k)) = \sum_{j=1}^{m} h_j(\vec{x}(t_k))$ et*
2. *μ_{k+1} suit une loi de Bernoulli généralisée à n issues possibles de probabilités $p_j(\vec{x}(t_k)) = \frac{h_j(\vec{x}(t_k))}{h_0(\vec{x}(t_k))}$.*

Démonstration. La preuve consiste à déterminer la loi du temps inter-réactions et les probabilités de déclenchement de chaque réaction à partir d'une hypothèse de loi exponentielle pour les temps de réactions des combinaisons de molécules. En théorie des probabilités, il est connu que si $(X_k)_{1\le k\le l}$ est une famille finie de l variables aléatoires exponentielles de paramètres $(\lambda_k)_{1\le k\le l}$ respectivement, alors (i) $\min_{1\le k\le l} X_k$ est une variable aléatoire exponentielle de paramètre $\sum_{k=1}^{l} \lambda_k$ et (ii) $\forall i \in \{1,\dots,l\}$, l'événement $\bigcap_{k=1,\dots,l;k\neq i}\{X_i < X_k\}$ a pour probabilité $\lambda_i/(\sum_{k=1}^{l} \lambda_k)$. Ainsi en raison de (i) le temps de réaction pour n'importe quelle combinaison de molécules réactifs de R_j suit la loi exponentielle de paramètre

$$c_j \cdot \mathrm{Card}\{\text{combinaisons de réactifs de } R_j \text{ reactants dans l'état } \vec{x}(t_k)\} = h_j(\vec{x}(t_k)).$$

Ensuite, d'après (ii) la probabilité que la prochaine réaction soit R_j est bien $h_j(\vec{x}(t_k))/h_0(\vec{x}(t_k))$. De plus, le temps inter-réaction, c'est-à-dire le temps de déclenchement pour n'importe quelle combinaison de réactifs de n'importe qu'elle réaction est aussi donné par (i), il suit donc la loi exponentielle de paramètre $h_0(\vec{x}(t_k))$. □

Voici une implémentation de l'algorithme de Gillespie. La fonction *propensity* correspond au calcul de la propension qui dépend de la constante stochastique et de la partie gauche de la réaction concernée. La fonction *exp_random* génère une valeur suivant la loi exponentielle de paramètre donné et la fonction *random* tire aléatoirement un entier (index de réaction) d'après la loi fournie en paramètre sous forme d'un vecteur de probabilités.

Algorithm 1: Alogrithme de Gillespie (SSA), méthode directe

Data: (α_{ij}), (β_{ij}), $t_{\max}$, (c_i), $\vec{x}_0$
Result: (t_k), (μ_k), (τ_k) and $(\vec{x}_k)$
$\forall i, j:\ s_{i,j} \leftarrow \beta_{i,j} - \alpha_{i,j}$;
$k \leftarrow 0$;
$t_k \leftarrow 0$;
$\vec{x}_k \leftarrow \vec{x}_0$;
while $t_k < t_{max}$ **do**
  for $i = j$ *to* m **do**
    $h_j \leftarrow$ propensity$(\vec{x}_k, c_j, \alpha_{j,\bullet})$;
  end
  $h_0 \leftarrow \sum_{j=1}^{m} h_j$;
  $\tau_k \leftarrow$ exp_random(h_0);
  $\mu_{k+1} \leftarrow$ random$((h_j/h_0)_j)$;
  $\vec{x}_{k+1} \leftarrow \vec{x}_k + S^{(\mu_{k+1})}$ (column μ_{k+1});
  $t_{k+1} \leftarrow t_k + \tau_k$;
  $k \leftarrow k+1$;
end

Les deux remarques suivantes permettent de mieux comprendre cette implémentation.

— Premièrement, il est important de comprendre que la loi exponentielle de paramètre c_j correspond à la loi du temps d'attente de *chacune* des combinaisons de réactifs possibles pour la réaction R_j. Ainsi une augmentation des quantités de réactifs aura pour effet d'augmenter le nombre de combinaisons possibles rendant la sélection de la réaction R_j plus probable. Par exemple si on considère la réaction $R:\ X + Y \rightarrow Z$ de constante stochastique c et des quantités d'espèces réactifs n_X et n_Y, alors le nombre de combinaisons de réactifs est $n_X \cdot n_Y$. Ainsi, d'après la proposition 2.6 le temps d'attente pour la prochaine occurrence d'une réaction R suit une loi exponentielle de paramètre $c \cdot n_X \cdot n_Y$. Ainsi il s'agit bien de calculer la *propension* $h = c \cdot \mathrm{Card}\{\text{combinaisons de réactifs}\}$ de la réaction. On peut remarquer que certaines réactions chimiques telles que la dimérisation $2X \rightarrow X_2$ conduit à des fonctions de propensions plus compliquées qu'une loi d'action de masse, dans cet exemple : $c \cdot n_X \cdot (n_X - 1)/2$. En utilisant une nouvelle fois la première partie de la proposition 2.6 on obtient alors que le temps inter-réactions suit la loi exponentielle de paramètre la somme des propensions de chaque réaction. On a donc obtenu la loi de τ_k en fonction des paramètres et des quantités d'espèces courantes.

— Deuxièmement, la dernière partie de la proposition 2.6 permet de déterminer les probabilités de sélection des réactions : la probabilité p_j que R_j soit la prochaine réaction sélectionnée est $p_j = h_j/(\sum_{k=1}^{m} h_k)$ où (h_j) sont les propensions de chaque réaction.

Grâce à l'algorithme de Gillespie on est donc en mesure de simuler un ensemble de trajectoires d'un réseau de réactions et d'en déduire empiriquement les trajectoires des $moments$. Cet algorithme a permis d'étudier stochastiquement plusieurs systèmes biologiques réels [ARM98, MA97]. Dans le cadre de cette thèse, si on souhaite confronter des séries temporelles expérimentales avec un réseau de réactions on peut alors comparer les trajectoires des moments obtenues à l'aide des expériences avec celles obtenues par l'algorithme de Gillespie. Cependant, cela nécessite de connaître les valeurs des constantes de réactions (c_j), ce que l'on souhaite éviter. De plus, l'inconvénient majeur de l'algorithme de Gillespie est qu'il est très coûteux puisqu'il simule chaque occurrence de réactions une à une. Il n'est donc plus utilisable dès que le nombre de molécules est trop grand ou que les durées de simulations sont trop importantes. Les variations exactes de cet algorithme (méthode $next\ reaction$ [Gil76], Gibson-Bruck [GB00]) ne permettent pas de s'affranchir significativement de cette limite tandis que les méthodes approchées (approximation de Poisson, méthode τ-leap [RPCG03, GP03, Wil06]) sont par nature inexactes et peuvent conduire à des résultats aberrants (quantités de molécules négatives). Enfin, on peut noter que l'approche consistant à contraindre les paramètres (c_j) en fonction d'informations sur les moments ne peut être appliquée puisque par nature les simulations ne fournissent pas d'expressions analytiques pour les trajectoires des moments.

2.5 Conclusion

La littérature en modélisation stochastique de la dynamique des réseaux de réactions repose essentiellement sur une description Markovienne des trajectoires. Elle présente des similitudes avec la dynamique déterministe car elle repose sur des lois cinétiques proches de la loi d'action de masse. Nous avons en particulier montré que dans certaines conditions, le calcul de l'espérance de sémantique stochastique correspond à la dynamique déterministe. Cependant en général, la sémantique stochastique repose sur des fondements physiques plus solides et fournit en général une information qualitativement plus riche.

Toutefois cette dynamique pose aussi des difficultés supplémentaires : le nombre d'états de la chaîne de Markov sous-jacente est grand voire infini par rapport à la taille du réseau. Le calcul explicite de la distribution des trajectoires par la résolution de l'équation maîtresse est donc en pratique impossible. L'approche de Monte-Carlo, incarnée par l'algorithme de Gillespie, permet de générer des trajectoires correctes vis à vis de la sémantique stochastique et d'estimer, suite à un grand nombre de simulations, les moments. Toutefois, cela nécessite de connaître les valeurs des paramètres cinétiques et ne fournit pas d'expression analytique des moments. En particulier il est impossible d'obtenir des contraintes de paramètres à partir de l'exploitation de trajectoires expérimentales puisqu'on ne possède dans ce cas d'aucune relation analytique.

Nous avons également présenté les méthodes d'approximation classiques de clôture des moments (dont l'approximation de bruit linéaire) qui fournissent des systèmes différentiels clos qui régissent tous les moments d'ordre au plus k, où k est fixé arbitrairement ($k = 2$ dans

le cas du LNA). On peut cette fois-ci obtenir des solutions analytiques pour les moments. Cependant, encore une fois la connaissance des lois et paramètres cinétiques est nécessaire. Une approche par contraintes est *a priori* envisageable mais elle n'apparaît alors pas très générale puisque les solutions analytiques dépendent de la topologie du réseau et des approximations utilisées pour les moments d'ordre supérieur. De plus il n'est pas évident de connaître la qualité des approximations de moments obtenues.

Lorsqu'on se compare aux approches par contraintes de flux en sémantique différentielles, on s'aperçoit qu'un élément clef est que le vecteur de flux s'obtient à l'aide des solutions stationnaires. Toutefois, dans le cadre d'une sémantique stochastique une trajectoire ne peut être "strictement équilibrée" car en raison des fluctuations elle ne peut pas se maintenir à l'équilibre. Cependant on peut tout de même exploiter la stationnarité au sens stochastique du terme, c'est-à-dire la convergence asymptotique de la loi de probabilité des trajectoires. C'est cette approche qui est envisagée dans cette thèse pour déterminer le vecteur de probabilités stationnaires en sémantique stochastique (chapitre 3) et le contraindre par la suite (chapitre 4 et 5).

Deuxième partie

Approximation de Bernoulli du régime stationnaire en dynamique stochastique et applications

Chapitre 3

Approximation de Bernoulli du régime stationnaire

L'approximation de Bernoulli décrite dans ce chapitre ainsi que ses applications font l'objet d'un article de revue en cours de rédaction. Le théorème central limite présenté dans ce chapitre (ainsi que ses applications) a été publié dans les actes du workshop SASB (SAS2014) [PSB14], où il a été démontré dans le cadre de la chaîne de Markov incluse. Ce chapitre permet de comprendre que ce cas correspond à celui de la limite thermodynamique ($p_0 \to 0$).

L'objectif de ce chapitre est d'introduire les éléments permettant d'aboutir à des méthodes par contraintes similaires aux contraintes de flux du FBA mais dans un cadre probabiliste. Il s'agit de pouvoir décrire l'état stationnaire stochastique par un paramètre similaire au vecteur de flux et qui peut-être analytiquement lié aux moments de la distribution des trajectoires. Pour cela, nous nous proposons d'approximer le régime stationnaire de la dynamique stochastique par un processus de tirage aléatoire des réactions selon une *loi de Bernoulli*. Dans l'espace des phases, cela correspond à un processus de *marche aléatoire* plus facile à étudier que le régime stationnaire du processus Markovien associé à la dynamique stochastique décrite dans le chapitre précédent.

Les principaux résultats de ce chapitre sont représentés sur le diagramme de la figure 3.1. Dans un premier temps, nous proposons une discrétisation $\vec{y}(k)$ ($k \in \mathbb{N}$) de la dynamique stochastique $\vec{x}(t)$ ($t \in \mathbb{R}^+$) reposant sur l'utilisation d'un pas de temps fixe δt. Cela permet de se ramener à l'étude plus simple de chaînes de Markov à temps discret sans changer la nature des résultats puisque, en particulier, les lois de $\vec{x}(k \cdot \delta t)$ et $\vec{y}(k)$ coïncident lorsque δt tend vers 0. Dans un second temps, nous proposons l'approximation de Bernoulli $\vec{z}(k)$ ($k \in \mathbb{N}$) du régime stationnaire de $\vec{y}(k)$. Nous montrerons que des expressions analytiques pour les moments de $\vec{z}(k)$ peuvent être déterminées ainsi qu'un théorème central limite. Enfin, nous étudierons la qualité de l'approximation de Bernoulli en ce qui concerne les moments d'ordre 1 et 2. Nous verrons en effet que l'on peut comparer les moments

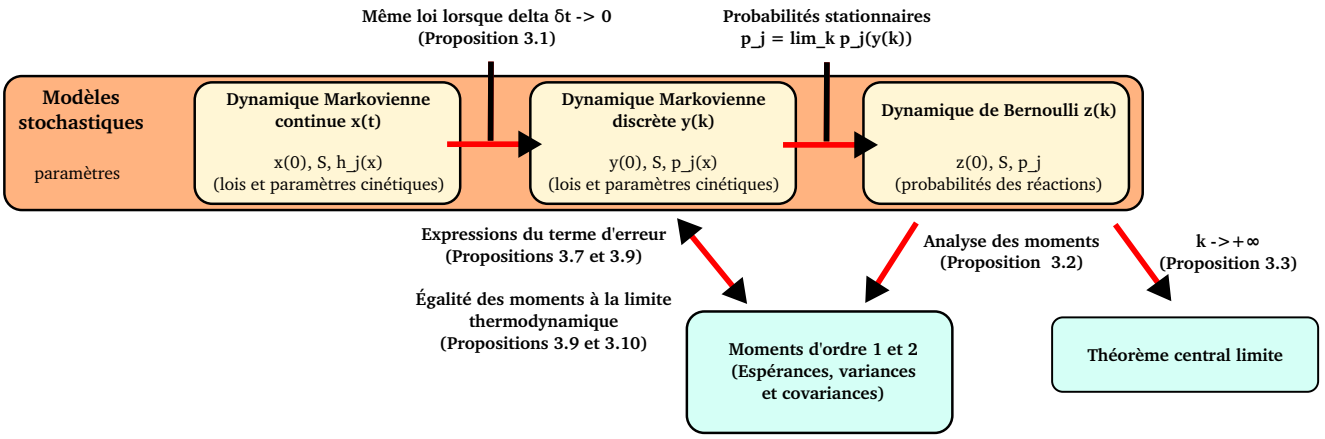


FIGURE 3.1 – **Diagramme représentant les résultats principaux du chapitre 3.**

d'ordre 1 et 2 de $\vec{y}(k)$ et $\vec{z}(k)$ par des égalités de type

$$\mathbb{E}\,\vec{y}(k) = \mathbb{E}\,\vec{z}(k) + \text{terme d'erreur} \tag{3.1}$$

$$\mathbb{C}\text{ov}\,\vec{y}(k) = \mathbb{C}\text{ov}\,\vec{z}(k) + \text{terme d'erreur} \tag{3.2}$$

où nous expliquerons dans quels cas les termes d'erreurs sont négligeables.

L'avantage est que l'approximation de Bernoulli $\vec{z}(k)$ est double. Premièrement elle est caractérisée par un paramètre $\vec{p}$, les *probabilités de tirage des réactions*, qui seront en vue de notre futur cadre de méthodes de contraintes l'analogue des flux stationnaires. Deuxièmement elle est plus simple à étudier que le processus initial $\vec{y}(k)$: nous obtiendrons des formules analytiques pour les moments de $\vec{z}(k)$ ainsi qu'un théorème central limite lorsque $k \to +\infty$.

3.1 Discrétisation de la dynamique stochastique

Nous commençons par définir la discrétisation à pas de temps fixe δt de la dynamique stochastique et nous prouvons qu'elle est correcte lorsque $\delta t \to 0$.

3.1.1 Définition

Afin d'obtenir des simulations de Gillespie plus efficaces, plusieurs discrétisations de la dynamique stochastique ont été proposées comme [San07, San08] qui repose sur l'uniformisation des chaînes de Markov continues. Nous proposons une discrétisation similaire, à pas de temps fixe δt, qui ne requiert aucune propriété sur les taux de transitions de la chaîne de Markov à temps continu sous-jacente et qui a la bonne propriété de converger vers le processus continu lorsque le pas de temps δt tend vers 0.

Pour simplifier, on ajoute virtuellement la réaction nulle $R_0 : \varnothing \to \varnothing$ au réseau de réaction de telle sorte qu'on considère qu'à chaque pas de temps δt une ou zéro réaction se produit. On considèrera les fonctions de probabilités de réactions suivantes (incluant la probabilité de la réaction nulle).

$$p_j(\vec{x}) = \begin{cases} \exp\left(-h_0(\vec{x})\,\delta t\right) & \text{si } j = 0, \\ \left(1 - p_0(\vec{x})\right)\frac{h_j(\vec{x})}{h_0(\vec{x})} & \text{sinon.} \end{cases} \tag{3.3}$$

En utilisant ces fonctions de probabilité, on peut définir le processus discrétisé $\vec{y}(k)$.

Définition 3.1 (Processus discrétisé)**.** Le *processus discrétisé* du processus stochastique continu $(\vec{x}(t))_{t\in\mathbb{R}^+}$ sur l'espace d'état Σ est la chaîne de Markov homogène $(\vec{y}(k))_{k\in\mathbb{N}}$ définie sur le même espace d'états Σ, avec même état initial $\vec{y}(0) = \vec{x}(0)$ et ayant comme probabilités de transitions

$$\forall j = 0, \ldots, m, \forall \vec{x}, \vec{x} + \vec{v}_j \in \Sigma, \quad \mathbb{P}\left[\vec{y}(k+1) = \vec{x} + \vec{v}_j \mid \vec{y}(k) = \vec{x}\right] = p_j(\vec{x}). \tag{3.4}$$

La discrétisation repose sur le principe suivant : soit on considère qu'aucune réaction ne s'est produite pendant le pas de temps, soit on considère qu'une seule réaction s'est produite. La probabilité qu'aucune réaction ne soit tirée peut se calculer facilement à partir de la chaîne de Markov $\vec{x}(t)$ et correspond à l'expression de la probabilité p_0. Les probabilités de déclenchement des réactions non nulles sont fixées comme proportionnelles aux probabilités de transitions de celles de $\vec{x}(t)$. La probabilité d'occurrence de plusieurs réactions dans le même pas de temps diminue lorsque le pas de temps est petit ce qui permet d'obtenir la convergence pour les petits pas de temps.

3.1.2 Validité

La validité de l'approximation lorsque le pas de temps δt est petit est donné par la proposition suivante.

Proposition 3.1. *Soit $(\vec{x}'(t))$ $(t \in \mathbb{R}^+)$ l'extension à $\mathbb{R}^+$ du processus discrétisé $(\vec{y}(k))$ $(k \in \mathbb{N})$, c'est-à-dire $\forall k \in \mathbb{N}, \vec{x}'(k\delta t) = \vec{y}(k)$ et $\vec{x}'(t)$ est constant sur chaque intervalle $[k\delta t, (k+1)\delta t[$. Alors, pour tout instant $t \in \mathbb{R}^+$, $\vec{x}'(t)$ converge en loi vers $\vec{x}(t)$ lorsque $\delta t \to 0$,*

$$\vec{x}'(t) \xrightarrow{\mathscr{L}} \vec{x}(t). \tag{3.5}$$

Démonstration. Tout comme $\vec{x}(t)$, le processus $\vec{x}'(t)$ peut-être défini à l'aide d'un algorithme de tirage de réactions utilisant un temps inter-réaction τ' et un numéro de prochaine réaction μ' dont les lois convergent vers celles de τ et μ de l'algorithme de Gillespie d'après le Lemme 3.1. □

Lemme 3.1. *Pour tout instant de réaction de $t_0 \in \delta t\mathbb{N}$ de $(\vec{x}'(t))$, on note l'état actuel $\vec{\chi} = \vec{x}'(t_0)$, la durée jusque la prochaine réaction non nulle $\tau' = \inf\{t' > 0 | \vec{x}'(t_0 + t') \neq \vec{\chi}\}$ et μ' le*

numéro de la réaction déclenchée au temps $(t_0 + \tau')$ *alors*

$$\forall t \geq 0, \mathbb{P}[\tau' > t] \to_{\delta t \to 0} \mathbb{P}[\tau > t],$$
$$\mathbb{P}[\mu' = j] = \mathbb{P}[\mu = j],$$

où τ *et* μ *sont respectivement la durée jusque la prochaine réaction et son numéro dans l'algorithme de Gillespie sachant que l'état courrant est* $\vec{x}(t_0) = \vec{\chi}$.

Démonstration. — Puisque les changements de valeurs de $\vec{x}'$ ont lieu nécessairement aux instants discrets $\delta t\mathbb{N}$ on a $\tau' \in \delta t\mathbb{N}$ donc par conséquent $\forall t \geq 0$, $\mathbb{P}[\tau' > t] = \mathbb{P}[\tau' > \lfloor t/\delta t \rfloor \delta t]$. D'après la définition de $(\vec{y}(k))$, ce dernier événement correspond à $\lfloor t/\delta t \rfloor$ succès d'un tirage de Bernoulli paramètre $\vec{p}_0(\vec{\chi})$. Donc, $\mathbb{P}[\tau' > t] = \exp(-h_0(\vec{\chi})\delta t)^{\lfloor t/\delta t \rfloor} = \exp(-h_0(\vec{\chi})\delta t \lfloor t/\delta t \rfloor)$. Ainsi on a l'encadrement suivant

$$\exp(-h_0(\vec{\chi})\delta t(t/\delta t + 1)) \leq \mathbb{P}[\tau' > t] \leq \exp(-h_0(\vec{\chi})\delta t(t/\delta t))$$

qui montre que $\lim_{\delta t \to 0} \mathbb{P}[\tau' > t] = \exp(-h_0(\vec{\chi})t) = \mathbb{P}[\tau > t]$.

— Pour obtenir la seconde égalité, on considère toute les valeurs possibles de τ' :

$$\begin{aligned}
\mathbb{P}[\mu' = j] &= \sum_{l \in \mathbb{N}^*} \mathbb{P}[\mu' = j, \tau' = l\delta t] \\
&= \sum_{l \in \mathbb{N}^*} \mathbb{P}[\tau' = l\delta t] \cdot \mathbb{P}[\mu' = j | \tau' = l\delta t] \\
&= \sum_{l \in \mathbb{N}^*} p_0(\vec{\chi})^{l-1}(1 - p_0(\vec{\chi})) \cdot \frac{h_j(\vec{\chi})}{h_0(\vec{\chi})} \\
&= \frac{1}{1 - p_0(\vec{\chi})}(1 - p_0(\vec{\chi}))\frac{h_j(\vec{\chi})}{h_0(\vec{\chi})} \\
&= h_j(\vec{\chi})/h_0(\vec{\chi}) = \mathbb{P}[\mu = j].
\end{aligned}$$

□

La proposition valable à tout instant est en particulier valable lors des multiples des pas de temps, ce qui donne le corollaire suivant.

Corollaire 3.1. *Pour tout* $k \in \mathbb{N}$, *le processus discret* $\vec{y}(k)$ *converge en loi vers* $\vec{x}(k\delta t)$ *lorsque* $\delta t \to 0$.

3.1.3 Illustration

À titre d'exemple, considérons l'exemple 1.1 de la dimérisation

$$\begin{aligned}
2X &\to Y \quad (c_1) \\
Y &\to 2X \quad (c_2),
\end{aligned}$$

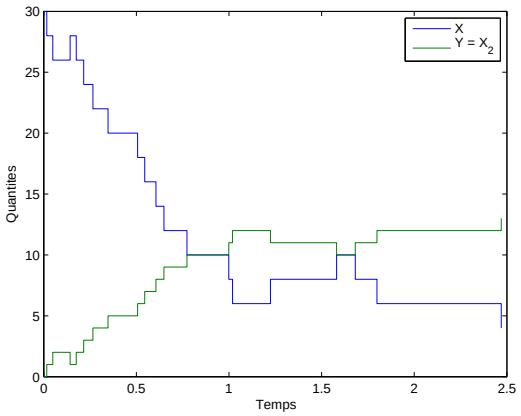


FIGURE 3.2 – **Simulation de Gillespie du réseau de dimérisation** pour $a_0 = 15$, $c_1 = c_2 = 0,1$.

où Y représente le dimère X_2. On choisit de considérer un état initial ne contenant aucun dimère et de la forme $(x_0, y_0) = (2a_0, 0)$ où a_0 est le nombre maximal de dimères pouvant se former. La chaîne de Markov $(x(t))$ $(t \in \mathbb{R})$ à temps continu décrivant la dynamique stochastique de ce réseau a donc pour espace d'états $\Sigma = \{(2a, a_0 - a) \ / \ a = 0, \ldots, a_0\}$ de cardinal $a_0 + 1$. Par simplicité on notera les états selon la valeur $a \in \{0, \ldots, a_0\}$. Les fonctions de propensions, c'est-à-dire les taux de transitions de la chaîne de Markov en fonction de l'état courant, sont

$$h_1(a) = c_1 a(2a - 1)/2, \tag{3.6}$$

$$h_2(a) = c_2(a_0 - a), \tag{3.7}$$

$$h_0(a) = h_1(a) + h_2(a). \tag{3.8}$$

L'algorithme de Gillespie permet de simuler la chaîne de Markov correspondante, la figure 3.2 représente un exemple possible de trajectoire pour $a_0 = 15$ et $c_1 = c_2 = 0,1$.

Si on souhaite discrétiser cette chaîne de Markov avec la méthode proposée, on se fixe un pas de temps δt et on considère la chaîne de Markov $(y(k))$ $(k \in \mathbb{N})$ à temps discret ayant même espace d'états, même état initial et pour probabilités de transitions

$$p_0(a) = \exp(-h_0(a)\delta t), \tag{3.9}$$

$$p_1(a) = (1 - p_0)h_1(a)/h_0(a), \tag{3.10}$$

$$p_2(a) = (1 - p_0)h_2(a)/h_0(a), \tag{3.11}$$

où p_0 est la probabilité de rester dans le même état, p_1 la probabilité de déclencher la première réaction et p_2 celle de déclencher la seconde réaction. Pour rappel $y(k)$ représente la valeur de $x(t)$ à l'instant $t = k\delta t$. La figure 3.3 représente une trajectoire possible de cette nouvelle chaîne de Markov à temps discret pour $\delta t = 0,01$.

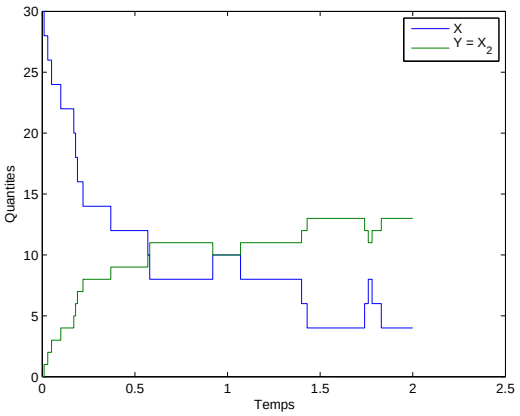


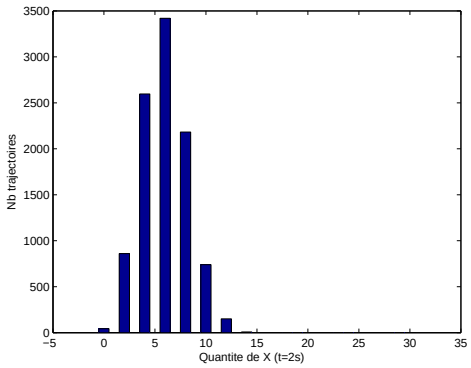
FIGURE 3.3 – **Simulation du processus discrétisé du réseau de dimérisation** pour $a_0 = 15$, $c_1 = c_2 = 0,1$ et $\delta t = 0,01$.

Pour obtenir une discrétisation correcte, le pas de temps δt doit être choisi petit devant le temps moyen nécessaire au déclenchement d'une réaction, c'est-à-dire $1/h_0$. Pour $a_0 = 15$ et $c_1 = c_2 = 0,1$, on s'aperçoit que ce temps moyen est maximum lorsque $a = a_0$ et vaut alors $2/(0,1 \times 30 \times 29) \simeq 0,023$. δt doit donc être choisi petit devant $0,023$.

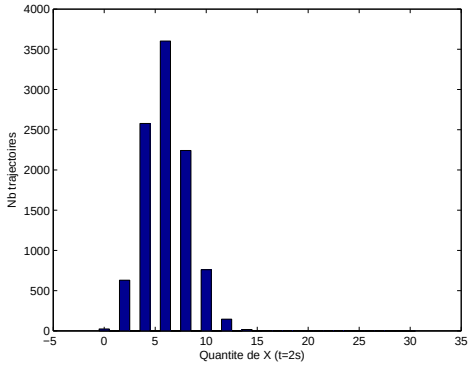
Pour comparer les chaînes de Markov $(x(t))$ et $(y(k))$, on se propose de comparer leurs distributions à l'état stationnaire. On s'aperçoit sur les figures 3.2 et 3.3 que les moyennes des trajectoires s'équilibrent au bout de 2s environ, on comparera donc les distributions de $(x(t))$ et $(y(k))$ à cet instant. La figure 3.4 représente cette comparaison pour différentes valeurs de δt. Lorsque δt est choisi trop grand ($\delta t = 0,5$ ou $\delta t = 0,3$) les distributions des deux processus sont visiblement très différentes. Le cas $\delta t = 0,1$ est un cas limite : la distribution semble avoir la bonne allure mais on se rend compte qu'elle est centrée autour de la valeur 8 alors que son maximum devrait plutôt se situer autour de la valeur 6. L'explication est que si le pas de temps est choisi trop grand, le processus discrétisé n'a pas le temps de *suivre* le processus continu, les trajectoires décroissent donc moins vite qu'elles ne le devraient. À partir de $\delta t = 0,01$ on obtient bien une bonne correspondance entre les deux distributions.

3.2 Dynamique de Bernoulli

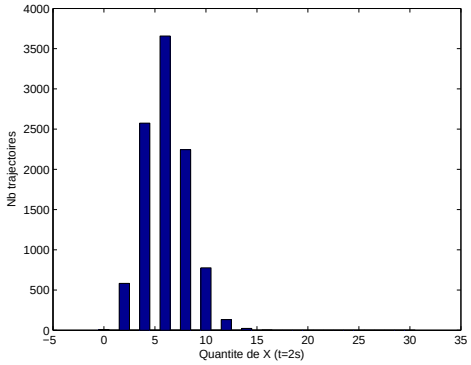
Nous introduisons maintenant la dynamique de Bernoulli d'un réseau de réaction. Cette dynamique a l'avantage d'être définie à l'aide d'un paramètre $\vec{p}$ qui peut-être mis en parallèle avec le vecteur de flux $\vec{f}$ des méthodes de flux à l'équilibre. De plus, nous montrerons que les expressions analytiques de ses moments peuvent être obtenues en fonction de $\vec{p}$ et de la matrice de stœchiométrie S ce qui permettra d'aboutir dans la suite aux méthodes par contraintes. Dans la suite de ce chapitre, nous établirons les relations entre ce processus et



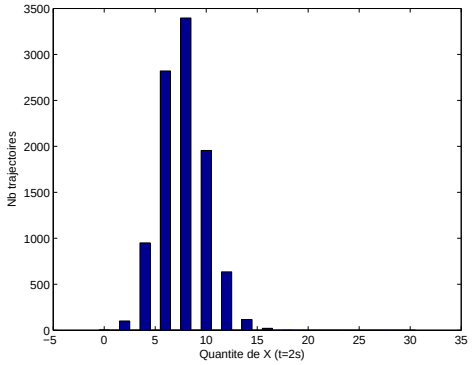
Processus continu



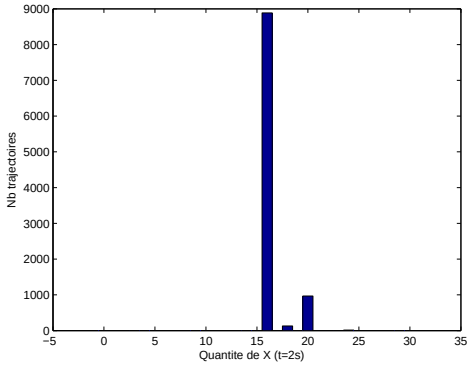
Processus discrétisé, $\delta t = 0,001$



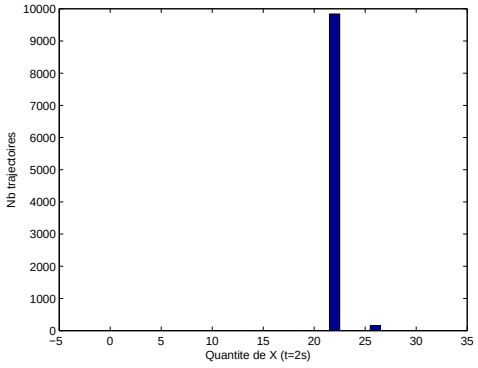
Processus discrétisé, $\delta t = 0,01$



Processus discrétisé, $\delta t = 0,1$



Processus discrétisé, $\delta t = 0,3$



Processus discrétisé, $\delta t = 0,5$

FIGURE 3.4 – **Comparaison des processus continu et discrétisé.** Comparaison de la distribution de la quantité de X après 2s pour 10 000 trajectoires.

le processus discrétisé défini dans la section précédente.

3.2.1 Définition

Nous commençons par donner un rappel de la loi de Bernoulli que nous généralisons à n résultats possibles.

Définition 3.2 (Variable de Bernoulli (généralisée))**.** Soit $\vec{p}$ un vecteur de probabilité de dimension n, on appelle *variable de Bernoulli* de paramètre $\vec{p}$ la variable aléatoire X à valeurs dans $\{1,\dots,n\}$ de loi

$$\mathbb{P}[X = k] = p_k. \tag{3.12}$$

On notera $\mathbb{B}(\vec{p})$ la loi de cette variable.

La loi de Bernoulli est au cœur de la définition de la dynamique qu'on nommera en conséquence dynamique de Bernoulli.

Définition 3.3 (Dynamique de Bernoulli)**.** Soit (n, m, α, β) un réseau de réaction de matrice de stœchiométrie $S = (\vec{v}_j)_{1\le j\le m}$, $\vec{z}(0) \in \mathbb{N}^n$ un état initial et $\vec{p}$ un vecteur de probabilité de taille n appelé *probabilités des réactions*. On appellera *dynamique de Bernoulli* le processus aléatoire discret $\vec{z}_k \in \mathbb{Z}^n$ défini par

$$\vec{z}(k) = \vec{z}(0) + \sum_{i=1}^{k}\sum_{j=1}^{m} \delta_j^{\mu_i}\vec{v}_j \qquad (k \in \mathbb{N}), \tag{3.13}$$

où les μ_i sont indépendants et identiquement distribués de loi $\mathbb{B}(\vec{p})$, δ_a^b est le symbole de Kronecker (valant 1 si $a = b$ et 0 sinon). On notera $\mathbb{B}(\vec{z}(0), \vec{p})$ la loi du processus $(z(k))$ $(k \in \mathbb{N})$.

De manière plus intuitive, la dynamique de Bernoulli consiste à tirer aléatoirement une réaction selon les probabilités $\vec{p}$ à chaque instant discret. On peut décrire ce processus de manière équivalente de la façon suivante

$$\vec{z}(k) = \vec{z}(0) + \sum_{i=1}^{k} S\vec{e}_{\mu_i}, \tag{3.14}$$

où $(\vec{e}_i)_{1\le i\le m}$ sont les vecteurs de la base canonique de $\mathbb{R}^m$.

3.2.2 Expression analytique des espérances et variances

L'avantage de la dynamique de Bernoulli est qu'elle permet une expression analytique des moments d'ordre 1 et 2, c'est-à-dire l'espérance $\mathbb{E}\,\vec{z}(k)$ et la matrice variance-covariance $\mathbb{C}\mathrm{ov}\,\vec{z}(k)$ en fonction de $\vec{p}$, $\vec{z}(0)$ et de la matrice de stœchiométrie S du réseau de réaction.

Proposition 3.2. *Soit $\vec{z}(k)$ la dynamique de Bernoulli ayant pour état initial $\vec{z}(0)$, de paramètre $\vec{p}$ d'un réseau de matrice de stœchiométrie S alors*

$$\mathbb{E}\,\vec{z}(k) = \vec{z}(0) + kS\vec{p}, \tag{3.15}$$

$$\mathbb{C}\mathrm{ov}\,\vec{z}(k) = kS(\mathrm{diag}\,\vec{p} - \vec{p}\vec{p}^t)S^t. \tag{3.16}$$

Démonstration. — Par définition de la dynamique de Bernoulli et linéarité de l'espérance,

$$\mathbb{E}\,\vec{z}(k) = \vec{z}(0) + \sum_{i=1}^{k}\sum_{j=1}^{m}\mathbb{E}(\delta_j^{\mu_i}\vec{v}_j). \tag{3.17}$$

Or

$$\mathbb{E}(\delta_j^{\mu_i}\vec{v}_j) = \sum_{l=1}^{m}\mathbb{P}[\mu_i = l]\delta_j^l\vec{v}_j = p_j\vec{v}_j. \tag{3.18}$$

En remarquant que $S\vec{p} = \sum_{j=1}^{m} p_j\vec{v}_j$ on obtient donc

$$\mathbb{E}\,\vec{z}(k) = \vec{z}(0) + \sum_{i=1}^{k} S\vec{p} = \vec{z}(0) + kS\vec{p}. \tag{3.19}$$

— La covariance étant invariante par translation, on peut supposer sans nuire à la généralité que $\vec{z}(0) = \vec{0}$. En utilisant l'équation (3.14), on obtient par linéarité de l'espérance

$$\mathbb{C}\mathrm{ov}(\vec{z}(k)) = \mathbb{E}(\vec{z}(k)\vec{z}^t(k)) - \mathbb{E}\,\vec{z}(k)\,\mathbb{E}\,\vec{z}^t(k) \tag{3.20}$$

$$= \mathbb{E}\left((\sum_{a=1}^{k} S\vec{e}_{\mu_a})(\sum_{b=1}^{k} S\vec{e}_{\mu_b})^t\right) - k^2 S\vec{p}\vec{p}^t S^t \tag{3.21}$$

$$= S\left(\sum_{a=1}^{k}\sum_{b=1}^{k}\mathbb{E}(\vec{e}_{\mu_a}\vec{e}_{\mu_b}^t) - k^2\vec{p}\vec{p}^t\right)S^t. \tag{3.22}$$

Étudions la valeur de $\mathbb{E}(\vec{e}_{\mu_a}\vec{e}_{\mu_b}^t)$ selon deux cas.

1. Lorsque $a \neq b$, μ_a et μ_b sont indépendants et on peut écrire

$$\mathbb{E}(\vec{e}_{\mu_a}\vec{e}_{\mu_b}^t) = \mathbb{E}\,\vec{e}_{\mu_a}\,\mathbb{E}\,\vec{e}_{\mu_b}^t = \vec{p}\vec{p}^t. \tag{3.23}$$

2. Dans le cas contraire, en notant $E_{i,j}$ la matrice contenant 1 aux coordonnées (i, j) et 0 ailleurs, on a

$$\mathbb{E}(\vec{e}_{\mu_a}\vec{e}_{\mu_a}^t) = \sum_{l=1}^{m}\mathbb{P}[\mu_a = l]E_{l,l} = \mathrm{diag}\,\vec{p}. \tag{3.24}$$

En remplaçant dans le calcul de la matrice de covariance $\mathbb{E}(\vec{e}_{\mu_a}\vec{e}_{\mu_b})$ par ces nouvelles expressions, on obtient

$$\mathbb{C}\mathrm{ov}(\vec{z}(k)) = S(k\,\mathrm{diag}\,\vec{p} + k(k-1)\vec{p}\vec{p}^t - k^2\vec{p}\vec{p}^t)S^t \tag{3.25}$$

$$= kS(\mathrm{diag}\,\vec{p} - \vec{p}\vec{p}^t)S^t. \tag{3.26}$$

□

3.2.3 Théorème central limite pour la dynamique de Bernoulli

En plus de pouvoir déterminer une expression analytique des moments d'ordre 1 et de 2 pour la dynamique de Bernoulli, on a également le résultat asymptotique suivant.

Proposition 3.3 (Théorème central limite)**.** *La dynamique de Bernoulli $\vec{z}(k)$ d'état initial $\vec{z}(0)$, de probabilités de réactions $\vec{p}$ et associée à un réseau de matrice stœchiométrique S converge en loi vers une loi normale*

$$\frac{1}{\sqrt{k}}\left(\vec{z}(k)-\left(\vec{z}(0)+kS\vec{p}\right)\right)\xrightarrow[k\to+\infty]{\mathscr{L}}\mathcal{N}\left(\vec{0},W(S,\vec{p})\right), \tag{3.27}$$

où $W(S,\vec{p})=S\left(\operatorname{diag}(\vec{p})-\vec{p}\vec{p}^{\,t}\right)S^t$.

On a donc déterminé que la distribution asymptotique de $\vec{z}(k)$ est une gaussienne dont le centre et la matrice variance-covariance sont déterminés analytiquement en fonction de S et $\vec{p}$. Intuitivement, la proposition montre que lorsque k est grand, $\vec{z}(k)$ est approximativement distribué selon la loi normale

$$\mathcal{N}\left(\vec{z}(0)+kS\vec{p},\ kS(\operatorname{diag}\vec{p}-\vec{p}\vec{p}^{\,t})S^t\right). \tag{3.28}$$

La démonstration repose sur l'utilisation du théorème central limite généralisé [VdV00] en dimension finie quelconque.

Lemme 3.2 (Théorème central limite généralisé)**.** *Soit $(\vec{X}_n)_{n\in\mathbb{N}*}$ un processus aléatoire à valeurs dans $\mathbb{R}^m$ indépendant et identiquement distribué, admettant comme moyenne $\vec{m}$ et comme matrice variance-covariance V, alors*

$$\frac{1}{\sqrt{k}}\left(\sum_{n=1}^{k}(\vec{X}_k-\vec{m})\right)\xrightarrow[k\to+\infty]{\mathscr{L}}\mathcal{N}(\vec{0},V). \tag{3.29}$$

Démonstration. (Proposition 3.3) On part de l'équation (3.14) et on applique le lemme 3.2 avec $\vec{X}_i=S\vec{e}_{\mu_i}$ en remarquant que $\vec{m}=\mathbb{E}\,\vec{X}_i=S\vec{p}$:

$$\frac{1}{\sqrt{k}}\left(\vec{z}(k)-\left(\vec{z}(0)+kS\vec{p}\right)\right)=\frac{1}{\sqrt{k}}\left(\sum_{i=1}^{k}S\vec{e}_{\mu_i}-kS\vec{p}\right) \tag{3.30}$$

$$=\frac{1}{\sqrt{k}}\sum_{i=1}^{k}(\vec{X}_i-\vec{m}) \tag{3.31}$$

$$\xrightarrow[k\to+\infty]{\mathscr{L}}\mathcal{N}(\vec{0},V) \tag{3.32}$$

où $V=\mathbb{C}\mathrm{ov}\,S\vec{e}_{\mu_i}$. On peut calculer cette matrice de covariance en utilisant la même méthode que dans le calcul de l'expression analytique des moments :

$$V=\mathbb{E}((S\vec{e}_{\mu_i})(S\vec{e}_{\mu_i})^t)-\mathbb{E}(S\vec{e}_{\mu_i})\,\mathbb{E}(S\vec{e}_{\mu_i})^t \tag{3.33}$$

$$=S\left(\mathbb{E}(\vec{e}_{\mu_i}\vec{e}^{\,t}_{\mu_i})\right)S^t-S\vec{p}\vec{p}^{\,t}S^t \tag{3.34}$$

$$=S\left(\operatorname{diag}\vec{p}-\vec{p}\vec{p}^{\,t}\right)S^t. \tag{3.35}$$

□

3.2.4 Interprétation en tant que marches aléatoires

Dans cette sous-section nous présentons une manière alternative de comprendre l'analyse de la dynamique de Bernoulli ainsi que son théorème central limite. En effet, si on reprend l'équation 3.14 :

$$\vec{z}(k) = \vec{z}(0) + \sum_{i=1}^{k} S\vec{e}_{\mu_i},$$

on s'aperçoit que la dynamique de Bernoulli est l'image par une transformation affine du processus stochastique $\vec{q}_k = \sum_{i=1}^{k} \vec{e}_{\mu_i}$. Ce processus compte toutes les occurrences de chaque réaction jusqu'à l'instant $k\delta t$, on l'appellera donc le *processus de comptage des réactions* (PCR). On peut alors redémontrer le théorème central limite en deux étapes. Premièrement on montre que le résultat est vrai pour le PCR (qui correspond au cas particulier $S = \mathbb{I}_n$) et ensuite, on utilise une transformation affine du PCR pour obtenir le cas général dans le cas général.

Nous remarquons que le PCR est une *marche aléatoire* [Fel74, Gut09] dans $\mathbb{R}^m$, puisque il possède des incréments $\vec{e}_{\mu_i}$ indépendants et identiquement distribués. En d'autres mots q_{k+1} est obtenu à partir de q_k en choisissant aléatoirement une dimension selon les probabilités $\vec{p}$ et ensuite en se déplaçant d'une unité dans cette direction. Ainsi, le théorème central limite multivarié [VdV00] implique que le PCR tend vers une loi normale, ce qui constitue un comportement typique des marches aléatoires [Fel74]. On peut ainsi obtenir la proposition suivante.

Proposition 3.4. *Le PCR $\vec{q}_k = \sum_{l=1}^{k} \vec{e}_{\mu_l}$ vérifie les propriétés suivantes :*

1. *$\vec{q}_k$ converge en loi vers une une loi normale multivariée ;*
2. *$\mathbb{E}(\vec{q}_k) = k\vec{p}$,*
3. *la matrice de variance-covariance de $\vec{q}_k$ est $\mathbb{C}\mathrm{ov}(\vec{q}_k) = k\left(\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t\right)$.*

Démonstration. Le premier point vient du théorème central limite multivarié tandis que les expressions des moments se déduisent de la proposition 3.2 pour $S = \mathbb{I}$. □

Maintenant que nous avons démontré que le PCR converge en loi vers la loi normale, on exploite alors d'après l'équation (3.14) le fait que la dynamique de Bernoulli $\vec{z}(k)$ est une transformation affine du PCR. Ainsi, le processus inclus est aussi une marche aléatoire où les incréments (les pas) possibles sont les images par cette transformation affine des vecteurs de la base canonique (voir la figure 3.5 pour un exemple).

La conséquence principale est que la dynamique de Bernoulli tend aussi vers une loi normale. Pour le démontrer nous utilisons la propriété essentielle de stabilité des lois normales par transformations affines.

Proposition 3.5. *Soit $\vec{x} \in \mathbb{R}^n$ un vecteur aléatoire gaussien, $\vec{z} \in \mathbb{R}^m$ un vecteur constant (i.e. non aléatoire) et A une matrice réelle de taille $m \times n$ alors le vecteur $\vec{y} = \vec{z} + A\vec{x}$ est aussi Gaussien de moyenne*

$$\mathbb{E}(\vec{y}) = \vec{z} + A\mathbb{E}(\vec{x})$$

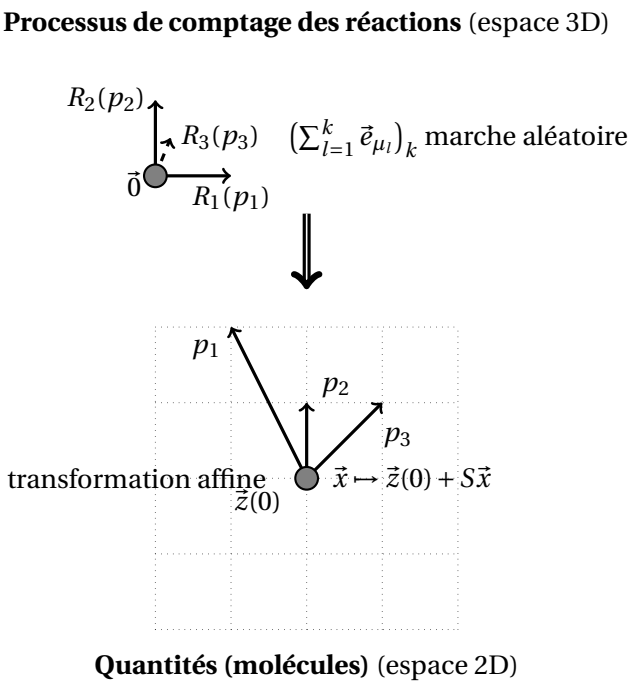


FIGURE 3.5 – **Illustration du comportement de marche aléatoire** sur l'exemple du réseau $\{X \to 2Y; \varnothing \to Y; \varnothing \to X + Y\}$. La sphère représente l'état courant du réseau et les flèches les nouveaux états possibles après l'occurrence de chacune des réactions qui a lieu avec probabilités p_i.

et de matrice de variance-covariance

$$\mathbb{C}\mathrm{ov}(\vec{y}) = A\mathbb{C}\mathrm{ov}(\vec{x})A^{\top}.$$

Ainsi, d'après les propositions 3.4 et 3.5, on aboutit à une nouvelle preuve du théorème central limite (Proposition 3.3) pour la dynamique de Bernoulli.

Cette manière alternative d'aboutir au même résultat est intéressante car elle permet de comprendre la dynamique de Bernoulli comme une transformation affine d'une marche aléatoire (le RCP) dans l'octant positif. Le loi limite gaussienne est alors l'image par cette transformation affine de la loi limite gaussienne du RCP.

3.3 Analyse stationnaire et validité de l'approximation

Dans cette dernière section, on montre comment doit être choisi le paramètre $\vec{p}$ de la dynamique de Bernoulli. On se servira pour cela des probabilités de réactions stationnaires $\vec{\mathfrak{p}}$ que l'on définira. Nous étudions ensuite la qualité de l'approximation obtenue en ce qui

concerne les moments d'ordre 1 et 2 c'est-à-dire qu'on obtiendra des relations de la forme

$$\mathbb{E}\,\vec{y}(k) = \mathbb{E}\,\vec{z}(k) + \text{terme d'erreur}$$
$$\mathbb{C}\text{ov}\,\vec{y}(k) = \mathbb{C}\text{ov}\,\vec{z}(k) + \text{terme d'erreur}$$

où on précisera le terme d'erreur. Nous montrerons que ce terme d'erreur peut-être négligé à la limite thermodynamique. Ces résultats d'approximation montrent aussi que le choix du paramètre $\vec{p} := \vec{\mathfrak{p}}$ dans la dynamique de Bernoulli est judicieux.

3.3.1 Probabilités de réactions stationnaires

Jusqu'à présent nous avons introduit la dynamique de Bernoulli mais nous n'avons toujours pas expliqué comment choisir son paramètre principal à savoir les probabilités de réactions $\vec{p}$. Pour cela, nous allons utiliser une approche de stationnarité en s'inspirant de la construction du FBA à partir des équations différentielles. En dynamique différentielle, on considère que les concentrations de réactifs n'évoluent plus et que leurs dérivées sont donc nulles. On peut alors définir le vecteur de flux stationnaires par $\vec{f} = \lim_{t\to+\infty} f(\vec{x}^*(t))$ où $\vec{x}^*(t)$ est la projection orthogonale de $\vec{x}(t)$ sur l'espace des réactifs. En dynamique stochastique, on considèrera une définition similaire.

Définition 3.4 (Probabilités de réactions stationnaires)**.** Soit $\vec{y}(k)$ la discrétisation de la dynamique stochastique d'un réseau de réactions et $\vec{p}(\cdot)$ la fonction de probabilités de transitions associée. Soit $\vec{y}^*(k)$ la projection orthogonale de $\vec{y}(k)$ sur l'espace des réactifs. Lorsque cette limite existe, on définira les *probabilités de réactions stationnaires* par

$$\vec{\mathfrak{p}} = \lim_{k\to+\infty} \mathbb{E}\,\vec{p}(\vec{y}^*(k)). \tag{3.36}$$

On remarque que $\vec{p}(\cdot)$ ne dépendant que des réactifs, on peut remplacer $\vec{y}^*$ par $\vec{y}$ dans la définition.

La possible non existence de la limite n'est pas un problème. En effet, même en dynamique différentielle certains systèmes n'admettent pas de régime stationnaire, autrement dit les concentrations de leurs réactifs ne tendent pas vers une constante. C'est par exemple le cas du système oscillant de Lotka-Volterra présenté dans le premier chapitre. À partir de maintenant on considèrera que le paramètre de la dynamique de Bernoulli est $\vec{p} := \vec{\mathfrak{p}}$. C'est ce paramètre qui jouera le rôle des flux stationnaires dans nos approches par contraintes. Ainsi en général, on ne connaîtra pas la valeur de ce paramètre et on tentera plutôt de le contraindre à l'aide d'informations sur les trajectoires.

Nous avons réussi à établir des relations analytiques entre $\mathbb{E}\,\vec{z}(k), \mathbb{C}\text{ov}\,\vec{z}(k)$ et $\vec{p} = \vec{\mathfrak{p}}$. Pour conclure il nous faut maintenant comprendre la relation entre $\mathbb{E}\,\vec{z}(k)$ (resp. $\mathbb{C}\text{ov}\,\vec{z}(k)$) et $\mathbb{E}\,\vec{y}(k)$ (resp. $\mathbb{C}\text{ov}\,\vec{y}(k)$). C'est l'objectif de la fin de ce chapitre.

3.3.2 Comparaison des espérances

Proposition 3.6. *Le terme d'erreur entre les espérances de la dynamique discrétisée $\vec{y}(k)$ et la dynamique de Bernoulli de paramètre $\vec{\mathfrak{p}}$ est donné par*

$$\mathbb{E}\,\vec{y}(k) = \mathbb{E}\,\vec{z}(k) + S\sum_{l=1}^{k}\left(\mathbb{E}(\vec{p}(\vec{y}(l-1)) - \vec{\mathfrak{p}}\right). \tag{3.37}$$

Démonstration. Si $(\mu_i)_{i\in\mathbb{N}}$ désigne les index des réactions tirées dans l'algorithme de simulation de $\vec{y}(k)$, alors on a pour tout $k \in \mathbb{N}$

$$\vec{y}(k+1) = \vec{y}(k) + S\vec{e}_{\mu_{k+1}}. \tag{3.38}$$

En utilisant l'espérance conditionnelle on obtient

$$\begin{aligned}
\mathbb{E}\,\vec{y}(k+1) &= \mathbb{E}\mathbb{E}(\vec{y}(k+1)|\vec{y}(k)) = \mathbb{E}\mathbb{E}(\vec{y}(k) + S\vec{e}_{\mu_{k+1}}|\vec{y}(k)) && (3.39)\\
&= \mathbb{E}(\vec{y}(k) + S\mathbb{E}(\vec{e}_{\mu_{k+1}}|\vec{y}(k))) && (3.40)\\
&= \mathbb{E}(\vec{y}(k) + S\vec{p}(\vec{y}(k))) && (3.41)\\
&= \mathbb{E}\,\vec{y}(k) + S\mathbb{E}\,\vec{p}(\vec{y}(k)), && (3.42)
\end{aligned}$$

on en déduit donc que pour tout $k \in \mathbb{N}$

$$\mathbb{E}\,\vec{y}(k) = \vec{y}(0) + \sum_{l=1}^{k} S\mathbb{E}\,\vec{p}(\vec{y}(l-1)). \tag{3.43}$$

D'autre part d'après l'équation 3.14

$$\vec{z}(k) = \vec{z}(0) + \sum_{l=1}^{k} S\vec{e}_{\mu_l}. \tag{3.44}$$

On obtient en appliquant l'espérance

$$\mathbb{E}\,\vec{z}(k) = \vec{z}(0) + \sum_{l=1}^{k} S\mathbb{E}\,\vec{e}_{\mu_l} = \vec{z}(0) + \sum_{l=1}^{k} S\vec{\mathfrak{p}}. \tag{3.45}$$

En comparant les deux expressions obtenues pour $\mathbb{E}\,\vec{y}(k)$ et $\mathbb{E}\,\vec{z}(k)$ on obtient le résultat souhaité. □

On peut remarquer que cette proposition justifie notre choix $\vec{p} := \vec{\mathfrak{p}}$ puisque alors le terme général de la série dans le terme d'erreur tend vers 0. Un autre choix aurait conduit à l'accumulation d'un nouveau terme d'erreur linéaire en k.

On note $\|\cdot\|$ la norme sup sur $\mathbb{R}^m$ et on utilisera la même notation sans ambiguïté pour la norme subordonnée matricielle associée. On rappelle que dans ce cas, la norme matricielle subordonnée correspond à la somme maximale des modules des éléments d'une ligne, c'est-à-dire que pour une matrice A quelconque de taille $n \times m$,

$$\|A\| = \max_{1\le i\le n}\sum_{j=1}^{m}|a_{i,j}|. \tag{3.46}$$

À partir de la dernière proposition, on peut obtenir une majoration grossière

Corollaire 3.2. *Le terme d'erreur est majoré par une fonction linéaire du temps*

$$\|\mathbb{E}\,\vec{y}(k) - \mathbb{E}\,\vec{z}(k)\| \leq k\|S\|. \tag{3.47}$$

Démonstration. Cette inégalité s'obtient par inégalité triangulaire et en utilisant l'inégalité $\|A\vec{u}\| \leq \|A\| \cdot \|\vec{u}\|$ vraie pour toute matrice A et tout vecteur $\vec{u}$. □

On remarque que la pente de la fonction linéaire majorant le terme d'erreur est égale à $\|S\|$ c'est-à-dire la somme maximale des coefficients stœchiométriques d'une ligne de la matrice. Cela revient à dire que le terme d'erreur est majoré par k fois la somme des coefficients de l'espèce globalement la plus impactée par les réactions.

Le corollaire suivant exploite le fait que $\mathbb{E}\,\vec{y}(k) \to \vec{\mathfrak{p}}$ (par définition de $\vec{\mathfrak{p}}$) et permet une compréhension plus fine du terme d'erreur.

Corollaire 3.3. $\forall \varepsilon > 0, \exists K > 0, \forall k \geq K$

$$\|\mathbb{E}\,\vec{y}(k) - \mathbb{E}\,\vec{z}(k)\| \leq \|S\| \sum_{l=1}^{K} \left\|\mathbb{E}(\vec{p}(\vec{y}(i-1))) - \vec{\mathfrak{p}}\right\| + (k-K)\varepsilon\|S\|. \tag{3.48}$$

Démonstration. Puisque $\mathbb{E}\,\vec{y}(k) \to \vec{\mathfrak{p}}$, pour tout $\varepsilon > 0$, il existe un entier $K > 0$ tel que pour tout $k \geq K$, $\|\mathbb{E}\,\vec{y}(k) - \vec{\mathfrak{p}}\| < \varepsilon$. En utilisant l'inégalité triangulaire et l'inégalité de la norme subordonnée on obtient le corollaire. □

Ainsi, la croissance du terme d'erreur au cours du temps est bornée par un nombre aussi petit qu'on le souhaite à partir d'un certain rang. Une autre manière de comprendre ce résultat est de se dire que l'on peut toujours borner le terme d'erreur par une fonction affine de pente aussi petite qu'on le souhaite.

Notion d'équilibre Le dernier corollaire permet de remarquer que le terme d'erreur provient surtout de l'écart entre $\mathbb{E}\,\vec{y}(k)$ et $\vec{\mathfrak{p}}$ aux premiers pas de temps. Cela invite naturellement à choisir les quantités initiales telles que cette différence soit petite. Pour cela on se propose de considérer la notion de *point d'équilibre*.

Définition 3.5 (Point d'équilibre, distance à l'équilibre)**.** Soit un vecteur de quantités de matière $\vec{x} \in \mathbb{N}^n$. On appellera *distance à l'équilibre* la quantité

$$d_e(\vec{x}) = \|\vec{p}(\vec{x}) - \vec{\mathfrak{p}}\|. \tag{3.49}$$

Lorsque $d_e(\vec{x}) = 0$, c'est-à-dire lorsque $\vec{p}(\vec{x}) = \vec{\mathfrak{p}}$, on dira que $\vec{x}$ est un *point d'équilibre*.

On notera qu'il n'existe pas toujours de point d'équilibre en raison des quantités de matière entières. Dans ce cas, on aura avantage à choisir des quantités de matières initiales ayant une faible distance à l'équilibre. On remarquera également que la notion de point d'équilibre est associée à la notion de point stationnaire dans les systèmes différentiels. Nous allons montrer qu'en choisissant comme point de départ un point d'équilibre, l'approximation de Bernoulli est exacte à la limite thermodynamique.

Limite thermodynamique Étudier un système à la limite thermodynamique consiste à faire tendre son volume à l'infini tout en préservant les concentrations de matière. Concrètement on se propose de définir une λ-*expansion* du système ($\lambda \in \mathbb{N}$), dans laquelle les quantités de matière initiales ainsi que le volume Ω sont multipliés par λ. Le tableau suivant résume quelques grandeurs impactées par cette λ-expansion.

Grandeur X	λ-expansion X^λ
c_j	$c_j^\lambda = c_j/\lambda^{\omega_j - 1}$
$\vec{x}(t)$	$\vec{x}^\lambda(t)$
$\vec{y}(k)$	$\vec{y}^\lambda(k)$
$\vec{z}(k)$	$\vec{z}^\lambda(k)$
$\vec{h}_j(\cdot)$	$\vec{h}_j^\lambda(\cdot) = \vec{h}_j(\cdot)/\lambda^{\omega_j - 1}$
$\vec{\mathfrak{p}}$	$\vec{\mathfrak{p}}^\lambda = \lim_{k\to+\infty} \vec{p}^\lambda(\vec{y}^\lambda(k))$

On note en particulier que les constantes stochastiques des réactions d'ordre ω sont divisées par $\lambda^{\omega-1}$ (voir le chapitre 2 pour la relation entre ces constantes de réactions stochastiques et le volume). Les processus $\vec{x}^\lambda, \vec{y}^\lambda, \vec{z}^\lambda$ correspondent aux dynamiques obtenues à partir de conditions initiales multipliées par λ et en utilisant les constantes stochastiques c_j^λ. Étudier le comportement à la limite thermodynamique revient alors à étudier le cas $\lambda \to +\infty$. Dans le cas qui nous concerne à présent, à savoir l'étude de l'espérance on a la proposition suivante

Proposition 3.7 (Limite thermodynamique, espérances)**.** *Soit $(\vec{y}(k))_{k\in\mathbb{N}}$ la dynamique discrétisée d'un système, $(\vec{y}^\lambda(k))_{k\in\mathbb{N}}$ la dynamique de sa λ-expansion, et soit $\vec{z}^\lambda(k)$ la dynamique de Bernoulli associée à $(\vec{y}^\lambda(k))_{k\in\mathbb{N}}$, on a*

$$\forall k \in \mathbb{N},\ \|\mathbb{E}\,\vec{y}^\lambda(k) - \mathbb{E}\,\vec{z}^\lambda(k)\| \le k\|S\| d_e^\lambda(\lambda\vec{y}_0) + o_{\lambda\to+\infty}(1). \tag{3.50}$$

Démonstration. On applique l'inégalité de norme subordonnée, l'inégalité triangulaire et l'inégalité de Jensen (la norme sup $\|\cdot\|$ étant une fonction convexe) à partir de la proposition 3.6

$$\|\mathbb{E}\,\vec{y}^\lambda(k) - \mathbb{E}\,\vec{z}^\lambda(k)\| = \left\| S\sum_{l=1}^{k}\left(\mathbb{E}(\vec{p}^\lambda(\vec{y}^\lambda(l-1)) - \vec{\mathfrak{p}}^\lambda\right)\right\| \tag{3.51}$$

$$\le \|S\| \sum_{l=1}^{k}\left(\mathbb{E}\left(\|\vec{p}^\lambda(\vec{y}^\lambda(l-1)) - \vec{\mathfrak{p}}^\lambda\|\right)\right). \tag{3.52}$$

On décompose ensuite l'expression

$$\|\vec{p}^\lambda(\vec{y}^\lambda(l-1)) - \vec{\mathfrak{p}}^\lambda\| = \|\vec{p}^\lambda(\vec{y}^\lambda(l-1)) - \vec{\mathfrak{p}}^\lambda\| \tag{3.53}$$

$$\le \|\vec{p}^\lambda(\vec{y}^\lambda(l-1)) - \vec{p}^\lambda(\vec{y}^\lambda(0))\| + \|\vec{p}^\lambda(\vec{y}^\lambda(0)) - \vec{\mathfrak{p}}^\lambda\| \tag{3.54}$$

$$\le \max_{\vec{w}\in B_{l-1}} \|\vec{p}^\lambda(\lambda\vec{y}(0) + \vec{w}) - \vec{p}^\lambda(\lambda\vec{y}(0))\| + d_e^\lambda(\vec{y}^\lambda(0)), \tag{3.55}$$

où B_{l-1} est l'ensemble *fini* de déplacements qu'on peut obtenir en effectuant $l-1$ réactions. Il s'en suit d'après le lemme 3.4 que chacun de ces majorants qui sont en nombre fini tendent vers 0 quand $\lambda \to +\infty$. On a donc

$$\|\vec{p}^{\lambda}(\vec{y}^{\lambda}(l-1)) - \vec{\mathfrak{p}}^{\lambda}\| \leq o_{\lambda\to+\infty}(1) + d_e^{\lambda}(\vec{y}^{\lambda}(0)). \tag{3.56}$$

En utilisant cette majoration on obtient donc

$$\|\mathbb{E}\,\vec{y}^{\lambda}(k) - \mathbb{E}\,\vec{z}^{\lambda}(k)\| \leq o_{\lambda\to+\infty}(1) + k\|S\| d_e^{\lambda}(\vec{y}^{\lambda}(0)). \tag{3.57}$$

□

La démonstration effectuée repose donc sur les deux lemmes suivants.

Lemme 3.3. *Soit (n, m, α, β) un réseau de réactions, $(c_j)_{j=1,\dots,m}$ les constantes stochastiques de réactions, on considère un état $\vec{x} \in \mathbb{N}^n$, un vecteur de $\vec{w} \in \mathbb{Z}^n$ un déplacement fixe et $\lambda \in \mathbb{N}$ un facteur d'expansion alors à la limite thermodynamique, on a les équivalents suivants*

$$\forall j = 1,\dots,m, \quad h_j^{\lambda}(\lambda\vec{x}+\vec{w}) \underset{\lambda\to+\infty}{\sim} \lambda c_j \prod_{i=1}^{n} \frac{x_i^{\alpha_{i,j}}}{\alpha_{i,j}!},$$
$$h_0^{\lambda}(\lambda\vec{x}+\vec{w}) \underset{\lambda\to+\infty}{\sim} \lambda \left(\sum_{j=1}^{m} c_j \prod_{i=1}^{n} \frac{x_i^{\alpha_{i,j}}}{\alpha_{i,j}!} \right).$$

Démonstration.

$$h_j^{\lambda}(\lambda\vec{x}+\vec{w}) = \frac{c_j}{\lambda^{\omega_j - 1}} \prod_{i=1}^{n} \binom{\lambda x_i + w_i}{\alpha_{i,j}}, \tag{3.58}$$

or $\binom{\lambda x_i + w_i}{\alpha_{i,j}}$ est un polynôme en λ donc asymptotiquement équivalent à son monôme dominant $(\lambda x_i)^{\alpha_{i,j}}/\alpha_{i,j}!$. En multipliant les équivalents on obtient alors le premier résultat puisque $\omega_j = \sum_{i=1}^{n} \alpha_{i,j}$. On remarque aussi que pour tout j, $h_j^{\lambda}(\lambda\vec{x}+\vec{w})$ est un polynôme de degré 1 en λ dont on vient de déterminer le monôme dominant, $h_0^{\lambda}(\lambda\vec{x}+\vec{w})$ est donc aussi un polynôme de degré 1 (les coefficients étant positifs on ne peut pas obtenir un polynôme constant) dont le monôme dominant est la somme des monômes dominants de $h_j^{\lambda}(\lambda\vec{x}+\vec{w})$ d'où le second équivalent. □

Lemme 3.4.

$$p_j(\lambda\vec{x}+\vec{w}) - p_j(\lambda\vec{x}) \to 0. \tag{3.59}$$

Démonstration. D'après le lemme 3.3, on sait que $\lim_{\lambda\to\infty} h_0^{\lambda}(\lambda\vec{x}+\vec{w}) = +\infty$ quelques que soient $\vec{x}$ et $\vec{w}$. Ainsi, $p_0^{\lambda}(\lambda\vec{x}+\vec{w}) = \exp(-\delta t h_0^{\lambda}(\lambda\vec{x}+\vec{w}))$ tend vers 0 ainsi que $p_0^{\lambda}(\lambda\vec{x})$ en prenant $\vec{w} = \vec{0}$. On a donc bien le résultat pour $j = 0$. Pour $j = 1,\dots,m$, on a

$$\frac{p_j^{\lambda}(\lambda\vec{x}+\vec{w})}{p_j^{\lambda}(\lambda\vec{x})} = \frac{(1 - p_0^{\lambda}(\lambda\vec{x}+\vec{w})) \times h_j^{\lambda}(\lambda\vec{x}+\vec{w}) \times h_0^{\lambda}(\lambda\vec{x})}{(1 - p_0^{\lambda}(\lambda\vec{x})) \times h_j^{\lambda}(\lambda\vec{x}) \times h_0^{\lambda}(\lambda\vec{x}+\vec{w})} \tag{3.60}$$

lorsque $p_j(\lambda\vec{x}) \neq 0$ (sinon prendre l'expression inverse pour les j concernés, et si $p_j(\lambda\vec{x}+\vec{w}) = 0$ également alors $p_j(\lambda\vec{x}+\vec{w}) - p_j(\lambda\vec{x}) = 0$). D'après le début de la preuve, $(1-p_0^\lambda(\lambda\vec{x}+\vec{w})) \to 1$ et $(1-p_0^\lambda(\lambda\vec{x})) \to 1$. D'après le lemme 3.3 on a également $h_j^\lambda(\lambda\vec{x}+\vec{w})/h_j^\lambda(\lambda\vec{x}) \to 1$ ainsi que $h_0^\lambda(\lambda\vec{x})/h_0^\lambda(\lambda\vec{x}+\vec{w}) \to 1$ car numérateurs et dénominateurs ont mêmes équivalents asymptotiques. On a donc $\frac{p_j^\lambda(\lambda\vec{x}+\vec{w})}{p_j^\lambda(\lambda\vec{x})} \to 1$ ce qui démontre le lemme. □

3.3.3 Comparaison des matrices de covariances

Proposition 3.8. *La covariance de $\vec{y}_k$ se décompose en 3 termes : la covariance du modèle équivalent de Bernoulli, un terme de "maintien d'équilibre" et un terme d'erreur initiale qui converge (si la convergence stationnaire est assez rapide) quand $k \to +\infty$.*

$$\mathbb{C}\mathrm{ov}\,\vec{y}(k) = \mathbb{C}\mathrm{ov}\,\vec{z}(k) + \sum_{l=0}^{k-1}(\mathbb{C}\mathrm{ov}(\vec{y}(l),\vec{p}(\vec{y}(l)))S^t + S\,\mathbb{C}\mathrm{ov}(\vec{p}(\vec{y}(l)),\vec{y}(l)))$$
$$+\sum_{l=0}^{k-1} S\left[\mathrm{diag}(\mathbb{E}\,\vec{p}(\vec{y}(l)) - \vec{\mathfrak{p}}) - (\mathbb{E}\,\vec{p}(\vec{y}(l))\,\mathbb{E}\,\vec{p}(\vec{y}(l))^t - \vec{\mathfrak{p}}\vec{\mathfrak{p}}^t)\right]S^t \quad (3.61)$$

Démonstration. Pour tout couple de vecteurs aléatoire $(\vec{x},\vec{y})$, on note $\mathbb{C}\mathrm{ov}(\vec{x},\vec{y}) = \mathbb{E}(\vec{x}\vec{y}^t) - \mathbb{E}\,\vec{x}\,\mathbb{E}\,\vec{y}^t$ et $\mathbb{C}\mathrm{ov}(\vec{x}) = \mathbb{C}\mathrm{ov}(\vec{x},\vec{x})$. On rappelle que la covariance est invariante par translation c'est à dire que pour tout couple de vecteurs non aléatoires $(\vec{a},\vec{b})$, $\mathbb{C}\mathrm{ov}(\vec{x}+\vec{a},\vec{y}+\vec{b}) = \mathbb{C}\mathrm{ov}(\vec{x},\vec{y})$. Pour simplifier les calculs, on supposera donc dans cette démonstration que $\vec{y}(0) = 0$ sans nuire à la généralité. Nous commençons par exprimer $\mathbb{C}\mathrm{ov}(\vec{y}(k+1))$ en fonction de $\mathbb{C}\mathrm{ov}(\vec{y}(k))$ à partir de la relation

$$\vec{y}(k+1) = \vec{y}(k) + S\vec{e}_{\mu_{k+1}} \quad (3.62)$$

où μ_{k+1} est l'index de la réaction choisie au temps discret $k+1$ dans l'algorithme discrétisé de Gillespie. On se servira de la propriété $\mathbb{E}(\vec{e}_{\mu_{k+1}}|\vec{y}(k)) = \vec{p}(\vec{y}(k))$. On a d'une part

$$\mathbb{E}(\vec{y}(k+1)\vec{y}(k+1)^t|\vec{y}(k)) = \vec{y}(k)\vec{y}(k)^t + \vec{y}(k)\,\mathbb{E}((S\vec{e}_{\mu_{k+1}})^t|\vec{y}(k)) \quad (3.63)$$
$$+\mathbb{E}(S\vec{e}_{\mu_{k+1}}|\vec{y}(k))\vec{y}(k)^t + \mathbb{E}(S\vec{e}_{\mu_{k+1}}(S\vec{e}_{\mu_{k+1}})^t|\vec{y}(k)) \quad (3.64)$$
$$= \vec{y}(k)\vec{y}(k)^t + \vec{y}(k)\vec{p}(\vec{y}(k))^t S^t \quad (3.65)$$
$$+ S\vec{p}(\vec{y}(k))\vec{y}(k)^t + S\,\mathrm{diag}\,\vec{p}(\vec{y}(k))S^t, \quad (3.66)$$

d'où

$$\mathbb{E}(\vec{y}(k+1)\vec{y}(k+1)^t) = \mathbb{E}\,\mathbb{E}(\vec{y}(k+1)\vec{y}(k+1)^t|\vec{y}(k)) \quad (3.67)$$
$$= \mathbb{E}(\vec{y}(k)\vec{y}(k)^t) + \mathbb{E}(\vec{y}(k)\vec{p}(\vec{y}(k))^t)S^t \quad (3.68)$$
$$+ S\,\mathbb{E}(\vec{p}(\vec{y}(k))\vec{y}(k)^t) + S\,\mathrm{diag}\,\mathbb{E}\,\vec{p}(\vec{y}(k))S^t. \quad (3.69)$$

D'autre part

$$\mathbb{E}\,\vec{y}(k+1) = \mathbb{E}\,\mathbb{E}(\vec{y}(k) + S\vec{e}_{\mu_{k+1}}|\vec{y}(k)) = \mathbb{E}\,y(k) + S\,\mathbb{E}\,\vec{p}(\vec{y}(k)), \quad (3.70)$$

ce qui permet d'obtenir

$$\mathbb{E}\,\vec{y}(k+1)\,\mathbb{E}\,\vec{y}(k+1)^t = \mathbb{E}\,\vec{y}(k)\,\mathbb{E}\,\vec{y}(k)^t + \mathbb{E}\,\vec{y}(k)\,\mathbb{E}\,\vec{p}(\vec{y}(k))^t S^t \tag{3.71}$$

$$+ S\,\mathbb{E}\,\vec{p}(\vec{y}(k))\,\mathbb{E}\,y(k)^t + S\,\mathbb{E}\,\vec{p}(\vec{y}(k))\,\mathbb{E}\,\vec{p}(\vec{y}(k))^t S^t. \tag{3.72}$$

En soustrayant les deux égalités on obtient finalement

$$\mathbb{C}\mathrm{ov}(\vec{y}(k+1)) = \mathbb{C}\mathrm{ov}(\vec{y}(k)) + \left[\mathbb{E}(\vec{y}(k)\vec{p}(\vec{y}(k))^t) - \mathbb{E}\,\vec{y}(k)\,\mathbb{E}\,\vec{p}(\vec{y}(k))^t\right] S^t \tag{3.73}$$

$$+ S\left[\mathbb{E}(\vec{p}(\vec{y}(k))\vec{y}(k)^t) - \mathbb{E}\,\vec{p}(\vec{y}(k))\,\mathbb{E}\,y(k)^t\right] \tag{3.74}$$

$$+ S\left[\mathrm{diag}\,\mathbb{E}\,\vec{p}(\vec{y}(k)) - \mathbb{E}\,\vec{p}(\vec{y}(k))\,\mathbb{E}\,\vec{p}(\vec{y}(k))^t\right] S, \tag{3.75}$$

$$= \mathbb{C}\mathrm{ov}(\vec{y}(k)) + \mathbb{C}\mathrm{ov}(\vec{y}(k), \vec{p}(\vec{y}(k)))S^t \tag{3.76}$$

$$+ S\,\mathbb{C}\mathrm{ov}(\vec{p}(\vec{y}(k)), \vec{y}(k)) \tag{3.77}$$

$$+ S\left[\mathrm{diag}\,\mathbb{E}\,\vec{p}(\vec{y}(k)) - \mathbb{E}\,\vec{p}(\vec{y}(k))\,\mathbb{E}\,\vec{p}(\vec{y}(k))^t\right] S^t. \tag{3.78}$$

Ainsi en sommant ces expressions on obtient

$$\mathbb{C}\mathrm{ov}(\vec{y}(k)) = \sum_{l=0}^{k-1} \Big(\mathbb{C}\mathrm{ov}(\vec{y}(l), \vec{p}(\vec{y}(l)))S^t \tag{3.79}$$

$$+ S\,\mathbb{C}\mathrm{ov}(\vec{p}(\vec{y}(l)), \vec{y}(l)) \tag{3.80}$$

$$+ S\left[\mathrm{diag}\,\mathbb{E}\,\vec{p}(\vec{y}(l)) - \mathbb{E}\,\vec{p}(\vec{y}(l))\,\mathbb{E}\,\vec{p}(\vec{y}(l))^t\right] S^t\Big). \tag{3.81}$$

Par ailleurs on avait montré dans la proposition 3.2 que

$$\mathbb{C}\mathrm{ov}\,\vec{z}(k) = kS(\mathrm{diag}\,\vec{\mathfrak{p}} - \vec{\mathfrak{p}}\vec{\mathfrak{p}}^t)S^t. \tag{3.82}$$

Par comparaison, on obtient bien le résultat voulu. □

Proposition 3.9 (Limite thermodynamique, covariances, variances)**.** *Soit* $(\vec{y}(k))_{k\in\mathbb{N}}$ *la dynamique discrétisée d'un système,* $(\vec{y}^\lambda(k))_{k\in\mathbb{N}}$ *la dynamique de sa* λ*-expansion, et soit* $\vec{z}^\lambda(k)$ *la dynamique de Bernoulli associée à* $(\vec{y}^\lambda(k))_{k\in\mathbb{N}}$*, on a*

$$\forall k \in \mathbb{N},\ \|\mathbb{C}\mathrm{ov}\,\vec{y}^\lambda(k) - \mathbb{C}\mathrm{ov}\,\vec{z}^\lambda(k)\| \le k(n+1)\|S\|\|S^t\|d_e^\lambda(\lambda\vec{y}_0) + o_{\lambda\to+\infty}(1). \tag{3.83}$$

Démonstration. En utilisant les mêmes arguments que dans la preuve de la proposition 3.6 que

$$\|\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{p}^\lambda(\vec{y}^\lambda(0))\| \le \max_{\vec{w}\in B_l}\|\vec{p}^\lambda(\lambda\vec{y}(0) + \vec{w}) - \vec{p}^\lambda(\lambda\vec{y}(0))\| \to 0 \tag{3.84}$$

où B_l représente tous les déplacements possibles en effectuant l réactions et où la limite vaut quand $\lambda \to +\infty$ en vertu du lemme 3.4. Cette remarque va nous permettre d'obtenir une majoration en partant de la proposition 3.8 :

$$\mathbb{C}\mathrm{ov}\,\vec{y}^\lambda(k) = \mathbb{C}\mathrm{ov}\,\vec{z}^\lambda(k) + \sum_{l=0}^{k-1}(\mathbb{C}\mathrm{ov}(\vec{y}^\lambda(l), \vec{p}^\lambda(\vec{y}^\lambda(l)))S^t + S\,\mathbb{C}\mathrm{ov}(\vec{p}^\lambda(\vec{y}^\lambda(l)), \vec{y}^\lambda(l)))$$

$$+ \sum_{l=0}^{k-1} S\left[\mathrm{diag}(\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{\mathfrak{p}}^\lambda) - (\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l))\,\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l))^t - \vec{\mathfrak{p}}^\lambda\vec{\mathfrak{p}}^{\lambda^t})\right] S^t. \tag{3.85}$$

En utilisant l'invariance par translation de la covariance et la remarque initiale, on a

$$\mathbb{C}\mathrm{ov}(\vec{y}^\lambda(l), \vec{p}^\lambda(\vec{y}^\lambda(l)) = \mathbb{C}\mathrm{ov}(\vec{y}^\lambda(l), \vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{p}^\lambda(\vec{y}^\lambda(0))) \underset{\lambda\to+\infty}{\to} 0, \tag{3.86}$$

montrant que la première somme est un $o(1)$ quand $\lambda \to +\infty$. Pour la seconde somme on remarque que

$$\|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{\mathfrak{p}}^\lambda\| \leq \|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{p}^\lambda(\vec{y}^\lambda(0))\| + \|\vec{p}^\lambda(\vec{y}^\lambda(0)) - \vec{\mathfrak{p}}^\lambda\| \tag{3.87}$$

$$\leq o_{\lambda\to+\infty}(1) + d_e^\lambda(\lambda\vec{y}(0)), \tag{3.88}$$

Le $o_{\lambda\to+\infty}(1)$ étant encore une fois conséquence de la remarque initiale. On a donc dans la seconde somme

$$\|\operatorname{diag}(\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{\mathfrak{p}}^\lambda)\| = \|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{\mathfrak{p}}^\lambda\| \leq o_{\lambda\to+\infty}(1) + d_e^\lambda(\lambda\vec{y}(0)) \tag{3.89}$$

et également, en utilisant l'identité $aa' - bb' = (a-b)a' + b(a'-b')$

$$\|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l))\,\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l))^t - \vec{\mathfrak{p}}^\lambda\vec{\mathfrak{p}}^{\lambda^t}\| \tag{3.90}$$

$$\leq \|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l)) - \vec{\mathfrak{p}}^\lambda\| \underbrace{\|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l))^t\|}_{\leq n} + \underbrace{\|\vec{\mathfrak{p}}^\lambda\|}_{\leq 1} \|\mathbb{E}\,\vec{p}^\lambda(\vec{y}^\lambda(l))^t - \vec{\mathfrak{p}}^{\lambda^t}\| \tag{3.91}$$

$$\leq o_{\lambda\to+\infty}(1) + (n+1)d_e^\lambda(\lambda\vec{y}(0)). \tag{3.92}$$

En appliquant les inégalités triangulaires, de normes d'algèbres et les inégalités ci-dessus à l'équation (3.85), on obtient le résultat cherché.

□

3.4 Conclusion

Dans les chapitres 1 et 2, nous avons dressé l'état de l'art de la modélisation dynamique des réseaux de réactions en adoptant le point de vue de la réfutation de réseaux à l'aide de systèmes de contraintes. Nous avons montré que cette approche est possible dans la dynamique différentielle mais qu'elle ne permet d'utiliser pour données que les pentes moyennes des trajectoires. Or, les techniques expérimentales fournissent de plus en plus de données à l'échelle de l'individu et l'on souhaiterait pouvoir exploiter les mesures de variances, covariances, etc. C'est pour cette raison que nous nous sommes intéressés à la sémantique stochastique. Cependant, on a montré dans le chapitre 2 que les méthodes existantes, les simulations de Monte-Carlo, l'approximation des moments par méthode des moments clos dont l'approximation de bruit linéaire ne permettent pas de fournir un analogue du vecteur de flux stationnaire qui puisse analytiquement être lié à des données sur les moments d'ordre 2.

Dans ce chapitre, nous avons donc proposé une nouvelle sémantique stochastique appelée *dynamique de Bernoulli*, obtenue lorsqu'on rend indépendant le tirage des réactions

dans la sémantique stochastique classique et en choisissant pour probabilités de tirages les probabilités de réactions stationnaires, c'est-à-dire associées à la distribution stationnaire du système. Cette dynamique présente de nombreux avantages. Premièrement, elle fait déjà apparaître clairement un vecteur $\vec{\mathfrak{p}}$ qui mesure les probabilités de tirages stationnaires des réactions et qui est tout à fait analogue à $\vec{f}$ mesurant les taux d'utilisation des réactions en sémantique déterministe. Deuxièmement, cette nouvelle dynamique est plus simple à étudier. Nous avons établi dans ce chapitre des expressions analytiques pour les moments d'ordre 1 et 2 de cette dynamique, en fonction du vecteur $\vec{\mathfrak{p}}$ et de la stœchiométrie. Cela correspond en différentiel au lien entre les pentes moyennes et les flux stationnaires $\vec{f}$. Cette relation analytique ouvre la voie aux méthodes par contraintes stochastiques développées dans le chapitre suivant et qui constituent un thème important de cette thèse. L'analyse asymptotique de la dynamique de Bernoulli est également effectuée : nous obtenons un théorème central limite et les paramètres de la loi limite sont analytiquement déterminés. Cela est parfaitement compatible avec les phénomènes de diffusion qu'on rencontre en sémantique stochastique classique (équation de diffusion de Langevin, approximation de bruit linéaire, théorème de Kurtz *etc*).

Enfin, une partie importante de ce chapitre est l'étude de la qualité de l'approximation proposée. On a donné deux expressions pour les termes d'erreurs de l'approximation en ce qui concerne la variance et l'espérance. En particulier pour l'espérance, nous avons démontré que l'incrément d'erreur à chaque tirage de réaction tend vers 0. De manière générale pour les moments d'ordre 1 et 2, nous avons conclu que si les systèmes sont proches de la limite thermodynamique avec un point initial proche de l'équilibre alors ces termes d'erreurs étaient négligeables. Ainsi, si on sait que plus on étudie un système près de la limite thermodynamique, plus la part de stochasticité dans les trajectoires est faible mais on sait aussi maintenant que cette stochasticité est de plus en plus proche de la dynamique de Bernoulli. Notre approximation est donc valable lorsque les réseaux se situent dans un juste milieu où les effets stochastiques ne sont pas négligeables (pas de trop grandes quantités de molécules) mais dans lequel il y a assez de molécules pour que la dynamique de Bernoulli soit valable. Dans la suite de la thèse, on supposera que cette approximation est valable.

Chapitre 4

Applications à la validation de modèles

Les applications présentées dans ce chapitre ont été publiées dans les actes du workshop international SASB de la conférence SAS2014 [PSB14].

Dans ce chapitre, nous montrons comment la dynamique de Bernoulli introduite dans le chapitre précédent, c'est-à-dire l'approximation de la dynamique stochastique discrète, permet de confronter des réseaux de réactions et des données. L'objectif principal recherché dans ce chapitre est la possibilité de *réfuter* un réseau de réactions à partir de données sous forme de séries temporelles des quantités de matières.

Le résultat principal de ce chapitre est l'introduction d'un *système de contraintes* associées aux données statistiques (moyennes, variances et covariances) des données à disposition et qui se caractérise par les aspects suivants.

— Une approche *stœchiométrique* dans laquelle des contraintes peuvent être obtenues sans *aucune information sur le type de lois* (action de masses, Michaelis-Menten, ...) régissant la dynamique du réseau et en particulier sans *aucune information sur la valeur des paramètres*, des constantes cinétiques. Elle est donc à la fois très générale et adaptée à l'étude de systèmes pour lesquels on possède peu d'informations cinétiques.

— Une extension naturelle des approches de contraintes de flux (*e.g.* le *Flux Balance Analysis*) à la dynamique stochastique permettant de prendre en compte les moments d'ordre 1 (espérances) et d'ordre 2 (variances et covariances). Elle en est une généralisation dans le sens où les contraintes sur les probabilités de réactions associées aux espérances correspondent en dynamique différentielle aux contraintes associées aux flux.

L'application principale de cette approche par contrainte est la *réfutation de modèle* qui s'obtient lorsque le système de contraintes obtenues à partir des informations statistiques expérimentales ne possède pas de solution. Une manière de lire ces résultats est de les voir comme le dual des résultats du chapitre précédent dans lequel des expressions précises pour les moments d'ordre 1 et 2 de la dynamique de Bernoulli ont été obtenues à partir de la

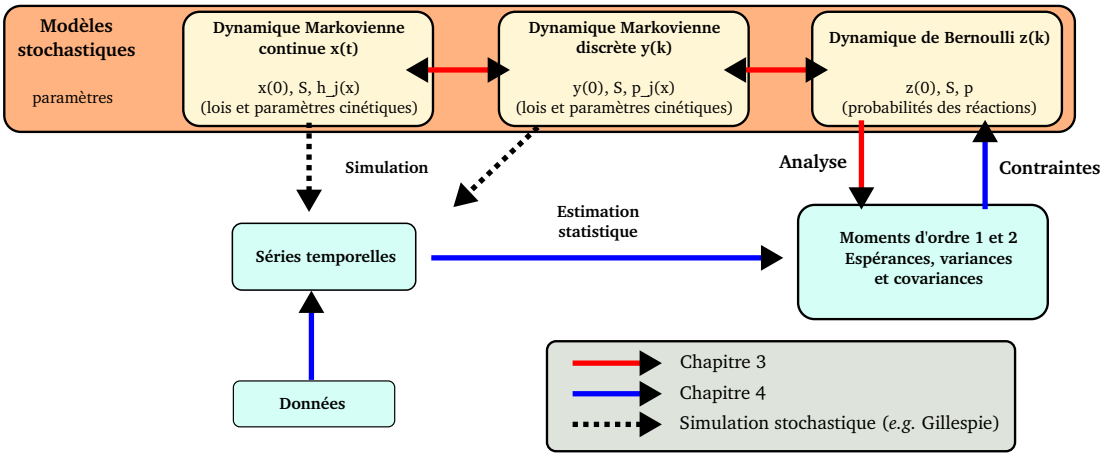


FIGURE 4.1 – **Diagramme résumant les principales approches décrites dans les chapitres 3 et 4**

connaissance de la stœchiométrie et des probabilités de réactions. De manière réciproque dans ce chapitre (Figure 4.1), on part d'informations sur les moments d'ordre 1 et 2 pour obtenir de l'information sur les probabilités de réactions.

Dans les exemples traités dans ce chapitre, on considèrera pour simplifier et sans nuire à la généralité que $\vec{p}_0 = 0$, c'est-à-dire qu'on se considère dans un état proche de la limite thermodynamique. Dans le cas général, le vecteur $\vec{p}_0$ est également une variable qui sera contrainte par les données à disposition.

4.1 Méthodes par contraintes pour les moments d'ordre 1 et 2

Nous commençons par introduire une correspondance entre des mesures sur les moments d'ordre 1 et 2 et des contraintes sur le vecteur de probabilités de réactions stationnaire $\vec{p}$. On présentera ensuite quelques exemples d'applications.

4.1.1 Table des contraintes

Soit $\vec{y}(l,k)$ $(1 \leq l \leq N, k \in \mathbb{N})$ une série de N trajectoires représentant les données expérimentales, $\vec{m}(k) = \frac{1}{N}\sum_{l=1}^{N} \vec{y}(l,k)$ la moyenne empirique, $c_{a,b}(k) = \frac{1}{N}\sum_{l=1}^{N}(y_a(l,k)-m_a(k))(y_b(l,k)-m_b(k))$ les variances-covariances empiriques.

Proposition 4.1. *Suppose que les conditions de validité de l'approximation de Bernoulli sont réalisées, c'est-à-dire que* $(\vec{y}(l,k))$ *sont des réalisations de la dynamique de Bernoulli de conditions initiales* $\vec{y}(0)$*, de matrice de stœchiométrie* S *et de probabilités de réactions* $\vec{p}$*. Alors les hypothèses décrites dans le tableau ci-dessous sur les* $\vec{m}(k)$ *et les* $c_{a,b}(k)$ *lorsque* $N \to +\infty$ *impliquent les contraintes sur* $\vec{p}$ *associées.*

	Observation	Contrainte (forme matricielle)	Contrainte (forme algébrique)	Type
(1a)	$\forall k,\ m_a(k) = y_a(0)$	$(S\vec{p})_a = 0$	$\sum_{i=1}^m s_{ai} p_i = 0$	linéaire
(1b)	$m_a(k) \leq y_a(0) + k\gamma$	$(S\vec{p})_a \leq \gamma$	$\sum_{i=1}^m s_{ai} p_i \leq \gamma$	linéaire
(1c)	$y_a(0) + k\gamma \leq m_a(k)$	$\gamma \leq (S^\top \vec{p})_a$	$\gamma \leq \sum_{i=1}^m s_{ai} p_i$	linéaire
(2a)	$c_{a,a}(k) \leq k\gamma$	$\left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{aa} \leq \gamma$	$\sum_{i=1}^m s_{ai}^2 p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{aj} p_i p_j \leq \gamma$	quadratique
(2b)	$k\gamma \leq c_{a,a}(k)$	$\gamma \leq \left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{aa}$	$\gamma \leq \sum_{i=1}^m s_{ai}^2 p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{aj} p_i p_j$	quadratique
(3a)	$c_{a,b}(k) \leq k\gamma$	$\left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{ab} \leq \gamma$	$\sum_{i=1}^m s_{ai} s_{bi} p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{bj} p_i p_j \leq \gamma$	quadratique
(3b)	$k\gamma \leq c_{a,b}(k)$	$\gamma \leq \left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{ab}$	$\gamma \leq \sum_{i=1}^m s_{ai} s_{bi} p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{bj} p_i p_j$	quadratique

TABLE 4.1 – Table de correspondance entre les mesures empiriques d'espérances et (co)variances et les contraintes linéaires et quadratiques associées sur les probabilités de réactions $\vec{p}$

Démonstration. On suppose que la dynamique de Bernoulli est valide c'est-à-dire que $\vec{y}(l,k)$ sont les réalisations de la dynamique de Bernoulli qu'on notera $\vec{y}(k)$ d'état initial $\vec{y}(0)$ de matrice de stœchiométrie S et de probabilités de réactions $\vec{p}$. Lorsque $N \to +\infty$, les estimateurs $\vec{m}(k)$ et $c_{a,b}(k)$ convergent respectivement vers $\mathbb{E}\,\vec{y}(k)$ et $(\mathbb{C}\mathrm{ov}\,\vec{y}(k))_{a,b}$. D'après les expressions de l'espérance et de la matrice de covariance de la dynamique de Bernoulli obtenus dans la proposition 3.2, les observations (1) (2) and (3) de la Table 4.1 impliquent bien les contraintes matricielles associées. Les contraintes algébriques sont une version scalaire de ce même résultat obtenues en calculant les coefficients correspondant dans la version matricielle. □

Analogie avec les méthodes d'équilibre de flux Nous soulignons immédiatement que ce tableau de contraintes peut-être lu comme une extension probabiliste des contraintes des méthodes d'équilibre de flux telles que le FBA. Dans ces approches, les flux $\vec{f}$ quantifient les taux d'utilisation de chaque réaction en état stationnaire, c'est-à-dire dans ce cas lorsque la pente de la trajectoire est constante. Cette stationnarité implique nécessairement un équilibre des flux dans lequel l'équilibre des réactifs du réseau est maintenu. Ainsi, les méthodes d'équilibre des flux reposent sur une *approche par contraintes* dans laquelle le vecteur de flux $\vec{f}$ est contraint dans un *cône de stationnarité*

$$\mathscr{C}\mathscr{F} = \left\{\vec{f} \in \mathbb{R}^m \mid (S^*)\vec{f} = \vec{0}\right\}, \tag{4.1}$$

où S^* est obtenue en ne conservant que les lignes des réactifs dans la matrice de stœchiométrie S. Dans notre approche, on peut aussi utiliser une contrainte d'espèces de réactifs constantes *moyenne* (1a). Cette contrainte, si elle est appliquée à chacun des réactifs du réseau implique également que les probabilités de réactions appartiennent au cône de stationnarité $\vec{p} \in \mathscr{C}\mathscr{F}$. De la même manière, les contraintes linéaires du tableau 4.1 ont un analogue différentiel correspondant à une contrainte du flux $\vec{f}$ obtenu en mesurant une pente sur les données. Ainsi les contraintes de flux obtenues à partir d'informations de pentes se retrouvent entièrement dans notre tableau sous forme de contraintes obtenues à partir des pentes moyennes des trajectoires. Cela est entièrement cohérent avec l'idée que les équations différentielles sont utilisées pour étudier le comportement moyen des réseaux.

Notre approche stochastique est originale car elle propose des contraintes supplémentaires obtenues grâce à des estimations des variances et des covariances. Toutefois, on note une différence théorique importante car les contraintes associées à ces moments d'ordre 2 sont *quadratiques* ce qui conduit un espace de solutions beaucoup plus complexe, non nécessairement connexe (et *a fortiori* non nécessairement convexe).

4.1.2 Effet d'un bruit blanc expérimental

Lorsqu'on considère des données expérimentales, celles-ci sont souvent bruitées par les erreurs de mesures. Dans cette sous-section, on étudie l'impact de ce bruit sur les estimations des espérances, variances et co-variances. Nous nous reposons sur une modélisation simple de ces erreurs en les assimilant à un bruit blanc indépendant.

Définition 4.1 (Bruit blanc indépendant)**.** On appelle *bruit blanc indépendant* de dimension n et de matrice de covariance C, un processus $(\vec{\psi}(t))_{t\geq 0}$ à valeurs dans $\mathbb{R}^n$ vérifiant

1. $\forall t \geq 0 \ , \mathbb{E}\psi(t) = \vec{0}$,
2. $\forall t \geq 0 \ , \mathbb{C}\mathrm{ov}\,\psi(t) = C$,
3. Pour tout $t, t' \geq 0$ si $t \neq t'$ alors $\psi(t)$ et $\psi(t')$ sont indépendants.

Cette définition est valable pour des temps discrets ou continus. Intuitivement, la matrice C représente la grandeur de l'erreur expérimentale. Les termes diagonaux de C représente l'erreur de la mesure sur chaque espèce tandis que les autres termes représentent les corrélation possibles lors des mesures (par exemple si le fait d'avoir une mesure sur-évaluée entraine la sur-évaluation d'une autre). Dans le cas le plus simple, on peut considérer que la matrice C est diagonale.

L'influence de ce signal de bruit sur les mesures d'espérances et de (co)variances est décrite dans le lemme suivant.

Lemme 4.1. *Soit $(\vec{x}(t))_t$ une trajectoire et $(\psi(t))_t$ un bruit blanc indépendant de dimension n et de matrice de covariance C et indépendant de $(\vec{x}(t))$ alors :*

$$\mathbb{E}(\vec{x}(t) + \vec{\psi}(t)) = \mathbb{E}(\vec{x}(t)), \tag{4.2}$$

$$\mathrm{Var}(x_a(t) + \psi_a(t)) = \mathrm{Var}(x_a(t)) + C_{a,a} \tag{4.3}$$

$$\mathbb{C}\mathrm{ov}(x_a(t) + \psi_a(t), x_b(t) + \psi_b(t)) = \mathbb{C}\mathrm{ov}(x_a(t), x_b(t)) + C_{a,b}. \tag{4.4}$$

Concrètement, si on connait la valeur de C ou une estimation, on peut en déduire une estimation du moment correspondant qui peut être utilisée pour déduire des contraintes à l'aide du tableau ci-dessus.

4.1.3 Exemples

Nous présentons maintenant quelques illustrations de systèmes de contraintes obtenus à l'aide de mesures de moments d'ordre 2. En particulier nous allons traiter le cas de

l'exemple introductif de la thèse.

4.1.3.1 Illustration de l'intérêt des contraintes de moments d'ordre 2

Considérons une nouvelle fois l'exemple (3) présenté dans le chapitre 1.

$$(\text{modèle } 1) \begin{cases} R_1 : & A & \rightarrow & B+C \\ R_2 : & B & \rightarrow & A+2D \end{cases} \qquad (\text{modèle } 2) \begin{cases} R_1 : & A & \rightarrow & B+C+2D \\ R_2 : & B & \rightarrow & A \end{cases}$$

On souhaite pouvoir discriminer ces deux réseaux de réactions en se basant sur plusieurs séries temporelles mesurant les quantités d'espèces C et D. Supposons que ces données soient celles représentées sur la figure 4.2. À partir de ces séries temporelles, il est possible comme nous l'avons vu d'estimer les espérances, variances et covariances. Ces estimations sont représentées sur la même figure 4.2. À partir de ces estimations nous allons comparer l'approche classique reposant sur les flux et l'utilisation de notre système de contraintes.

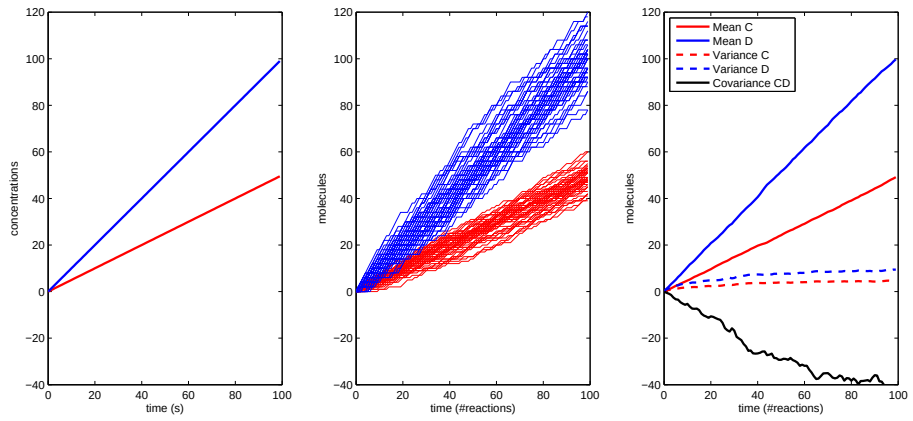


FIGURE 4.2 – **Donnnées de départ pour discriminer les modèles 1 ou 2**. Les données ont pour forme un ensemble de trajectoires dont on a estimé les moments d'ordre 1 et 2.

Méthode classique : contraindre les flux L'approche différentielle classique reposant sur les flux ne peut pas utiliser directement plusieurs trajectoires différentes puisque le système différentiel est sensé être déterministe et ne fournir qu'une seule trajectoire solution. Le seul moyen d'utiliser plusieurs séries temporelles est de considérer que le système différentiel représente le comportement moyen du système et utiliser alors en tant que donnée de base la moyenne empirique des trajectoires. À partir des moyennes empiriques, on peut remarquer que le taux de production de l'espèce D est environ le double du taux de production de l'espèce C. On rappelle alors que l'équation $d\vec{x}/dt = S\vec{f}$ permet de lier les pentes, les flux et les coefficients stœchiométriques. Ainsi on obtient les contraintes suivantes sur $\vec{f}$ pour chacune des deux deux hypothèses.

Modèle 1	Modèle 2	Contrainte
$f_1 = f_2$	$f_1 = f_2$	A et B sont équilibrés
$f_1 = f_2$	$f_1 = f_1$	$2d[C]/dt = d[D]/dt$.

Si un de ces systèmes d'équations ne possède pas de solution alors le réseau de réactions correspondant peut être réfuté. Cependant, on observe pour chacun des systèmes que le vecteur de flux $\vec{f} = (1,1)^t$ satisfait à la fois l'équilibre des réactifs A and B et le rapport double de production moyen de D par rapport à C. Il est même possible de proposer des valeurs d'équilibre pour les concentrations ($[A]_0 = [B]_0 = 1$) ainsi que des constantes cinétiques ($k_1 = k_2 = 1$) qui conduisent à l'équilibre dans lequel D est deux fois plus produit que C. Ainsi ces deux systèmes ne peuvent être distinguées uniquement à l'aide de contraintes de flux reposant sur des mesures de pentes moyennes. Ce résultat était attendu puisque la figure 1 montre que les deux réseaux ont le même comportement moyen.

Méthode probabiliste : contraindre les probabilités de réactions De façon similaire, on peut supposer que les trajectoires sont issues d'un processus stochastique et tenter l'approche par contraintes sur les probabilités de réactions stationnaires $\vec{p}$ avec l'avantage de pouvoir utiliser les moments d'ordre deux. En ce qui concerne les moyennes on peut réutiliser les contraintes correspondantes à celles utilisées dans la tentative de réfutation par les flux. On se propose d'exploiter la pente négative de la covariance entre C et D ce qui conduit au système suivant.

Les contraintes obtenues à partir de la figure 4.2 sont

modèle 1	modèle 2	constraintes
$p_1 + p_2 = 1, 0 \leq p_1, p_2 \leq 1$	$p_1 + p_2 = 1, 0 \leq p_1, p_2 \leq 1$	$\vec{p}$ est un vecteur de probabilité
$p_1 = p_2$	$p_1 = p_2$	A et B sont équilibrés
$p_1 = p_2$	$p_1 = p_1$	$2dm_C(t)/dt = dm_D(t)/dt$
$-2p_1p_2 < 0$	$2p_1(1-p_1) < 0$	$c_{C,D}(k) \leq \alpha k$ avec $\alpha < 0$.

Cette fois on observe que le second système de contraintes n'admet pas de solutions et qu'on peut donc réfuter le réseau de réactions correspondant. Lorsqu'on simule par l'algorithme de Gillespie ces deux réseaux on voit bien (figure 1) que c'est au niveau de la covariance entre C et D que les deux réseaux se distinguent et c'est pour cela qu'on a pu réfuter le second grâce à une contrainte liée à cette mesure de covariance. On peut noter que dans cette nouvelle méthode, on a ajouté la contrainte que la somme des probabilités de réactions vaut 1. Toutefois, ce n'est pas cette contrainte qui différencie les deux approches puisque l'ajout d'une contrainte de type $f_1 + f_2 = \mathrm{constante}$ dans la méthode classique n'aurait pas changé sa conclusion. On rappelle l'avantage dans les deux approches de ne reposer sur aucune information concernant les lois et paramètres cinétiques, seule l'information stœchiométrique du réseau est utilisée.

4.1.3.2 Exemple d'un réseau métabolique jouet

Nous proposons l'exemple d'un réseau métabolique jouet (Figure 4.3) pour illustrer une autre application de notre approche par contraintes. Cette fois-ci, on va montrer comment des informations sur les moments de certaines sorties peut nous donner d'autres informations sur les sorties. Le réseau de réactions consiste en une réaction produisant un métabolite A qui représente le point d'entrée du système. Le fait que la réaction ne possède pas de réactif permet de modéliser une quantité constante de l'espèce en entrée qui produit A (cela peut être réalisé expérimentalement par exemple). Ce métabolite peut ensuite être transformé en quatre différents métabolites intermédiaires (B, C, D, E) qui seront utilisés pour produire quatre métabolites à la sortie (O_1, O_2, O_3, O_4). Nous faisons ensuite les hypothèses suivantes (obtenues par exemples par observations expérimentales) sur le système :

(H_0) un état stationnaire a été atteint dans lequel les quantités de métabolites internes (A, B, C, D, E) sont en moyenne constants et les espèces de sorties O_1, O_2, O_3 et O_4 s'accumulent,

(H_1) la variance de O_1 est bornée : $Var(y_{O_1}(k)) \leq k \cdot 0.2$,

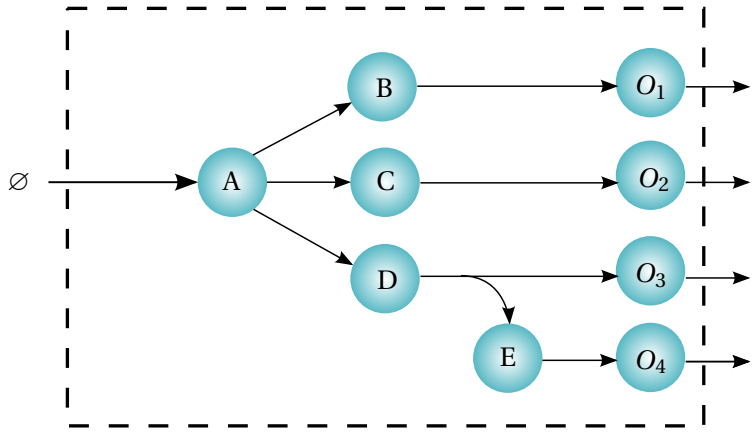
(H_2) la covariance entre O_1 et O_2 satisfait $Cov(y_{O_1}(k), y_{O_2}(k)) \leq -0.01k$,

Les valeurs numériques proposées (0.2, -0.01 et $1/2$) sont purement arbitraires et sont choisies à titre d'illustration.

L'hypothèse H_0 conduit à l'ensemble de contraintes suivants sur le vecteur de probabilités de réactions $\vec{p}$

$$
\begin{array}{ll}
0 \leq p_i \leq 1, i = 1, \ldots, 8, & (p_i \text{ sont des probabilités}) \\
\sum_{i=1}^{8} p_i = 1 & (\vec{p} \text{ est un vecteur de probabilités}) \\
p_8 = p_5 + p_6 + p_7 & A \text{ constant en moyenne} \\
p_5 = p_1 & B \text{ constant en moyenne} \\
p_6 = p_2 & C \text{ constant en moyenne} \\
p_7 = p_3 + p_4 & D \text{ constant en moyenne} \\
p_3 = p_4 & E \text{ constant en moyenne}
\end{array}
$$

Ces contraintes sont linéairement indépendantes, donc le système a deux degrés de liberté et nous décidons de nous considérer le couple (p_1, p_2) en tant que variable, sachant que les autres composantes de $\vec{p}$ peuvent être calculées à l'aide du système suivant (obtenu à partir



	A	B	C	D	E	O_1	O_2	O_3	O_4
$R1: B \to O_1$	0	-1	0	0	0	1	0	0	0
$R2: C \to O_2$	0	0	-1	0	0	0	1	0	0
$R3: D \to O_3 + E$	0	0	0	-1	1	0	0	1	0
$R4: E \to O_4$	0	0	0	0	-1	0	0	0	1
$R5: A \to B$	-1	1	0	0	0	0	0	0	0
$R6: A \to C$	-1	0	1	0	0	0	0	0	0
$R7: A \to D$	-1	0	0	-1	0	0	0	0	0
$R8: \varnothing \to A$	1	0	0	0	0	0	0	0	0

FIGURE 4.3 – **Un exemple de réseau métabolique jouet** illustratif pour les méthodes de contraintes

du précédent par réduction de Gauss) :

$$\begin{aligned}
p_3 &= 1/6 - p_1/2 - p_2/2,\\
p_4 &= 1/6 - p_1/2 - p_2/2,\\
p_5 &= p_1,\\
p_6 &= p_2,\\
p_7 &= 1/3 - p_1 - p_2,\\
p_8 &= 1/3.
\end{aligned}$$

De manière générale, si on a m réactions et n' métabolites équilibrés en moyenne, le nombre de degrés de liberté sera $m - n' - 1$. Puisque les probabilités de réactions doivent être dans le segment $[0,1]$, on doit également être attentif à ne considérer que les valeurs (p_1, p_2) pour lesquelles les expressions ci-dessus des autres probabilités $p_i (i = 3, \ldots, 8)$ sont dans le domaine correct. L'hypothèse (H_0) se résume alors à

$$(H_0): \; p_1 + p_2 \leq 1/3.$$

En utilisant notre tableau de contraintes, on convertit les hypothèses (H_1) et (H_2) en contraintes sur $\vec{p}$ ce qui donne

$$(H_1) : p_1 - p_1^2 \leq 0.2,$$
$$(H_2) : -p_1 p_2 \leq -0.01.$$

Les valeurs possibles pour (p_1, p_2) assujetties à ces contraintes sont représentées dans la figure 4.4. L'association des trois contraintes conduit à un ensemble de solutions $\mathscr{S}$ pour les valeurs possibles de (p_1, p_2) compte tenu des données, on obtient ainsi des informations sur les valeurs possibles de ces paramètres. La figure permet de souligner que l'espace de solutions n'est pas nécessairement connexe, c'est par exemple le cas pour la contrainte (H_1).

Une fois cette espace de possibilités pour (p_1, p_2) obtenu on peut soit essayer d'en déterminer une valeur particulière en utilisant par exemple un critère d'optimisation, c'est l'approche utilisée dans le FBA. On peut aussi, et c'est l'approche qu'on préfèrera dans cette thèse, envisager l'ensemble des valeurs possibles afin d'en obtenir des bornes. Par exemple on peut s'intéresser à la quantité $\bar{\rho}_{1,2} = p_1/p_2$ correspondant à la mesure du taux de production de O_1 par rapport à celui de O_2. Cette quantité sera ré-introduite plus loin dans la dernière section de ce chapitre car, on le verra, elle est facilement accessible par l'expérience. Dans la figure 4.4 on a représenté les valeurs extrémales de p_1/p_2 dans $\mathscr{S}$. Sous les hypothèses (H_0), (H_1) et (H_2) les valeurs possibles pour p_1/p_2 sont donc le segment $[0.11, 7.6]$. Ainsi on aboutit au résultat suivant

Résultat 4.1. *Dans l'exemple jouet proposé, supposons que le système soit dans un état stationnaire dans lequel les réactifs sont en quantités moyennes constantes (H_0) et qu'on a observé les propriétés (H_1) (concernant la variance de O_1) et (H_2) (concernant la covariance entre O_1 et O_2) alors le rapport de taux de production entre O_1 et O_2 $\bar{\rho}_{1,2} = p_1/p_2$ est situé dans l'intervalle $\bar{\rho}_{1,2} \in [0.11, 7.6]$.*

Notons encore une fois comme ce résultat a été obtenu sans la connaissance des lois et des paramètres cinétiques.

4.2 Ellipsoïdes de confiance

Nous présentons maintenant une application du théorème central limite introduit dans le chapitre 3 qui consiste à déterminer des régions de confiance pour la trajectoire de la dynamique de Bernoulli. Ceci permet de confronter des trajectoires données avec une dynamique de Bernoulli dont les paramètres S et $\vec{p}$ sont connus.

4.2.1 Définition

Dans la théorie des probabilités, la valeur de la réalisation d'une variable aléatoire réelle dont ont connait la distribution est souvent estimée à l'aide d'intervalles de confiance. Les

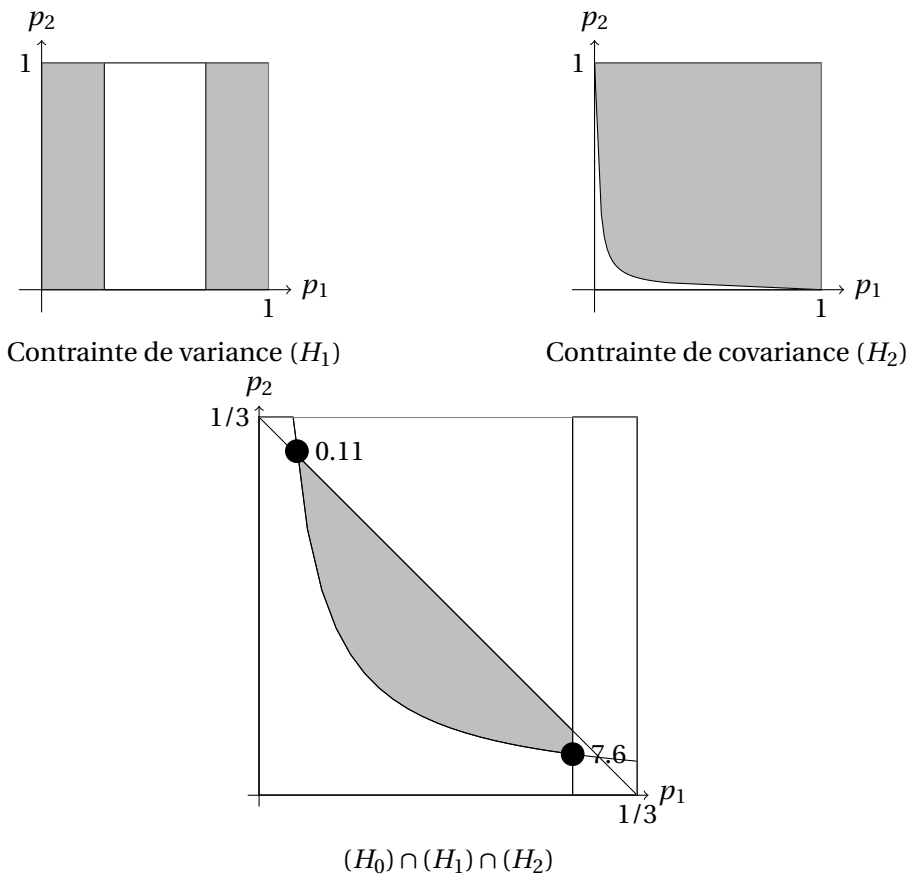


FIGURE 4.4 – **Illustration de contraintes de variances et co-variances**. Les zones grises correspondent à l'ensemble des couples (p_1, p_2) valides en fonction des contraintes indiquées ainsi que de la contrainte (H_0) d'équilibre des réactifs (permettant de se ramener à un espace plan). Les points noirs correspondent aux valeurs extrêmes de p_1/p_2.

ellipsoïdes de confiance sont une généralisation des intervalles de confiance en dimension supérieure lorsque le vecteur aléatoire considéré est gaussien c'est-à-dire, suit une loi normale multivariée. Or, nous savons que la dynamique de Bernoulli se comporte asymptotiquement comme un vecteur gaussien, on peut donc en déterminer des *ellipsoïdes de confiance*.

Proposition-Définition 4.1 (Ellipsoïdes de confiance)**.** *Soit S de taille $m \times n$ la matrice de stœchiométrie d'un réseau, $\vec{p}$ un vecteur de probabilité de réactions strictement positives, $\vec{z}_0$ des quantités de matière initiales et , $\alpha \in]0,1]$ une tolérance d'erreur. On considère t_α l'unique solution de l'équation*

$$\frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\vec{x} \in B_n(\vec{0}, t_\alpha)} \exp\left(-\frac{\|\vec{x}\|^2}{2}\right) d\vec{x} = 1 - \alpha \tag{4.5}$$

où $B_n(\vec{0}, t_\alpha)$ est la boule centrée de $\mathbb{R}^n$ de rayon t_α et $\|\cdot\|$ est la norme euclidienne. Supposons que $\ker W(S, \vec{p}) = \{\vec{0}\}$ *et considérons sa* racine carrée $V \in GL_n(\mathbb{R})$ *vérifiant $W(S, \vec{p}) = VV^t$.*

Alors le sous-ensemble

$$\mathscr{E}(S,\vec{p},\vec{z}_0,\alpha,k)=\left\{\vec{z}\in\mathbb{R}^n \quad | \quad \left\|\frac{1}{\sqrt{k}}V^{-1}\left(\vec{z}-\vec{z}_0-kS^t\vec{p}\right)\right\|\leq t_\alpha\right\} \tag{4.6}$$

ne dépend pas du choix de V et constitue dans $\mathbb{R}^n$ une ellipsoïde *non dégénérée qu'on appellera α*-ellipsoïde de confiance.

Démonstration. Premièrement, on remarque que $\mathscr{E}(S,\vec{p},\vec{z}_0,\alpha,k)$ est une ellipsoïde non dégénérée de $\mathbb{R}^n$ car c'est l'image de la boule $B_n(\vec{0},t_\alpha\sqrt{k})$ par l'application affine inversible $\vec{z}\mapsto V\vec{z}+\vec{z}_0+kS^t\vec{p}$. Maintenant démontrons que le choix de V n'est pas important. Si $W(S,\vec{p})=VV^t=UU^t$ avec $U,V\in GL_n(\mathbb{R})$ alors $U^{-1}V$ est un automorphisme unitaire :

$$(U^{-1}V)(U^{-1}V)^t=(U^{-1}V)^t(U^{-1}V)=\mathbb{I}_n.$$

Ainsi $U^{-1}V$ est isométrique pour la norme euclidienne $\|\cdot\|$ et pour tout $\vec{z}\in\mathbb{R}^n$:

$$\begin{aligned}\left\|\frac{1}{\sqrt{k}}V^{-1}\left(\vec{z}-\vec{z}_0-kS^t\vec{p}\right)\right\|\leq t_\alpha \Leftrightarrow &\left\|U^{-1}V\left(\frac{1}{\sqrt{k}}V^{-1}\left(\vec{z}-\vec{z}_0-kS^t\vec{p}\right)\right)\right\|\\ &=\left\|\frac{1}{\sqrt{k}}U^{-1}\left(\vec{z}-\vec{z}_0-kS^t\vec{p}\right)\right\|\\ &\leq t_\alpha.\end{aligned}$$

□

Remarquons qu'une matrice V adaptée peut être calculée grâce à la décomposition de Choleski [HJ12] ou la décomposition spectrale. Cette définition se comprend intuitivement comme la volonté de déterminer une transformation affine de $\vec{z}(k)$ en loi normale multivariée centrée réduite. La proposition suivante montre que cette définition est correcte.

Proposition 4.2. *La dynamique de Bernoulli $\vec{z}(k)$ appartient à l'ellipsoïde de confiance $\mathscr{E}(S,\vec{p},\vec{z}_0,\alpha,k)$ avec une probabilité tendant vers $1-\alpha$ lorsque k tend vers l'infini.*

$$\mathbb{P}\left(\vec{z}(k)\in\mathscr{E}(S,\vec{p},\vec{z}_0,\alpha,k)\right)\underset{k\to+\infty}{\longrightarrow}1-\alpha. \tag{4.7}$$

Démonstration. La proposition 3.3 nous indique que $\frac{1}{\sqrt{k}}\left(\vec{z}(k)-\left(\vec{z}_0+kS^t\vec{p}\right)\right)$ converge en distribution vers $G\sim\mathscr{N}(\vec{0},W(S,\vec{p}))$. Le théorème de Mann-Wald [MW43] et la proposition 3.5 impliquent que $\frac{1}{\sqrt{k}}V^{-1}\left(\vec{z}(k)-\left(\vec{z}_0+kS^t\vec{p}\right)\right)$ converge en loi vers G' de distribution

$$G'\sim\mathscr{N}\left(\vec{0},V^{-1}W(S,\vec{p})(V^{-1})^t\right)=\mathscr{N}\left(\vec{0},\mathbb{I}_n\right).$$

Ainsi, le théorème porte-manteau (la mesure de la frontière de $B_n(\vec{0},t_\alpha)$ étant nulle) :

$$\begin{aligned}\mathbb{P}\left(\vec{z}(k)\in\mathscr{E}(S,\vec{p},\vec{z}_0,\alpha)_k\right)&=\mathbb{P}\left(\frac{1}{\sqrt{k}}V^{-1}\left(\vec{z}(k)-\vec{z}_0-kS^t\vec{p}\right)\in\mathscr{B}_n(\vec{0},t_\alpha)\right)\\ &\underset{k\to+\infty}{\longrightarrow}\mathbb{P}\left(G'\in B_n(\vec{0},t_\alpha)\right)=1-\alpha.\end{aligned}$$

□

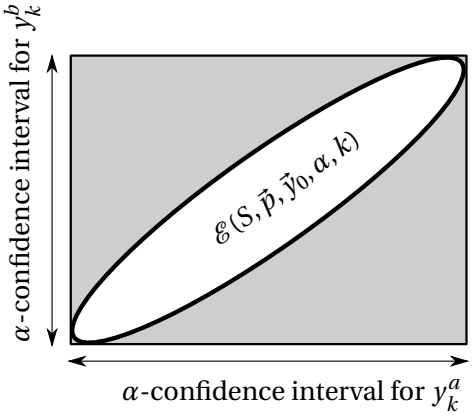


FIGURE 4.5 – **Comparaison entre l'ellipsoïde de confiance et la multiplication des intervalles de confiance pour deux espèces.** La figure illustre l'avantage de l'utilisation de méthodes multivariées. Le rectangle obtenu par multiplication des interfaces est plus grand que l'α-ellipsoïde de confiance. La zone grise correspond aux prédictions exclues par la méthode multivariée grâce à l'utilisation des corrélations entre les deux espèces.

Ces ellipsoïdes de confiance montrent que l'approche stochastique permet de prendre en compte à la fois les variances des quantités d'espèce mais aussi les corrélations inter-espèces. En effet l'ellipsoïde est centrée sur l'espérance de $\vec{z}(k)$ tandis que sa forme, ses dimensions et ses axes sont déterminées par la matrice de covariance. De plus les avantages de l'approche multivariée est très clair dans ce cas : les α-intervalles de confiance pour chaque espèce $(y^a(k))_{1\leq a\leq n}$ sont obtenus en projetant l'ellipsoïde sur chacun des axes de $\mathbb{R}^n$. Cela est aussi équivalent à supprimer toutes les lignes de la matrice de stœchiométrie sauf celle de l'espèce dont on cherche à calculer l'intervalle de confiance.

4.2.2 Étude des cas dégénérés de la loi limite

Jusqu'à présent nous avons montré que le processus $(\vec{y}(k))_k$ est approximativement une marche aléatoire entièrement déterminée par S et les probabilités de réactions stationnaires $\vec{\mathfrak{p}}$. Après normalisation, cette marche aléatoire se réparti asymptotiquement selon une loi normale de matrice de variance-covariance $W(S, \vec{\mathfrak{p}})$. Cependant, cela ne signifie évidemment pas que la marche aléatoire s'étend dans toutes les dimensions de l'espace. Une manière mathématique d'exprimer ce fait est que la matrice $W(S, \vec{\mathfrak{p}})$ est symétrique, positive mais pas nécessairement définie. Lorsqu'elle n'est pas définie, on dira que la marche aléatoire est dégénérée car elle ne se diffuse pas dans tout l'espace, mais seulement dans un sous-espace vectoriel strict. L'étude des cas dégénérés, est intéressante pour au moins deux raisons : elle offre une meilleure compréhension de la dynamique du système étudié et permet aussi de travailler de manière équivalente dans un espace de plus petite dimension plus simple où la marche aléatoire est non dégénérée.

Nous montrons qu'il existe seulement deux causes de dégénérescence. La première rai-

son est liée à la notion de $P\text{-}invariants$ [Mur89, Wil12], correspondant à des lois de conservation de matière qui restreignent la trajectoire dans un sous-espace affine de l'espace des phases. La loi limite est alors contrainte de la même manière ce qui produit la dégénérescence. Une seconde raison plus subtile et pouvant apparaître même en l'absence de P-invariants est liée à la dégénérescence du processus de comptage de réactions qui peut être transportée dans l'espace des phases dans des conditions que nous préciserons.

4.2.2.1 P-invariants

Soit (n, m, α, β) un réseau de réaction et $S = \beta - \alpha$ sa matrice de stœchiométrie. Une solution $\vec{z}$ non nulle à l'équation linéaire $S^t \vec{z} = \vec{0}$ est appelée un $P\text{-}invariant$ du système. Les P-invariants sont importants car ils correspondent à des $lois\ \ de\ \ conservation$ dans le réseau [Wil12]. En effet, si $\vec{z}$ est un P-invariant alors ses composantes sont aussi les coordonnées dans la $base\ \ duale$ d'une $forme\ \ linéaire$ φ conservée par les trajectoires du système, c'est-à-dire que

$$\forall t, \varphi(\vec{x}(t)) = \varphi(\vec{x}(0)). \tag{4.8}$$

Puisque cette forme linéaire est non nulle ($\varphi \neq 0$) , cela signifie que la trajectoire est incluse dans un hyperplan affine défini par l'équation $\varphi(\vec{z}) = \varphi(\vec{x}(0))$. Cela conduit nécessairement à une matrice de variance-covariance non définie, y compris pour la loi normale limite. Plus généralement si $\dim \ker S^t = k$ alors la trajectoire est incluse dans un sous-espace affine de $n - k$.

Élimination des P-invariants En présence de P-invariants on a généralement intérêt à considérer un réseau de plus petite taille, c'est-à-dire avec moins d'espèces, mais sans perdre d'informations. Pour cela il suffit d'éliminer une espèce qui apparaît avec un coefficient non nul dans le P-invariant. Cela revient à supprimer une ligne de la matrice de stœchiométrie. Il n'y a pas de perte d'information car grâce à l'équation (4.8) il est toujours possible de calculer la quantité de cette espèce en fonction des quantités des espèces non éliminées. Formellement, si $\vec{z}$ est un P-invariant avec (par exemple) $z_1 \neq 0$ alors la quantité $x_1(t)$ est fonction de $(x_2, \cdots, x_n)(t)$:

$$\forall t, x_1(t) = \frac{1}{z_1}\left(z_1 x_1(0) - \sum_{i=2}^{n} z_i (x_i(0) - x_i(t)\right). \tag{4.9}$$

Ainsi, en répétant ce processus on aboutit à un sous-système sans P-invariants, utilisant un sous-ensemble des espèces de base mais dont les quantités des espèces éliminées peuvent toujours être déduites par des formules explicites.

4.2.2.2 Cas du processus de comptage

La seconde source de dégénérescence est due à la possibilité dans certaines marches aléatoires que l'ensemble des points atteignables après n pas soit toujours inclus dans un

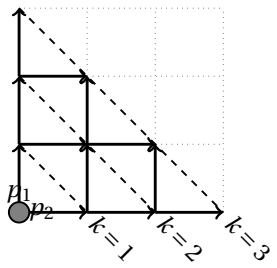


FIGURE 4.6 – **Illustration d'une marche aléatoire dégénérée** pour le réseau de réactions $\{\varnothing \to X; \varnothing \to Y\}$. Après k réactions, les quantités de matière sont nécessairement situées dans l'hyperplan d'équation $X + Y = k$.

sous-espace affine. Le cas du processus de comptage des réactions $\vec{q}(k) = \left(\sum_{l=1}^{k} \vec{e}_{\mu_l}\right)_k$ en est un exemple car $\vec{q}(k)$ est toujours inclus dans l'hyperplan d'équation $\sum_{i=1}^{n} z_i = k$. On rappelle que le processus de comptage est un cas particulier de trajectoire discrète qui est obtenu en choisissant un réseau tel que $S = \mathbb{I}$ et des quantités initiales nulles.

Cette dégénérescence peut, dans des conditions que nous préciserons, se transporter à la dynamique de Bernoulli qui est l'image par une application affine du processus de comptage des réactions. De manière similaire aux P-invariants, connaître l'équation de cet hyperplan permet d'éliminer une des espèces sans perdre d'information sous réserve que la valeur de k est donnée. En effet, une différence importante est que les P-invariants sont valables à tout instant k tandis qu'ici l'équation de l'hyperplan dépend de k.

4.2.2.3 Caractérisation

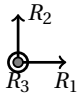
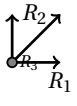
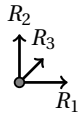
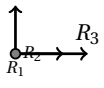
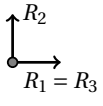
Nous donnons maintenant une proposition qui caractérise les cas dégénérés de la loi limite gaussienne, c'est-à-dire qui détermine en fonction de la matrice de stœchiométrie S et du vecteur $\vec{p}$. La proposition inclue les deux cas dégénérés présentés ci-dessus et montre qu'il n'en existe pas d'autres. La démonstration repose essentiellement sur la réduction orthogonale des matrices symétriques réelles.

Proposition 4.3. *Soit S la matrice de stœchiométrie du réseau. Soit $\vec{p}$ un vecteur de probabilité de réactions strictement positif. Notons $\vec{u} = (1, \cdots, 1)^t \in \mathbb{R}^n$, le vecteur ayant toutes ses composantes égales à* 1*. Alors, un et un seul des cas suivants est vrai.*

1. *Il existe un P-invariants et alors la loi limite est dégénérée* $\dim \ker W(S, \vec{p}) > 0$.
2. *S^t est injective et l'équation $S^t \vec{z} = \vec{u}$ a une unique solution $\vec{\eta}$ et alors* $\ker W(S, \vec{p}) = \mathrm{vect}(\vec{\eta})$.
3. *S^t est injective et l'équation $S^t \vec{z} = \vec{u}$ n'a pas de solution et alors* $\ker W(S, \vec{p}) = \{\vec{0}\}$.

Le premier cas apparaît lorsque S^t n'est pas injective, ce qui équivaut par dualité au cas S non surjective. D'après l'équation (3.14) $(\vec{z}(k))$ appartient à l'image de S donc on comprend que s'il existe un P-invariant alors on aboutit à un cas dégénéré. Dans le cas où il n'y a pas de P-invariant et qu'il existe une solution $\vec{\eta}$ à l'équation mentionnée, $\vec{z}(k)$ est inclus dans l'hyperplan affine dirigé par l'hyperplan vectoriel d'équation $\sum_{i=1}^{n} \eta_i z_i = 0$ et passant par le point $\vec{z}_0 + kS^t \vec{p}$. Le dernier cas est le cas régulier, non dégénéré.

Le tableau suivant représente quelques exemples de réseaux de réactions simples qui illustrent les trois cas.

Réseau							
S^t injective ?	oui	non	oui	oui	oui	oui	oui
Solution à $S^t\vec{z} = \vec{u}$?	oui	non	non	non	oui	non	oui
Dégénéré ?	oui	oui	non	non	oui	non	oui

Démonstration. La preuve traite les trois cas successivements.

Premièrement, on remarque que puisque $W(S,\vec{p}) = S\left(\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t\right)S^t$, on a $\ker S^t \subset \ker W(S,\vec{p})$ ce qui explique le premier cas.

D'après la proposition 3.4, nous savons que $\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t$ est une matrice symétrique positive. On peut calculer son noyau en remarquant que $p_l \neq 0$ pour tout l, car $\vec{p}$ est un vecteur de probabilité strictement positif. Dans ce cas l'équation $\left(\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t\right)\vec{x} = \vec{0}$ implique que $x_l = \sum_{k=1}^m p_k x_k$ pour tout $l \in \{1,\dots,m\}$. Cela implique que $\vec{x} \in \mathrm{vect}(\vec{u})$ où $\vec{u} = \begin{pmatrix}1 & 1 & \cdots & 1\end{pmatrix}^t$. Réciproquement, $\left(\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t\right)\vec{u} = \vec{0}$ car les composantes de $\vec{p}$ ont pour somme 1.

Ainsi, il existe une base orthogonale ayant $\vec{u}$ comme premier vecteur et dans laquelle $\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t$ est diagonale. D'un point de vue formel, il existe $P \in \mathcal{O}(m)$ (ensemble des matrices orthogonales réelles de taille $m \times m$) tel que $P^t\begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}^t = \vec{u}$ et

$$\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t = P^t\,\mathrm{diag}(0, \underbrace{\lambda_2}_{>0}, \dots, \underbrace{\lambda_m}_{>0})P.$$

Pour les deux derniers cas nous allons tout d'abord montrer le résultat intermédaire suivant : pour tout vecteur $\vec{x}$, $\vec{x} \in \ker W(S,\vec{p})$ si et seulement si $\exists \mu \in \mathbb{R} : S^t\vec{x} = \mu\vec{u}$. En effet, supposons qu'il existe $\vec{x}$ tel que $W(S,\vec{p})\vec{x} = \vec{0}$, alors $\vec{x}^t SP^t DPS^t\vec{x} = 0$ et donc $\left(PS^t\vec{x}\right)^t D\left(PS^t\vec{x}\right) = 0$. On peut réécrire cette dernière égalité sous la forme $\sum_{i=2}^m \lambda_i \left(PS^t\vec{x}\right)_i^2 = 0$. Puisque $\lambda_i > 0$ pour $i > 1$, on a $PS^t\vec{x} \in \mathrm{vect}\begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}^t$. Ainsi, $S^t\vec{x} = P^tPS^t\vec{x} \in \mathrm{vect}(\vec{u})$. Réciproquement, si $S^t\vec{x} \in \mathrm{vect}(\vec{u})$, la relation $\ker\left(\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^t\right) = \mathrm{vect}(\vec{u})$ implique que $W(S,\vec{p})\vec{x} = \vec{0}$. Ainsi, nous avons démontré le résultat intermédiaire.

Supposons maintenant que S^t est injective et considérons l'équation $S^t\vec{z} = \vec{u}$. Puisque $\ker S^t = \{0\}$, ce système d'équations possède au plus une solution. Deux cas apparaissent donc :

— S'il n'y a aucune solution, considérons $\vec{x} \in \ker W(S,\vec{p})$ alors la valeur μ associée à $\vec{x}$ est nulle puisque sinon $\frac{\vec{x}}{\mu}$ est une solution de l'équation. Donc, $\mu = 0$ et $S^t\vec{x} = \vec{0}$. Puisque S^t est injective nous avons donc $\vec{x} = \vec{0}$ montrant que $\dim\ker W(S,\vec{p}) = 0$.

— Si le système a une unique solution $\vec{\eta}$, lorsque $\mu \neq 0$ on a $S^t\vec{x} = \mu\vec{u}$ si et seulement si $S^t\frac{\vec{x}}{\mu} = \vec{u}$ si et seulement si $\vec{x} = \mu\vec{\eta}$. Le cas $\mu = 0$ correspond au vecteur nul appartenant toujours au noyau. Cela montre finalement que $\ker W(S,\vec{p}) = \mathrm{vect}(\vec{\eta})$.

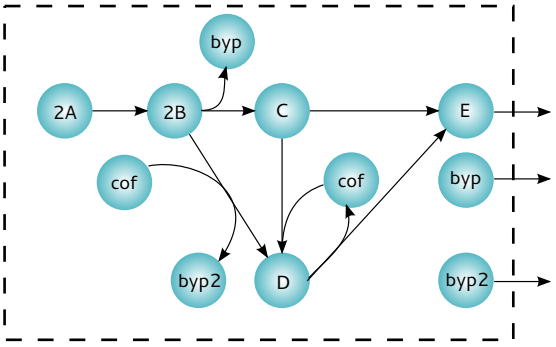


FIGURE 4.7 – **Retour sur l'exemple du réseau métabolique introductif** dont les réactions et la matrice de stœchiométrie ont été introduits dans le chapitre 1.

□

4.2.3 Application à la réfutation de modèles

Dans cette section, on suppose que les probabilités de réaction $\vec{p}$ sont connues et on en déduit une *région de confiance* pour la trajectoire $\vec{z}(k)$. Dans les cas où $\vec{p}$ n'est pas connu, on peut restreindre l'ensemble de ses valeurs possibles en utilisant l'approche par contraintes présentée dans la section 1 et qui sera étendue dans la section 3 de ce chapitre.

Illustrons ces régions de confiance à l'aide de l'exemple d'un réseau modélisant une voie métabolique présenté dans [PPW+03]. Nous avons légèrement modifié ce système en donnant des noms différents aux bi-produits des réactions 2 et 3 afin de les distinguer. Cela conduit au réseau de réaction représenté dans la figure 1.1. Nous supposons abritrairement que $\vec{z}_0 = \vec{0}$ et $\vec{p} = (0.3, 0.2, 0.1, 0.2, 0.1, 0.1)^t$. Cette valeur particulière de $\vec{p}$ garantie un équilibre en moyenne des quantités de métabolites B, C, et D. Afin de se concentrer uniquement sur les sorties du système on considère la matrice de stœchiométrie réduite S' considérant uniquement les espèces $(E, byp, byp2)$. Maintenant supposons que nous pouvons mesurer les quantités de E, byp and $byp2$ dans trois cellules individuelles différentes après $k = 100$ pas de temps δt :

$$\vec{o}_1 = (40, 15, 5)^t \quad \vec{o}_2 = (23, 19, 11)^t \quad \vec{o}_3 = (35, 25, 15)^t. \tag{4.10}$$

On s'intéresse à un problème de *validation de modèle* en tentant de déterminer lesquelles de ces cellules ont des quantités consistantes par rapport au réseau proposé. On se fixe alors $ut_\alpha = 3$ correspondant à une tolérance d'erreur de

$$1 - \frac{1}{(2\pi)^{\frac{3}{2}}} \int_{\vec{x} \in B_3(\vec{0},3)} \exp\left(-\frac{\|\vec{x}\|^2}{2}\right) d\vec{x} \simeq 0.0292909 \simeq 2.9\% \tag{4.11}$$

et nous calculons l'ellipsoïde de confiance correspondante. D'après les résultats précédents,

la matrice de variance-covariance de la loi limite (normalisée) est

$$W(S',\vec{p}) = \begin{pmatrix} 0.21 & -0.06 & -0.03 \\ -0.06 & 0.16 & -0.02 \\ -0.03 & -0.02 & 0.09 \end{pmatrix} \tag{4.12}$$

dont on peut calculer la décomposition de Choleski

$$V = \begin{pmatrix} 0.4583 & 0 & 0 \\ -0.1309 & 0.378 & 0 \\ -0.0655 & -0.0756 & 0.2828 \end{pmatrix}. \tag{4.13}$$

Comme cette loi limite n'est pas dégénérée, nous pouvons alors déterminer l'équation de l'ellipsoïde de confiance d'ordre α pour $\vec{z}(k)$.

$$\mathscr{E}(\alpha,100): \quad \underbrace{\left\| \frac{1}{\sqrt{100}} \underbrace{\begin{pmatrix} 2.1822 & 0 & 0 \\ 0.7559 & 2.6458 & 0 \\ 0.7071 & 0.7071 & 3.5355 \end{pmatrix}}_{V^{-1}} \left(\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} - 100 \begin{pmatrix} 0.3 \\ 0.2 \\ 0.1 \end{pmatrix} \right) \right\|}_{err(\vec{z})} \leq t_\alpha. \tag{4.14}$$

On applique alors cette équation pour déterminer si les points proposés (4.10) appartiennent à l'ellipsoïde, en supposant que k est assez grand pour que la loi limite gaussienne soit valide. Le calcul de l'erreur quadratique pour les points donnés donne $err(\vec{o}_1) \simeq 2.66$, $err(\vec{o}_2) \simeq 1.73$ and $err(\vec{o}_3) \simeq 3.2$ ce qui montre que les deux points $\vec{o}_1$ et $\vec{o}_2$ appartiennent à l'ellipsoïde de confiance $\mathscr{E}(\alpha,100)$ tandis que ce n'est pas le cas de $\vec{o}_3$. On peut alors conclure, sous réserve de bonne convergence de la loi limite et de bonne approximation par la dynamique de Bernoulli qu'avec une probabilité $1-\alpha \simeq 97\%$ les données concernant la cellule 3 ne sont pas consistantes avec le réseau de réactions et les probabilités de réactions proposées. Une remarque intéressante est que la valeur moyenne $E(\vec{z}(100)) = (30,20,10)^t$ n'était pas difficile à calculer mais n'aurait pas permis seule d'effectuer cette analyse. En effet, l'ellipsoïde de confiance repose sur la capacité d'utiliser les variances et covariances pour comprendre comment la trajectoire se répartie autour de sa valeur moyenne.

4.3 Contraintes par l'exploitation d'un rapport de taux de production

Dans cette dernière section, nous présentons un résultat de convergence en probabilité du rapport de taux de production de deux espèces d'un réseau. Cette convergence est démontrée à l'aide des ellipsoïdes introduites précédemment. L'avantage de ce résultat de convergence est qu'il permet de compléter notre tableau de contraintes en utilisant des informations sur un rapport de pentes entre les trajectoires de deux espèces. Contrairement à ce qui a été présenté jusqu'à présent, cette contrainte peut s'obtenir à l'aide d'un seul

individu, c'est-à-dire à l'aide d'une mesure effectuée sur une seule trajectoire (vectorielle) $\vec{z}(k)$. De plus, la quantité mesurée est expérimentalement intéressante car elle permet de s'affranchir de problèmes d'échelles de temps et de mesures quantitatives.

4.3.1 Convergence du rapport de taux de production

Notre second résultat de convergence concerne les rapports de taux de production. Sur un intervalle de temps discret $[0, T]$ on définit le *taux moyen de production* ou encore le *taux de synthèse* de l'espèce X_i comme le rapport

$$\frac{z(T)_a - z(0)_a}{T}. \tag{4.15}$$

Nous avons vu dans le chapitre 3 que $\vec{z}(k)$ sert à approximer $\vec{y}(k)$ qui est elle même une discrétisation à pas de temps fixe δt de la chaîne de Markov $\vec{x}(t)$ de la dynamique stochastique. La discrétisation étant valide lorsque δt est petit, on se retrouve avec de grandes valeurs de T. On préfèrera utiliser la notion de rapport de taux de production.

Définition 4.2 (Rapport de taux de production)**.** Le rapport de taux de production entre deux espèces X_a and X_b est défini comme

$$\rho_{a,b}(k) = \frac{(z(k)_a - z(0)_a)/T}{(z(k)_b - z(0)_b)/T} = \frac{z(k)_a - z(0)_a}{z(k)_b - z(0)_b} \qquad (k > 0). \tag{4.16}$$

On voit immédiatement l'intérêt de ce rapport qui est de faire disparaître les considérations temporelles. Cette quantité possède également les avantages suivants.

— Premièrement, il est facile à mesurer expérimentalement, en effet, on a vu (Proposition 3.2) qu'en moyenne les trajectoires du processus de Bernoulli sont des droites, la quantité mesurée correspond dont à la mesure d'un rapport de pentes entre la trajectoire de X_a et celle de X_b.

— Deuxièmement, la quantité est invariante par homothétie ce qui permet d'éviter les problèmes liées à l'échelle de mesure des données. En effet, dans beaucoup de méthodes expérimentales (western blots, Southern blots et autres techniques reposant sur l'électrophorèse) les données obtenues sont exprimées en unité arbitraire, c'est-à-dire que les quantités de matières sont connues de manière relative. On peut alors mesurer directement $\rho_{a,b}(k)$ sur ces données contrairement au taux moyen de production (4.15).

— Enfin une dernière bonne raison d'utiliser ce rapport est justement le résultat théorique de convergence ci-après. On montre en effet que $\rho_{a,b}(k)$ converge vers une constante ce qui permet d'une part de mesurer s'abstenir de s'abstenir de connaître k en mesurant plutôt la limite et d'autre part de pouvoir effectuer cette mesure **sur une seule réalisation** du processus stochastique $\vec{z}(k)$.

Voici maintenant le résultat de convergence qui nous intéresse.

Proposition 4.4. *Pour tout $a, b \in \{1, \dots, n\}$, si $(S\vec{p})_b \neq 0$ alors le rapport de taux de production $\rho_{a,b}(k)$ entre les espèces X_a et X_b converge en probabilité vers $\bar{\rho}_{a,b} = (S\vec{p})_a/(S\vec{p})_b$:*

$$\forall \varepsilon > 0, \lim_{k \to +\infty} \mathbb{P}\left(\left|\rho_{a,b}(k) - (S\vec{p})_a/(S\vec{p})_b\right| > \varepsilon\right) = 0. \tag{4.17}$$

La convergence en probabilité se lit comme une bonne prédiction de $\rho_{a,b}(k)$ puisque la probabilité que $\rho_{a,b}(k)$ appartienne à n'importe quel segment non réduit à un point, aussi petit qu'on le souhaite, et contenant $\bar{\rho}_{a,b}$ tend vers 1. Ainsi la proposition établit que $\rho_{a,b}(k)$ est un *estimateur consistant* de $\bar{\rho}_{a,b}$ qui permet lui même d'obtenir des informations sur $\vec{p}$. Cette remarque permettra d'étendre notre approche par contrainte en y ajoutant des contraintes liées à la mesure de $\bar{\rho}_{a,b}$ via cet estimateur.

Dans le reste de cette section, nous présentons une preuve qui a l'avantage de reposer sur une bonne interprétation géométrique. Dans un premier temps, nous donnons une interprétation géométrique du domaine des valeurs de $\vec{z}(k)$ vérifiant $\rho_{a,b}(k) \in [\bar{\rho}_{a,b} - \varepsilon, \bar{\rho}_{a,b} + \varepsilon]$. Dans un second temps, nous utilisons une sur-approximation reposant sur les ellipsoïde de confiance pour prouver que $\vec{z}(k)$ appartient asymptotiquement à cette région avec une probabilité arbitrairement choisie.

Interprétation géométrique L'ensemble des valeurs de $\vec{z}(k)$ correspondant à un rapport de taux de production $\rho_{a,b}(k)$ donné a une interprétation géométrique simple. Puisque $\rho_{a,b}(k)$ ne dépend que des espèces X_a et X_b, on se place naturellement dans un espace à deux dimensions. On considère la matrice de stœchiométrie S' de taille $2 \times m$ obtenue en ne conservant que les lignes de S concernant les espèces X_a et X_b, c'est-à-dire les lignes numérotées a et b. On considèrera sans nuire à la généralité que $a < b$. Ainsi, la dynamique de Bernoulli $\vec{z}(k)$ initiée en $(z(0)_a, z(0)_b)^t$ associée à S' et $\vec{p}$ est un vecteur de dimension 2 qui correspond à la projection orthogonale sur le plan vect$\{\vec{e}_a, \vec{e}_b\}$ de la dynamique de Bernoulli initiée en $\vec{z}(0)$ associée à S et $\vec{p}$, c'est-à-dire sans l'élimination des autres espèces. Ce qui est important est que du point de vue du calcul de $\rho_{a,b}(k)$, cela ne fait donc aucune différence de considérer cette nouvelle dynamique de Bernoulli réduite. Pour $\mu \neq 0$, on considère $\mathscr{D}_\mu$ l'ensemble des valeurs de $(z(k)_a, z(k)_b)$ telles que $\rho_{a,b} = \mu$. La nature géométrique de $\mathscr{D}_\mu$ est une droite passant par $(z(0)_a, z(0)_b)$ privée de ce même point et de pente μ^{-1}. Ainsi, l'ensemble $\mathscr{C}(r,s)$ des points $(z(k)_a, z(k)_b)$ tels que $\rho_{a,b} \in [r,s]$ est un cône formé par l'union des droites $\mathscr{D}_\mu$ pour $\mu \in [r,s]$. D'un point de vue formel, on a donc la remarque suivante.

Remarque 4.1. Soit $r, s \in \mathbb{R}$ et le cône

$$\mathscr{C}(r,s) = \bigcup_{r \leq \mu \leq s} \left\{\vec{z} \in \mathbb{R}^2 \mid \exists t \in \mathbb{R}^* : \vec{z} = \vec{z}_0 + t\vec{f}_\mu\right\}, \tag{4.18}$$

où $\vec{f}_\mu = (\mu, 1)^t$. Alors,

$$r \leq \rho_{a,b}(k) \leq s \Leftrightarrow \vec{z}(k) \in \mathscr{C}(r,s). \tag{4.19}$$

Sur approximation de $\mathscr{E}$ Nous avons montré que l'ellipsoide de confiance est asymptotiquement une région de confiance d'ordre α, nous introduisons maintenant une sur approximation de $\mathscr{E}$ qui est suffisante pour démontrer la proposition 4.4. La sur-approximation est la plus petite boule qui contient $\mathscr{E}$.

Définition 4.3 (Boule d'approximation)**.** Soit S une matrice de stœchiométrie de taille $n \times m$, $\vec{p}$ un vecteur de probabilités strictement positif, $\alpha \in]0,1]$ une tolérance d'erreur et $\vec{z}(0)$ des quantités d'espèces initiales. On considère également t_α l'unique solution de l'équation (4.5). Supposons que $W(S,\vec{p})$ est définie et considérons alors son rayon spectral $\lambda = \max \mathrm{Sp}\left(W(S,\vec{p})\right) > 0$. On définit la *boule d'approximation*

$$\mathscr{B}(S,\vec{p},\vec{z}(0),\alpha,k) = B_n\left(\vec{z}(0) + kS^t\vec{p}, t_\alpha\sqrt{k\lambda}\right). \tag{4.20}$$

Proposition 4.5. *La boule d'approximation est une sur-approximation de l'ellipsoïde de confiance :*

$$\forall k \in \mathbb{N}^*, \quad \mathscr{E}(S,\vec{p},\vec{z}(0),\alpha,k) \subseteq \mathscr{B}(S,\vec{p},\vec{z}(0),\alpha,k). \tag{4.21}$$

De plus, $\mathscr{B}(S,\vec{p},\vec{z}_0,\alpha,k)$ est la plus petite boule qui contient $\mathscr{E}(S,\vec{p},\vec{z}_0,\alpha,k)$.

Démonstration. Puisque $W(S,\vec{p})$ est symétrique réelle, on considère la décomposition spectrale $W(S,\vec{p}) = P\Lambda P^t$ où P est une matrice unitaire. En raison de la proposition 4.1 on peut choisir $V = P\Lambda^{1/2}$ pour définir l'ellipsoïde de confiance. Ainsi on a $V^{-1} = \Lambda^{-1/2}P^t$. Maintenant, considérons $\vec{z} \in \mathscr{E}(S,\vec{p},\vec{z}(0),\alpha,k)$. En remarquant que $\vec{z} \mapsto P^t\vec{z}$ est une isométrie on a

$$\left\|\vec{z} - (\vec{z}(0) + kS^t\vec{p})\right\| = \left\|\Lambda^{1/2}\Lambda^{-1/2}P^t(\vec{z} - (\vec{z}(0) + kS^t\vec{p}))\right\| \leq \||\Lambda^{1/2}\||\cdot\left\|\Lambda^{-1/2}P^t(\vec{z} - (\vec{z}(0) + kS^t\vec{p}))\right\|,$$

où $\||\cdot\||$ est la norme d'opérateur induite par $\|\cdot\|$. Dans ce cas, $\||\Lambda^{1/2}\|| = \sqrt{\max \mathrm{Sp}\left((\Lambda^{1/2})^t\Lambda^{1/2}\right)} = \sqrt{\lambda}$. Ainsi,

$$\left\|\vec{z} - (\vec{z}(0) + kS^t\vec{p})\right\| \leq t_\alpha\sqrt{\lambda k}$$

montrant que $\vec{z} \in B_n(\vec{z}(0) + kS^t\vec{p}, t_\alpha\sqrt{\lambda k}) = \mathscr{B}(S,\vec{p},\vec{z}(0),\alpha,k)$. De plus, considérons $\vec{z}_\lambda$ un vecteur propre unitaire pour la valeur propre λ, alors on peut calculer que $\vec{z}(0) + kS^t\vec{p} + t_\alpha\sqrt{k\lambda}\vec{z}_\lambda$ et $\vec{z}(0) + kS^t\vec{p} - t_\alpha\sqrt{k\lambda}\vec{z}_\lambda$ appartiennent à $\mathscr{E}(S,\vec{p},\vec{z}(0),\alpha,k)$. Ainsi, une boule contenant l'ellipsoïde de confiance doit nécessairement contenir le segement $[\vec{z}(0) + kS^t\vec{p} + t_\alpha\sqrt{k\lambda}\vec{z}_\lambda,\ \vec{z}(0) + kS^t\vec{p} - t_\alpha\sqrt{k\lambda}\vec{z}_\lambda]$ de longueur $2t_\alpha\sqrt{k\lambda}$ ce qui montre la minimalité de $\mathscr{B}(S,\vec{p},\vec{z}_0,\alpha,k)$. □

Nous pouvons maintenant obtenir une preuve de la proposition 4.4.

Preuve de la Proposition 4.4. Considérons $\varepsilon > 0$ et $\alpha > 0$. On veut montrer l'existence de $K \in \mathbb{N}^*$ tel que

$$\forall k > K,\ \mathbb{P}\left(\left|\rho_{a,b}(k) - \bar{\rho}_{a,b}\right| > \varepsilon\right) \leq \alpha.$$

Pour prouver cette existence, on considère le cône $\mathscr{C} = \mathscr{C}\left(\bar{\rho}_{a,b} - \varepsilon, \bar{\rho}_{a,b} + \varepsilon\right)$. En utilisant la remarque 4.1 on a

$$\mathbb{P}\left(\left|\rho_{a,b}(k) - \bar{\rho}_{a,b}\right| > \varepsilon\right) = \mathbb{P}\left(\vec{z}(k) \notin \mathscr{C}\right).$$

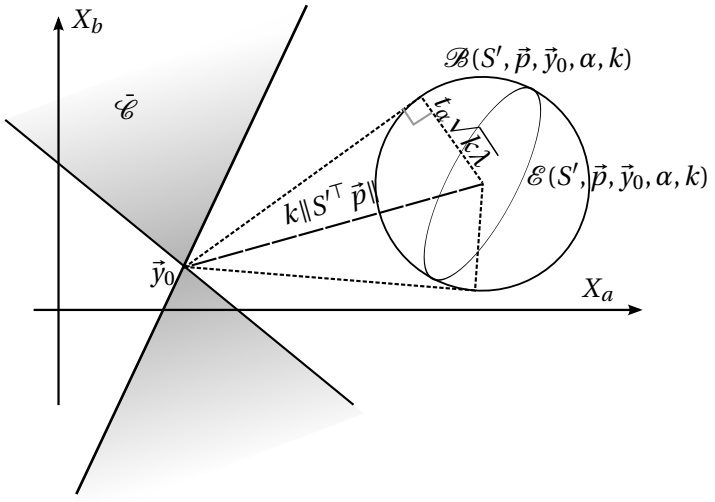


FIGURE 4.8 – **Une preuve géométrique de la consistance de l'estimateur $\rho_{a,b}(k)$.** Pour k assez grand, l'ellipsoïde de confiance d'ordre α est incluse dans $\mathscr{C}$ parce que le rayon de la sur-approximation est dominé par la distance entre $\vec{z}(0)$ et $E(\vec{z}(k))$ lorsque k croît. La région grise correspond au complémentaire de $\mathscr{C}$.

Pour simplifier les écritures, on notera $\mathscr{B}(\alpha, k) = \mathscr{B}(S, \vec{p}, \vec{z}(0), \alpha, k)$ et $\mathscr{E}(\alpha, k) = \mathscr{E}(S, \vec{p}, \vec{z}(0), \alpha, k)$. Puisque $(S\vec{p})_b \neq 0$ on a $\|S'\vec{p}\| \neq 0$ donc la distance entre $\vec{z}(0)$ et le centre de la boule de sur-appproximation $\mathscr{B}(\alpha, k)$ domine le rayon de $\mathscr{B}(\alpha, k)$ lorsque k tend vers l'infini, c'est-à-dire en notation de Landau $t_\alpha\sqrt{k\lambda} = o\left(k\|S'^t\vec{p}\|\right)$ (figure 4.8). Par conséquent, il existe un entier $K_1 \in \mathbb{N}^*$ tel que $\forall k > K_1, \mathscr{B}(\alpha, k) \subseteq \mathscr{C}$. Ensuite, grâce à la proposition 4.2 on sait qu'il existe un entier $K_2 \in \mathbb{N}^*$ tel que $\forall k > K_2, \mathbb{P}\left(\vec{z}(k) \in \mathscr{E}(\alpha, k)\right) \geq 1 - \alpha$. En conclusion si on choisit $K = \max(K_1, K_2)$ on a d'après la proposition 4.5

$$\forall k > K,\ \mathbb{P}(\vec{z}(k) \notin \mathscr{C}) \leq \mathbb{P}\left(\vec{z}(k) \notin \mathscr{B}(\alpha, k)\right) \leq \mathbb{P}(\vec{z}(k) \notin \mathscr{E}(\alpha, k)) \leq \alpha.$$

□

4.3.2 Nouveau tableau de contraintes

À titre d'application du résultat de convergence du rapport de taux de productions, nous pouvons ajouter de nouvelles lignes à notre table des contraintes introduite en début de chapitre. En effet, le théorème de convergence des rapports de taux de production établit lui aussi une relation analytique entre S, $\vec{p}$ et le rapport $\rho_{a,b}(k)$ lui-même, on peut donc en déduire de nouvelles contraintes. Contrairement à ce qui précédait cette nouvelle contrainte peut-être obtenue à l'aide d'une mesure sur un seul individu. On exploite ici le fait que le rapport de taux de productions, *a priori* aléatoire et donc variable pour chaque individu converge en réalité vers une valeur constante non aléatoire.

	Observation	Contrainte (forme matricielle)	Contrainte (forme algébrique)	Type
(1a)	$\forall k,\ m_a(k) = y_a(0)$	$(S\vec{p})_a = 0$	$\sum_{i=1}^m s_{ai} p_i = 0$	linéaire
(1b)	$m_a(k) \leq y_a(0) + k\gamma$	$(S\vec{p})_a \leq \gamma$	$\sum_{i=1}^m s_{ai} p_i \leq \gamma$	linéaire
(1c)	$y_a(0) + k\gamma \leq m_a(k)$	$\gamma \leq (S^\top \vec{p})_a$	$\gamma \leq \sum_{i=1}^m s_{ai} p_i$	linéaire
(2a)	$c_{a,a}(k) \leq k\gamma$	$\left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{aa} \leq \gamma$	$\sum_{i=1}^m s_{ai}^2 p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{aj} p_i p_j \leq \gamma$	quadratique
(2b)	$k\gamma \leq c_{a,a}(k)$	$\gamma \leq \left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{aa}$	$\gamma \leq \sum_{i=1}^m s_{ai}^2 p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{aj} p_i p_j$	quadratique
(3a)	$c_{a,b}(k) \leq k\gamma$	$\left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{ab} \leq \gamma$	$\sum_{i=1}^m s_{ai} s_{bi} p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{bj} p_i p_j \leq \gamma$	quadratique
(3b)	$k\gamma \leq c_{a,b}(k)$	$\gamma \leq \left(S(\operatorname{diag}\vec{p} - \vec{p}\vec{p}^t)S^t\right)_{ab}$	$\gamma \leq \sum_{i=1}^m s_{ai} s_{bi} p_i - \sum_{1\leq i,j\leq m} s_{ai} s_{bj} p_i p_j$	quadratique
(4a)	$\lim_k \rho_{a,b}(k) = \gamma$	$(S^t\vec{p})_a = \gamma(S^t\vec{p})_b$	$\sum_{i=1}^m (s_{ia} - \gamma s_{ib}) p_i = 0$	linéaire
(4b)	$\lim_k \rho_{a,b}(k) \leq \gamma$	$(S^t\vec{p})_a \leq \gamma(S^t\vec{p})_b$	$\sum_{i=1}^m (s_{ia} - \gamma s_{ib}) p_i \leq 0$	linéaire
(4c)	$\gamma \leq \lim_k \rho_{a,b}(k)$	$(S^t\vec{p})_a \geq \gamma(S^t\vec{p})_b$	$\sum_{i=1}^m (s_{ia} - \gamma s_{ib}) p_i \geq 0$	linéaire

TABLE 4.2 – Nouvelle table de contraintes permettant de prendre en compte des mesures de rapport de taux de production

4.4 Conclusion

Dans le chapitre 3, nous avons introduit la dynamique de Bernoulli, une approximation de la dynamique stochastique en état stationnaire. Nous avons montré qu'il est possible d'analyser cette dynamique tout d'abord en obtenant des expressions analytiques des moments d'ordre 1 et 2 en fonction de la stœchiométrie S et des réactions de probabilité $\vec{p}$. Nous avons aussi déterminer le comportement asymptotique gaussien de cette dynamique et nous avons explicité les paramètres de la loi limite. Ce chapitre nous a permis de présenter plusieurs exemples d'applications de cette approximation.

L'application la plus importante est l'approche par contraintes qui correspond à une extension probabiliste des méthodes de contraintes de flux. Alors que les méthodes de flux traitent des données moyennes, c'est-à-dire à l'échelle de la population, notre approche permet de considérer l'échelle de la cellule individuelle en intégrant non seulement les informations de moyennes mais aussi les moments d'ordre deux (variances et co-variances croisées) des sorties. Ainsi, nous pouvons intégré la variabilité intrinsèque des productions à l'échelle de l'individu. Nous avons également résolu l'exemple de réfutation proposé en introduction de la thèse où il était nécessaire de prendre en compte les informations de covariances des données à disposition. Un autre exemple nous a permis d'explorer d'autres pistes d'application de l'approches par contraintes : nous avons vu que dans certain cas les contraintes obtenues à l'aide d'informations sur certaines sorties du système pouvait nous aider à contraindre le comportement d'autres sorties. Toute ces approches sont d'autant plus intéressantes qu'elles n'ont jamais eu besoin d'utiliser d'autres informations que la stœchiométrie du réseau.

Cependant, les nouvelles contraintes obtenues pour les moments d'ordre deux ne sont plus linéaires mais quadratiques (et non nécessairement quadratiques positives) et ne peuvent plus être résolues grâce à la programmation linéaire (par exemple par l'utilisation de l'algorithme du simplexe [DOW+55]). Ainsi, notre nouvelle approche par contraintes ouvre

de nouvelles perspectives d'applications des méthodes informatiques de traitement des contraintes quadratiques et d'optimisation quadratique. Par exemple une possibilité intéressante est d'utiliser certains algorithmes de géométrie algorithmiques tels que [BEH+05] pour obtenir des calculs exacts de l'ensemble des solutions d'un jeu de contraintes.

Un autre cadre d'application est celui de l'obtention d'ellipsoïdes de confiance. Pour cela, nous avons eu besoin d'étudier plus en détails la loi limite gaussienne de la dynamique de Bernoulli. Nous avons obtenu un théorème qui caractérise entièrement les cas dégénérés de la loi limite. Cette approximation permet de répondre à des problèmes de confrontation de données (trajectoires) avec une loi de Bernoulli de paramètres connus. Ces ellipsoïdes de confiance nous ont permis aussi de démontrer un théorème de convergence sur le rapport de taux de production de deux sorties d'un réseau. Ce résultat est important car il permet d'étoffer notre tableau de contraintes. En effet, on est maintenant capable à l'aide d'un rapport de taux de production mesuré chez un $un\ seul\ individu$ d'obtenir des contraintes sur les probabilités de réaction $\vec{p}$. De plus nous avons vu que ce rapport de taux de productions est facile à mesurer d'un point de vue expérimental.

Le dernier chapitre de cette thèse fournit un cadre plus formel à toutes ces méthodes. On développera un langage logique qui permet d'encoder soit des mesures faites sur les moments d'ordre 1 et 2, soit des hypothèses sur ces moments et de les traduire de façon automatique en ensembles de contraintes linéaires, quadratiques ou polynomiales. La résolution de ces systèmes permettra alors de réfuter des réseaux ou de vérifier formellement les propriétés de certaines sorties à partir d'informations sur d'autres sorties.

Chapitre 5

Vérification de propriétés asymptotiques sur les réseaux stationnaires

Les travaux présentés dans ce chapitre ont été publiés dans les actes du workshop international Bioinformatics and Artificial Intelligence *(BAI) de la conférence IJCAI2015 [PBS15].*

Dans les deux chapitres précédents, nous avons proposé d'utiliser une source de Bernoulli pour modéliser le comportement stationnaire stochastique d'un réseau de réaction. Ce modèle a l'avantage de posséder un petit nombre de paramètres, les probabilités de réaction, qui agrègent les valeurs des constantes cinétiques et des quantités de matière à l'équilibre. Ces probabilités sont l'analogue des flux à l'équilibre du FBA et peuvent être analytiquement reliées aux valeurs de l'espérance et des covariances. Grâce à ce lien, nous avons pu convertir des informations expérimentales en systèmes de contraintes sur les probabilités de réactions dont la résolution permet éventuellement de rejeter le réseau de réaction obtenu.

Dans ce chapitre, nous formalisons cette approche grâce au langage de la logique. Nous nous proposons d'exprimer les informations expérimentales sous forme d'une formule logique qui pourra ou non être $satisfaisable$ vis à vis du réseau de réactions considéré. Lorsqu'elle n'est pas satisfaisable le réseau de réaction peut être réfuté sous réserve de validité de l'approximation de Bernoulli. De plus la logique permet d'exprimer non seulement des observations mais aussi des questions biologiques comme par exemple :

— est-ce que le fait de savoir que le taux de production de A est en moyenne inférieur à x permet de dire que le taux de production moyen de B est aussi inférieur à x ?

— est-ce que le fait de savoir que le taux de production de A est en moyenne inférieur à x et que A et B sont anti-corrélés permet de dire que le taux de production moyen de C est supérieur à y ?

On montrera ensuite comment les modèles de ces formules correspondent aux solutions

de systèmes de contraintes polynomiales et comment cette correspondance peut-être obtenue de façon automatique. On peut alors d'une part vérifier la satisfaisabilité de formules correspondant à des observations et d'autre part vérifier la validité de formules correspondant à des hypothèses. Lorsque les formules ne sont pas valides, nous pouvons également proposer un contre-exemple, c'est-à-dire un choix de constantes stochastiques qui aboutit à un régime stationnaire ne satisfaisant pas la formule testée.

L'introduction de cette logique a donc un triple objectif :

— **formaliser les données** expérimentales asymptotiques observées dans un langage compréhensible par l'homme et la machine,

— **formaliser des hypothèses biologiques** dans le même langage sur les moments d'ordre 2 des sorties,

— **vérifier de manière automatique** la satisfaisabilité ou la véracité des formules par rapport à un réseau de réaction donné.

5.1 Syntaxe et sémantique

On définit dans un premier temps la syntaxe de notre logique. Nous définissons le langage des formules à l'aide de règles d'induction. Intuitivement les formules seront des égalités ou inégalités polynomiales sur les moments d'ordre 1 ou 2. On introduira ensuite la sémantique des formules qui se définit comme l'espace des vecteurs de probabilités $\vec{p}$ pour lesquels la dynamique de Bernoulli associée satisfait la formule considérée.

5.1.1 Syntaxe

Définition des termes Le but à terme de la syntaxe est d'exprimer des propriétés asymptotiques sur $(\vec{y}(k))_{k\in\mathbb{N}}$ et en particulier on souhaite comparer asymptotiquement des expressions polynomiales des espérances et des variances-covariances de la trajectoire $(\vec{y}(k))$ en état stationnaire. Ces expressions polynomiales seront les *termes* de cette logique. On notera dans cette section $\mathscr{C} = \{X_1, \ldots, X_n\}$ l'ensemble fini des symboles d'espèces chimiques. L'algèbre des *termes* est définie par induction structurelle comme le plus petit ensemble $\mathscr{T}$ vérifiant les propriétés suivantes :

$$\forall X, Y \in \mathscr{C},\ \mathbb{E}(X) \in \mathscr{T},\ \mathrm{Var}(X) \in \mathscr{T},\ \mathbb{C}\mathrm{ov}(X, Y) \in \mathscr{T}, \tag{5.1}$$

$\forall \lambda \in \mathbb{Q},\ \forall T_1, T_2 \in \mathscr{T},\ \lambda \in \mathscr{T},\ (\lambda \cdot T_1) \in \mathscr{T}, (T_1 + T_2) \in \mathscr{T}, (T_1 \times T_2) \in \mathscr{T}$. Pour le moment, $\mathbb{E}$, Var et $\mathbb{C}$ov sont simplement des symboles de fonctions, c'est-à-dire des éléments syntaxiques, dont on définira la sémantique dans la suite.

Exemple 5.1. $(\mathrm{Var}(X_1) + \mathbb{C}\mathrm{ov}(X_3, X_4))$ et $((3 \cdot \mathbb{E}(X_1)) \times \mathrm{Var}(X_2))$ sont des termes.

Définition des formules Nous pouvons maintenant définir la syntaxe des formules qui consistent en une comparaison de deux termes, c'est-à-dire une comparaison de deux expressions polynomiales des espérances, variances et covariances de $(\vec{y}(k))_k$. Afin d'alléger la définition formelle de la syntaxe et par la suite de la sémantique, on introduira comme seules *formules atomiques* la comparaison d'une expression polynomiale avec 0 (qui permet par la suite d'exprimer la comparaison de deux expression polynomiales) :

$$\text{Formules atomiques } : \mathscr{AF} = \{(T \geq 0) \ / \ T \in \mathscr{T}\}. \tag{5.2}$$

Les formules consistent en une association de formules atomiques à l'aide des connecteurs logiques usuels (et, ou, non). Formellement, l'ensemble des *formules* est défini par induction structurelle comme le plus petit ensemble $\mathscr{F}$ vérifiant : $\mathscr{AF} \subset \mathscr{F}$ and $\forall F_1, F_2 \in \mathscr{F}, \neg F_1 \in \mathscr{F}, (F_1 \vee F_2) \in \mathscr{F}, (F_1 \wedge F_2) \in \mathscr{F}$. Les formules atomiques $\mathscr{AF}$ ainsi que ces trois opérateurs logiques sont suffisants pour exprimer les comparaisons usuelles ainsi que d'autres connecteurs logiques couramment utilisés, nous les introduisons en tant que notations : $\forall T_1, T_2 \in \mathscr{T}, \forall F_1, F_2 \in \mathscr{F}, (T > 0) \equiv \neg((-1 \cdot T) \geq 0), (T_1 \geq T_2) \equiv ((T_1 + (-1 \cdot T_2)) \geq 0), (T_1 > T_2) \equiv ((T_1 + (-1 \cdot T_2)) > 0), (F_1 \rightarrow F_2) \equiv (\neg F_1 \vee F_2)$ et $(T_1 = T_2) \equiv ((T_1 \geq T_2) \wedge (T_2 \geq T_1))$.

Exemple 5.2. $\mathbb{E}(X_1) \geq (3 \cdot \mathbb{E}(X_2))$ est une formule.

5.1.2 Sémantique

Utilisation de l'approximation de la dynamique de Bernoulli La sémantique de notre logique repose en grande partie sur le sens donné aux symboles de fonctions correspondants aux moments (espérances, variances, covariances) du processus stochastique $(\vec{y}(k))_{k \in \mathbb{N}}$. On souhaite en effet utiliser une interprétation cohérente avec la dynamique stochastique discrétisée définie au chapitre 3, c'est-à-dire correcte vis-à-vis de l'algorithme de Gillespie. On se repose pour cela sur l'approximation de Bernoulli décrite dans le chapitre 3 décrivant correctement, sous certaines conditions, le régime stationnaire du système. Concrètement, cela signifie qu'on considère dans la suite sur $\mathbb{E}\,\vec{y}(k) = \mathbb{E}\,\vec{z}(k)$ et $\mathbb{C}\mathrm{ov}\,\vec{y}(k) = \mathbb{C}\mathrm{ov}\,\vec{z}(k)$ où $\vec{z}(k)$ est la dynamique de Bernoulli de même état initial que $\vec{y}(k)$ et de paramètre $\vec{\mathfrak{p}}$. On rappelle que ce paramètre est un vecteur de probabilités de dimension m appelé *vecteur des probabilités de réactions stationnaires* représentant les probabilités d'occurrences de chaque réaction dans le régime stationnaire. On notera $\mathscr{P}_m$ l'ensemble des vecteurs de probabilités de dimension m c'est-à-dire les vecteurs $\vec{u} \in \mathbb{R}^m$ vérifiant $\forall i \in \{1, \ldots, m\}, 0 \leq u_i \leq 1$ and $\sum_{i=1}^m u_i = 1$. Ainsi $\vec{\mathfrak{p}} \in \mathscr{P}_m$.

On pourra se reporter au chapitre 3 pour plus de détails sur cette approximation et son domaine de validité. L'analyse de la dynamique de Bernoulli nous a permis (proposition 3.2)

d'en obtenir une expression des moments d'ordre un et deux :

$$\mathbb{E}\, y^a(k) = y^a(0) + k \sum_{j=1}^{m} s_{ja}\mathfrak{p}_j, \tag{5.3}$$

$$\operatorname{Var} y^a(k) = k \sum_{j=1}^{m} s_{ja}^2 \mathfrak{p}_i - k \sum_{1\le j,l\le m} s_{ja}s_{la}\mathfrak{p}_j\mathfrak{p}_l, \tag{5.4}$$

$$\mathbb{C}\mathrm{ov}(y^a(k), y^b(k)) = k \sum_{j=1}^{m} s_{ja}s_{jb}\mathfrak{p}_j - k \sum_{1\le j,l\le m} s_{ja}s_{lb}\mathfrak{p}_j\mathfrak{p}_l. \tag{5.5}$$

Ainsi, on obtient des expressions linéaires en k pour les moments d'ordre 1 et 2 à condition de connaître la valeur du triplet $(S, \vec{y}(0), \vec{\mathfrak{p}})$. Cela nous conduit à définir un environnement d'évaluation pour les moments.

Définition 5.1 (Contexte, interprétation)**.** Un *contexte* est un couple $C = (S, \vec{y}(0))$ où $\vec{y}(0) \in \mathbb{N}^n$ représente les conditions initiales et S une matrice de stœchiométrie de taille $m \times n$. Une *interprétation* est un triplet $I = (S, \vec{y}(0), \vec{p})$ où $(S, \vec{y}(0))$ est un contexte et $\vec{p} \in \mathscr{P}_m$ est un vecteur de probabilités de réactions.

Évaluation des termes Lorsque qu'un contexte est fourni, les termes peuvent être évaluées en tant que polynômes multivariés en la variable de temps k et les variables de probabilités de réactions $\vec{p} = (p_i)_{0\le i\le m}$. L'évaluation des symboles de fonctions (les feuilles de la syntaxe des termes) lorsque $\vec{p} := \vec{\mathfrak{p}}$ correspond alors aux polynômes $\mathbb{Q}[k]$ donnés dans les équations (5.3), (5.4) et (5.5).

Définition 5.2 (Évaluation des termes)**.** L'*évaluation* $[T]_C$ d'un terme dans le contexte $C = (S, \vec{y}(0))$ est le polynôme de $\mathbb{Q}[k, p_1, \ldots, p_m]$ défini par induction structurelle par

$$[\mathbb{E}(X_a)]_C = y^a(0) + k \sum_{j=1}^{m} s_{ja}p_j,$$

$$[\mathrm{Var}(X_a)]_C = k\left(\sum_{j=1}^{m} s_{ja}^2 p_i - \sum_{1\le j,l\le m} s_{ja}s_{la}p_jp_l\right),$$

$$[\mathbb{C}\mathrm{ov}(X_a, X_b)]_C = k\left(\sum_{j=1}^{m} s_{ja}s_{jb}p_j - \sum_{1\le j,l\le m} s_{ja}s_{lb}p_jp_l\right),$$

$[c]_C = c$ lorsque c est une constante rationnelle, $[(\lambda \cdot T)]_C = \lambda \cdot [T]_C$, $[(T_1 + T_2)]_C = [T_1]_C + [T_2]_C$, $[(T_1 \times T_2)]_C = [T_1]_C[T_2]_C$.

Le choix de la sémantique est justifié par la proposition suivante qui énonce que $[T]_C$ correspond bien à l'interprétation mathématique usuelle d'une expression polynomiale des moments de $\vec{y}(k)$ lorsque $\vec{p} = \vec{\mathfrak{p}}$. La preuve est directe par induction structurelle.

Proposition 5.1. *Soit un réseau de réaction de matrice stœchiométrique S, d'état initial $\vec{y}(0)$ et de probabilités stationnaires de réactions $\vec{\mathfrak{p}}$ et soit un terme T (i.e. une expression polynomiale des moments d'ordre 1 et 2). On note $u(k)$ l'interprétation mathématique usuelle de*

T en tant que polynôme d'espérances, variances et covariances de $\vec{y}(k)$, alors sous réserve de validité de l'approximation de Bernoulli, on a $[T]_C(k,\vec{\mathfrak{p}}) = u_k$.

Évaluation et modèles des formules

Définition 5.3 (Évaluation des formules)**.** L'évaluation $[F]_C$ d'une formule F dans le contexte $C = (S, \vec{x}_0)$ est le sous-ensemble de $\mathscr{P}_m$ défini par induction structurelle par

$$[(T \geq 0)]_C = \{\vec{p} \in \mathscr{P}_m : \mathrm{dom}_k([T]_C) \geq 0\}, \tag{5.6}$$

où $\mathrm{dom}_k(P) \in \mathbb{Q}[p_1, \ldots, p_m]$ est le polynôme associé au monôme de plus haut degré en k dans le polynôme multivarié $P \in \mathbb{Q}[k, p_1, \ldots, p_m]$, $[\neg F]_C = \mathscr{P}_m \setminus [F]_C$, $[(F_1 \vee F_2)]_C = [F_1]_C \cup [F_2]_C$, $[(F_1 \wedge F_2)]_C = [F_1]_C \cap [F_2]_C$.

Ainsi, l'évaluation des formules atomiques $(T \geq 0)$ est l'ensemble des vecteurs de probabilités $\vec{p} \in \mathscr{P}_m$ tels que $k \mapsto [T]_C(k, \vec{p}) \in \mathbb{Q}[k]$ est asymptotiquement positif. En effet, le comportement asymptotique (positif ou négatif à l'infini) d'un polynôme est entièrement déterminé par le signe de son coefficient dominant.
À partir de cette définition, nous sommes en mesure de définir les modèles d'une formule. Une interprétation $I = (S, \vec{y}(0), \vec{p})$ est un *modèle* d'une formule F, noté

$$I \models F, \quad \text{si } \vec{p} \in [F]_{(S,\vec{y}(0))}. \tag{5.7}$$

Une formule F est *valide*, noté $\models F$, si toute interprétation est un modèle de F. Une formule F est *valide dans le contexte* $C = (S, \vec{y}(0))$, noté $C \models F$, si $\forall \vec{p} \in \mathscr{P}_m$, $(S, \vec{y}(0), \vec{p}) \models F$. Une formule F est *satisfaisable dans le contexte* $C = (S, \vec{y}(0))$, s'il existe $\vec{p} \in \mathscr{P}_m$ tel que $(S, \vec{x}_0, \vec{p}) \models F$.

En choisissant cette définition des modèles, on déduit que les modèles $(S, \vec{y}(0), \vec{p})$ des formules atomiques correspondent aux réseaux de réactions de matrice stœchiométrique S, d'état initial $\vec{y}(0)$ et de probabilités stationnaires de réactions $\vec{p}$ qui satisfont *asymptotiquement* (lorsque $k \to +\infty$)), la propriété énoncée dans la formule. Cela est formellement énoncé dans la proposition suivante.

Proposition 5.2. *L'interprétation $I = (S, \vec{y}(0), \vec{p})$ est un modèle de $F = (T \geq 0)$ si et seulement si*

$$\exists K \in \mathbb{N}, \quad \forall k \geq K, \quad [T]_{(S,\vec{y}(0))}(k, \vec{p}) \geq 0. \tag{5.8}$$

Ainsi, d'après la proposition 5.1, dire qu'une interprétation $I = (C, \vec{p})$ est un modèle d'une formule de comparaison $(F \leq G)$ signifie que pour les réseaux correspondants à cette interprétation, l'inégalité entre expressions polynomiales des moments est toujours vraie à partir d'un rang K.

Démonstration. Notons $f(x) = [T]_C(x, \vec{p})$ définie pour $x \in \mathbb{R}$. La fonction f est un polynôme dans $\mathbb{Q}[x]$ donc il a un ensemble restreint de comportements asymptotiques possibles en

termes de signe : soit f est constant, soit $\lim_{+\infty} f = +\infty$, soit $\lim_{+\infty} f = -\infty$. Si $I \models F$ alors par définition $\mathrm{dom}_k([T]_C) \geq 0$ ce qui signifie que soit f est une constante positive soit f est non constante avec un coefficient dominant positif. Dans les deux cas, l'équation (5.8) est vérifiée. Réciproquement, si l'équation (5.8) est vérifiée alors soit f est constant soit $\lim_{+\infty} f = +\infty$, donc $\mathrm{dom}_k([T]_C) \geq 0$, donc $I \models F$. □

Nos définitions permettent de distinguer les formules valides qui sont toujours vraies indépendamment du système étudié (par exemple $(7 \geq 5)$ ou $(\mathbb{E}(X_1) \geq 2\mathbb{E}(X_1))$) et les formules valides dans un contexte qui représentent des propriétés conséquence de la topologie et de la stœchiométrie du réseau considéré. En effet les formules valides dans un contexte correspondent à des propriétés asymptotiques quel que soit le vecteur de probabilité choisi et en particulier pour le vecteur $\vec{\mathfrak{p}}$. Une conséquence importante de la dernière proposition est que si une propriété asymptotique F est observée, et qu'on suppose que le système atteint un régime stationnaire alors la formule F doit être satisfaisable (avec comme vecteur de probabilités de réaction $\vec{\mathfrak{p}}$). Cette dernière remarque est très importante car *elle permet de réfuter* un contexte $(S, \vec{y}(0))$, et donc de rejeter S si on connait $\vec{y}(0)$, dans le cas où la formule F codant des observations expérimentales du régime stationnaire n'est pas satisfaisable dans le contexte considéré. Ceci nous amène naturellement à traiter le problème de la satisfaisabilité des formules dans un contexte donné.

5.2 Satisfaisabilité et validité

Nous avons défini la validité et la satisfaisabilité dans un contexte d'une formule. La prochaine étape est de concevoir un algorithme pour déterminer si une formule est valide ou satisfaisable dans un contexte donné. Le lemme suivant montre que ces deux notions se déduisent l'une de l'autre ce qui nous permet de nous concentrer sur le problème de la satisfaisabilité.

Lemme 5.1. — *Une formule F est valide dans un contexte C si et seulement si $[F]_C = \mathscr{P}_m$.*

— *Une formule F est satisfaisable dans un contexte C si et seulement si $[F]_C \neq \varnothing$.*

— *Dans un contexte C, une formule F est valide (resp. satisfaisable) si et seulement si $\neg F$ n'est pas satisfaisable (resp. pas valide).*

Complexité théorique La proposition suivante montre que décider la satisfaisabilité dans un contexte est NP-difficile. La preuve repose sur un réduction à partir de 3-SAT.

Proposition 5.3. *Le problème de décision suivant, noté $\mathscr{F}$-SAT, est NP-difficile.*
Le problème $\mathscr{F}$-SAT
Instance : *n (nombre d'espèces), m (nombre de réactions), S ($n \times m$ matrice stœchiométrique), $\vec{y}(0)$ (quantités initiales), F (une formule).*
Question : *Existe-t-il $\vec{p} \in \mathscr{P}_m$ tel que $(S, \vec{x}_0, \vec{p}) \models F$?*

Ainsi, il n'existe pas d'algorithme polynomial qui permet de vérifier la satisfaisabilité d'une formule pour un réseau (et même la validité en vertu du lemme 5.1) en temps polynomial (à moins que P = NP).

Démonstration. La preuve est obtenue par réduction en temps polynomial du problème de décision suivant.
Le problème 3-SAT [GJ02]
Instance : n (nombre de variables), une formule de la logique propositionnelle sous forme normale conjonctive (CNF) $\varphi = \bigwedge_{i=1}^{r}(l_i^1 \vee l_i^2 \vee l_i^3)$, où $l_i^{1/2/3}$ sont des littéraux.
Question : Existe-t-il une valuation qui satisfait φ ?
Nous allons construire une réduction en temps polynomial de 3-SAT. Considérons une formule sous forme CNF $\varphi = \bigwedge_{i=1}^{r}(l_i^1 \vee l_i^2 \vee l_i^3)$ et notons $\{x_1, \dots, x_n\}$ les n variables de φ. Pour toute formule d'entrée de 3-SAT φ, nous associons une formule de notre logique $F = \bigwedge_{i=1}^{r}(G_i^1 \vee G_i^2 \vee G_i^3)$ définie par $G_i^q = (\mathbb{E}(X_k) > 0)$ si $l_i^q = x_k$ et $G_i^q = \neg(\mathbb{E}(X_k) > 0)$ if $l_i^q = \neg x_k$. Cette formule F se construit en temps polynomial par rapport à la taille de φ. On prouve ensuite que φ est satisfaisable si et seulement F est satisfaisable.

— Soit $v : \{x_1, \dots, x_n\} \to \{\top, \bot\}$ une valuation satisfaisant φ. Construisons un modèle de F. Pour cela, on considère un réseau de réaction à n espèces et $n+1$ réactions $\{R_0 : \varnothing \to \varnothing, R_i : \varnothing \to X_i (i = 1 \dots n)\}$ de matrice de stœchiométrie S. On considère les probabilités de réactions $p_k = 1/n$ si $v(x_k) = \top$, $p_k = 0$ si $v(x_k) = \bot$ et $p_0 = 1 - \sum_{k=1}^{n} p_k$. On définit également les conditions initiales $\vec{y}(0) = \vec{0}$. Considérons l'interprétation $I = (S, \vec{y}(0), \vec{p})$. Alors $I \models (\mathbb{E}(X_k) > 0) \Leftrightarrow p_k > 0 \Leftrightarrow v(x_k) = \top$ et $I \models \neg(\mathbb{E}(X_k) > 0) \Leftrightarrow p_k \leq 0 \Leftrightarrow v(x_k) = \bot$. Par conséquent, $I \models F$.
— La réciproque est triviale, si $I = (S, \vec{x}_0, \vec{p}) \models F$ alors on définit la valuation $v(x_k) = \top$ si $I \models (\mathbb{E}(X_k) > 0)$ et $v(y(k)) = \bot$ si $I \models \neg(\mathbb{E}(X_k) > 0)$. Par définition de $\models$ il suit que v satisfait φ.

□

Une procédure de décision de $\mathscr{F}$-SAT Nous avons prouvé que la décision de $\mathscr{F}$-SAT est NP-difficile, néanmoins il est toujours intéressant de concevoir un algorithme pour résoudre ce problème. En effet il est toujours possible de trouver des méthodes rapides en pratiques et lentes sur quelques rares instances du problème. Nous proposons l'algorithme 1 suivant pour déterminer $\exists ? \vec{p} \in \mathscr{P}_m,\ (S, \vec{x}_0, p) \models F$.

Dans l'étape 2, trouver $\vec{p} \in [(T_u^q \geq 0)]_C$ (resp. $[\neg(T_u^q \geq 0)]_C$), consiste à trouver une solution $\vec{p}$ telle que $(\mathrm{dom}_k[T]_{(S,\vec{y}(0))})(\vec{p}) \geq 0$ (resp. < 0), c'est-à-dire trouver une solution d'une inégalité polynomiale. Ainsi, l'étape 2 consiste à trouver une solution d'un système de n_u contraintes polynomiales. Décider si une telle solution existe est décidable comme l'a démontré [Tar51]. Une solution peut être obtenue grâce aux récents outils du domaine du model-checking comme les SMT-solvers tels que `dReal` [GKC13].
Notre algorithme a deux sources de complexité. Premièrement, convertir F en DNF dans l'étape 1 peut-être intense en calculs puisque la taille de la DNF peut être exponentielle par rapport à la taille de F et ainsi r peut être très grand. Cependant, dans les cas d'utilisation

Algorithm 2: Deciding $\mathscr{F}$-SAT

Data: Un contexte $C = (S, \vec{x}_0)$, une formule F
Result: $\vec{p}$ tel que $(C, \vec{p}) \models F$ ou UNSAT
Étape 1 : Convertir F en forme normale disjonctive (DNF);

$$F = \bigvee_{u=1}^{r} F_u = \bigvee_{u=1}^{r} (G_u^1 \wedge \cdots \wedge G_u^{n_u})$$

for $u = 1$ *to* r **do**
  Essayer Étape 2 : trouver $\vec{p} \in [F_u]_{(S,\vec{y}(0))}$;
  if $\vec{p}$ *est trouvé* **then**
    renvoyer $\vec{p}$;
  end
end
renvoyer UNSAT;

courants les formules sont assez simples et contiennent peu de clauses. Deuxièmement, résoudre l'étape 2, c'est-à-dire trouver une solution d'un système de contraintes polynomiales peut être lourd en calculs.

Les termes sans multiplications conduisent à des contraintes quadratiques Nous avons proposé une procédure pour déterminer la satisfaisabilité d'une formule F en se reposant sur la résolution de systèmes d'inégalités polynomiales. Résoudre de tels systèmes est difficile en général. Pour simplifier cet aspect, on peut restreindre l'expressivité de la logique en considérant que les formules ne peuvent décrire que des expressions linéaires des espérances et covariances. Formellement, on définit un ensemble restreint de termes $\mathscr{T}_{\text{lin}} \subset \mathscr{T}$ où la règle de construction par multiplication a été enlevée : $\forall T_1, T_2 \in \mathscr{T},\ (T_1 \times T_2) \in \mathscr{T}$. On définit ensuite $\mathscr{F}_{\text{lin}} \subset \mathscr{F}$ de la même manière que $\mathscr{F}$ mais en utilisant exclusivement les termes dans $\mathscr{T}_{\text{lin}}$.

Proposition 5.4. *Soit un contexte C et un terme linéaire $T \in \mathscr{T}_{lin}$ alors trouver $\vec{p} \in \mathscr{P}_m$, $\vec{p} \in [(T \geq 0)]_C$ peut être réalisé en résolvant une inéquation quadratique en les variables (p_i).*

Démonstration. La preuve consiste à démontrer par induction structurelle sur les termes que pour tout $T \in \mathscr{T}_{\text{lin}}$, le degré total des variables (p_i) dans $[T]_C$ est au plus 2. Cette induction fonctionne car les feuilles $\mathbb{E}(\cdot)$ ont pour sémantique des polynômes de degré total au plus 1 en les (p_i) et les feuilles $\text{Var}(\cdot)$ et $\mathbb{C}\text{ov}(\cdot,\cdot)$ ont pour sémantique des polynômes de degré total au plus 2 en les (p_i). De plus, sommer des termes ou les multiplier par un scalaire ne fait pas augmenter le degré total en les (p_i). □

Puisque la multiplication des termes n'intervient pas dans la preuve de la proposition 5.3, le problème de satisfaisabilité pour $\mathscr{F}_{\text{lin}}$ est lui aussi NP-difficile. Cependant, l'utilisation de l'algorithme de décision proposé s'avère plus simple car la seconde source de complexité, à

savoir la résolution de systèmes d'inéquations polynomiales, est maintenant réduit à la résolution d'un système d'inéquations quadratiques. Il existe plusieurs outils qui permettent d'aborder de telles résolutions, par exemple le problème du *second-order cone programming* (SOCP) [AG03] peut être résolu par des méthodes de points intérieurs en utilisant des outils tels que `CPLEX` ou `Gurobi`.

5.3 Exemple

À titre d'illustration, on se propose toujours de résoudre formellement l'exemple introductif consistant à réfuter l'un des modèles (équations (3)) du premier chapitre. On rappelle que l'utilisation de la covariance des données était nécessaire pour réfuter le modèle 2 et par conséquent la réfutation reposant sur les méthodes de flux classiques ne fonctionnait pas. On considère que les données expérimentales à disposition sont représentées sur la figure la première ligne de la Figure 1 (ou au choix sur la Figure 4.2). Aussi, nous supposons qu'un état stationnaire a été atteint et qu'en conséquence les productions et consommations de A et B sont équilibrées en moyenne. On peut encoder les informations disponibles sous cette forme

$$F = (\mathbb{E}(A) = 1000) \wedge (\mathbb{E}(B) = 1000) \wedge (\mathbb{E}(D) \geq 2\mathbb{E}(C)) \wedge (\mathbb{C}\mathrm{ov}(C, D) < 0). \tag{5.9}$$

À partir de là, on souhaite discriminer les deux réseaux de réactions proposés, on introduit donc deux contextes $C_1 = (S_1, \vec{y}(0))$ et $C_2 = (S_2, \vec{y}(0))$ associés à chaque réseau de réactions en vue de vérifier la satisfaisabilité de F dans chacun de ces contextes. Dans ce cas, nous connaissons les quantités initiales $\vec{y}(0) = (1000, 1000, 0, 0)$ et F est déjà sous forme DNF. En appliquant la sémantique des formules, on obtient un système de contraintes polynomiales pour chacun des contextes.

formules atomiques	contexte C_1	contexte C_2
$(\mathbb{E}(A) = 1000)$	$p_2 - p_1 = 0$	$p_2 - p_1 = 0$
$(\mathbb{E}(B) = 1000)$	$p_1 - p_2 = 0$	$p_1 - p_2 = 0$
$(\mathbb{E}(D) \geq 2\mathbb{E}(C))$	$p_2 \geq p_1$	$0 \geq 0$
$(\mathbb{C}\mathrm{ov}(C, D) < 0)$	$-2p_1p_2 < 0$	$2p_1(1 - p_1) < 0.$

Comme prévu les contraintes sont quadratiques au plus car la multiplication des termes n'est pas utilisée dans F. Le premier système de contraintes admet une (unique) solution $\vec{p} = (1/2, 1/2)$ tandis que le second system n'admet pas de solution. Par conséquent, F est satisfaisable dans le contexte C_1 mais pas dans le contexte C_2.

$$\exists \vec{p} \in \mathscr{P}_m, (C_1, \vec{p}) \models F \qquad \nexists \vec{p} \in \mathscr{P}_m, (C_2, \vec{p}) \models F.$$

Ainsi, les propriétés stationnaires observées et encodées dans F ne peuvent pas être obtenues en utilisant le second réseau de réactions (modèle 2) qui peut être réfuté.

5.4 Conclusion

Nous avons introduit une logique dont la syntaxe permet d'exprimer des propriétés sur les moments asymptotiques d'ordre 1 et 2 d'un réseau de réactions en régime stationnaire. La sémantique des formules est obtenue grâce aux formules analytiques des moments d'ordre 1 et 2 de la dynamique de Bernoulli obtenues dans le chapitre 3. Un modèle d'une formule est formé d'un réseau de réaction, de quantités initiales et de probabilités de réactions qui satisfont asymptotiquement la formule lorsqu'on y remplace les moments par leur expression analytique. Lorsqu'une formule, codant des informations expérimentales, n'est pas satisfaisable dans un contexte alors ce contexte et en particulier le réseau de réaction peut-être réfuté. Ainsi nous aboutissons encore une fois à une méthode de réfutation reposant sur l'utilisation des moments d'ordre 1 et 2 mais effectuée cette fois-ci dans un cadre logique très formel.

Après avoir introduit la logique nous avons démontré que le problème $\mathscr{F}$-SAT, essentiel pour réfuter des réseaux de réactions, est NP-difficile. Nous avons tout de même proposé un algorithme élémentaire reposant sur l'utilisation de la DNF et la résolution de systèmes de contraintes polynomiales. D'un point de vue informatique cela créé de nombreuses perspectives soit d'applications de solveurs existants et efficaces pour répondre à des questions de réfutation, soit pour le développement de solveurs spécifiques pour les instances de $\mathscr{F}$-sat. Une perspective d'approfondissement est par exemple d'étudier plus précisément quelles instances de $\mathscr{F}$-SAT peuvent être résolues rapidement et comment. Cela nécessite de tester notre algorithme de résolution sur des couples de réseaux de réactions et jeux de données, en utilisant différentes tailles et des formules (et donc des contraintes) de différentes formes.

Conclusion

Contexte

Dans le vaste paysage dessiné par la recherche en informatique, les résultats de cette thèse trouvent leur place dans le domaine de la *bio-informatique*, c'est-à-dire de l'informatique au service de la biologie et plus précisément dans le cas présent *la biologie des systèmes*. Cette branche de la biologie a pour but d'étudier le *comportement dynamique* de systèmes vivants complexes, intégrant des agents à diverses échelles spatiales (gènes, protéines, cellules, tissus, *etc*) et fonctionnant à des échelles de temps variées.

Dans ce domaine les modèles sont souvent décrits dans le langage des *réseaux de réactions*, c'est-à-dire par un ensemble fini de *règles* de transformation décrivant les actions élémentaires possibles du système en termes de consommations et de productions d'espèces. D'un point de vue informatique, un réseaux de réactions n'est donc pas très différent de la donnée d'un jeu fini d'instructions permettant de modifier l'état du système. Pour obtenir une modélisation dynamique, c'est-à-dire une description de l'évolution temporelle de l'état du système, soit les quantités de chacune de ses espèces, il convient de définir une *sémantique* de ces réseaux. La littérature scientifique décrit aujourd'hui deux grandes classes de sémantiques : les *sémantiques déterministes* reposant sur les équations différentielles et les *sémantiques probabilistes* reposant sur les chaînes de Markov.

La modélisation dynamique par équations différentielles est la méthode historique, similaire aux équations de mouvements en physique, est la plus répandue dans le monde des biologistes et des modélisateurs. Dans ce cas, les trajectoires sont entièrement déterminées par l'état initial du système. Toutefois, la biologie moderne montre, de plus en plus souvent, l'existence de comportements aléatoires dans le vivant. La modélisation probabiliste également appelée *modélisation stochastique*, qui décrit non plus une trajectoire unique mais décrit une loi de probabilité sur un ensemble de trajectoires possibles, devient de plus en plus pertinente.

Dans tous les cas, il importe de déterminer à la base le bon réseau de réactions, ce qui constitue déjà un problème difficile. Il est souvent construit *à la main* par le biologiste-modélisateur en fonction de ses connaissances ou préconceptions sur le système étudié, de la question biologique posée mais aussi souvent en comblant, automatiquement ou manuellement, un manque de connaissance grâce à l'intégration de réactions, de méca-

nismes, issus ou inspirés d'espèces voisines. C'est par exemple le cas dans l'utilisation de programmes de reconstruction automatique de réseaux métaboliques qui à partir de connaissances initiales incomplètes construisent un réseau fonctionnel en intégrant des réactions supplémentaires à partir d'une données de réactions possibles. Il est donc important de posséder les moyens de *valider* ou au contraire de *réfuter* ces réseaux à partir de données temporelles expérimentales. Un bon moyen serait de comparer les trajectoires expérimentales avec les trajectoires différentielles ou la loi des trajectoires stochastiques. Malheureusement cette approche nécessite de bien connaître les lois d'évolutions de ces sémantiques ainsi que leur paramètre. Cela constitue un problème difficile même pour les petits modèles : l'inférence des paramètres ne passe pas à l'échelle et on ne peut pas traiter le cas des grands réseaux.

Dans le cas de la sémantique différentielle, l'*analyse des flux à l'équilibre* fournit un ensemble de méthodes pour étudier le comportement *stationnaire* d'un réseau de réactions. L'avantage est que ces méthodes passent à l'échelle car elles reposent sur une simple information stœchiométrique qui conduit à un système de contraintes linéaires, facile à manipuler d'un point de vue informatique. Nous avons montré que ces méthodes permettent d'obtenir des contraintes sur les flux à partir d'informations sur les *pentes stationnaires* des trajectoires. À titre d'application, on peut alors réfuter certains réseaux sur la seule base d'informations sur les pentes des trajectoires si les contraintes dérivées sont incompatibles. Cette réfutation peut être vue comme une variante du *flux variability analysis* dans laquelle on détermine qu'aucun vecteur de flux ne permet de produire les pentes stationnaires observées. **La question traitée dans cette thèse est de démontrer qu'une méthode similaire peut être appliquée dans le cadre de la sémantique probabiliste**.

Résultats

L'approximation de Bernoulli En premier lieu nous nous sommes attachés à déterminer un *analogue probabiliste du vecteur de flux* qui pourrait être relié aux *moments la trajectoire* et en particulier aux espérances, variances et co-variances croisées des sorties des réseaux étudiés. En effet, les méthodes de modélisation stochastiques classiques à savoir l'équation chimique maîtresse, l'algorithme de simulation stochastique de Monte-Carlo associé mais aussi les méthodes usuelles dérivées pour l'approximation des moments (la *méthode des moments clos* et *l'approximation de bruit linéaire* ne nous fournissent pas un tel vecteur équivalent. De plus toutes ces méthodes reposent, comme en sémantique déterministe, sur une bonne connaissance des paramètres stochastiques dont on sait qu'elle est difficile à déterminer lorsque la taille des réseaux croît. Nous avons donc décidé de s'inspirer de l'*hypothèse stationnaire* de l'analyse de l'équilibre des flux pour aboutir à cet analogue. Pour cela, nous nous reposons non plus sur la notion de *solution stationnaire* des équations différentielles mais sur la *stationnarité de la distribution* de la chaîne de Markov associée. Plus, on considère le cas où les probabilités de tirages aléatoires des réactions se stabilisent asymptotiquement en moyenne nous fournissant ainsi un *vecteur de probabilités de réactions stationnaires* $\vec{\mathfrak{p}}$. À partir de ce vecteur, nous construisons un processus stochastique

approximatif pour les trajectoires appelé *dynamique de Bernoulli* dans lequel les réactions sont tirées indépendamment selon les probabilités $\vec{\mathfrak{p}}$. Nous notons immédiatement que le vecteur de flux et $\vec{\mathfrak{p}}$ ont même dimensions et significations à savoir le *taux moyen* d'utilisation des réactions. Le vecteur des réactions de probabilités stationnaires est donc un bon candidat pour l'analogue probabiliste recherché. De plus, nous montrons deux propriétés intéressantes de la dynamique de Bernoulli.

1. La possibilité d'obtenir par le calcul des *expressions analytiques des moments d'ordre un et deux*, c'est-à-dire des *espérances, variances et covariances* en fonction de la matrice de stœchiométrie et du vecteur $\vec{\mathfrak{p}}$.
2. L'existence d'un *théorème central limite* qui permet de bien comprendre la nature asymptotique normale de ce processus et dont on a aussi déterminer des expressions analytiques pour les paramètres de la loi gaussienne limite.

Enfin nous avons également étudié la qualité de l'approximation fournie par la dynamique de Bernoulli. Nous nous sommes intéressés à la comparaison entre les espérances et la matrice de covariance de la dynamique de Bernoulli. Dans les deux cas, nous avons montré qu'une erreur se cumule à chaque réaction mais cette erreur tend asymptotiquement vers 0 lorsque le paramètre de la dynamique de Bernoulli est le vecteur $\mathfrak{p}$ (ce qui justifie l'utilisation de ces probabilités). De plus nous avons démontré que si les quantités initiales sont proches d'un point d'équilibre autour duquel les probabilités de tirages sont environ $\vec{\mathfrak{p}}$ alors tous les termes d'erreurs tendent vers 0 à la *limite thermodynamique*. Autrement dit, plus on se rapproche de la limite thermodynamique plus les phénomènes stochastiques sont petits comparés à la taille des populations mais plus ils se comportent comme un tirage de Bernoulli.

Applications, approches par contraintes Dans les cas où la dynamique de Bernoulli est valide, on obtient une description linéaire par rapport au temps des moments d'ordre un et deux. Les pentes de ces droites sont entièrement analytiquement déterminées par les probabilités de réactions $\vec{\mathfrak{p}}$ et la matrice de stœchiométrie. **Le résultat central de cette thèse est donc la possibilité d'obtenir des contraintes sur $\vec{\mathfrak{p}}$ à l'aide d'estimations des espérances, des variances et des co-variances croisées**. Ces mesures sont obtenues à l'aide d'estimateur statistiques classiques appliqués à un ensemble de trajectoires. Notre approche nous permet donc vraiment de traiter des données avec *trajectoires multiples*, issues de *plusieurs individus*, plutôt que de ne considérer que la trajectoire moyenne d'une population. De manière élégante, les contraintes associées à la mesure de la pente des espérances sont identiques aux contraintes de flux obtenues à l'aide de mesure de pentes moyennes. Ainsi notre système de contraintes doit se lire comme une extension des approches classiques par contraintes dans laquelle on est aussi capable d'intégrer des informations concernant les variances, c'est-à-dire la variabilité des données, et les co-variances, c'est-à-dire les corrélations inter-espèces. Toujours dans le cadre de la dynamique de Bernoulli et de cette approche par contraintes, nous avons aussi démontré que le rapport de taux de production entre deux espèces converge en probabilités vers une constante (non aléatoire) qui s'exprime comme un rapport de probabilités de production/consommation des deux espèces. Ce résultat permet d'obtenir d'obtenir d'autres contraintes sur $\vec{\mathfrak{p}}$ mais cette fois ci

sans l'aide d'estimateur. Autrement dit on a démontré mathématiquement qu'on est capable d'obtenir certaines informations sur les probabilités de réactions de la dynamique de Bernoulli en utilisant une mesure asymptotique sur *une seule trajectoire*. Un avantage majeur de ce taux de productions est qu'il est aussi très facile à mesurer expérimentalement car il suffit que les trajectoires expérimentales soient données en unités arbitraires, aussi bien pour l'échelle de temps que pour l'échelle quantitative. Enfin, une application secondaire de nos résultats asymptotiques sur la dynamique de Bernoulli est la possibilité de construire des ellipsoïdes de confiance dans l'espace des phases. Ces ellipsoïdes peuvent être utilisées pour construire un test statistique de cohérence entre un modèle et une trajectoire donnée, lorsqu'on connaît les probabilités de réactions stationnaires $\vec{\mathfrak{p}}$ (cette connaissance peut s'obtenir partiellement à l'aide de notre système de contraintes). Ce test permet donc de valider/réfuter un modèle par rapport à une trajectoire donnée.

Formalisation logique Nous avons introduit un langage logique qui permet d'exprimer les propriétés des moments d'ordre un et deux dans un formalisme compréhensible par la machine. Des formules telles que

$$(\mathbb{E}\,A \leq 3\,\mathbb{E}\,B) \wedge (\mathbb{C}\mathrm{ov}(B,C) < 0) \tag{5.10}$$

peuvent alors être automatiquement transformées en ensemble de systèmes de contraintes qui définissent *l'espace des paramètres* valides pour cette formule, c'est-à-dire l'ensemble des probabilités de réactions stationnaires qui satisfont cette formule pour le réseau considéré. A partir d'une formule φ, encodant une propriété ou une observation, nous avons alors montré qu'on peut déterminer

1. *sa validité* lorsque toutes les probabilités stationnaires $\vec{\mathfrak{p}}$ vérifient φ,
2. *sa possibilité* lorsqu'il existe un vecteur $\vec{\mathfrak{p}}$ vérifiant φ et
3. *son impossibilité* lorsqu'il n'existe pas de probabilités stationnaires φ

Dans ce dernier cas, si la formule φ encode une observation qu'on considère juste, on aboutit à une *réfutation* du réseau de réactions considéré. Un test de validité particulièrement intéressant est celui d'une formule de *conséquence* de forme $A \to B$ qui revient à tester si les vecteurs $\vec{\mathfrak{p}}$ satisfaisant A satisfont aussi B, c'est-à-dire que l'espace solution de A est inclus dans celui de B. Si on prouve la validité de $A \to B$ et qu'on observe expérimentalement A alors on peut en déduire que le régime stationnaire vérifie la propriété B. De telles formules permettent donc de tester des hypothèses sur un réseau sachant certaines observations. Dans tous les cas, sous réserve de validité de la dynamique de Bernoulli, ces réfutations ou ces test d'hypothèses s'obtiennent sans connaissances sur les lois ou les paramètres de la dynamique du réseau. Ce sont en effet des vérifications reposant sur la stœchiométrie. D'un point de vue théorique, on a montré que savoir si une formule est valide (et respectivement non satisfaisable) est un problème NP-difficile. Toutefois, notre procédure pour déterminer la F des formules, qui repose sur la résolution de systèmes de contraintes quadratiques ouvre de nouvelles perspectives d'application pour les méthodes de résolution de contraintes quadratiques.

Perspectives

Nous présentons maintenant quelques pistes de continuation des travaux présentés dans cette thèse.

Application à des données réelles en biologie ou en écologie Afin d'appliquer les méthodes par contraintes présentée dans les chapitres introductifs et décrite dans le chapitre 4 et plus formellement dans le chapitre 5, il est nécessaire d'obtenir en premier lieu des données se présentant sous la forme de multiples séries temporelles où chaque trajectoire est une réalisation du système dynamique aléatoire étudié. Par exemple, si on étudie le comportement aléatoire d'une population de N cellules, on doit obtenir N séries temporelles quantitatives. À partir de ces séries, on peut ensuite en estimer la trajectoire moyenne, la trajectoire des variances, des co-variances, *etc.*
En biologie plusieurs méthodes expérimentales permettent de quantifier le comportement dynamique d'une population d'individus, comme par exemple l'électrophorèse ou des mesures reposant sur la radioactivité [LRML+14]. Récemment, la métabolomique quantitative [DAH07] a montré qu'il est possible grâce à la spectroscopie de masse de quantifier à large échelle une grande diversité de métabolites. Toutefois, ces expériences ne peuvent être interprétées qu'en tant que mesure moyenne de populations d'individus hétérogènes. D'autre part, le domaine du *single-cell biology* étudiant les cellules à l'échelle individuelle est en essor. Le domaine de la modélisation stochastique s'appuie par exemple sur plusieurs études biologiques à l'échelle individuelle chez les micro-organismes [MA97, ARM98] et même chez l'homme [GZRI+06]. Même si ces études apparaissent dans un premier temps comme des cas isolés on note une volonté de développer ces méthodes à plus large échelle. On pourra par exemple se référer à [HZ11] présentant plusieurs méthodes permettant d'appliquer la spectroscopie de masse à l'échelle de l'individu et permettant d'aller vers une métabolomique quantitative *single-cell*. On peut donc espérer à l'avenir obtenir de plus en plus facilement des données à l'échelle individuelle et à large échelle.
D'un point de vue méthodologique, nous avons présenté dans le chapitre 1 la méthode du *flux balance analysis* (FBA) et ses variantes qui permettent d'étudier des réseaux de réactions de grande taille sans connaître au préalable les paramètres et lois cinétiques et qui permettent d'intégrer des informations de mesures à l'échelle de la population. Les travaux menés dans cette thèse ont ainsi permis de développer des pistes pour intégrer à l'aide d'une approche par contraintes, des données à l'échelle de l'individu. Un autre cadre d'application des méthodes introduites dans cette thèse est celui de l'écologie. En effet, l'utilisation de réseaux de réactions et d'équations différentielles pour modéliser le comportement dynamique de systèmes écologiques est courante (voir par exemple [BTRB12]). En raison de phénomènes cycliques abiotiques, par exemple le cycle des saisons, on obtient souvent des données temporelles répétant périodiquement un certain motif (par exemple [BEO+11]). On peut alors interpréter chacune de ces périodes de temps comme la réalisation d'un certain modèle dynamique aléatoire, ayant une moyenne donnée (le motif périodique) mais aussi une certaine variabilité. On peut alors obtenir, de manière plus simple qu'en biologie *single cell*, des trajectoires individuelles où les individus représentent cette fois ci les diffé-

rentes occurrences du phénomène écologique périodique observé.

Applications à la résolution de systèmes de contraintes Dans les chapitres 4 et 5 de cette thèse, nous avons établi une relation entre des propriétés asymptotiques aléatoires d'un réseau de réaction (sous réserve de validité de l'approximation de Bernoulli) et la résolution de systèmes d'équations et d'inéquations polynomiales. Plus précisément, dans le chapitre 5 nous avons démontré, toujours dans le même cadre d'approximation, que toute propriété s'exprimant comme un polynômes d'espérances, variances et co-variances peut être vérifiée en déterminant des systèmes de contraintes polynomiales admettent ou non une solution. Lorsqu'on se restreint aux propriétés linéaires (ne faisant pas intervenir de multiplications) sur les moments d'ordre 2 ou moins, nous avons démontré que les systèmes de contraintes obtenus sont quadratiques. On sait que cette correspondance entre propriétés et contraintes existe en sémantique différentielle dans laquelle des propriétés sur les pentes correspondent à des contraintes linéaires sur les flux. Cette thèse a permis de montrer que les propriétés sur les moments d'ordre au plus 2 correspondent à des contraintes quadriques. S'il existe des algorithmes très efficaces pour résoudre les programmes linéaires [DOW+55], la résolution des *programmes quadratiques* est un problème plus difficile. À l'heure actuelle les solveurs de contraintes usuels tels que `CPLEX` ou `Gurobi` ne permettent de résoudre que les problèmes quadratiques positifs ce qui ne constitue qu'un sous-ensemble des contraintes que l'on peut obtenir à partir de formules linéaires sur les moments d'ordre au plus 2. Par ailleurs, dans la cas général (*i.e.* si on tolère la multiplication) les contraintes obtenues sont polynomiales et il faut s'orienter vers d'autres types d'outils (par exemple les solveurs SMT tels que `dREAL` [GKC13]). Ainsi, la correspondance entre biologie et contraintes permet de progresser sur deux fronts puisque d'une part les avancées dans le domaine de la résolution de contraintes pourront s'appliquer à la vérification de propriétés biologiques et d'autre part la biologie fourni des instances de programmes quadratiques ou polynomiaux pour la communauté de résolution de contraintes. En effet, on constate par exemple dans le cadre de programmation logique (programmation par ensemble réponse (ASP)) que les instances de problèmes biologiques sont particulièrement difficiles.

Cadre général des sources probabilistes Enfin, je souhaite terminer cette thèse en la replaçant dans un cadre plus général. En effet, les résultats présentés ont pour fondement la dynamique de Bernoulli qui consiste à générer une liste réactions, en tirant aléatoirement et indépendamment les réactions avec une probabilité fixée. On peut ensuite obtenir la trajectoire du réseau de réactions, en comptant le nombre de réactions tirées pour chaque type de réactions (processus de comptage de réactions) et en considérant son image par une transformation affine déterminée par la stœchiométrie (voir le chapitre 3) pour plus de détails.
En d'autres termes, la dynamique du réseau est déterminée par un mot aléatoire w sur l'ensemble fini des symboles de réactions $\Sigma = \{R_1, \dots, R_m\}$. Le processus de comptage des réactions peut alors être facilement étudié dans le domaine de l'analyse d'algorithmes en moyenne [FS09] à l'aide d'une fonction génératrice multivariée

$$F(l_1, \dots, l_m) = \sum_{w \in \Sigma^*} p(w) l_1^{|w|_1} \times \cdots l_m^{|w|_m},$$

où $p(w)$ est la probabilité d'obtenir le mot w parmi les mots même taille et $|w|_j$ est le nombre de lettres R_j dans le mot w. Par exemple, dans le cas des sources de Bernoulli on peut démontrer que cette fonction s'écrit

$$F(l_1, \ldots, l_m) = \frac{1}{1 - (p_1 l_1 + \cdots + p_1 l_m)}$$

où $(p_j)_{1 \leq j \leq m}$ sont les probabilités de réactions. A partir de cette fonction génératrice, il est possible par des opérations de dérivations décrites dans [FS09] d'obtenir des expressions analytiques pour les moments du processus de comptage. On peut aussi obtenir, comme il a été fait dans cette thèse, des théorèmes central limites à partir de l'étude de ces fonctions génératrices. Voici un exemple de proposition que l'on peut établir et qui est très proche du résultat qu'on a obtenu sur les processus de comptage des réactions dans le chapitre 3.

Proposition 5.5 (puissances)**.** *Soit X_n un processus aléatoire discret à valeurs dans $\mathbb{N}$, dont la fonction génératrice s'écrit $F_{(X_n)}(u, z) = \sum_n p_n(u) z^n$. Supposons que uniformément dans un voisinage complexe de $u = 1$, $p_n(u)$ s'écrit*

$$p_n(u) = A(u) \cdot B(u)^n,$$

A et B sont analytiques en 1, $A(1) = B(1) = 1$ et que B satisfait la condition de variabilité $\mathfrak{v}(B) = B''(1) + B'(1) - B'(1)^2 \neq 0$, alors l'espérance et la variance de X_n satisfont

$$E(X_n) = n \cdot \mathfrak{m}(B) + \mathfrak{m}(A)$$
$$Var(X_n) = n \cdot \mathfrak{v}(B) + \mathfrak{v}(A).$$

De plus, après normalisation X_n tend en loi vers une loi normale avec une vitesse en $O(n^{-1/2})$:

$$\mathbb{P}\left\{ \frac{X_n - E(X_n)}{\sqrt{Var(X_n)}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-w^2/2} dw + O(n^{-1/2}).$$

Un autre résultat connu de même type est le théorème des quasi-puissances de Hwang [FS09]. En général, l'idée est qu'on peut obtenir une analyse détaillée d'un processus en s'intéressant seulement à l'étude de sa fonction génératrice. Si la fonction génératrice pour les sources de Bernoulli peut être facilement obtenue, comme énoncé ci-dessus, c'est également le cas pour des sources plus complexes telles que les chaînes de Markov ou les modèles de Markov cachés qui peuvent être obtenues à l'aide de la matrice de transition de la chaîne de Markov impliquée. Ainsi, nous pourrions penser, grâce à ces résultats théoriques, développer des méthodes par contraintes où les réactions sont choisies en utilisant des sources plus complexes que la source de Bernoulli.

Bibliographie

[AFLB+12] Geoffroy Andrieux, Laurent Fattet, Michel Le Borgne, Ruth Rimokh, and Nathalie Théret. Dynamic regulation of tgf-b signaling by tif1-γ : A computational approach. *PLoS ONE*, 7(3) :e33761, 03 2012.

[AG03] Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical programming*, 95(1) :3–51, 2003.

[AK11] David F Anderson and Thomas G Kurtz. Continuous time markov chain models for chemical reaction networks. In *Design and analysis of biomolecular circuits*, pages 3–42. Springer, 2011.

[Alo06] Uri Alon. *An introduction to systems biology : design principles of biological circuits*. CRC press, 2006.

[ARM98] Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4) :1633–1648, 1998.

[Bar58] Anthony F Bartholomay. Stochastic models for chemical reactions : I. theory of the unimolecular reaction process. *The bulletin of mathematical biophysics*, 20(3) :175–190, 1958.

[Bar59] Anthony F Bartholomay. Stochastic models for chemical reactions : Ii. the unimolecular rate constant. *The bulletin of mathematical biophysics*, 21(4) :363–373, 1959.

[BEH+05] Eric Berberich, Arno Eigenwillig, Michael Hemmer, Susan Hert, Lutz Kettner, Kurt Mehlhorn, Joachim Reichel, Susanne Schmitt, Elmar Schömer, and Nicola Wolpert. Exacus : Efficient and exact algorithms for curves and surfaces. In *Algorithms–ESA 2005*, pages 155–166. Springer, 2005.

[BEO+11] Nicholas J Bouskill, Damien Eveillard, Gregory O'Mullan, George A Jackson, and Bess B Ward. Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. *Environmental microbiology*, 13(4) :872–886, 2011.

[BHP+15] Sergiy Bogomolov, Thomas A Henzinger, Andreas Podelski, Jakob Ruess, and Christian Schilling. Adaptive moment closure for parameter inference of biochemical reaction networks. In *Computational Methods in Systems Biology*, pages 77–89. Springer, 2015.

[BPSC10] Robert Bellé, Sylvain Prigent, Anne Siegel, and Patrick Cormier. Model of cap-dependent translation initiation in sea urchin : a step towards the eukaryotic translation regulation network. *Mol. Reprod. Dev.*, 77(3) :257–64, Mar 2010.

[BTRB12] Nicholas J Bouskill, Jinyun Tang, William J Riley, and Eoin L Brodie. Trait-based representation of biological nitrification : model development, testing, and predicted community composition. *Frontiers in microbiology*, 3, 2012.

[CFS06] Laurence Calzone, François Fages, and Sylvain Soliman. Biocham : an environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics*, 22(14) :1805–1807, 2006.

[CKL15] Luca Cardelli, Marta Kwiatkowska, and Luca Laurenti. Stochastic analysis of chemical reaction networks using linear noise approximation. In *Computational Methods in Systems Biology*, pages 64–76. Springer, 2015.

[CSRS+04] Holger Conzelmann, Julio Saez-Rodriguez, Thomas Sauter, Eric Bullinger, Frank Allgöwer, and Ernst Dieter Gilles. Reduction of mathematical models of signal transduction networks : simulation-based approach applied to egf receptor signalling. *Systems biology*, 1(1) :159–169, 2004.

[DAH07] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1) :51–78, 2007.

[Del40] Max Delbrück. Statistical fluctuations in autocatalytic reactions. *The Journal of Chemical Physics*, 8(1) :120–124, 1940.

[DFF+07] Vincent Danos, Jérôme Feret, Walter Fontana, Russell Harmer, and Jean Krivine. Rule-based modelling of cellular signalling. In *CONCUR 2007–Concurrency Theory*, pages 17–41. Springer, 2007.

[DL04] Vincent Danos and Cosimo Laneve. Formal molecular biology. *Theoretical Computer Science*, 325(1) :69–110, 2004.

[DOW+55] George B Dantzig, Alex Orden, Philip Wolfe, et al. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2) :183–195, 1955.

[EE03] Johan Elf and Måns Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, 13(11) :2475–2484, 2003.

[EP00] JS Edwards and BO Palsson. The escherichia coli mg1655 in silico metabolic genotype : its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10) :5528–5533, 2000.

[FDK+09] Jérôme Feret, Vincent Danos, Jean Krivine, Russ Harmer, and Walter Fontana. Internal coarse-graining of molecular systems. *Proceedings of the National Academy of Sciences*, 106(16) :6453–6458, 2009.

[Fel74] William Feller. *Introduction to Probability Theory and Its Applications, Vol. II POD*. John Wiley & sons, 1974.

[FPI+12] Maxime Folschette, Loïc Paulevé, Katsumi Inoue, Morgan Magnin, and Olivier Roux. Concretizing the process hitting into biological regulatory networks. In *Computational methods in systems biology*, pages 166–186. Springer, 2012.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. cambridge University press, 2009.

[GB00] Michael A Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9) :1876–1889, 2000.

[Gil76] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4) :403–434, 1976.

[Gil77] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25) :2340–2361, 1977.

[Gil00] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113 :297, 2000.

[Gil01] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115 :1716, 2001.

[Gil07] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58 :35–55, 2007.

[GJ02] Michael R Garey and David S Johnson. *Computers and intractability*, volume 29. wh freeman, 2002.

[GKC13] Sicun Gao, Soonho Kong, and Edmund M Clarke. dreal : An smt solver for nonlinear theories over the reals. In *Automated Deduction–CADE-24*, pages 208–214. Springer, 2013.

[GLO05] Chetan Gadgil, Chang Hyeong Lee, and Hans G Othmer. A stochastic analysis of first-order reaction networks. *Bulletin of mathematical biology*, 67(5) :901–946, 2005.

[GP03] Daniel T Gillespie and Linda R Petzold. Improved leap-size selection for accelerated stochastic simulation. *The Journal of Chemical Physics*, 119 :8229, 2003.

[Gut09] Allan Gut. *Stopped random walks : Limit theorems and applications*. Springer, 2009.

[GZRI+06] Naama Geva-Zatorsky, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, Yuvalal Liron, Paz Polak, Galit Lahav, et al. Oscillations and variability in the p53 system. *Molecular systems biology*, 2(1), 2006.

[Hel08] Volkhard Helms. *Principles of computational cell biology*. Wiley, 2008.

[Hes08] Joao Hespanha. Moment closure for biochemical networks. In *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*, pages 142–147. IEEE, 2008.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HO01] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2) :159–195, 2001.

[HZ11] Matthias Heinemann and Renato Zenobi. Single cell metabolomics. *Current Opinion in Biotechnology*, 22(1) :26–31, 2011.

[K+01] Hiroaki Kitano et al. *Foundations of systems biology*. MIT press Cambridge, 2001.

[Kau69] Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3) :437–467, 1969.

[Kau93] Stuart A. Kauffman. *The origins of order : Self organization and selection in evolution*. Oxford university press, 1993.

[KT75] Samuel Karlin and Howard M Taylor. A first course in stochastic processes. *Academic Press, New York*, 1975.

[Kur72] Thomas G Kurtz. The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics*, 57(7) :2976–2978, 1972.

[Lot10] Alfred J Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3) :271–274, 1910.

[LRML+14] Sébastien Laurent, Adrien Richard, Odile Mulner-Lorillon, Julia Morales, Didier Flament, Virginie Glippa, Jérémie Bourdon, Pauline Gosselin, Anne Siegel, Patrick Cormier, and Robert Bellé. Modelization of the regulation of protein synthesis following fertilization in sea urchin shows requirement of two processes : a destabilization of eIF4E :4E-BP complex and a great stimulation of the 4E-BP-degradation mechanism, both rapamycin-sensitive. *Front Genet*, 5 :117, 2014.

[MA97] Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3) :814–819, 1997.

[McQ67] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of applied probability*, 4(3) :413–478, 1967.

[MLN15] Guillaume Madelaine, Cédric Lhoussaine, and Joachim Niehren. Structural simplification of chemical reaction networks preserving deterministic semantics. In *Computational Methods in Systems Biology*, pages 133–144. Springer, 2015.

[MM13] Leonor Michaelis and Maud L Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369) :352, 1913.

[MMB03] Carmen G Moles, Pedro Mendes, and Julio R Banga. Parameter estimation in biochemical pathways : a comparison of global optimization methods. *Genome research*, 13(11) :2467–2474, 2003.

[Mur89] T. Murata. Petri nets : Properties, analysis and applications. *Proceedings of the IEEE*, 77(4) :541–580, 1989.

[MW43] Henry B Mann and Abraham Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3) :217–226, 1943.

[Nås03a] Ingemar Nåsell. An extension of the moment closure method. *Theoretical population biology*, 64(2) :233–239, 2003.

[Nås03b] Ingemar Nåsell. Moment closure and the stochastic logistic model. *Theoretical population biology*, 63(2) :159–168, 2003.

[OTP10] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis ? *Nature biotechnology*, 28(3) :245–248, 2010.

[PAK13] Loïc Paulevé, Geoffroy Andrieux, and Heinz Koeppl. Under-approximating cut sets for reachability in large scale automata networks. In *Computer aided verification*, pages 69–84. Springer, 2013.

[PBS15] Vincent Picard, Jérémie Bourdon, and Anne Siegel. A logic for checking the probabilistic steady-state properties of reaction networks. In *Proceedings of the 2015 international workshop Bioinformatics and Artificial Intelligence (BAI) at the International Joint Conference on Articficial Intelligence (IJCAI 2015)*, 2015.

[PMR11] Loïc Paulevé, Morgan Magnin, and Olivier Roux. Refining dynamics of gene regulatory networks in a stochastic π-calculus framework. In *Transactions on computational systems biology xiii*, pages 171–191. Springer, 2011.

[PMR12] Loïc Paulevé, Morgan Magnin, and Olivier Roux. Static analysis of biological regulatory networks dynamics using abstract interpretation. *Mathematical Structures in Computer Science*, 22(04) :651–685, 2012.

[PPW+03] Jason A Papin, Nathan D Price, Sharon J Wiback, David A Fell, and Bernhard O Palsson. Metabolic pathways in the post-genome era. *Trends in biochemical sciences*, 28(5) :250–258, 2003.

[PRP04] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. Genome-scale models of microbial cells : evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11) :886–897, 2004.

[PSB14] Vincent Picard, Anne Siegel, and Jérémie Bourdon. Multivariate normal approximation for the stochastic simulation algorithm : limit theorem and applications. In *SASB-5th International Workshop on Static Analysis and Systems Biology*, 2014.

[RC07] Adrien Richard and Jean-Paul Comet. Necessary conditions for multistationarity in discrete dynamical systems. *Discrete Applied Mathematics*, 155(18) :2403–2413, 2007.

[RGZL08] Ovidiu Radulescu, Alexander N Gorban, Andrei Zinovyev, and Alain Lilienbaum. Robust simplifications of multiscale biochemical networks. *BMC systems biology*, 2(1) :86, 2008.

[RPCG03] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie. Stiffness in stochastic chemically reacting systems : The implicit tau-leaping method. *The Journal of Chemical Physics*, 119(24) :12784, 2003.

[San07] Werner Sandmann. Stochastic simulation of biochemical systems via discrete-time conversion. In *Proceedings of the 2nd Conference on Foundations of Systems Biology in Engineering*, pages 267–272, 2007.

[San08] Werner Sandmann. Discrete-time stochastic modeling and simulation of biochemical networks. *Computational biology and chemistry*, 32(4) :292–297, 2008.

[SDZ02] Ilya Schmulevich, Edward R Dougherty, and Wei Zhang. From boolean to probabilistic boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*, 90(11), 2002.

[SFR14] Sylvain Soliman, François Fages, and Ovidiu Radulescu. A constraint solving approach to model reduction by tropical equilibration. *Algorithms for Molecular Biology*, 9(1) :24, 2014.

[SH06] Abhyudai Singh and Joao Pedro Hespanha. Lognormal moment closures for biochemical reactions. In *Decision and Control, 2006 45th IEEE Conference on*, pages 2063–2068. IEEE, 2006.

[Sno98] El Houssine Snoussi. Necessary conditions for multistationarity and stable periodicity. *Journal of Biological Systems*, 6(01) :3–9, 1998.

[Tar51] Alfred Tarski. A decision method for elementary algebra and geometry. *Rand report*, 1951.

[Tho73] René Thomas. Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3) :563–585, 1973.

[VdV00] A. W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VK92] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[Vol27] Vito Volterra. *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C. Ferrari, 1927.

[WGSP12] EWJ Wallace, DT Gillespie, KR Sanft, and LR Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET systems biology*, 6(4) :102–115, 2012.

[Whi57] P Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 268–281, 1957.

[Wil06] Darren Wilkinson. *Stochastic Modelling for Systems Biology (Chapman & Hall/CRC Mathematical & Computational Biology)*. Chapman and Hall/CRC, April 2006.

[Wil09] Darren J Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2) :122–133, 2009.

[Wil12] Darren J Wilkinson. *Stochastic modelling for systems biology*, volume 44. CRC press, 2012.

[WZJC04] Ruiqi Wang, Tianshou Zhou, Zhujun Jing, and Luonan Chen. Modelling periodic oscillation of biological systems with multiple timescale networks. *Systems Biology*, 1(1) :71–84, 2004.

Table des figures

1 **Réseaux distinguables par leurs moments d'ordre 2** Dynamiques différentielles et stochastiques pour le modèle 1 (première ligne) et le modèle 2 (seconde ligne) décrits dans Eq. (3). Les courbes rouges (resp. bleues) représentent les quantités de C (resp. D). La première colonne représente la solution de l'équation différentielle issue de la loi d'action de masse. La seconde colonne représente 50 générations de trajectoires aléatoires obtenues à l'aide d'un algorithme de simulation stochastique. Les paramètres cinétiques et stochastiques ont été fixés à 1 et il y a $1000A$ et $1000B$ initialement. La troisième colonne représente les estimations des espérances, des variances et des covariances croisées entre C et D. Ces estimations sont obtenues à partir des 50 trajectoires de la seconde colonne. On constate que seule la trajectoire des covariances de C et D permettent de distinguer ces réseaux. 13

1.1 **Premier exemple de réseau de réactions.** Un réseau métabolique ([PPW+03]), sa modélisation en tant que réseau de réactions et la matrice de stœchiométrie associée. Dans la représentation graphique, les sorties sont mises en valeurs à droite parfois en les dédoublant 22

1.2 **Modèle de synthèse protéique à deux voies chez l'oursin.** Le but de ce modèle est d'introduire le cas particulier de la synthèse de la cyclineB qui serait synthétisée par deux voies, c'est-à-dire deux mécanismes distincts. Le mécanisme de la seconde voie est inspiré de celui d'une espèce marine voisine. 37

1.3 **Quelques méthodes d'analyse reposant sur un système de contraintes des flux à l'équilibre.** (image extraite de [PRP04]) 40

1.4 **Comparaison des taux d'utilisation des deux voies dans le modèle de synthèse protéique chez l'oursin.** Ce graphique représente des bornes pour les valeurs possibles de f_{26} en fonction de celles de f_{25} et d'après les hypothèses que nous avons formulées. 41

2.1 **Comparaison entre loi d'action de masse et équation maîtresse pour l'exemple des désintégrations par paires.** En rouge, la loi d'action de masse prédit une décroissance inverse. En bleu, l'espérance issue de l'équation maîtresse prédit une décroissance exponentielle. Il est donc faux en général d'énoncer que la loi d'action de masse décrit les valeurs moyennes des quantités de matière. . . 54

2.2 **Loi de probabilité du temps de vie d'un système de molécules se désintégrant par paires.** Cette information ne peut-être obtenue par la lois d'action de masse. 55

3.1 **Diagramme représentant les résultats principaux du chapitre 3.** 68

3.2 **Simulation de Gillespie du réseau de dimérisation** pour $a_0 = 15$, $c_1 = c_2 = 0,1$. 71

3.3 **Simulation du processus discrétisé du réseau de dimérisation** pour $a_0 = 15$, $c_1 = c_2 = 0,1$ et $\delta t = 0,01$. 72

3.4 **Comparaison des processus continu et discrétisé.** Comparaison de la distribution de la quantité de X après 2s pour 10 000 trajectoires. 73

3.5 **Illustration du comportement de marche aléatoire** sur l'exemple du réseau $\{X \to 2Y; \varnothing \to Y; \varnothing \to X + Y\}$. La sphère représente l'état courant du réseau et les flèches les nouveaux états possibles après l'occurrence de chacune des réactions qui a lieu avec probabilités p_i. 78

4.1 **Diagramme résumant les principales approches décrites dans les chapitres 3 et 4** 90

4.2 **Donnnées de départ pour discriminer les modèles 1 ou 2**. Les données ont pour forme un ensemble de trajectoires dont on a estimé les moments d'ordre 1 et 2. 93

4.3 **Un exemple de réseau métabolique jouet** illustratif pour les méthodes de contraintes 96

4.4 **Illustration de contraintes de variances et co-variances**. Les zones grises correspondent à l'ensemble des couples (p_1, p_2) valides en fonction des contraintes indiquées ainsi que de la contrainte (H_0) d'équilibre des réactifs (permettant de se ramener à un espace plan). Les points noirs correspondent aux valeurs extrêmes de p_1/p_2. 98

4.5 **Comparaison entre l'ellipsoïde de confiance et la multiplication des intervalles de confiance pour deux espèces.** La figure illustre l'avantage de l'utilisation de méthodes multivariées. Le rectangle obtenu par multiplication des interfaces est plus grand que l'α-ellipsoïde de confiance. La zone grise correspond aux prédictions exclues par la méthode multivariée grâce à l'utilisation des corrélations entre les deux espèces. 100

4.6 **Illustration d'une marche alétoire dégénérée** pour le réseau de réactions $\{\varnothing \to X; \varnothing \to Y\}$. Après k réactions, les quantités de matière sont nécessairement situées dans l'hyperplan d'équation $X + Y = k$. 102

4.7 **Retour sur l'exemple du réseau métabolique introductif** dont les réactions et la matrice de stœchiométrie ont été introduits dans le chapitre 1. 104

4.8 **Une preuve géométrique de la consistance de l'estimateur $\rho_{a,b}(k)$.** Pour k assez grand, l'ellipsoïde de confiance d'ordre α est incluse dans $\mathscr{C}$ parce que le rayon de la sur-approximation est dominé par la distance entre $\vec{z}(0)$ et $E(\vec{z}(k))$ lorsque k croît. La région grise correspond au complémentaire de $\mathscr{C}$. 109

Résumé

L'étude de la dynamique des réseaux de réactions est un enjeu majeur de la biologie des systèmes. Cela peut-être réalisé de deux manières : soit de manière déterministe à l'aide d'équations différentielles, soit de manière probabiliste à l'aide de chaînes de Markov. Dans les deux cas, un problème majeur est celui de la détermination des lois cinétiques impliquées et l'inférence de paramètres cinétiques associés. Pour cette raison, l'étude directe de grands réseaux de réactions est impossible. Dans le cas de la modélisation déterministe, ce problème peut-être contourné à l'aide d'une analyse stationnaire du réseau. Une méthode connue est celle de l'analyse des flux à l'équilibre (FBA) qui permet d'obtenir des systèmes de contraintes à partir d'informations sur les pentes moyennes des trajectoires. Le but de cette thèse est d'introduire une méthode analogue dans le cas de la modélisation probabiliste. Les résultats de la thèse se divisent en trois parties. Tout d'abord on présente une analyse stationnaire de la modélisation probabiliste reposant sur une approximation de Bernoulli. Dans un deuxième temps, cette dynamique approximée nous permet d'établir des systèmes de contraintes à l'aide d'informations obtenues sur les moyennes, les variances et les co-variances des trajectoires du système. Enfin, on présente plusieurs applications à ces systèmes de contraintes telles que la possibilité de réfuter des réseaux de réactions à l'aide d'informations de variances ou de co-variances et la vérification formelle de propriétés logiques sur le régime stationnaire du système.

Abstract

A major goal in systems biology is to inverstigate the dynamical behavior of reaction networks. There exists two main dynamical frameworks : the first one is the deterministic dynamics where the dynamics is described using odinary differential equations, the second one is probabilistic and relies on Markov chains. In both cases, one major issue is to determine the kinetic laws of the systems together with its kinetic parameters. As a consequence the direct study of large biological reaction networks is impossible. To deal with this issue, stationnary assumptions have been used. A widely used method is flux balance analysis, where systems of constraints are derived from information on the average slopes of the system trajectories. In this thesis, we construct a probabilistic analog of this stationnary analysis. The results are divided into three parts. First, we introduce a stationnary analysis of the probabilistic dynamics which relies on a Bernoulli approximation. Second, this approximated dynamics allows us to derive systems of constraints from information about the means, variances and co-variances of the system trajectories. Third, we present several applications of these systems of constraints such as the possibility to reject reaction networks using information from experimental variances and co-variances and the formal verification of logical properties concerning the stationnary regime of the system.