Prédiction de la structure de contrôle de bactéries par optimisation sous incertitude

Marouane Ait El Faqir

N° d'ordre NNT : 2016LYSEC36

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Ecole Centrale de Lyon

Ecole Doctorale ED 160
ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE

Spécialité de doctorat :
Automatique

Soutenue publiquement le 22/11/2016, par :
Marouane Ait El Faqir

Prédiction de la structure de contrôle de bactéries par optimisation sous incertitude

---

Devant le jury composé de :

M. Luc DUGARD	Directeur de recherche CNRS GIPSA Lab	Président, rapporteur
M. Didier DUMUR	Professeur des Universités CentraleSupélec	Rapporteur
M. Julien HUILLERY	Maître de Conférences, École Centrale de Lyon	Examinateur
M. Marc DINH	Docteur, Ingénieur d'études, INRA - Centre Jouy-en-Josas	Examinateur

Résumé

L'approche de la biologie des systèmes vise à intégrer les méthodologies appliquées dans la conception et l'analyse des systèmes technologiques complexes, au sein de la biologie afin de comprendre les principes de fonctionnement globaux des systèmes biologiques.

La thèse s'inscrit dans le cadre de la biologie des systèmes et en particulier dans la prolongation d'une méthode issue de ce cadre : la méthode Resource Blance Analysis (RBA).

Nous visons dans cette thèse à augmenter le pouvoir prédictif de la méthode via un travail de modélisation tout en gardant un bon compromis entre représentativité des modèles issus de ce cadre et leur résolution numérique efficace.

La thèse se décompose en deux grandes parties : la première vise à intégrer les aspects thermodynamiques et cinétiques inhérents aux réseaux métaboliques. La deuxième vise à comprendre l'impact de l'aspect stochastique de la production des enzymes sur le croissance de la bactérie.

Des méthodes numériques ont été élaborées pour le résolution des modèles ainsi établis dans les deux cas déterministe et stochastique.

Mots clés : Optimisation stochastique, Méthodes du Premier Ordre, Optimisation déterministe, Approximation Stochastique, RBA

Abstract

In order to understand the global functioning principals of biological systems, system biology approach aims to integrate the methodologies used in the conception and the analysis of complex technological systems, within the biology.

This PhD thesis fits into the system biology framework and in particular the extension of the already existing method Resource Balance Analysis (RBA).

We aim in this PhD thesis to improve the predictive power of this method by introducing more complex model. However, this new model should respect a good trade-off between the representativity of the model and its efficient numerical computation.

This PhD thesis is decomposed into two major parts. The first part aims the integration of the metabolic network inherent thermodynamical and kinetic aspects. The second part aims the comprehension of the impact of enzyme production stochastic aspect on the bacteria growth.

Numerical methods are elaborated to solve the obtained models in both deterministic and stochastic cases .

Keywords : Stochastic optimization, First Oder Methods, deterministic optimisation, Stochastic Approximation, RBA

Remerciement

Je tiens à remercier chaleureusement mes directeurs de thèse Messieurs Gérard SCORLETTI et Vincent FROMION pour la confiance qu'ils m'ont accordée pour mener cette thèse. Je les remercie également pour leurs bons conseils.

J'exprime également ma reconnaissance envers mes encadrants Messieurs Julien HUILLERY et Marc DINH et les remercie pour leur soutien et leur présence durant les moments difficiles mais également durant les moments agréables de cette thèse.

Mon passage au laboratoire Ampère était plein d'apprentissage où j'ai eu l'occasion de connaître au quotidien le métier de la recherche.

Je remercie aussi les membres du laboratoire Ampère et en particulier ceux attachés à l'Ecole Centrale de Lyon.

Je remercie également mes collègues du bureau Sylvain TORU et Olivier POIRION : merci pour la bonne ambiance !

Finalement, je dédie ce travail à mes parents, ma famille et mes amis

Liste des symboles

$\mathbb{E}[.]$ Espérance mathématique

$\mathbb{E}_{X\sim\mathbb{P}(\lambda)}F(X)$ Espérance mathématique de $F(X)$ étant donné que X a une densité de probabilité continue $\mathbb{P}$ de paramètre λ

$\mathbb{E}\left[X|Y\right]$ espérance de X conditionnée par Y

$\langle ., . \rangle$ produit scalaire

$\|.\|$ norme quelconque en dimension finie

$\|.\|_*$ norme duale de la norme $\|.\|$

$\|.\|_2$ norme euclidienne

$\|.\|_1$ norme l_1

$O(1)$ une constante donnée

π_X projection euclidienne sur l'ensemble X

f^* valeur optimale d'un problème d'optimisation

x^* solution optimale d'un problème d'optimisation

$domf$ domaine de définition de la fonction f

$\partial f(x)$ sous-différentiel de la fonction f en x

$Conv(X)$ enveloppe convexe de l'ensemble X

$Lin(\{\eta_1, \ldots, \eta_k\})$ espace vectoriel engendré par la famille de vecteur $\{\eta_1, \ldots, \eta_k\}$

$\mathcal{O}$ oracle d'ordre 1

Prob$\{A\}$ probabilité de l'événement A

$\lceil x \rceil$ l'entier juste au dessus de x

$dist(x, A)$ distance du point x à l'ensemble A

$\bar{X}$ adhérence topologique de l'ensemble X

N nombre d'itérations

N_e nombre de fois où un problème d'optimisation stochastique a été résolu numériquement par un algorithme donné

$\omega(.)$ fonction génératrice de distance associée à un algorithme donné

Table des figures

Liste des tableaux

Table des algorithmes

Table des matières

Chapitre 1

Introduction générale

La progression dans la compréhension du fonctionnement des systèmes biologiques constitue sans aucun doute un des acquis majeurs de la seconde moitié du 20 ème siècle, en particulier, mais non exclusivement, à travers les progrès sans précédent réalisés par la biologie moléculaire qui s'est concentrée à démonter au mieux les processus élémentaires des cellules. Cette approche est souvent qualifiée de réductionniste car elle s'attache avant tout à caractériser (finement) le fonctionnement des processus individuels de la cellule comme la transcription, la traduction ou encore la réplication pour ne citer qu'eux. Après plus d'une cinquantaine d'années et des progrès sans précédent, le caractère réductionniste de l'approche trouve ses limites, non pas dans le démontage des processus où beaucoup reste à faire, mais dans la difficulté qu'il y a à les intégrer ensemble et faire émerger un modèle global de la cellule. De fait, et face à cette limitation intrinsèque, Kitano fait la remarque suivante dans son article publié en 2002 (voir [37]) :

> « *Identifying all the genes and proteins in an organism is like listing all the parts of an airplane. While such a list provides a catalog of the individual components, by itself it is not sufficient to understand the complexity underlying the engineered object. We need to know how these parts are assembled to form the structure of the airplane. This is analogous to drawing an exhaustive diagram of gene regulatory networks and their biochemical interactions. Such diagrams provide limited knowledge of how changes to one part of a system may affect other parts, but to understand how a particular system functions, we must first examine how the individual components dynamically interact during operation.* »

Pour continuer, il insiste sur le fait que nous ne pourrons pas appréhender le puzzle du vivant en ne nous focalisant que sur les pièces : il faut décrypter le système dans sa globalité et voir la cellule comme un système intégrant différents processus essentiels à la vie d'une cellule. Cette vision systémique (et globale) mène naturellement à se poser certaines questions, Kitano toujours dans [37] en identifie deux importantes :

(*i*) la nature de la structure du système global c'est-à-dire par exemple la nature des réseaux d'interaction entre les gènes, celle des réactions biochimiques ou encore des mécanismes permettant de réguler telles ou telles interactions, etc. ;

(*ii*) la nature des mécanismes contrôlant l'état de la cellule et lui garantissant de préserver son fonctionnement normal malgré la présence de perturbations et d'incertitudes dans son environnement.

Face à cela, il constate qu'il existe une certaine similarité entre la structure et l'organisation des processus d'une cellule et celles des processus présents dans les systèmes technologiques

complexes. Afin de convaincre le lecteur, il fait un parallèle, une analogie entre cellule et Boeing 777. Celui-ci de fait contient environ 150000 sous-systèmes différents, sous la forme de modules organisés via un système complexe de contrôle incluant à peu près 1000 calculateurs embarqués qui assurent l'automatisation de toutes les fonctions de l'avion.
Il insiste qu'au delà de cette apparente ressemblance, la cellule et l'avion doivent posséder tous les deux certaines propriétés importantes comme par exemple la robustesse (immunité, adaptabilité) vis-à-vis des dysfonctionnements et des changements environnementaux (pour ne citer qu'eux).
L'ensemble de ces éléments conduit Kitano à fonder une nouvelle approche du vivant qu'il appelle **Biologie des Systèmes** et qui se propose d'utiliser les méthodologies utilisées pour l'analyse et la conception des systèmes technologiques pour les appliquer aux systèmes biologiques. Cela devrait permettre de dégager, s'ils existent, des principes génériques et globaux de fonctionnement du vivant.
Sans surprise, cette proposition de Kitano trouve un écho au sein de la communauté automaticienne, qui peut être illustré à travers cette phrase extraite d'un article écrit par J. C. Doyle (voir [23]) :

> « *Systems-level approaches in biology have a long history but are just now receiving renewed mainstream attention, whereas systems-level design has consistently been at the core of modern engineering, motivating its most sophisticated theories in control, information, and computation. The hidden nature of complexity and discipline fragmentation within engineering have been barriers to a dialog with biology.* »

La thèse qui est présentée ici se situe clairement dans ce contexte général de la biologie des systèmes, et plus spécifiquement dans le prolongement d'une méthode récement introduite, trouvant ses racines dans la biologie des systèmes, et qui s'appelle la méthode **Resource Balance Analysis** (RBA). Elle a été introduite dans une série d'article [29, 28, 30] et a été très récemment validée expérimentalement dans [31]. La méthode RBA consiste à prédire la répartition parcimonieuse des ressources au sein de la bactérie par rapport à son milieu, en modélisant la bactérie sous la forme d'une interconnexion entre plusieurs sous-systèmes représentant les différents processus essentiels (voir figure 1.1).

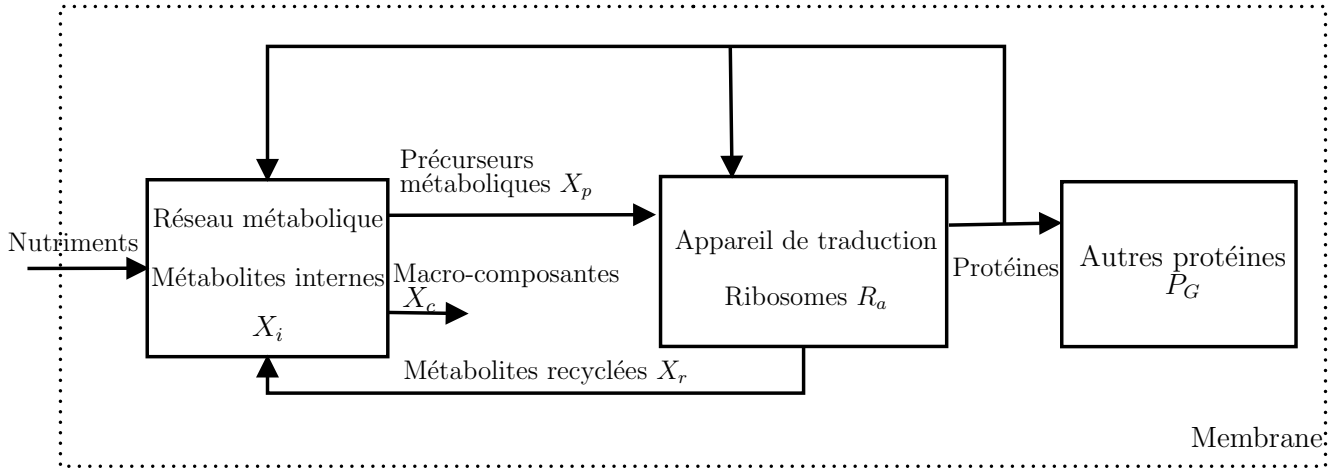

FIGURE 1.1 – Vision systémique de la bactérie

Le point remarquable c'est que le problème RBA peut se formuler, en adoptant des hypothèses raisonnables (comme le démontre sa validation récente), comme un problème d'optimisation qui peut être du point de vue numérique efficacement résolu. En effet, la

résolution actuelle du problème RBA fait appel à de l'optimisation linéaire qui constitue un avantage important en termes d'outils associés à l'optimisation, compte tenu de la très grande dimension théorique des systèmes biologiques (très rapidement, plusieurs milliers de variables de décision).
Dans ce contexte, l'objectif central de cette thèse est, en exploitant au plus près les méthodes et outils attachés à l'optimisation (convexe) d'accroître le pouvoir prédictif de la méthode RBA. A cette fin, nous avons travaillé suivant deux axes : d'un côté, du point de vue de la modélisation, nous souhaitions affiner la description des enzymes en prenant en compte explicitement les aspects thermodynamiques et cinétiques inhérents au fonctionnement des réseaux métaboliques ; de l'autre côté, mesurer les conséquences de la stochasticité sur la répartition des ressources au sein de la bactérie.
Au cours de notre développement, la question centrale est de savoir comment complexifier la modélisation des RBA sans pour autant perdre l'efficacité de sa résolution numérique. L'un des enjeux de cette thèse est de garantir au modèle RBA un bon compromis entre représentativité vis-à-vis de la réalité et efficacité de résolution numérique. Cela s'inscrit clairement dans un courant de pensée qui a été fécond dans les années 90 en Automatique et qui est résumé ici par une citation extraite du livre de Nesterov ([48]) :

> « *In my experience, if an optimization model is created without taking into account the abilities of numerical schemes, the chances that it will be possible to find an acceptable numerical solution are close to zero. In any field of human activity, if we create something, we know in advance why we are doing so and what we are going to do with the result. And only in numerical modelling is the situation still different.* »

Ce mémoire de thèse se compose de deux grandes parties. La première partie est consacrée à la prise en compte, au sein des RBA, de l'aspect thermodynamique et cinétique des réseaux métaboliques ainsi qu'à la recherche de la formulation d'un problème d'optimisation associé au modèle RBA admettant des méthodes de résolutions efficaces.

Modèle RBA et extension thermodynamique et cinétique (chapitre 2) : au cours de ce chapitre nous commençons par rappeler la méthode RBA ainsi que les aspects liés à son efficacité de résolution numérique. Nous passons ensuite aux aspects thermodynamiques et cinétiques et leur intégration au sein des RBA, ce qui constitue la contribution de ce chapitre. A l'issue de ce chapitre un modèle RBA étendu contenant ces phénomènes physiques est formulé.

Formulation du problème d'optimisation pour le modèle RBA étendu (chapitre 3) : dans ce chapitre, nous étudions la possibilité de ramener le problème d'optimisation associé au modèle RBA étendu à un problème d'optimisation dont la résolution est efficace. L'optimisation issue de la programmation géométrique est un bon candidat : cependant la présence de deux contraintes non convexes irréductibles ne permet pas de s'y ramener. Une relaxation convexe est alors proposée, dont l'intérêt sera discuté à travers un exemple numérique. La contribution ce chapitre est donc d'explorer les potentialités de l'optimisation géométrique pour le problème d'optimisation associé au modèle RBA étendu.

La deuxième partie de cette thèse est consacrée à l'intégration et à la gestion des incertitudes stochastiques au sein du cadre RBA. L'objectif est de prendre en compte l'impact de la nature stochastique de l'expression des gènes sur le taux de croissance d'une population de cellules bactériennes et sur l'allocation des ressources.

Modélisation stochastique (chapitre 4) : l'objectif de ce chapitre est de proposer une modélisation de la nature stochastique de l'expression des gènes, de formuler le problème d'optimisation stochastique associé et de quantifier l'impact de la modélisation stochastique par rapport à la modélisation déterministe pour une classe particulière de réseaux métaboliques admettant une solution analytique dans les deux cas.

Problèmes d'optimisation déterministe de grande dimension et leur résolution numérique (chapitre 5) : dans ce chapitre, nous proposons une bibliographie sur les méthodes de résolutions de problèmes d'optimisation convexe de grande dimension. Ceci permet de donner une vue globale sur les motivations derrière l'utilisation des méthodes permettant de gérer efficacement du point de vue numérique l'aspect de la grande dimension inhérent aux applications biologiques. Nous proposons les méthodes les plus adaptées dans le cas des problèmes d'optimisation de grande dimension, leur cadre théorique ainsi que leur complexité algorithmique. Nous introduisons également une analyse de convergence systématique basée sur la théorie de dissipativité issue de l'automatique, ce qui constitue la contribution du chapitre.

Méthodes pour l'optimisation stochastique (chapitre 6) dans ce chapitre nous faisons une bibliographie sur les méthodes les plus efficaces pour résoudre les problèmes d'optimisation stochastique. La définition d'une solution numérique pour un problème d'optimisation stochastique est présentée ainsi que l'efficacité algorithmique de certaines approches de résolution numérique. Un comparatif sur la complexité algorithmique des ces approches est dressé à la fin du chapitre. La contribution de ce chapitre est d'évaluer la pertinence de ces algorithmes dans le contexte de modèles stochastiques de taille réaliste. Elle a permis de proposer des améliorations suffisamment substantielles pour résoudre des problèmes de taille significative, qui ont fait l'objet d'un article de conférence soumis à l'*IEEE Conference on Decision and Control* en 2016, voir l'annexe du document de thèse, chapitre 8.

Le document se termine sur des conclusions et des perspectives.

Première partie
Extension de la méthode RBA (Resource Balance Analysis)

Chapitre 2

Modèle RBA et extension thermodynamique et cinétique

Dans ce chapitre, nous commençons en Section 2.1 et 2.2 par rappeler les éléments essentiels de la méthode *Resource Balance Analysis* (RBA) introduite dans [29, 28, 30] et récemment validée expérimentalement dans [31]. Nous étudions ensuite la possibilité d'augmenter le pouvoir prédictif de la méthode en raffinant la modélisation des activités enzymatiques, en cherchant en particulier à y inclure explicitement les contraintes dites thermodynamiques et cinétiques : la Section 2.3 détaille ces contraintes thermodynamiques et cinétiques alors que la Section 2.4 les intègre dans le modèle RBA. La Section 2.5 conclut le chapitre.

2.1 Approche système en biologie : aspects fondamentaux du modèle RBA

Dans la vue systémique de la cellule (voir figure 1.1) sur laquelle la méthode RBA est basée, le réseau métabolique, composé de protéines (les enzymes en l'occurrence), transforme les nutriments importés à l'intérieur par d'autres protéines (les transporteurs en l'occurence) de la cellule pour produire l'energie ou encore les précurseurs métaboliques nécessaires à la synthèse des différentes composantes de la cellule. Divers processus cellulaires contribuent à construire les composantes de la cellule que sont par exemple les protéines, l'ADN, les ARN ribosomaux, les ARN messagers, la paroi cellulaire, en utilisant les précurseurs fournis par le réseau métabolique. Ces processus cellulaires sont de nature très diverse et sont par exemple des enzymes (protéines), transformant des métabolites substrats en métabolites produits ou encore des macromolécules complexes comme les ribosomes qui permettent de produire les protéines en assemblant des acides aminés selon l'ordre déterminé par un ARN messager (la séquence codante). Le point central de l'approche RBA est de noter que beaucoup de ressources de la cellule doivent être partagées par les différents processus présents dans la cellule et qu'à ce titre, et afin de réaliser l'ensemble des fonctions nécessaires à sa survie ou à sa croissance, la cellule doit à chaque instant décider comment allouer/répartir chaque ressource (pris dans un sens large) à sa disposition aux différents processus.
L'approche RBA proposée dans [29, 28, 30] a intégré un certain nombre de processus cellulaires, dont on rappelle ici les principaux :

– **l'appareil de traduction des protéines :** son rôle principal est de synthétiser les

protéines. Ce processus est basé sur un complexe moléculaire constitué de protéines et d'ARN : le ribosome. La concentration des ribosomes (actifs) sera notée R_a ;

– **le réseau métabolique :** son rôle est de transporter, de casser et de transformer les nutriments extra-cellulaires afin de produire de l'énergie et les précurseurs métaboliques nécessaires à tous les processus cellulaires. Ces précurseurs métaboliques sont consommés par les processus cellulaires afin de produire toutes sortes de composantes cellulaires (protéines, ADN, ARN ribosomaux, etc.). Les composantes principales du réseau métaboliques sont des protéines spéciales : les enzymes, qui catalysent les réactions biochimiques transformant des substrats en produits. Dans la suite, le réseau métabolique sera décrit par les éléments suivants :

  (*i*) N_m enzymes $\mathcal{E} := (\mathcal{E}_1, \ldots, \mathcal{E}_{N_m})$ aux concentrations $E := (E_1, \ldots, E_{N_m})$ et de flux enzymatiques associés $\nu := (\nu_1, \ldots, \nu_{N_m})$;

  (*ii*) N_i métabolites internes $\mathbb{X}_i := (\mathbb{X}_{i1}, \ldots, \mathbb{X}_{i_{N_i}})$ aux concentrations $\bar{X}_i := (\bar{X}_{i1}, \ldots, \bar{X}_{i_{N_i}})$;

  (*iii*) N_p précurseurs métaboliques consommés durant la synthèse des protéines $\mathbb{X}_p := (\mathbb{X}_{p_1}, \ldots, \mathbb{X}_{p_{N_p}})$;

  (*iv*) N_r métabolites recyclés produits durant la synthèse des protéines, $\mathbb{X}_r := (\mathbb{X}_{r1}, \ldots, \mathbb{X}_{r_{N_r}})$;

  (*v*) les macro composantes comme la paroi cellulaire, la membrane, l'ADN, etc, de dimension N_C, et seront notées $\mathbb{X}_C := (\mathbb{X}_{C_1}, \ldots, \mathbb{X}_{C_{N_C}})$.

  Le réseau métabolique est caractérisé par sa matrice de stœchiométrie S décrivant la répartition des métabolites au sein du réseau métabolique, conformément au principe de conservation de masse, ainsi que leurs relations avec les flux enzymatiques. Cette matrice est de taille $(N_i + N_p + N_r) \times N_m$. En régime permanent la loi de conservation de masse au sein du réseau métabolique est donnée sous la forme du système linéaire suivant :

$$S_I \nu = 0 \tag{2.1}$$

  où S_I est la matrice stœchiométrique relative aux métabolites internes ;

– **les autres composantes cellulaires :** l'ensemble des protéines, de dimension N_G, n'appartenant ni à l'appareil de traduction ni au réseau métabolique, sera noté $\mathbb{P}_G := (\mathbb{P}_{G_1}, \ldots, \mathbb{P}_{G_{N_G}})$ et leur concentration $P_G := (P_{G_1}, \ldots, P_{G_{N_G}})$.

La méthode RBA se focalise sur la phase de croissance des bactéries. Il s'agit d'une phase spécifique où le nombre de bactéries augmente à travers un cycle de reproduction assez simple consistant pour chaque bactérie à accroître dans une première phase sa (bio)masse puis lorsque celle-ci a doublé à se diviser. Cette phase est exponentielle par nature et est décrite par exemple en décrivant le volume de la population des bactéries par $V(t) = V_0 \exp(\mu t)$, où V_0 est le volume initial à $t = 0$ et μ correspond au taux de croissance (exponentiel) de la population.

Dans ce contexte, si $P_j(t)$ est la concentration de $\mathbb{P}_j$, l'une des composantes présentes dans le cytoplasme de la bactérie, alors comme par définition de la concentration, on a :

$$P_j(t) := \frac{n_j(t)}{V(t)} \tag{2.2}$$

où $n_j(t)$ est le nombre de protéines de $\mathbb{P}_j$ dans le volume $V(t)$, alors la variation de la

concentration $P_j(t)$ par rapport au temps est donnée par :

$$\frac{dP_j(t)}{dt} = \underbrace{\frac{dn_j(t)}{dt}\frac{1}{V(t)}}_{\text{production}=p_j(t)} - \underbrace{\frac{dV(t)}{dt}\frac{n_j(t)}{V^2(t)}}_{\text{Dilution}} = p_j(t) - \mu P_j(t). \tag{2.3}$$

De façon remarquable, il a été montré expérimentalement qu'en régime de croissance exponentielle, les bactéries maintiennent **la concentration de toutes les composantes cellulaires à une valeur constante** à travers des mécanismes complexes de régulation non discutés ici (mode dit equilibré). Ainsi pour maintenir la concentration $P_j(t)$ constante au sein de chaque bactérie malgré la variation de volume due à la croissance, la bactérie doit produire à chaque instant la quantité suivante de cette composante : $p_j = \mu P_j$. On déduit de cela que si α_{kj} molécules du précurseur métabolique $\mathbb{X}_{pk}$ sont utilisées pendant la synthèse de la composante $\mathbb{P}_j$, alors un flux du métabolite $\mathbb{X}_{pk}$ égal à $\mu\alpha_{kj}P_j$ est consommé pour maintenir la concentration P_j constante au taux de croissance μ. Sur cette base, nous pouvons donc maintenant décrire les contraintes que les différents processus cellulaires doivent satisfaire pour qu'il y ait croissance exponentielle de la population bactérienne :

(C_1) **Contraintes sur la capacité du réseau métabolique :** la capacité du réseau doit permettre de :

(a) produire suffisamment de précurseurs métaboliques pour la croissance de la cellule. Autrement dit, le flux de synthèse des N_p précurseurs métaboliques doit être plus important que le flux consommé durant la synthèse des composants cellulaires.
(C_{1a}) : pour tout $i \in \{1, \ldots, N_p\}$,

$$-\sum_{j=1}^{m} S_{pij}\nu_j + \mu\left(\sum_{j=1}^{m} C_{M_{ij}}^{M_p} E_j + C_{R_i}^{M_p} R_a \sum_{j=1}^{N_G} C_{G_{ij}}^{M_p} P_{G_j}\right) - \nu_Y \leqslant 0 \tag{2.4}$$

où S_p est la sous-matrice stœchiométrique de S décrivant la partie du réseau métabolique liée aux précurseurs métaboliques, ν_Y correspond au flux d'échange avec l'environnement tels que la diffusion des métabolites à travers la membrane. $C_{M_{ij}}^{M_p}, C_{R_i}^{M_p}, C_{G_{ij}}^{M_p}$ sont des cœfficients positifs correspondant respectivement au nombre de précurseurs métaboliques $\mathbb{X}_{p_i}$ requis : (i) pour la synthèse de la $j^{\text{ème}}$ enzyme appartenant au réseau métabolique ; (ii) pour la synthèse d'un ribosome ; (iii) pour la synthèse de la $j^{\text{ème}}$ protéine appartenant à $\mathbb{P}_G$ (voir figure 1.1) ;

(b) maintenir la concentration $\bar{X}_c$ de l'ensemble des macro-composantes $\mathbb{X}_c$ constante.
(C_{1b}) : pour tout $i \in \{1, \ldots, N_c\}$,

$$-\sum_{j=1}^{m} S_{cij}\nu_j + \mu\bar{X}_{c_i} \leqslant 0 \tag{2.5}$$

où S_c est la sous-matrice stœchiométrique de S relative aux macro-composantes ;

(c) absorber tous les métabolites recyclés produits lors de la synthèse des différentes composantes cellulaires.
(C_{1c}) : pour tout $i \in \{1, \ldots, N_r\}$

$$\sum_{j=1}^{m} S_{rij}\nu_j + \mu\left(\sum_{j=1}^{m} C_{M_{ij}}^{M_r} E_j + C_{R_i}^{M_r} R_a + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_r} P_G\right) \leqslant 0 \tag{2.6}$$

où S_r est la matrice stœchiométrique spécifique aux métabolites recyclés $\mathbb{X}_r$, $C^{M_r}_{M_{ij}}, C^{M_r}_{R_i}, C^{M_r}_{G_{ij}}$ sont des cœfficients positifs correspondant respectivement au nombre du $i^{\text{ème}}$ métabolite recyclé $\mathbb{X}_{r_i}$ produit durant la synthèse : (i) de la $j^{\text{ème}}$ enzyme ; (ii) d'un ribosome ; (iii) de la $j^{\text{ème}}$ protéine appartenant à $\mathbb{P}_G$;

(d) satisfaire la loi de la conservation de masse.
(C_{1d}) : pour tout $i \in \{1, \ldots, N_i\}$,

$$\sum_{j=1}^{m} S_{Iij}\nu_j = 0 \tag{2.7}$$

où S_I est la matrice stœchiométrique relative aux métabolites internes ;

(e) il reste à ajouter une contrainte, celle liée à la capacité maximale de production des enzymes. Cette contrainte qui sera étudiée en détail dans la suite de ce chapitre, a été traitée dans [29, 28, 30] à travers l'introduction de l'hypothèse suivante :
Hypothèse 1. *Le couplage entre la concentration E_j de l'enzyme $\mathcal{E}_j$ et le flux ν_j passant à travers cet enzyme est décrit par la contrainte :*

$$(C_{1e}) \quad |\nu_j| \leqslant k_{E_j} E_j, \tag{2.8}$$

où $k_{E_j} > 0$ est une constante strictement positive correspondant à l'efficacité enzymatique (maximale) de l'enzyme $\mathcal{E}_j$.

Remarque : l'objectif central de ce chapitre est de proposer une modélisation plus fine du couplage (2.8) en prenant typiquement en compte les facteurs thermodynamiques et cinétiques reliant flux ν_j et concentration E_j au sein d'un réseau métabolique.

(C_2) **Contraintes sur la capacité de l'appareil de traduction :** l'appareil de traduction doit être capable de maintenir constante la concentration de toutes les protéines à μ donné.

$$\mu \left(\sum_{j=1}^{m} C^R_{M_j} E_j + C^R_R R_a + \sum_{j=1}^{N_G} C^R_{G_j} P_{G_j} \right) - k_T R_a \leqslant 0 \tag{2.9}$$

où $C^R_{M_j}, C^R_R, C^R_{G_j}$ sont des nombres positifs correspondant respectivement au nombre total en acides aminés : (i) pour la $j^{\text{ème}}$ enzyme ; (ii) par ribosome ; (iii) pour la $j^{\text{ème}}$ protéine appartenant à $\mathbb{P}_G$. k_T désigne l'efficacité de la traduction (en nombre d'acides aminés traduits par unité de temps).

(C_3) **Contrainte de densité :** la cellule doit gérer sa densité intracellulaire pour assurer la diffusion adéquate de toutes les composantes cellulaires (protéines, métabolites, ions, etc.)

$$\sum_{j=1}^{m} C^D_{M_j} E_j + C^D_R R_a + \sum_{j=1}^{N_G} C^D_{G_j} P_{G_j} - \bar{D} \leqslant 0 \tag{2.10}$$

où $\bar{D}$ est la densité moyenne exprimée en équivalent acides aminés. Les cœfficients $C^D_{M_j}$ et $C^D_{G_j}$ sont respectivement égaux à $C^R_{M_j}$ et $C^R_{G_j}$; C^D_R correspond à la densité d'un ribosome en équivalent acides aminés.

2.2 Problème RBA sous forme de problème d'optimisation linéaire

Avant d'aborder cette partie, nous allons présenter la terminologie et les propriétés de base de la classe de problèmes d'optimisation convexe.

On utilise en général la notation suivante :

$$\begin{aligned} &\min_{x\in X} f_0(x) \\ \text{tel que : } &\begin{cases} f_i(x) \leqslant 0, & i = 1,\dots,m \\ g_j(x) = 0, & j = 1,\dots,p \end{cases} \end{aligned} \tag{2.11}$$

où $X \subset \mathbb{E}$ est convexe et $\mathbb{E}$ est un espace vectoriel normé de dimension n finie ($\mathbb{R}^n$ en général), n est le nombre des variables de décision $x = (x_1, \dots, x_n)$. f_i , $i = 0, \dots, m$, g_j $j = 1, \dots, p$ sont des fonctions de $\mathbb{E}$ dans $\mathbb{R}$ définies sur X. (2.11) désigne dans ce contexte, un problème d'optimisation cherchant à trouver un point x^* qui minimise la fonction f_0, appelée la fonction objectif, et qui appartient à l'ensemble $\mathcal{D}$ défini par

$$\mathcal{D} := X \bigcap_{i=1}^{m} \{x | f_i(x) \leqslant 0\} \bigcap_{j=1}^{p} \{x | g_j(x) = 0\}.$$

Les conditions $f_i(x) \leqslant 0, \quad i = 1, \dots, m$, $g_j(x) = 0, \quad j = 1, \dots, p$, $x \in X$, sont appelées contraintes du problème (2.11). Le point x^* est dit solution optimale du problème (2.11). On dit que la $i^{\text{ème}}$ contrainte inégalité $f_i(x) \leqslant 0$, est convexe si la fonction f_i est convexe sur X. On dit que la $i^{\text{ème}}$ contrainte égalité $g_i(x) = 0$ est convexe, si la fonction g_i est affine sur X. Si les ensembles X, $\{x | f_i(x) \leqslant 0\}$ $i = 1, \dots, m$, $\{x | g_j(x) = 0\}$ $j = 1, \dots, p$ sont convexes (c'est-à-dire les fonctions f_i sont convexes et les fonctions g_j sont affines) alors $\mathcal{D}$ est convexe. Si en plus f_0 est une fonction convexe, alors le problème (2.11) est dit problème d'optimisation convexe. Le problème (2.11) est un problème d'optimisation linéaire si les fonctions f_i, $i = 0, \dots, m$, g_j, $j = 1, \dots, p$ sont affines et l'ensemble X est un polyèdre. Le problème (2.11) est dit de faisabilité si f_0 est constante et dans ce cas on le met sous la forme suivante :

$$\begin{aligned} &\text{Trouver} \quad x \in X. \\ &\text{tel que : } \begin{cases} f_i(x) \leqslant 0, & i = 1,\dots,m \\ g_i(x) = 0, & i = 1,\dots,p \end{cases} \end{aligned} \tag{2.12}$$

Si un tel point existe, alors on dit que le problème admet une solution faisable. Concernant l'existence de la solution optimale x^* de (2.11), on a le théorème fondamental suivant.

Théorème 1. *[Weierstrass][42] Si $\mathcal{D}$ est un compact non vide de $\mathbb{E}$ et si la restriction de la fonction f_0 à $\mathcal{D}$ $\tilde{f}_0 : \mathcal{D} \to \mathbb{R}$ est semi-continue inférieurement (ou continue) alors le problème (2.11) admet au moins une solution.*

Le théorème précédent suppose que $\mathcal{D}$ est compact. En général, $\mathcal{D}$ n'est pas toujours compact. Pour cela, il existe une autre version du théorème précédent basée sur la notion de fonction coercive.

Définition 1. *La fonction $\tilde{f}_0$ est dite coercive si elle vérifie*

$$\lim_{\|x\|\to+\infty} \tilde{f}_0(x) = +\infty.$$

On a le théorème suivant.

Théorème 2. *[45] Si $\mathcal{D}$ est une partie fermée non vide de $\mathbb{E}$ et si $\tilde{f}_0$ est semi-continue inférieurement (ou continue) et coercive alors le problème admet au moins une solution.*

Définition 2. *Le point $x \in \mathcal{D}$ est une solution locale du problème (2.11) s'il existe un voisinage V de x dans $\mathbb{E}$ tel que*

$$et \quad \tilde{f}_0(x) \leqslant \tilde{f}_0(y), \quad \forall y \in \mathcal{D} \cap V$$

Une des propriétés fondamentales dont bénéficie le problème d'optimisation convexe (2.11) est la suivante.

Théorème 3. *[33] Si le problème (2.11) est convexe alors toute solution locale du problème (2.11) est globale.*

Cela est d'une utilité importante, car si un algorithme est capable de déterminer une solution locale (Karush ?Kuhn ?Tucker (KKT), etc.) du problème (2.11), le théorème 3 assure que celle-ci sera forcément globale. Autrement, il suffit de calculer une solution locale pour résoudre le problème. Contrairement à la classe d'optimisation convexe, en optimisation non-convexe, il n'existe en général pas de test local pour l'optimalité globale et cela fait la difficulté intrinsèque de cette classe de problème d'optimisation [34]. Tout cela rend la classe des problèmes d'optimisation convexe attractive du fait qu'elle bénéficie de méthodes de résolution numérique efficaces [9] comme par exemple les méthodes de points intérieurs. De plus, il existe des sous-classes de problèmes d'optimisation convexe dites bien structurées à savoir l'optimisation linéaire, quadratique, semi-définie, géométrique, etc. ayant des structures particulières, possédant des propriétés spécifiques. Ces sous-classes ont été les plus touchées par les dernières avancées en optimisation et en méthodes numériques rendant leur résolution plus mûre et plus maniable du point de vue numérique.

Après avoir présenté les contraintes structurelles issues de la méthodologie RBA, nous précisons que leur satisfaction se traduit de façon pratique par l'obtention d'un taux de croissance maximal de la population des bactéries en régime exponentiel qui est compatible avec l'ensemble des contraintes. En effet le processus de maximisation du taux de croissance en régime permanent peut être modélisé sous la forme du problème d'optimisation suivant :

$$\begin{aligned} \mu^* = \max_{E\geqslant 0,\ R_a\geqslant 0,\ \nu} & \mu \\ \text{tel que : } & (C_{1a}), (C_{1b}), (C_{1c}), (C_{1d}), (C_{1e}), (C_2), (C_3). \end{aligned} \tag{2.13}$$

La formulation (2.13) du problème n'est pas convexe. Néanmoins, sa résolution passe par celle d'une séquence de problèmes d'optimisation (de faisabilité) convexe linéaire (2.14) : pour $\mu \geqslant 0$,

$$\begin{aligned} \text{Trouver} \quad & E \geqslant 0, R_a \geqslant 0, \nu \\ \text{tels que : } & (C_{1a}), (C_{1b}), (C_{1c}), (C_{1d}), (C_{1e}), (C_2), (C_3). \end{aligned} \tag{2.14}$$

En effet, il a été démontré dans [28] que ce problème dispose de la propriété suivante.

Proposition 1. *[28]*

1. *Si pour P_G et pour $\mu^+ > 0$, le problème (2.14) est faisable alors pour tout $\mu \in [0, \mu^+]$, le problème (2.14) est faisable.*
2. *Pour tout $P_G \geqslant 0$, il existe μ_0 fini tel que le problème (2.14) est faisable et pour tout $\mu > \mu_0$, le problème (2.14) est infaisable.*

μ_0 n'est rien d'autre que μ^* solution optimale du problème (2.13). De ce point de vue, la résolution de (2.13) se ramène à la résolution itérative de (2.14). En effet il suffit d'appliquer le schéma de bissection de l'algorithme 1 page 13. Cela représente un avantage

Algorithme 1 Schéma itératif de la méthode de bissection

Initialisation :
choisir une précision ϵ et prendre $\underline{\mu}$ pour lequel (2.14) est faisable (on peut prendre $\underline{\mu} = 0$ vue la proposition 1) et $\overline{\mu}$ pour lequel (2.14) est infaisable (on a $\mu^* \in [\underline{\mu}, \overline{\mu}]$).
Prendre $k = 0$.
Tant que :
$\overline{\mu} - \underline{\mu} > \epsilon$, prendre $k = k + 1$, $\mu_k = (\underline{\mu} + \overline{\mu})/2$ et résoudre (2.14) pour μ_k.
Si (2.14) est non faisable, prendre $\overline{\mu} = \mu_k$, **sinon** prendre $\underline{\mu} = \mu_k$.
Fin de tant que.

important pour ce qui est de la performance numérique de la méthode RBA : la résolution numérique du problème (2.13) se ramène à celle d'une séquence de problèmes convexes (2.14) pour différentes valeurs de μ qui peuvent, comme nous l'avons déjà mentionné, être résolus de façon (très) efficace, en particulier lorsqu'il s'agit de problème linéaire [48].

De plus on sait qualitativement qu'il y a limitation du taux de croissance, la deuxième propriété intéressante est liée à la quantification de cette limitation du taux de croissance μ étant donné un milieu nutritif. En effet, la résolution de (2.13) montre que la bactérie peut croître au taux de croissance $\mu \leqslant \mu^*$ où μ^* est la valeur optimale de (2.13) c'est-à-dire que μ^* est le taux maximal dont la bactérie est capable étant donné un milieu nutritif. Cette remarque fondamentale en biologie révèle le fait que le taux de croissance est ici limité par les contraintes du problème (2.14). Cela indique aussi que le taux satisfait un compromis concernant la distribution des ressources en général et en particulier au niveau des protéines en s'assurant par exemple d'une bonne répartition de celles-ci entre le réseau métabolique et l'appareil de traduction.

Maintenant que nous avons passé en revue les principaux éléments de la méthode RBA, nous allons présenter notre premier axe de contribution au développement de cette méthode. Dans un premier temps nous renforçons le pouvoir de prédiction de la méthode en intégrant les aspects thermodynamiques et cinétiques présents lors des réactions biochimiques au sein du réseau métabolique, puis nous intégrerons ce nouveau couplage au sein de la méthode RBA.

2.3 Prise en compte de la cinétique et de la thermodynamique du réseau métabolique

Au sein du réseau métabolique, les réactions sont catalysées par les enzymes et sont régies par des contraintes thermodynamiques et cinétiques [21, 22]. Plus spécifiquement

ces aspects interviennent au niveau du couplage entre le flux ν_j (flux de réaction catalysée par une enzyme $\mathbb{E}_j$) et la concentration E_j. Notre objectif est de prendre en compte un couplage plus réaliste que celui de la contrainte (C_{1e}) (voir hypothèse 1) en prenant les aspects thermodynamiques et cinétiques en compte. Pour déterminer le lien entre flux métabolique et concentration de substrats, de produits et d'enzymes, nous rappelons quelques aspects de la cinétique enzymatique.

2.3.1 Rappels sur la modélisation de la cinétique des enzymes

La majorité des réactions chimiques à l'intérieur de la cellule sont catalysées par des enzymes. Les enzymes accélèrent la vitesse de réaction dans le sens direct et dans le sens inverse sans être consommées durant la réaction. De plus, les enzymes ont un comportement sélectif : les enzymes en général accélèrent seulement des réactions spécifiques. Le modèle de l'action des enzymes a été établi il y a longtemps (voir [22] pour les détails) et repose sur un modèle cyclique où l'enzyme se lie à l'élément réactif (le substrat) en formant un complexe enzyme-réactif puis ce complexe subit une transformation qui finit par libérer l'enzyme ainsi que le produit. L'enzyme est alors disponible de nouveau pour un autre cycle.

Considérons la réaction suivante où $\mathbb{S}$ est le substrat de concentration S, $\mathbb{E}$ est une enzyme de concentration E et $\mathbb{P}$ est le produit de concentration P. k_1, k_{-1}, k_2 sont des constantes de vitesse des réactions décrites sur le schéma réactionnel suivant :

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P. \tag{2.15}$$

Cette réaction illustre le mécanisme de liaison entre l'enzyme $\mathbb{E}$ et le substrat $\mathbb{S}$ et la libération du produit $\mathbb{P}$. On remarque ici que dans ce schéma, seule la réaction de liaison entre enzyme et substrat est réversible, alors que la réaction de libération du produit $\mathbb{P}$ est irréversible, rendant la réaction irréversible.

Enzyme irréversible. Afin d'obtenir un modèle de cette réaction, il est fait appel à l'hypothèse de Briggs et Haldane [22] qui consiste à supposer que la formation du complexe ES atteint rapidement son régime permanent. L'équation de cinétique enzymatique concernant le complexe ES s'écrit sous la forme suivante :

$$\frac{dES}{dt} = k_1 E.S - (k_{-1} + k_2)ES. \tag{2.16}$$

Selon cette hypothèse, on obtient à l'équilibre :

$$k_1 E.S - (k_{-1} + k_2)ES = 0. \tag{2.17}$$

Notons la quantité totale d'enzymes E_T, on a :

$$E_T = E + ES.$$

Ainsi la concentration du complexe enzyme-substrat peut être écrite sous la forme :

$$ES = \frac{E_T.S}{(k_{-1} + k_2)/k_1 + S}.$$

Par définition, le flux enzymatique[1] ν correspondant au modèle (2.15) est donné par $\nu = k_2 ES$. Par la suite on obtient le modèle de Michaelis et Menten de la cinétique de l'enzyme E dans le cas irréversible :

$$\nu = \frac{k_2 E_T.S}{(k_{-1}+k_2)/k_1 + S}. \tag{2.18}$$

L'efficacité enzymatique dans ce cas est par définition donnée par

$$k_E = \frac{k_2.S}{(k_{-1}+k_2)/k_1 + S}.$$

On remarque ici que contrairement à ce que suppose l'hypothèse 2.8 page 10, l'efficacité enzymatique ci-dessus n'est pas constante puisque c'est une fonction de S.

Enzyme réversible. On peut étendre sans difficulté le calcul précédent au schéma réactionnel réversible suivant :

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E + P. \tag{2.19}$$

La principale différence est que le flux enzymatique au niveau du produit met en jeu un flux direct et un flux inverse cette fois-ci :

$$\nu = \underbrace{k_2 ES}_{\text{flux direct}=J^+} - \underbrace{k_{-2} E.P}_{\text{flux inverse}=J^-}. \tag{2.20}$$

Dans ce cas, on a :

$$\frac{dES}{dt} = k_1 E.S + k_{-2} E.P - (k_{-1}+k_2)ES = 0,$$

ce qui donne :

$$ES = E\left(\frac{S}{K_1} + \frac{P}{K_2}\right),$$

où $K_1 = (k_{-1}+k_2)/k_1$ et $K_2 = (k_{-1}+k_2)/k_{-2}$ sont les affinités des réactifs S et P pour l'enzyme $\mathbb{E}$. Ainsi on obtient :

$$E_T = E\left(1 + \frac{S}{K_1} + \frac{P}{K_2}\right).$$

En revenant à l'équation (2.20) on obtient le modèle de Michaelis et Menten de la cinétique de l'enzyme E dans le cas réversible :

$$\nu = E_T \frac{k_2 S/K_1 - k_{-1} P/K_2}{1 + S/K_1 + P/K_2}. \tag{2.21}$$

L'efficacité de l'enzyme $\mathbb{E}$ dans ce cas est donc égale à :

$$k_E = \frac{k_2 S/K_1 - k_{-1} P/K_2}{1 + S/K_1 + P/K_2}.$$

1. le flux ν représente la vitesse de la réaction (2.15) et donc homogène à un $Mole.s^{-1}$. Il peut être écrit conformément à la loi d'action de masse : $\nu = \frac{dP}{dt} = -\frac{dS}{dt}$.

Cette expression peut aussi être donnée en fonction de l'énergie de Gibbs de la réaction (2.19) (enthalpie) [22]. En effet, l'énergie de Gibbs est égale par définition à [2] :

$$\Delta_r G' = \Delta_r G'^0 + RT \ln\left(\frac{P}{S}\right), \tag{2.22}$$

où R est la constante des gaz parfaits, T est la température et $\Delta_r G'^0$ est l'énergie standard de Gibbs de la réaction. Or à l'équilibre, on a $\nu = 0$ et donc

$$k_2 S/K_1 - k_{-1} P/K_2 = 0 \Rightarrow \frac{k_2}{k_{-1}}\frac{K_2}{K_1} = \frac{P}{S}.$$

De plus, on sait qu'à l'équilibre chimique, l'énergie de Gibbs de la réaction est nulle [22], ce qui conduit à déduire que

$$\Delta_r G' = \Delta_r G'^0 + RT \ln\left(\frac{P}{S}\right) = 0.$$

En passant par la relation de Haldane [22], relation fondamentale dans la suite de notre développement car elle met en évidence l'impact du facteur thermodynamique (enthalpie) sur l'efficacité de l'enzyme,

$$K_{eq} = e^{(-\Delta_r G'^0/RT)}, \tag{2.23}$$

où $K_{eq} = \frac{k_2}{k_{-1}}\frac{K_2}{K_1}$ est la constante d'équilibre de la réaction, on obtient :

$$\nu = \frac{E_T k_2 S/K_1 (1 - e^{(\Delta_r G'^0/RT)} P/S)}{1 + S/K_1 + P/K_2},$$

qui peut se réécrire sous la forme suivante :

$$\nu = E_T k_2 \frac{S/K_1}{1 + S/K_1 + P/K_2}(1 - e^{(\Delta_r G'/RT)}). \tag{2.24}$$

La nouvelle formule, en fonction de l'énergie de Gibbs de la réaction (2.19), de l'efficacité d'enzyme est alors la suivante :

$$k_E = k_2 \frac{S/K_1}{1 + S/K_1 + P/K_2}(1 - e^{(\Delta_r G'/RT)}).$$

On remarque que cette efficacité est affectée par les concentrations du substrat et du produit ainsi que le facteur thermodynamique, caractérisé par l'énergie de Gibbs (l'enthalpie) de la réaction. Ainsi on en conclut que l'efficacité de l'enzyme est soumise à deux contraintes structurelles de natures différentes. L'une est de nature cinétique concernant les concentrations des réactifs et l'autre est de nature thermodynamique concernant l'énergie de la réaction.

Les réseaux métaboliques comprennent des centaines, voire des milliers de réactions enzymatiques. Malgré les progrès réalisés ces dernières années, il est encore illusoire de penser qu'on sait à une telle échelle (celle qui nous intéresse) identifier non seulement le schéma

2. Pour une vision détaillée voir [21, 22].

réactionnel attaché à chaque réaction du réseau mais aussi les paramètres qui lui sont associés. Cette situation étant classique, elle a donné lieu à des travaux qui proposent des solutions permettant de contourner en partie ces difficultés à travers l'utilisation de modèles empiriques dits « *generalized rate law* » des réactions enzymatiques [60]. Ici, nous avons choisi la loi empirique proposée dans l'article [40] et présentée dans la suite. Cela a été motivé par un certain nombre de raisons. En effet et contrairement aux autres lois empiriques classiquement utilisées dans la littérature, le modèle du flux métabolique proposé dans [40] n'est pas local, permet de représenter des réactions irréversibles et réversibles, respecte les contraintes thermodynamiques et enfin intègre en partie les contraintes cinétiques des réactions. Enfin, cette loi est donnée sous forme d'un modèle modulaire permettant de distinguer les différents facteurs contrôlant le flux métabolique (voir [60, 40] pour plus de détails et de discussions). Dans la suite nous allons présenter et intégrer au sein du cadre RBA ce modèle empirique.

2.3.2 Formulation générale du couplage entre concentrations enzymatiques et flux métaboliques

Dans cette section, nous rappelons les bases du modèle proposé dans l'article [40] permettant de représenter l'efficacité enzymatique et prenant explicitement en compte les contraintes cinétiques et thermodynamiques inhérentes aux réactions biochimiques. On se place dans le cadre le plus général, celui d'une réaction métabolique réversible. On considère que la $j^{\text{ème}}$ réaction métabolique à la forme générique suivante :

$$\underbrace{\sum_{l' \in \mathcal{L}'} m^+_{l'j} X_{l'}}_{\mathbb{S}} \leftrightharpoons \underbrace{\sum_{l \in \mathcal{L}} m^-_{lj} X_l}_{\mathbb{P}}, \tag{2.25}$$

où $\mathcal{L}'$ représente l'ensemble des indices des métabolites substrats, c'est-à- dire si $l \in \mathcal{L}'$ alors X_l est un substrat. $\mathcal{L}$ est défini de façon similaire et représente l'ensemble des indices des métabolites produits. $\mathbb{S}$ et $\mathbb{P}$ représentent l'ensemble des substrats et des produits respectivement. On note également q le nombre total des métabolites du réseau métabolique en question. $m^+_{l'j}$ est le cœfficient stœchiometrique du métabolite $\mathbb{X}_{l'}$ s'il s'agit d'un substrat et vaut 0 sinon. La concentration du métabolite $\mathbb{X}_i$ est notée par $\bar{X}_i$. m^-_{lj} est le cœfficient stœchiometrique du métabolite $\mathbb{X}_l$ s'il s'agit d'un produit et vaut 0 sinon. On associe à la réaction (2.25), son flux métabolique ν_j et sa constante de vitesse de réaction directe et inverse k^+_j et k^-_j.

La loi « générique » proposée dans l'article [40]

Le modèle de l'activité enzymatique proposé dans l'article [40] a la forme suivante :

$$\nu_j = E_j f_r \frac{T_r}{D_r + D_r^{reg}}, \tag{2.26}$$

où les différents termes sont définis comme suit.

Le numérateur T_r correspond au « terme thermodynamique » et est une fonction des concentrations des métabolites $\mathbb{X}_i$ et des constantes de vitesse de réaction k^+_i et k^-_i, $i = 1, \ldots, q$:

$$T_r = k^+_j \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}} \right)^{m^+_{lj}} - k^-_j \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}} \right)^{m^-_{lj}},$$

où K_{mlj} est la constante de Michaelis et Menten de l'enzyme $\mathbb{E}_j$ par rapport au $l^{\text{ème}}$ réactif $\mathbb{X}_l$.

Le dénominateur D_r représente la façon dont la liaison entre les métabolites $\mathbb{X}_i$ et l'enzyme $\mathbb{E}_j$ s'établit. Cinq modes (indépendants les uns par rapport aux autres) de liaison ont été introduits dans [40]. On a considéré le mode dit « Direct » vu son interprétation physique d'un côté et son degré de commodité numérique de l'autre (ce point sera éclairé ultérieurement). Ce mode correspond à trois états de l'enzyme $\mathbb{E}_j$: soit tous les substrats sont liés à l'enzyme, soit tous les produits sont liés à l'enzyme, ou bien l'enzyme est libre auquel cas on peut écrire D_r comme

$$D_r = 1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}} \right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}} \right)^{m_{lj}^-},$$

où l'on ira voir [40] pour les détails.

Le terme f_r représente un terme de régulation et particulièrement la régulation allostérique. Dans ce modèle, deux types de régulations sont considérés, la régulation complète et la régulation partielle. Le régulateur enzymatique peut se lier à l'enzyme dans n'importe quel état : enzyme relié avec des métabolites ou libre. Une fois le régulateur lié à l'enzyme, il peut renforcer ou diminuer l'état de l'enzyme (libre/lié). Si le régulateur a la capacité d'activer/d'inhiber complètement cet état de l'enzyme, alors il s'agit de régulateur à activité/inhibition complète. La formule explicite du terme f_r se trouve dans [40]. Néanmoins, f_r a la propriété importante suivante :

$$f_r \leqslant 1. \tag{2.27}$$

Sachant qu'en l'absence de la régulation on a $f_r = 1$. Ceci montre que la régulation a pour effet de diminuer le flux métabolique.

Le terme D_r^{reg} représente la régulation non allostérique. Cela peut correspondre par exemple à une inhibition compétitive de l'activité de l'enzyme : un inhibiteur entre en compétition avec le substrat pour le site actif de l'enzyme de façon à perturber la réaction. La formule explicite du terme se trouve dans [40]. On retient seulement que

$$D_r^{reg} \geqslant 0. \tag{2.28}$$

A l'image des hypothèses que nous avons faites lorsqu'on a obtenu le modèle de Michaelis-Menten, le modèle générique (2.26) est obtenu sur la base de deux hypothèses importantes à savoir :

i) l'hypothèse du régime permanent : lors de la réaction enzymatique (2.25), l'enzyme $\mathbb{E}_j$ va se lier aux substrats en formant un complexe substrat-produit. On supposera toujours que la concentration de ce complexe au fur et à mesure de la réaction est constante ;

ii) la quantité d'enzymes disponible lors de la réaction est négligeable par rapport aux concentrations des métabolites.

D'autre part et par simplicité du modèle, on suppose que la majorité des enzymes est activée : les phénomènes de coopérativité sont négligés.

Formulation explicite des contraintes thermodynamiques et cinétiques pour un flux de signe connu

Afin de simplifier la présentation, on suppose dans la suite de cette section que le signe du flux enzymatique est connu et en accord avec la thermodynamique (supposé positif dans la suite). Je reviendrai sur ce point en fin de section et dans la section suivante. On introduit ici une quantité idéale, appelée flux métabolique idéal, noté ν_j^{ideal} et définie par

$$\nu_j^{ideal} = E_j \frac{T_r}{D_r}, \tag{2.29}$$

et qui peut se réécrire suivant la formule (2.26) comme

$$\nu_j^{ideal} = E_j \frac{k_j^+ \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} - k_j^- \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}. \tag{2.30}$$

Au regard des inégalités (2.27) et (2.28), et suivant l'hypothèse de positivité du flux enzymatique, on déduit l'inégalité suivante :

$$0 \leqslant \nu_j \leqslant \nu_j^{ideal}. \tag{2.31}$$

Par conséquent, on a obtenu une borne supérieure sur le flux enzymatique ν_j :

$$0 \leqslant \nu_j \leqslant E_j \frac{k_j^+ \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} - k_j^- \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}. \tag{2.32}$$

qui nous conduit à modifier le k_{E_j} constant proposé dans l'hypothèse 1 page 10 par

$$k_{E_j} = \frac{k_j^+ \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} - k_j^- \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}.$$

Ainsi, on est passé d'une efficacité enzymatique constante (dans le cas de l'hypothèse 1), à une efficacité enzymatique dépendant des concentrations des métabolites et prenant en compte les facteurs thermodynamiques et cinétiques dans les réactions métaboliques.
Afin de préciser la nature de la borne supérieure que nous venons de présenter, nous allons rappeler les notions de flux direct et inverse :

Définition 3. *Quand la réaction (2.25) est irréversible, le flux métabolique ne peut prendre qu'un seul sens, le sens direct de la réaction. Quand la réaction est réversible, elle peut être vue comme la combinaison de deux réactions irréversibles qui s'effectuent simultanément. Quand le flux métabolique va dans le sens direct de la réaction, il s'agit du flux direct qu'on note J_j^+. Dans le cas inverse, il s'agit de flux inverse qu'on note J_j^-. Le flux métabolique correspond à la différence entre le flux direct et le flux inverse de deux réactions irréversibles fictives, soit*

$$\nu_j =: J_j^+ - J_j^-.$$

On peut identifier dans le modèle (2.26), les flux directs et inverses en isolant les contributions positives et négatives :

$$J_j^+ \approx f_r E_j \frac{k_j^+ \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-} + D_r^{reg+}} \leqslant E_j \frac{k_j^+ \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}, \tag{2.33}$$

$$J_j^- \approx f_r E_j \frac{k_j^- \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-} + D_r^{reg-}} \leqslant E_j \frac{k_j^- \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}. \tag{2.34}$$

Il faut noter que l'approximation faite au niveau de (2.33) et (2.34) est due au fait que dans le cas de réaction irréversible, la constante de vitesse de réaction correspondant au sens inverse de la réaction irréversible est par définition négligeable devant la constante de vitesse de réaction correspondant au sens direct.

Du point de vue thermodynamique, quand la réaction (2.25) est à l'équilibre chimique, son énergie de Gibbs est nulle ($\Delta_r G' = 0$), les flux direct J_j^+ et inverse J_j^+ sont égaux et le flux métabolique est donc égal à zéro puisque par définition $\nu_j = J_j^+ - J_j^- = 0$. Une réaction est dite thermodynamiquement favorable si son énergie de Gibbs est négative : $\Delta_r G' < 0$. Cette propriété est quantifiée par la favorabilité thermodynamique : $-\Delta_r G'$ qui permet de relier les flux direct et inverse par la relation suivante :

$$\frac{J_j^+}{J_j^-} = \exp(-\Delta_r G'/RT), \tag{2.35}$$

En effet, par définition de l'énergie de Gibbs on a :

$$\Delta_r G' = \Delta_r G'^0 + RT \ln \left(\frac{\prod_{l=1}^{q} \bar{X}_l^{m_{lj}^-}}{\prod_{l=1}^{q} \bar{X}_l^{m_{lj}^+}} \right), \tag{2.36}$$

où $\Delta_r G'^0$ est l'énergie de Gibbs standard. On a $\Delta_r G'^0 = -RT \ln K_{jeq}$, où R est la constante des gaz parfaits, T la température absolue et K_{jeq} la constante d'équilibre de la réaction (2.25). Puisqu'à l'équilibre, on a $\nu_j = 0$ alors (il s'agit de la relation de Haldane vue précédemment) :

$$K_{jeq} := \left(\frac{\prod_{l=1}^{q} \bar{X}_l^{m_{lj}^-}}{\prod_{l=1}^{q} \bar{X}_l^{m_{lj}^+}} \right)_{\text{à l'équilibre}} = \frac{k_j^+}{k_j^-} \prod_{l=1}^{q} \frac{{K_{mlj}}^{m_{lj}^-}}{{K_{mlj}}^{m_{lj}^+}}, \tag{2.37}$$

ainsi en combinant (2.36) et (2.37) on obtient directement (2.35).

A partir de (2.35) on obtient la formule pratique suivante :

$$\nu_j := J_j^+ - J_j^- = J_j^+(1 - \exp(\Delta_r G'/RT)).$$

Cela nous permet d'aboutir à la formule modulaire suivante, mettant en évidence la contribution du facteur thermodynamique et cinétique sur le flux métabolique, en supposant que la réaction (2.25) est thermodynamiquement favorable :

$$\nu_j = J_j^+ - J_j^- \leqslant \underbrace{E_j}_{\text{concentration enzymatique totale}} \ldots \underbrace{\frac{k_j^+ \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{m_{lj}^-}}}_{\text{facteur cinétique}} \ldots \underbrace{(1 - e^{\Delta_r G'/RT})}_{\text{facteur thermodynamique}}, \tag{2.38}$$

La formule (2.38) montre que le flux métabolique est impacté par deux types de facteurs ou contraintes inhérents à la réaction métabolique et au réseau métabolique en question. Il s'agit du facteur cinétique (avec les constantes de vitesse de réaction et les concentrations des métabolites), du facteur thermodynamique (avec la favorabilité thermodynamique de la réaction). Sous l'hypothèse que la réaction (2.25) est thermodynamiquement favorable, c'est-à- dire que $\nu_j \geqslant 0$, l'inégalité (2.38) est une contrainte structurelle supplémentaire à intégrer dans la méthode RBA. Cette contrainte fixe le domaine des concentrations d'enzyme et des métabolites admissibles du point de vue thermodynamique et cinétique. Pour rendre cette contrainte plus exploitable, on va écrire le facteur thermodynamique en fonction des concentrations des métabolites et autres paramètres caractéristiques à la réaction en question. La tâche est simple, il suffit de constater que :

$$\Delta_r G' = RT \left(\sum_{l=1}^{q} (n_{lj} \ln(\bar{X}_l)) - \ln(\underbrace{K_{jeq}}_{e^{(-\Delta_r G'^0/RT)}}) \right),$$

où n_{lj} est le cœfficient stœchiometrique de la $l^{\text{ème}}$ métabolite X_l tel que

$$n_{lj} = \begin{cases} -m_{lj}^+ & \text{si } X_l \text{ est un substrat} \\ m_{lj}^- & \text{si } X_l \text{ est un produit} \end{cases} \tag{2.39}$$

Ainsi on obtient la nouvelle version de la contrainte (2.38) :

$$\nu_j \leqslant k_j^+ E_j \frac{1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{-m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{n_{lj}}}. \tag{2.40}$$

Comme nous avons supposé dans cette section que le signe du flux était constant, c'est-à-dire ici $\nu_j \geqslant 0$, la contrainte (2.40) n'est pas tout à fait prête à être intégrée dans le problème (2.14). Dans la section suivante, nous allons montrer comment cette contrainte sera intégrée dans le cas de flux métaboliques à signe quelconque.

Formulation des contraintes thermodynamiques et cinétiques dans le cas d'un flux métabolique à signe quelconque

Dans le cas du réseau métabolique et sous l'hypothèse de la réversibilité des enzymes, le signe des flux est aussi un degré de liberté et donc il peut être dans le sens opposé à celui de la convention « substrat-produit » prise dans la configuration (2.25). Dans ce cas la contrainte thermo-cinétique (2.40) obtenue dans la section précédente n'est pas valide (car basée sur le fait que $\nu_j \geqslant 0$). Comme on ne peut pas déterminer a priori dans le cadre RBA les signes des flux métaboliques, on ne peut pas traiter au cas par cas ces flux. Il est alors pertinent de pallier cette difficulté dans sa généralité. Par rapport à ce qui précède, un flux est négatif si et seulement s'il prend le sens opposé à celui de la convention (2.25) (des substrats vers les produits). Ce qui est identiquement équivalent à prendre le sens direct de la configuration :

$$\underbrace{\sum_l m_{lj}^- X_l}_{\mathbb{P}} \underset{k_j^+}{\overset{k_j^-}{\rightleftharpoons}} \underbrace{\sum_l m_{lj}^+ X_l}_{\mathbb{S}}. \tag{2.41}$$

Sous cette configuration le flux en question est bel et bien positif. Et inversement, si ce flux était négatif sous la configuration (2.41), il serait positif sous la configuration (2.25) par définition. L'idée est qu'au lieu de basculer entre les signes, on garde le signe positif et on bascule entre les configurations (2.25) et (2.41).

Plus spécifiquement on suppose que toutes les enzymes de la voie métabolique sont constituées de deux enzymes fictifs que l'on note $\mathbb{E}_j^+$ et $\mathbb{E}_j^-$. $\mathbb{E}_j^+$ fait passer le flux $\overline{\nu}_j$ selon la configuration (2.25) et $\mathbb{E}_j^-$ fait passer le flux $\underline{\nu}_j$ selon la configuration (2.41). Le fonctionnement de ces deux enzymes fictives est mutuellement exclusif du fait que le flux métabolique prend un sens unique à la fois, les signes de $\overline{\nu}_j$ et $\underline{\nu}_j$ sont opposés par construction comme on l'a mentionné plus haut et la concentration totale d'enzyme est $E_j = E_j^+ = E_j^-$. Le flux effectif ν_j est alors par construction :

$$\nu_j := \nu_j^+ - \nu_j^-, \tag{2.42}$$

où

$$\begin{aligned} \nu_j^+ &= \max\{0, \overline{\nu}_j\}, \\ \nu_j^- &= \max\{0, \underline{\nu}_j\}. \end{aligned} \tag{2.43}$$

et les flux $\overline{\nu}_j, \underline{\nu}_j$ sont tels que :

$$\begin{aligned} (C_4^+): \quad \overline{\nu}_j &\leqslant E_j^+ k_j^+ \frac{1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{-m_{lj}^+} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K_{mlj}}\right)^{n_{lj}}}, \\ (C_4^-): \quad \underline{\nu}_j &\leqslant E_j^- k_j^- \frac{1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n'_{lj}}}{K'_{jeq}}}{1 + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{-m_{lj}^-} + \prod_{l=1}^{q} \left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{n'_{lj}}}, \end{aligned} \tag{2.44}$$

où n'_{lj} est le coefficient stoechiométrique, positif pour les métabolites de $\mathbb{S}$ et négatifs pour ceux de $\mathbb{P}$ (autrement dit, il est « l'homologue » de n_{lj} au cas de la configuration (2.25) voir (2.39)). K'_{mlj} l'affinité du l^{eme} réactif par rapport à la réaction (2.41), et K'_{jeq} est la constante d'équilibre de la réaction (2.41).

Ce modèle de l'enzyme est illustré sur la figure 2.1 et il conduit pour intégrer les aspects thermodynamiques et cinétiques au sein de la méthodologie RBA, à remplacer les contraintes liées à (C_{1e}), définies page 10, par le couple de contraintes définies par (2.44) et (2.42).

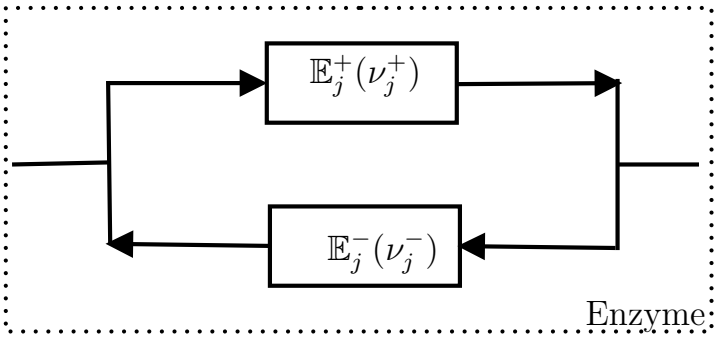

FIGURE 2.1 – Fonctionnement de l'enzyme

2.4 Définition du problème RBA étendu

L'ajout des contraintes thermodynamiques et cinétiques au problème RBA classique conduit à formuler le problème RBA étendu en modifiant la forme initiale du problème RBA classique en substituant aux contraintes (C_{1e}), les contraintes (C_4^+) et (C_4^-) et en ajoutant les contraintes (C_5) (a) et (b). Cela conduit à formuler le problème RBA étendu à travers ce problème d'optimisation : Etant donné P_G

$$\max_{\mu\geqslant 0, R_a\geqslant 0, \nu_j^-, \nu_j^+, \underline{\nu}_j, \overline{\nu}_j, \bar{X}_l\geqslant 0, E_j\geqslant 0,\ \nu_j} \mu$$

tel que :

$$(C_{1a}): \quad \mu\left(\sum_{j=1}^{m} C_{M_{ij}}^{M_p} E_j + C_{R_i}^{M_p} R_a + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_p} P_{G_j}\right) \leqslant \sum_{j=1}^{m} S_{pij}\nu_j + \nu_Y, i \in \{1,\ldots,N_p\}$$

$$(C_{1b}): \quad -\sum_{j=1}^{m} S_{cij}\nu_j + \mu\bar{X}_{c_i} \leqslant 0, i \in \{1,\ldots,N_c\}$$

$$(C_{1c}): \quad \mu\left(\sum_{j=1}^{m} C_{M_{ij}}^{M_r} E_j + C_{R_i}^{M_r} R_a + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_r} P_G\right) \leqslant -\sum_{j=1}^{m} S_{rij}\nu_j - \nu_Y, i \in \{1,\ldots,N_r\}$$

$$(C_{1d}): \quad \sum_{j=1}^{m} S_{Iij}\nu_j = 0, i \in \{1,\ldots,N_i\}$$

$$(C_2): \quad \mu\left(\sum_{j=1}^{m} C_{M_j}^{R} E_j + C_R^R R_a + \sum_{j=1}^{N_G} C_{G_j}^{R} P_{G_j}\right) \leqslant k_T R_a,$$

$$(C_3): \quad \tfrac{1}{D}\left(\sum_{j=1}^{m} C_{M_j}^{D} E_j + C_R^D R_a + \sum_{j=1}^{N_G} C_{G_j}^{D} P_{G_j}\right) \leqslant 1,$$

$$(C_4^+): \quad \overline{\nu}_j \leqslant E_j k_j^+ \frac{1-\frac{\prod_{l=1}^{q}\bar{X}_l^{n_{lj}}}{K_{jeq}}}{1+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{-m_{lj}^+}+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{n_{lj}}}, j \in \{1,\ldots,m\}$$

$$(C_4^-): \quad \underline{\nu}_j \leqslant E_j k_j^- \frac{1-\frac{\prod_{l=1}^{q}\bar{X}_l^{n'_{lj}}}{K'_{jeq}}}{1+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{-m_{lj}^-}+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{n'_{lj}}}, j \in \{1,\ldots,m\}$$

$$(C_5): \quad \nu_j = \nu_j^+ - \nu_j^-, j \in \{1,\ldots,m\}$$

$$(a): \quad \nu_j^+ = \max\{0, \overline{\nu_j}\}, j \in \{1,\ldots,m\}$$

$$(b): \quad \nu_j^- = \max\{0, \underline{\nu_j}\}, j \in \{1,\ldots,m\} \tag{2.45}$$

On note qu'à l'exception des flux qui sont de signe quelconque, le reste des variables de décision sont positives.

Pour simplifier le problème, nous éliminons, sans perte de généralité, les variables $\overline{\nu_j}, \underline{\nu_j}$ ainsi que les contraintes $(a), (b)$. Dans ce cas, l'ensemble des concentrations des métabolites $\bar{X}_l$ telles que $1 - \frac{\prod_{l=1}^{q}\bar{X}_l^{n_{lj}}}{K_{jeq}} < 0$ font que $\nu_j^+ < 0$ (voir (C_4^+)), or ν_j^+ est par définition un flux métabolique positif, ainsi cet ensemble de concentration de métabolite ne fait pas partie de l'ensemble des concentrations faisables pour le problème RBA étendu. Le même raisonnement est valable pour ν_j^-. Ainsi les contraintes $(a), (b)$ et les variables $\overline{\nu_j}, \underline{\nu_j}$ seront désormais éliminées du problème et les contraintes (C_4^+) et (C_4^-) deviendront :

$$(C_4^+): \quad \nu_j^+ \leqslant E_j k_j^+ \frac{\max\{0, 1-\frac{\prod_{l=1}^{q}\bar{X}_l^{n_{lj}}}{K_{jeq}}\}}{1+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{-m_{lj}^+}+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{n_{lj}}}, j \in \{1,\ldots,m\} \tag{2.46}$$

$$(C_4^-): \quad \nu_j^- \leqslant E_j k_j^- \frac{\max\{0,1-\frac{\prod_{l=1}^{q} \bar{X}_l^{n'_{lj}}}{K'_{jeq}}\}}{1+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{-m_{lj}^-}+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{n'_{lj}}}, j \in \{1,\ldots,m\} \tag{2.47}$$

avec $\nu_j^+ \geqslant 0$ et $\nu_j^- \geqslant 0$. On remarque que les contraintes (2.46) et (2.47) ne sont pas a priori convexes. En effet, si on prend le cas de la contrainte (C_4^+) (le cas de (C_4^-) étant similaire), on remarque qu'elle peut s'écrire sous la forme

$$\nu_j^+\left(1+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{-m_{lj}^+}+\prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{n_{lj}}\right)-k_j^+E_j\max\left\{0,1-\frac{\prod_{l=1}^{q}\bar{X}_l^{n_{lj}}}{K_{jeq}}\right\}\leqslant 0.$$

En effet, cette contrainte fait intervenir des termes sous forme de produit mixte entre les différents variables de décision à savoir $\nu_j, E_j, \bar{X}_l$ et ces produits ne définissent pas a priori une fonction convexe. Ainsi, l'intégration de ces contraintes au sein du programme (2.14) initial conduit dans la forme que nous avons choisie à un problème d'optimisation formalisé de façon non convexe.

2.5 Conclusion

L'utilisation du modèle générique de l'activité enzymatique proposé dans [40] permet de formuler sous la forme d'un problème d'optimisation une extension du problème RBA classique, en y intégrant explicitement les contraintes thermodynamiques et cinétiques que doivent respecter l'ensemble des réactions enzymatiques du réseau métabolique. Cela permet d'enrichir les entités pouvant être prédites par les RBA classiques, en y ajoutant explicitement l'ensemble des concentrations des métabolites présents dans la bactérie. Cela implique que la résolution du problème d'optimisation associé au problème RBA étendu permettrait de prédire en ce qui concerne le réseau métabolique, en plus des concentration des divers enzymes, la répartition fine et optimale des concentrations des métabolites conduisant implicitement à maximiser la production de la biomasse en utilisant une quantité totale minimale d'enzyme .

Le fait que l'intégration de ces nouvelles contraintes structurelles au sein des RBA classiques conduise à un problème d'optimisation dont la formulation à ce stade est non convexe est critique. En effet, si le caractère non convexe du problème est confirmé, cela réduira drastiquement l'intérêt de cette nouvelle formulation au regard de la dimension des problèmes qu'il s'agit de résoudre (quelques milliers de variables de décision). Dans cette perspective, il est clair que l'hypothèse simplificatrice proposée dans [29, 28, 30] (hypothèse 1 page 10) qui consiste à représenter l'efficacité de chaque enzyme par un simple paramètre constant peut être interprétée comme une première relaxation convexe du problème que nous venons d'établir.

En fait, le caractère convexe ou non convexe du problème RBA étendu est moins immédiat qu'il n'y parait car les nouvelles contraintes, en particulier (C_4^+) et (C_4^-) correspondent à des contraintes classiquement considérées en optimisation dite géométrique (qui conduit après un changement de variables adéquate à transformer les contraintes géométriques en des contraintes convexes). La question est donc de savoir si les RBA étendus peuvent ou ne peuvent pas rentrer dans le cadre de l'optimisation géométrique dite mixte ou plus généralement de l'optimisation convexe. Ces questions sont traitées au chapitre suivant.

Chapitre 3

Formulation du problème d'optimisation pour le modèle RBA étendu

L'intégration des aspects thermodynamiques et cinétiques au sein du cadre RBA permet à la méthode de gagner en pouvoir de prédiction. Cependant, comme nous allons le voir, cela se fait au détriment de son efficacité de résolution numérique. Les RBA classiques aboutissent à la résolution itérative de problèmes de faisabilité linéaires [29, 28, 30]. La classe de problèmes d'optimisation linéaire bénéficie du privilège d'être résolue efficacement par les méthodes de points intérieurs et cela pour des milliers de variables de décision [15, 70]. Malheureusement, le problème RBA étendu formulé dans le chapitre précédent ne possède pas a priori ces propriétés de convexité et de linéarité. Comme il s'agit ici aussi de manipuler de l'ordre de quelques milliers de variables de décision, la nécessité de se ramener au cadre de l'optimisation convexe à cette échelle est essentielle si on veut préserver le pouvoir prédictif de la méthode. En effet à cette échelle seuls les problèmes d'optimisation convexes peuvent être résolus efficacement [48].
Dans la première partie de ce chapitre, nous explorons la possibilité de reformuler le problème d'optimisation associé au modèle RBA étendu comme un problème d'optimisation géométrique-linéaire mixte [70, 14], les problèmes d'optimisation de ce type pouvant être résolus efficacement. Il s'agit a priori de la classe de problèmes d'optimisation « convexe » qui présente la structure la plus proche de celle du problème d'optimisation associé au RBA étendu. Malheureusement, un exemple montre que les RBA étendus sont en fait non convexes et ceci conduira dans une seconde partie à explorer des pistes proposant des relaxations convexes du problème RBA étendu.

3.1 Classe de problèmes d'optimisation géométriques : terminologie et propriétés principales

Dans cette section on va présenter une classe d'optimisation qui n'est pas convexe dans sa formulation initiale mais que l'on peut mettre sous une forme convexe par simple changement de variable et par un simple passage au logarithme. Dans ce sens on considère, par abus de langage, cette classe comme convexe : on peut toujours la mettre, d'une façon équivalente, sous sa forme convexe lors de la résolution numérique comme nous le verrons

dans la suite de ce chapitre. Pour cela on introduit d'abord quelques définitions nécessaires (voir [14] pour plus de détails).

Définition 4 (Fonctions monomiales et posynomiales)**.** *[14] Une fonction* $f : \mathbb{R}_+^{n*} \to \mathbb{R}$ *définie par*

$$f(x) = cx_1^{a_1}x_2^{a_2}\dots x_n^{a_n}, \tag{3.1}$$

où $c > 0$ *et* $a_i \in \mathbb{R}$ *est dite fonction monomiale. La somme de plusieurs fonctions monomiales (K monomiales) est la fonction définie par*

$$f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} \dots x_n^{a_{nk}}, \tag{3.2}$$

où $c_k > 0$ *pour tout k, est dite fonction posynomiale.*

Problème d'optimisation géométrique

Définition 5 (Problème d'optimisation géométrique)**.** *Le problème d'optimisation de la forme :*

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f_0(x) \\ &\quad f_i(x) \leqslant 1, \quad i = 1,\dots,m \\ \textit{tel que :}\ &g_i(x) = 1, \quad i = 1,\dots,p \\ &x_i > 0, \end{aligned} \tag{3.3}$$

où $f_0,\dots,f_m$ *sont des fonctions posynomiales et* $g_1,\dots,g_p$ *sont des fonctions monomiales, est dit un problème d'optimisation géométrique (PG). Les contraintes inégalités sont dites contraintes posynomiales et les contraintes égalités sont dites des contraintes monomiales. Les variables de décision intervenant dans les fonctions posynomiales sont dites variables posynomiales. Les variables intervenant dans les contraintes monomiales sont dites variables monomiales.*

Afin de mettre le problème (3.3) sous sa forme convexe, on va procéder à un pré-traitement. On va effectuer le changement de variable suivant :

$$x_i = e^{y_i}, \quad i = 1,\dots,n. \tag{3.4}$$

Ainsi la fonction monomiale de (3.1) devient :

$$\begin{aligned} f(x) &= f(e^{y_1},\dots,e^{y_n}) \\ &= ce^{a_1y_1}\dots^{a_ny_n} \\ &= e^{a^Ty+b} \end{aligned}$$

où $b = \ln(c)$, $a' = [a_1,\dots,a_n]$ et $y' = [y_1,\dots,y_n]$. D'une façon similaire, avec le même changement de variable (3.4) la fonction posynomiale de (3.2) devient :

$$f(x) = \sum_{k=1}^{K} e^{a_k^Ty+b_k},$$

où $a_k = (a_{1k}, \ldots, a_{nk})$ et $b_k = \ln(c_k)$. Ainsi, après changement de variable, la fonction monomiale (3.1) est devenue une exponentielle d'une fonction affine, la fonction posynomiale (3.2) est devenue une somme d'exponentielle de fonction affine. Cette formulation des fonctions monomiales et posynomiales est convexe. Par conséquent, en appliquant ce traitement à ses contraintes monomiales et posynomiales, le problème (3.3) peut être exprimé d'une façon équivalente sous la nouvelle forme (encore non convexe du fait des contraintes égalités non affines en y)

$$\begin{aligned} &\min_{y\in\mathbb{R}^n} \sum_{k=1}^{K} e^{a_{0k}^T y + b_{0k}} \\ \text{tel que : } &\sum_{k=1}^{K_i} e^{a_{ik}^T y + b_{ik}} \leqslant 1, \quad i = 1, \ldots, m \\ &e^{h_i^T y + b} = 1, \quad i = 1, \ldots, p \end{aligned} \tag{3.5}$$

où les $a_{ik} \in \mathbb{R}^n$, $i = 0, \ldots, m$, contiennent les exposants des inégalités posynomiales de (3.3) et les $h_i \in \mathbb{R}^n$, $i = 1, \ldots, p$ contiennent les exposants des inégalités monomiales de (3.3).

La fonction objectif ainsi que les contraintes inégalités sont sous la forme d'une somme de fonctions composées d'exponentielles et de fonctions affines en y. Cette forme est convexe car la composée d'une fonction exponentielle avec une fonction affine est convexe et la somme de fonctions convexes est convexe. Seules les contraintes égalités restent non convexes. En appliquant le logarithme de part et d'autre de ces égalités, on obtient des fonctions affines en y donc convexes. Par conséquent nous pouvons nous ramener à une version convexe du problème (3.5) qui est la suivante :

$$\begin{aligned} &\min_{y\in\mathbb{R}^n} \sum_{k=1}^{K} e^{a_{0k}^T y + b_{0k}} \\ \text{tel que : } &\sum_{k=1}^{K_i} e^{a_{ik}^T y + b_{ik}} \leqslant 1, \quad i = 1, \ldots, m \\ &h_i^T y + b = 0, \quad i = 1, \ldots, p \end{aligned} \tag{3.6}$$

Problème d'optimisation géométrique-linéaire mixte Nous présentons maintenant une variante du problème (3.3) possédant des contraintes dont le membre de droite est une fonction affine :

$$\begin{aligned} &\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^{n'}} f_0(x) \\ \text{tel que : } &f_i(x) \leqslant e_i(z), \quad i = 1, \ldots, m \\ &g_i(x) = 1, \quad i = 1, \ldots, p \end{aligned} \tag{3.7}$$

c'est-à-dire les e_i sont des fonctions affines. Ce problème a la structure particulière de présenter des contraintes mixtes avec des fonctions posynomiales et affines (les contraintes inégalités). On remarque que les variables de décision sont partitionnées de façon à ce que l'on ait d'un côté des variables qui interviennent uniquement dans les fonctions posynomiales et monomiales (variables posynomiales x) et de l'autre côté, les variables intervenant uniquement dans la partie affine (e_i) : les z qu'on appellera dans ce cas des variables

« linéaires ». On appelle cette contrainte, contrainte géométrique linéaire mixte. Cette partition fait que si on élimine les variables x on se ramène à un problème d'optimisation convexe et si on élimine les variables z on se ramène au problème d'optimisation géométrique sous sa forme posynomiale (3.3). Pour déterminer la forme convexe de cette classe particulière de problèmes d'optimisation, on considère d'abord les contraintes mixtes et on procède à un changement de variable uniquement pour les variables x et d'une façon similaire au cas de la classe d'optimisation géométrique présentée plus haut. Les variables z resteront intactes. En effet après être passé aux nouvelles variables y via le changement variable $e^y = x$ on obtient,

$$f_i(e^y) \leqslant e_i(z), \quad i = 1, \ldots, m \tag{3.8}$$

ce qui est une contrainte convexe vu la convexité des fonctions $f_i(e^y)$. Concernant les contraintes monomiales (contraintes d'égalité) le calcul de leurs formes convexes est similaire à ce qu'on a présenté plus haut. Ainsi, à partir du problème (3.7), on se ramène (similairement à ce qui précède et en gardant les mêmes notations) d'une façon équivalente au problème d'optimisation convexe suivant :

$$\begin{aligned} &\min_{y\in\mathbb{R}^n, z\in\mathbb{R}^{n'}} \sum_{k=1}^{K} e^{a_{0k}^T y + b_{0k}} \\ &\text{tel que : } \begin{array}{l} f_i(e^y) \leqslant e_i(z), \quad i = 1, \ldots, m \\ h_i^T y + b = 0, \quad i = 1, \ldots, p \end{array} \end{aligned} \tag{3.9}$$

Pour plus de détails sur les propriétés et les exemples d'application de l'optimisation géométrique, voir [14].

La classe de problèmes d'optimisation géométrique est assez répandue en applications (voir [14]) et a bénéficié des récentes avancées des méthodes de résolution numérique rendant sa résolution efficace. En effet, les problèmes de cette classe peuvent être résolues pour un nombre de variables de décision pouvant aller jusqu'à 1000 et avec un nombre de contraintes de l'ordre de 10000 [14], ceci grâce aux méthodes de points intérieurs pouvant résoudre les problèmes d'optimisation géométrique, avec la précision désirée, en un nombre d'itérations polynômial en fonction du nombre des variables de décision et de contraintes. Dans le cas où la dimension du problème n'est pas extrêmement grande (inférieure à 10^6 variables de décision par exemple), les méthodes de points intérieurs restent les plus performantes à nos jours (en terme de nombre maximal d'itérations et d'opérations arithmétiques requises pour la résolution du problème avec une certaine précision) [48, 47, 9]. La complexité de telles méthodes appliquées sur la classe de problème (3.7) est donnée dans le théorème suivant.

Théorème 4. *[51] Pour calculer une solution avec une précision ϵ au problème (3.9) avec une méthode du type points intérieurs (méthode des barrières), il suffit d'un nombre d'itérations n'excédant pas*

$$O(1)(k+m)^{1/2} \ln \frac{(k+m)Cste}{\epsilon}$$

et un nombre total d'opérations arithmétiques n'excédant pas

$$O(1)k(m+n)(n+k)(m+k)^{1/2} \ln \frac{(m+k)Cste}{\epsilon}.$$

où k est le nombre des exposants total intervenant dans (3.9), m le nombre de contraintes, n le nombre de variable de décision et $Cste$ une constante liée à la structure du problème (3.9).

La complexité polynomiale de (3.9) (en termes de nombre d'itérations et d'opérations arithmétiques) sous-entend l'efficacité de résolution numérique (rapidité) car elle n'explose pas en fonction de la taille du problème (nombre de contraintes et de variables de décision). De plus pour augmenter la précision de résolution numérique d'un digit, le terme argument du logarithme, nous montre que la complexité (nombre d'itérations et nombre d'opérations arithmétiques) est simplement multipliée par un nombre constant. Cela fait l'un des aspects essentiels qui font l'efficacité des méthodes de points intérieurs. Ceci dit, il ne faut pas que la taille du problème (n, m) soit extrêmement grande auquel cas l'efficacité des méthodes de points intérieurs n'est pas suffisante. On reviendra sur ce point dans la seconde partie de cette thèse.

3.2 Le problème RBA étendu n'est pas un problème géométrique-mixte, il est de plus non convexe en général

La section précédente a précisé la classe des problèmes géométriques qui pouvaient se ramener à la résolution de problème convexe. L'analyse détaillée du problème défini en (2.45) page 24, nous indique que le problème RBA étendu ne rentre pas dans le cadre général de la section précédente car de fait, certaines contraintes mélangent des variables de décision attachées au problème linéaire (les ν_j typiquement) avec des variables de décision attachées aux contraintes géométriques.
Néanmoins, même si cela semble indiquer que le problème considéré n'est pas convexe, il faut comme dans tous les problèmes d'optimisation, nous assurer que cette non convexité n'est pas due à la formulation que nous avons choisie et qu'elle est de fait intrinsèque au problème qu'on cherche à résoudre. La façon la plus directe pour montrer cela est d'exhiber un exemple démontrant la non convexité du problème. C'est ce que nous allons maintenant faire en démontrant que la prise en compte dans le problème RBA étendu de la thermodynamique et de la cinétique enzymatique conduit à définir un problème non convexe.
Pour cela, on définit un réseau constitué d'une interconnexion entre trois réactions métaboliques :

$$\begin{aligned} X_1 &\overset{E_1}{\rightleftharpoons} X_2, \\ X_2 &\overset{E_2}{\rightleftharpoons} X_3, \\ X_4 &\overset{E_3}{\rightleftharpoons} X_3, \end{aligned} \tag{3.10}$$

où X_1 et X_4 sont deux métabolites externes de concentration connue et où le réseau respecte les contraintes stœchiométriques suivantes :

$$\begin{aligned} \nu_T &= \nu_1 - \nu_2, \\ \nu &= \nu_2 + \nu_3, \\ \nu &= 5\nu_T. \end{aligned} \tag{3.11}$$

Pour ce réseau, nous voulons optimiser le flux ν avec des contraintes sur les concentrations (les ressources) d'enzymes :

$$E_1 + 0.1\,E_2 + E_3 \leqslant 1,$$

et où l'on suppose que les concentrations externes sont fixées et égales à $X_1 = 0.17$ et $X_4 = 0.755$. Par ailleurs, pour cet exemple numérique, nous prendrons tous les paramètres (les constantes de vitesse, constantes d'équilibre et constantes de Michaelis et Menten) des trois réactions égaux à 1. En se basant sur le chapitre 2, on définit ici un sous problème du problème RBA étendu, consistant à ne considérer que les contraintes liées au réseau métabolique et que nous écrivons sous la forme suivante (de façon compacte) :

$$\begin{aligned}
&\text{maximize} && \nu \\
&\nu_1, \nu_2, \nu_3, \nu_T, \nu, \\
&E_1 \geqslant 0, E_2 \geqslant 0, E_3 \geqslant 0, \\
&X_2 \geqslant 0, X_3 \geqslant 0 \\
&\text{tel que} \\
&\qquad \nu_T = \nu_1 - \nu_2 \\
&\qquad \nu = \nu_2 + \nu_3 \\
&\qquad \nu = 5\nu_T \\
&\qquad 0 \leqslant sign(X_1 - X_2)\nu_1 \leqslant sign(X_1 - X_2)\tfrac{X_1 - X_2}{1 + X_1 + X_2} E_1 \\
&\qquad 0 \leqslant sign(X_2 - X_3)\nu_2 \leqslant sign(X_2 - X_3)\tfrac{X_2 - X_3}{1 + X_2 + X_3} E_2 \\
&\qquad 0 \leqslant sign(X_4 - X_3)\nu_3 \leqslant sign(X_4 - X_3)\tfrac{X_4 - X_3}{1 + X_3 + X_4} E_3 \\
&\qquad E_1 + 0.1E_2 + E_3 \leqslant 1
\end{aligned}$$

On résoud ce problème à l'aide de la fonction `fmincon` de Matlab en initialisant l'algorithme de minimisation en 1000 valeurs initiales, chacune des composantes étant choisie uniformément dans l'intervalle $[0, 1]$. On obtient alors deux optima locaux (qui sont isolés) et qui sont reportés dans la table 3.1.

ν^*	0.2869	0.2702
ν_1	0	0.0540
ν_2	-0.0574	0
ν_3	0.3442	0.2702
ν_T	0.0574	0.0540
E_1	0	0.3719
E_2	1.0659	0
E_3	0.8934	0.6281
X_2	0	0
X_3	0.0569	0

TABLE 3.1 – Deux optima locaux

Ces deux optima locaux sont représentés sur la figure 3.1 (les échelles et les couleurs ne sont pas les mêmes sur les deux figures). Ces figures ont été obtenues en « griddant » les

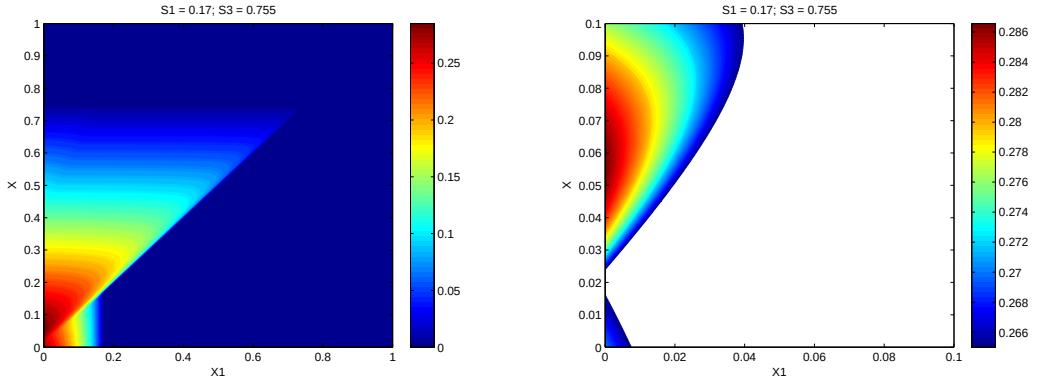

FIGURE 3.1 – Illustration de deux optima locaux

valeurs des concentrations des métabolites internes. Le problème devient alors un problème de programmation linéaire dont l'optimum global est connu.
Ce sous-problème démontre qu'intrinsèquement, le problème que nous cherchons à traiter est non convexe et donc nous n'avons pas besoin d'explorer davantage la possibilité de le transformer en un problème convexe équivalent. En revanche, on peut s'interroger sur la possibilité de faire une relaxation convexe de ce problème, c'est-à-dire de le résoudre de façon approchée mais en appelant la résolution de problèmes d'optimisation convexe. C'est l'objet de la prochaine section.

3.3 Relaxation convexe des RBA étendus

D'une certaine façon, la non convexité du problème RBA étendu est liée à plusieurs difficultés. Nous traitons dans un premier temps celles liées aux contraintes (C_4^+) et (C_4^-) car nous allons montrer que l'on peut les approcher de façon très satisfaisantes par des contraintes convexes. Nous traiterons ensuite celles liées aux contraintes qui mélangent les variables de décision attachées au problème linéaire (les ν_j) avec des variables de décision attachées aux contraintes géométriques.

3.3.1 Approximation convexe des contraintes (C_4^+) et (C_4^-)

On se focalise sur la contrainte (C_4^+), la démarche pour (C_4^-) étant similaire. On remarque que le dénominateur a une forme posynomiale. La question est peut-on se ramener à des contraintes posynomiales ? Une première étape est de remonter le dénominateur vers la partie gauche des contraintes en le multipliant par ν_j^+. Maintenant, si le terme $\max\{0, 1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n'_{lj}}}{K'_{jeq}}\}$ avait une forme monomiale ou bien une forme $1/P$ avec P une fonction posynomiale alors il suffirait de diviser l'inégalité par ce terme et par le produit $E_j^+ k_j^+$ et ainsi on se ramènerait au rapport d'une fonction posynomiale au numérateur par une fonction monomiale au dénominateur ou bien au produit de deux fonctions posynomiales. Dans les deux cas, on obtiendrait une fonction posynomiale et la contrainte serait sous la forme d'une fonction posynomiale inférieure à 1. Il s'agit du type de contrainte d'un problème d'optimisation géométrique sous sa forme posynomiale comme on l'a présenté

dans la section précédente. Malheureusement le terme $\max\{0, 1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n'_{lj}}}{K'_{jeq}}\}$ n'est pas posynomial. Néanmoins, on va l'approcher par une fonction posynomiale à l'aide d'une interpolation.

Pour ce faire, on se fixe un domaine admissible pour les concentrations des métabolites $\bar{X}_l$, $l = 1, \ldots, q$ où q est le nombre total des métabolites du réseau en question. Ce domaine sera défini par l'hyperrectangle suivant :

$$H = I_1 \times \ldots \times I_q, \tag{3.12}$$

où $I_i = [\bar{X}_i^{min}, \bar{X}_i^{max}]$, $\bar{X}_i^{min}$ et $\bar{X}_i^{max}$ correspondent aux ordres de grandeur des concentrations minimales et maximales du point de vue biologique pour un réseau métabolique donné. Ensuite, on pose :

$$Z = \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}. \tag{3.13}$$

Au lieu d'interpoler directement le terme $\max\{0, 1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}\}$ sur l'hyperrectangle H, on va interpoler simplement le terme $\max\{0, 1 - Z\}$ sur un simple intervalle $\mathcal{D}$ de $\mathbb{R}$ défini par :

$$\mathcal{D} = [\mathcal{D}_{min}, \mathcal{D}_{max}], \tag{3.14}$$

où $\mathcal{D}_{min}$ et $\mathcal{D}_{max}$ sont les valeurs optimales des deux problèmes d'optimisation géométrique suivants :

$$\mathcal{D}_{min} := \min_{\bar{X}_1, \ldots, \bar{X}_q} \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}} \tag{3.15}$$
$$\text{tel que : } \begin{array}{l} \bar{X}_l^{-1}.\bar{X}_{lmin} \leqslant 1 \quad l = 1, \ldots, q, \\ \bar{X}_l.\bar{X}_{lmax}^{-1} \leqslant 1 \quad l = 1, \ldots, q, \end{array}$$

$$\mathcal{D}_{max} := \min_{\bar{X}_1, \ldots, \bar{X}_q} \left(\frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}} \right)^{-1} \tag{3.16}$$
$$\text{tel que : } \begin{array}{l} \bar{X}_l^{-1}.\bar{X}_{lmin} \leqslant 1 \quad l = 1, \ldots, q, \\ \bar{X}_l.\bar{X}_{lmax}^{-1} \leqslant 1 \quad l = 1, \ldots, q, \end{array}$$

L'avantage de cette technique est de ramener l'interpolation d'une fonction à plusieurs variables (ce qui est en général une tâche nécessitant un effort numérique très important) à celle d'une fonction à une seule variable sur un intervalle borné, ce qui ne pose aucune difficulté numérique et permet d'avoir des résultats plus stables numériquement. Le choix de notre fonction interpolante comme on l'a expliqué est soit une fonction monomiale, soit une fonction de la forme $1/P$ avec P posynomiale. On choisit le deuxième choix car il est plus riche en degrés de liberté que celui d'un simple monomial. Ce problème peut être posé formellement sous la forme suivante :

Formulation du problème 1. *Soit la fonction θ définie par :*

$$\begin{array}{rccl} \theta: & \mathcal{D} \subset \mathbb{R} & \to & \mathbb{R} \\ & Z & \mapsto & \max\{0, 1-Z\}. \end{array} \tag{3.17}$$

étant donné $n \in \mathcal{N}$, quelle est la fonction qui représente la meilleure approximation de θ sur $\mathcal{D}$ au sens des moindres carrées et ayant la structure particulière suivante :

$$\hat{\theta}_n(Z) = \frac{1}{1 + Pos_n(Z)}.$$

Cette meilleure approximation sera notée $\hat{\theta}_n$. Pos_n est une fonction posynomiale définie comme suit :

$$Pos_n(x) = \sum_{k=1}^{n} c_k x_1^{a_1} \dots x_n^{a_n}. \tag{3.18}$$

Afin de résoudre ce problème on applique la procédure suivante :

1) on génère un ensemble de données à partir de la fonction à interpoler θ :

$$\left(Z^{(i)}, \theta^{(i)}\right), \quad i = 1, \dots, N,$$

où $\theta^{(i)} = \max\{\epsilon, 1 - Z^{(i)}\}$ et $Z^{(i)} \in \mathcal{D} \subset \mathbb{R}^+$;

2) on fixe un nombre n de monômes pour la fonction posynômiale Pos_n et on cherche les c_k, a_k pour $k = 1, \dots, n$ tels que :

$$\hat{\theta}_n(Z^{(i)}) \approx \theta^{(i)}, \quad i = 1, \dots, N.$$

Ce problème peut se formuler sous forme du problème de moindres carrés suivant :

$$\begin{aligned} \min_x \ & \sum_{i=1}^{N} (\hat{\theta}_n(Z^{(i)}) - \theta^{(i)})^2 \\ \text{tel que : } & c_k \geqslant 0, \quad k = 1, \dots, n, \end{aligned} \tag{3.19}$$

où $x = (c_1, \dots, c_n, a_1, \dots, a_n)$. La solution optimale x^* de ce problème donne les paramètres de la fonction $\hat{\theta}_n$ représentant la meilleure approximation de la fonction θ au sens des moindres carrés. Pour remonter à une bonne approximation du terme $\max\{0, 1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}\}$ sur le domaine H, on remplace dans $\hat{\theta}_n$, Z par $Z = \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}$. Dans la pratique l'approximation reste satisfaisante tant que $\bar{X}_i \in I_i$, $i = 1, \dots, q$. On pose $\alpha^+(X) = \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}$, la fonction interpolant $\max\{0, 1 - \frac{\prod_{l=1}^{q} \bar{X}_l^{n_{lj}}}{K_{jeq}}\}$ sur H est $\hat{\theta}_n(\alpha^+(\bar{X}))$. On rappelle que cette fonction a par construction la forme $1/P$ avec P posynomiale.

Discussion : Cette approche consiste, dans un premier temps, à approcher la fonction affine par morceaux θ par la fonction $\hat{\theta}_n$. L'avantage est double. D'un côté, l'approche est assez générique car la solution du problème (3.19) permet une bonne approximation de la fonction $\theta \circ \alpha^+$ pour tout n_{lj}, q et K_{jeq} sur le domaine H. D'un autre côté, le problème est plus simple que celui d'approcher directement la fonction non linéaire $\theta \circ \alpha$ ce qui nous évitera les problèmes numériques liés à la complexité de ces fonctions.

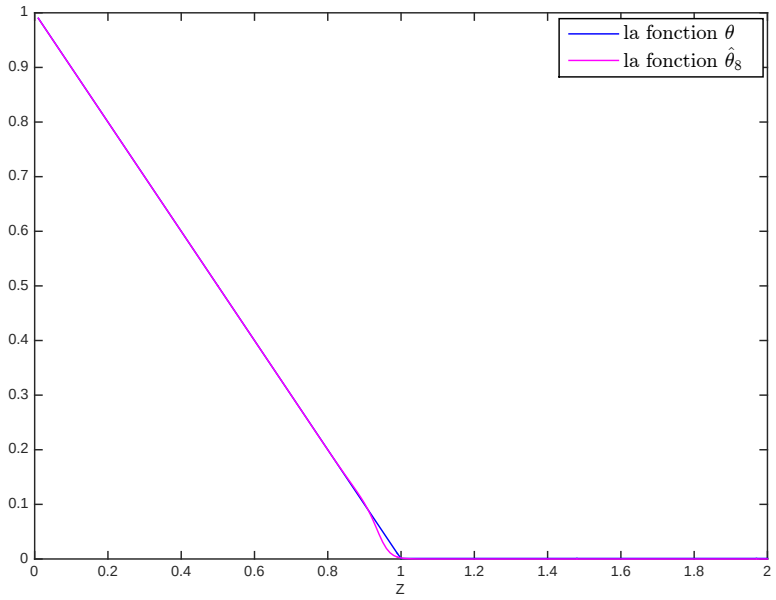

FIGURE 3.2 – Interpolation de la fonction θ par $\hat{\theta}_8$

Mise en œuvre sur un exemple numérique. On met en œuvre cette procédure d'approximation pour approcher la fonction, définie sur $\mathbb{R}^2_+$, qui à (x, y) associe $\max\{0, 1 - \frac{y^2}{100x^3}\}$. On a $\alpha^+(x, y) = \frac{y^2}{100x^3}$. On considère l'intervalle $\mathcal{D} = [10^{-3}, 2]$ comme intervalle de validité pour la variable Z. Cet intervalle a été échantillonné uniformément avec un pas de 0.1, soit 20 point $Z^{(i)}$ $(i = 1, \ldots, 20)$. Le problème (3.19) peut se résoudre par exemple avec la méthode de Levenberg-Marquardt [43, 41] ou par la méthode Trust-Region [20]. Pour cet exemple numérique, ce problème a été résolu sous Matlab avec la macro `lsqcurvefit` en choisissant la méthode Trust-Region, pour un nombre de monômes $n = 8$ de la fonction interpolante $\hat{\theta}_n$ (voir figure 3.2). Sur la figure 3.3, l'erreur due à cette interpolation atteint, en valeur absolue, une valeur maximale de 0.02 alors que la fonction à approximer atteint une valeur maximale de 1 (voir figure 3.2). On constate alors, que l'ordre de grandeur de l'erreur est négligeable devant l'ordre de grandeur des données à approcher ce qui nous permet de conclure sur la bonne qualité de cette approximation. On a pris pour intervalles de validité des concentrations x et y, le même que celui de Z et on a tracé en 3D les deux fonctions $\theta \circ \alpha^+ : (x, y) \mapsto \max\{0, 1 - \frac{y^2}{100x^3}\}$ et $\hat{\theta}_8 \circ \alpha^+$. Le résultat d'interpolation est affiché sur la figure 3.4, l'erreur d'approximation est affichée sur la figure 3.5. On remarque la même chose concernant la qualité d'approximation en dimension 1 qu'en dimension 3. La qualité d'approximation se conserve. On note aussi l'aspect générique de cette approche car une fois le problème (3.19) résolu, cette solution reste valable pour approcher toute fonction $(X_1, \ldots, X_q) \mapsto \max\{\epsilon, 1 - \frac{\prod_{l=1}^{q} X_l^{n_{lj}}}{K_{jeq}}\}$ $\forall q, n_{lj}, K_{jeq}$.

A ce stade, on peut appliquer la même procédure sur la contrainte (C_4^-). Au final on obtient

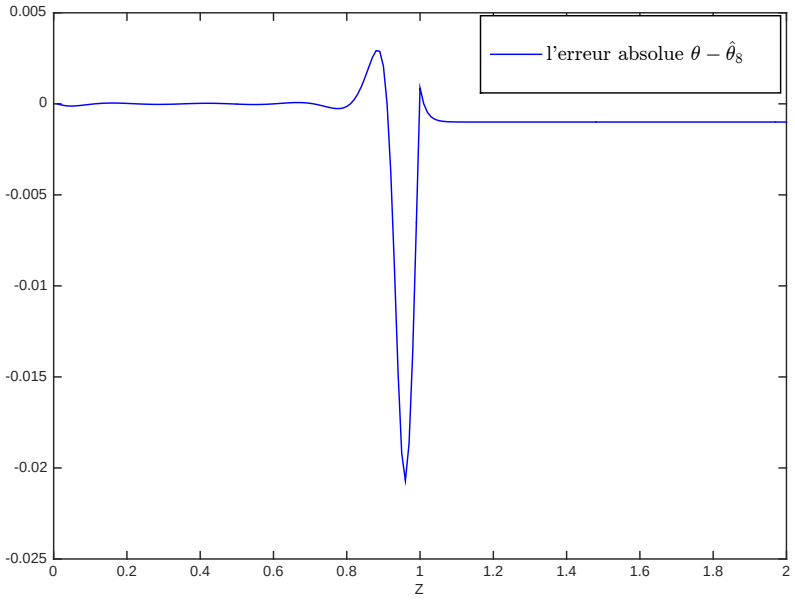

FIGURE 3.3 – Erreur absolue $\theta - \hat{\theta}_8$ sur le domaine $\mathcal{D} = [0, 2]$

l'approximation suivante des contraintes (C_4^+) et (C_4^-) qu'on notera $(\tilde{C}_4^+)$ et $(\tilde{C}_4^-)$:

$$(\tilde{C}_4^+) : \quad \frac{\nu_j^+}{E_j k_j^+}\left(1 + \prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{-m_{lj}^+} + \prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K_{mlj}}\right)^{n_{lj}}\right)\left(1 + \hat{\theta_n}(\alpha^+(\bar{X}))\right) \leqslant 1, j \in \{1, \ldots, m\}$$

$$(\tilde{C}_4^-) : \quad \frac{\nu_j^-}{E_j k_j^-}\left(1 + \prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{-m_{lj}^-} + \prod_{l=1}^{q}\left(\frac{\bar{X}_l}{K'_{mlj}}\right)^{n'_{lj}}\right)\left(1 + \hat{\theta_n}(\alpha^-(\bar{X}))\right) \leqslant 1, j \in \{1, \ldots, m\}$$

où $\bar{X} := (\bar{X}_1, \ldots, \bar{X}_q)$, $\alpha^-(\bar{X})$ représente l'homologue de $\alpha^+(\bar{X})$ dans le cas de la contrainte $(\tilde{C}_4^-)$. On obtient ainsi deux contraintes sous la forme posynomiale. Ainsi on peut passer à leurs versions convexes par un simple changement de variable et par un passage au logarithme comme on l'a expliqué dans la première section de ce chapitre. Dans cette version convexe, chacune des variables de décision de ces deux contraintes correspondra à une nouvelle variable $y_i > 0$ et interviendra dans ces deux contraintes sous une forme exponentielle : e^{y_i}. On va appeler les variables correspondant aux y_i (dans notre cas il s'agit des $\bar{X}_l$, E_j, ν_j^+, et ν_j^-) des variables posynomiales. Elles interviendront sous forme d'exponentielle de nouvelles variables y_i dans la version convexe des contraintes $(\tilde{C}_4^+)$ et $(\tilde{C}_4^-)$.

3.3.2 Approximation des contraintes mélangeant les variables de décision linéaires et géométriques

Nous passons maintenant à le seconde source de non convexité c'est-à-dire aux contraintes qui ont une structure similaire aux contraintes mixtes du problème d'optimisation (3.7), seulement les variables appartenant au problème linéaire, ici ν_j^+ et ν_j^- sont malheureusement aussi des variables posynomiales. En effet, elles interviennent aussi dans les contraintes posynomiales $(\tilde{C}_4^+)$ et $(\tilde{C}_4^-)$). Dans ce cas, on a en général des contraintes qui ne sont pas convexes car après être passé au changement de variable exponentielle, les ν_j

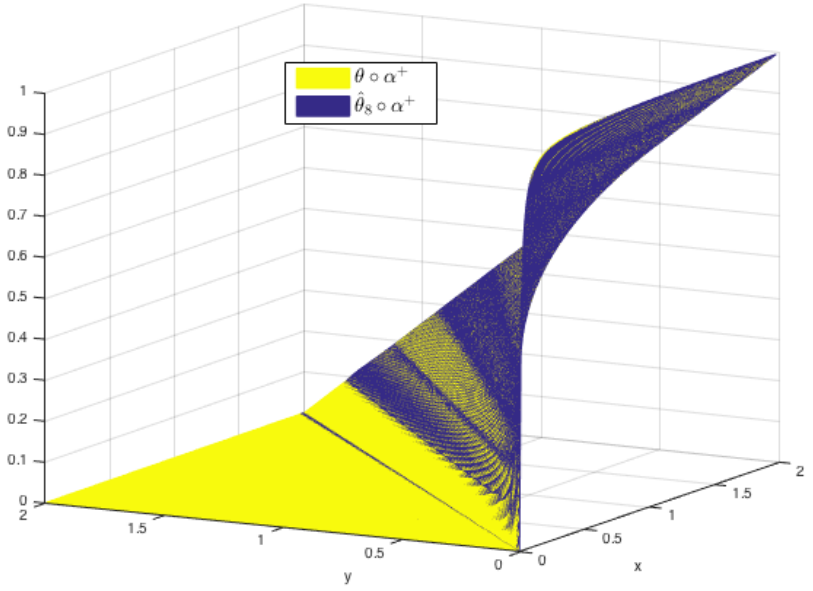

FIGURE 3.4 – Résultat de l'interpolation de $\theta \circ \alpha^+$ par $\hat{\theta}_8 \circ \alpha^+$

seront remplacés par une différence d'exponentielles (contrainte (C_5)) qui s'avère ne pas définir une contrainte convexe.

Pour pallier cette difficulté, l'idée est de remplacer les ν_j^+ et ν_j^- par des nouvelles variables μ_j^+ et μ_j^- au niveau des contraintes $(\tilde{C}_4^+)$ et $(\tilde{C}_4^-)$ de la façon suivante :

$$\begin{aligned} \mu_j^+ \nu_j^+ &= 1, \\ \mu_j^- \nu_j^- &= 1. \end{aligned} \tag{3.20}$$

Ces deux nouvelles variables sont des variables posynomiales car elles interviennent dans les fonctions posynomiales des contraintes $(\tilde{C}_4^+)$ et $(\tilde{C}_4^-)$. Ainsi, les variables ν_j^+ et ν_j^- n'interviendront plus au niveau des contraintes $(\tilde{C}_4^+)$ et $(\tilde{C}_4^-)$ et lors du passage à la forme convexe de ces deux contraintes, seules μ_j^+ et μ_j^- seront remplacées par des exponentielles. Quant aux ν_j^+ et ν_j^-, elles resteront intactes et sont considérées maintenant comme des variables affines. Ainsi, les contraintes (C_{1a}) et (C_{1c}) deviennent de la même nature que les contraintes mixtes du problème (3.7) et par conséquent peuvent être mises sous la forme convexe similairement. Notons que chaque occurrence de ν_j sera maintenant remplacée par $\nu_j^+ - \nu_j^-$, ce qui élimine la contrainte (C_5). Notons également que l'obtention de forme convexe de ces contraintes (C_{1a}) et (C_{1c}) s'est faite via l'introduction des égalités (3.20). Le calcul de cette forme convexe se répercutera sur le système (3.20) de façon à remplacer les μ_j^+ et μ_j^- par des exponentielles. Ceci donnera lieu au système de contraintes égalités avec des fonctions exponentielles, ce qui n'est pas convexe. L'approximation de ce système de contrainte est présentée maintenant.

Ici on ne va considérer que la première contrainte de (3.20), la stratégie présentée étant similaire pour la deuxième contrainte. Cette contrainte peut être divisée en deux contraintes inégalités :

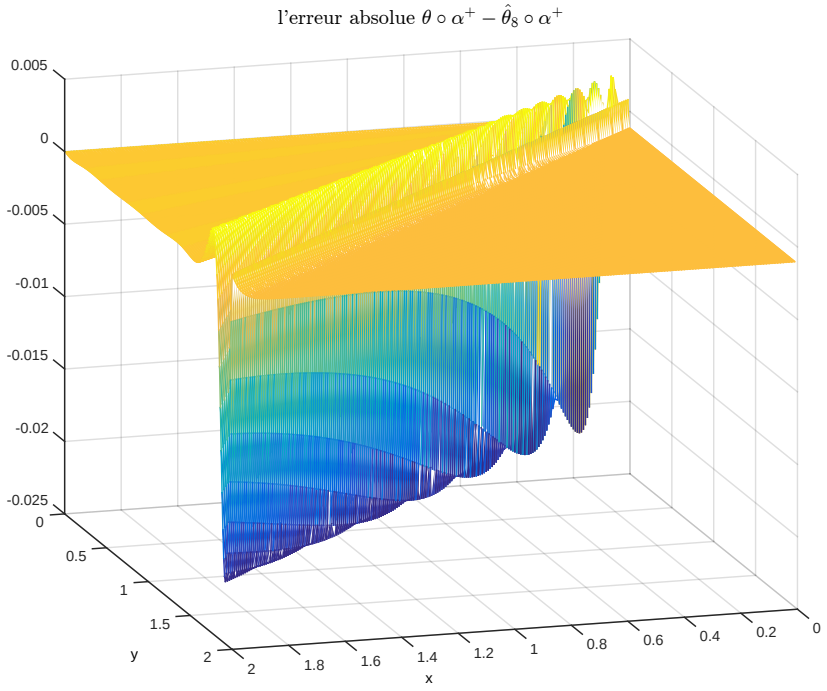

FIGURE 3.5 – Erreur absolue $\theta \circ \alpha^+ - \hat{\theta}_8 \circ \alpha^+$

$$\mu_j^+ \nu_j^+ \geqslant 1,$$
$$\mu_j^+ \nu_j^+ \leqslant 1$$

La première inégalité peut être écrite sous la forme :

$$\frac{1}{\mu_j^+} \leqslant \nu_j^+.$$

On rappelle que la variable μ_j^+ est une variable posynomiale et ν_j^+ est linéaire par conséquent cette contrainte est une contrainte géométrique-linéaire mixte (car elle fait intervenir d'une façon affine le terme monomial μ_j^{+-1} et la variable linéaire ν_j^+) qu'on peut mettre immédiatement sous sa forme convexe (voir section 3.1 page 29) en effectuant le changement de variable suivant :

$$\mu_j^+ = e^{s_j^+}, \tag{3.21}$$

ainsi notre contrainte se met sous la forme convexe suivante :

$$e^{-s_j^+} \leqslant \nu_j^+.$$

Maintenant nous passons à la deuxième inégalité à savoir :

$$\mu_j^+ \nu_j^+ \leqslant 1.$$

Avant d'effectuer le changement de variable exponentiel (3.21), nous allons nous débarrasser du terme bilinéaire $\mu_j^+ \nu_j^+$ en le remplaçant par son enveloppe convexe. Pour cela nous allons d'abord rapeller la définition suivante :

Définition 6. *[1][enveloppe convexe] L'enveloppe convexe d'une fonction f sur un ensemble convexe C, $conv_C f$, est la fonction qui vérifie les conditions suivantes :*

i) $conv_C f$ est convexe sur C ;

ii) $conv_C f(x) \leqslant f(x)$, pour tout $x \in C$;

iii) si g est une fonction convexe sur C qui satisfait l'inégalité :

$$g(x) \leqslant f(x), \quad \forall x \in C,$$

alors on a

$$g(x) \leqslant conv_C f(x) \quad \forall x \in C.$$

Autrement dit, l'enveloppe convexe d'une fonction sur C est la plus grande fonction convexe sur C bornant inférieurement la fonction en question. Ainsi le remplacement de la fonction en question par son enveloppe convexe au niveau d'une contrainte d'un problème d'optimisation donnera l'approximation convexe la plus petite possible au sens de l'inclusion. L'enveloppe convexe du terme bilinéaire $\mu_j^+\nu_j^+$ est donnée dans la proposition suivante :

Proposition 2. *[1] On considère que $(\mu_j^+, \nu_j^+) \in \Omega_j^+ := [m_j^+, M_j^+] \times [n_j^+, N_j^+]$. L'enveloppe convexe du terme $\mu_j^+\nu_j^+$ est donnée par*

$$conv_{\Omega_j^+}\mu_j^+\nu_j^+ = \max\{m_j^+\nu_j^+ + n_j^+\mu_j^+ - m_j^+n_j^+, M_j^+\nu_j^+ + N_j^+\mu_j^+ - M_j^+N_j^+\}.$$

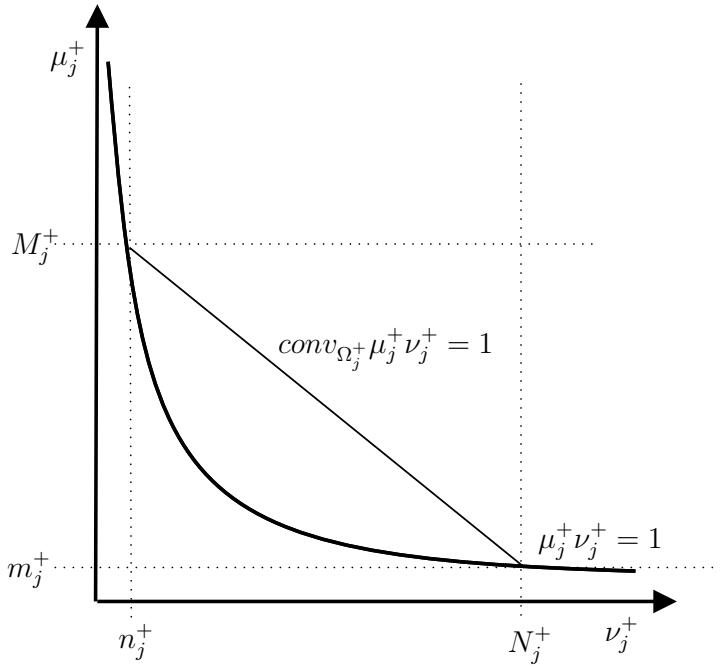

FIGURE 3.6 – $\mu_j^+\nu_j^+$ et son enveloppe convexe $conv_{\Omega_j^+}\mu_j^+\nu_j^+$ sur Ω_j^+.

Ainsi l'approximation convexe qu'on propose pour la contrainte

$$\mu_j^+\nu_j^+ \leqslant 1,$$

est la suivante :

$$\begin{aligned} &m_j^+\nu_j^+ + n_j^+\mu_j^+ - m_j^+n_j^+ \leqslant 1, \\ &M_j^+\nu_j^+ + N_j^+\mu_j^+ - M_j^+N_j^+ \leqslant 1. \end{aligned} \tag{3.22}$$

Remarque : Sachant que notre contrainte initiale est (3.20) et que par hypothèse $\nu_j^+ \in [n_j^+, N_j^+]$ alors $\mu_j^+ \in [m_j^+, M_j^+]$ avec $m_j^+ = 1/N_j^+$ et $M_j^+ = 1/n_j^+$. Ainsi si on choisit $N_j^+ = 1/n_j^+$, on a $m_j^+ = n_j^+$ et $M_j^+ = N_j^+$. Par conséquent, il est facile de voir que les deux contraintes (3.22) sont devenues redondantes. Par suite, nous adoptons le choix de $N_j^+ = 1/n_j^+$ et notre approximation convexe (3.22) devient :

$$m_j^+ \nu_j^+ + n_j^+ \mu_j^+ - m_j^+ n_j^+ \leqslant 1. \tag{3.23}$$

Une interprétation géométrique de cette approximation est donnée dans la figure 3.6.
La contrainte (3.23) est géométrique-linéaire mixte (car elle fait intervenir d'une façon affine le terme monomial μ_j^+ et la variable affine ν_j^+). Après application du changement de variable (3.21), la contrainte sera sous la forme de la somme d'une exponentielle et d'une variable linéaire c'est-à-dire :

$$m_j^+ \nu_j^+ + n_j^+ e^{s_j^+} - m_j^+ n_j^+ \leqslant 1,$$

ce qui présente une contrainte inégalité convexe.

D'une façon similaire, l'approximation convexe qu'on propose pour la contrainte

$$\mu_j^- \nu_j^- = 1,$$

est la suivante :

$$\begin{aligned} & e^{-s_j^-} \leqslant \nu_j^-, \\ & m_j^- \nu_j^- + n_j^- e^{s_j^-} - m_j^- n_j^- \leqslant 1, \end{aligned} \tag{3.24}$$

où le changement de variable pris en compte dans ce cas est :

$$\mu_j^- = e^{s_j^-}. \tag{3.25}$$

En conclusion, la relaxation convexe des contraintes égalités (3.20) est la suivante :

$$\begin{aligned} &(C_{\mu\,1}^+): \quad m_j^+ \nu_j^+ + n_j^+ e^{s_j^+} - m_j^+ n_j^+ \leqslant 1, \quad j \in \{1, \ldots, m\}, \\ &(C_{\mu\,2}^+): \quad e^{-s_j^+} \leqslant \nu_j^+ \quad j \in \{1, \ldots, m\}, \\ &(C_{\mu\,1}^-): \quad m_j^- \nu_j^- + n_j^- e^{s_j^-} - m_j^- n_j^- \leqslant 1, \quad j \in \{1, \ldots, m\}, \\ &(C_{\mu\,2}^+): \quad e^{-s_j^-} \leqslant \nu_j^-. \quad j \in \{1, \ldots, m\}. \end{aligned} \tag{3.26}$$

3.3.3 Approximation du reste des contraintes

Concernant les contraintes (C_2), le fait de diviser par $k_T R_a$ nous ramène à une inégalité posynomiale dont le membre de gauche est une fonction posynomiale et celui de droit vaut 1. La contrainte (C_3) est sous une forme de contrainte posynomiale. La contrainte (C_{1d}) présente une fonction affine des variables linéaires (les ν_j) et est donc convexe. La contrainte (C_{1b}) présente également une fonction affine en les ν_j et est donc convexe. Ainsi et à ce stade on constate que le nouveau problème RBA est déjà sous sa forme posynomiale sauf pour les deux égalités (3.20). Pour les contraintes sous forme posynomiale et les contraintes mixtes, on va procéder au changement de variable exponentielle comme cela a été expliqué dans la section précédente sur les problèmes d'optimisation géométrique.

Contraintes	Type de contraintes
(C_{1a}), (C_{1c}), (3.23), (3.24)	contraintes mixtes géométriques linéaires
(C_{1b}), (C_{1d})	contraintes affines
(C_2), (C_3), $(\tilde{C}_4^+)$ et $(\tilde{C}_4^+)$	contraintes posynomiales

Table 3.2 – Nature de chaque contrainte du problème RBA

Variables	Type de variables	Changement de variable correspondant
μ	variable posynomiale	e^s, $s \in \mathbb{R}$
E_j, $j = 1, \ldots, m$	variables posynomiales	e^{e_j}, $e_j \in \mathbb{R}$, $j = 1, \ldots, m$
ν_j^+, $j = 1, \ldots, m$	variables linéaires	pas de changement de variable à effectuer
ν_j^-, $j = 1, \ldots, m$	variables linéaires	pas de changement de variable à effectuer
R_a	variable posynomiale	e^{r_a}, $r_a \in \mathbb{R}$
$\bar{X}_i$, $i = 1, \ldots, q$	variables posynomiales	$e^{\bar{x}_i}$, $\bar{x}_i \in \mathbb{R}$, $i = 1, \ldots, q$
μ_j^+, $i = 1, \ldots, m$	variables posynomiales	$e^{s_i^+}$, $s_i^+ \in \mathbb{R}$, $i = 1, \ldots, m$
μ_j^-, $i = 1, \ldots, m$	variables posynomiales	$e^{s_i^-}$, $s_i^+ \in \mathbb{R}$, $i = 1, \ldots, m$

Table 3.3 – Nature et changement de variable correspondant à chaque variable dans le problème RBA

3.3.4 Une relaxation convexe pour les RBA étendus

Après avoir passé en revue toutes les contraintes du problème RBA étendu, on va le mettre explicitement sous sa forme convexe en se basant sur toutes les considérations précédentes. D'après ce qui précède, les tableaux 3.2 et 3.3 résument la nature de chaque variable et chaque contrainte de notre problème.

Suivant l'ensemble de ces considérations, on obtient une relaxation convexe du probleme RBA étendu sous la forme d'un problème d'optimisation géométrique mixte donné par (3.27) où toutes les variables de décision sont positives.

$$\max_{s,r_a,\nu_j^-,\nu_j^+,\bar{x}_l,e_j,s_j^+,s_j^-} s$$

tel que :

$$(C_{1a}): \quad e^s\left(\sum_{j=1}^{m} C_{M_{ij}}^{M_p} e^{e_j} + C_{R_i}^{M_p} e^{r_a} + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_p} P_{G_j}\right) \leqslant \sum_{j=1}^{m} S_{pij}(\nu_j^+ - \nu_j^-) + \nu_Y, i \in \{1,\ldots,N_p\}$$

$$(C_{1b}): \quad e^s \bar{X}_{c_i} \leqslant \sum_{j=1}^{m} S_{cij}(\nu_j^+ - \nu_j^-), i \in \{1,\ldots,N_c\}$$

$$(C_{1c}): \quad e^s\left(\sum_{j=1}^{m} C_{M_{ij}}^{M_r} e^{e_j} + C_{R_i}^{M_r} e^{r_a} + \sum_{j=1}^{N_G} C_{G_{ij}}^{M_r} P_{G_j}\right) \leqslant -\sum_{j=1}^{m} S_{rij}(\nu_j^+ - \nu_j^-), i \in \{1,\ldots,N_r\}$$

$$(C_{1d}): \quad \sum_{j=1}^{m} S_{Iij}(\nu_j^+ - \nu_j^-) = 0, i \in \{1,\ldots,N_i\}$$

$$(C_2): \quad \frac{e^s}{k_T e^{r_a}}\left(\sum_{j=1}^{m} C_{M_j}^{R} e^{e_j} + C_R^R e^{r_a} + \sum_{j=1}^{N_G} C_{G_j}^R P_{G_j}\right) \leqslant 1,$$

$$(C_3): \quad \frac{1}{D}\left(\sum_{j=1}^{m} C_{M_j}^{D} e^{e_j} + C_R^D e^{r_a} + \sum_{j=1}^{N_G} C_{G_j}^D P_{G_j}\right) \leqslant 1,$$

$$(\tilde{C}_4^+): \quad \frac{1}{e^{s_j^+} e^{e_j} k_j^+}\left(1 + \prod_{l=1}^{q}\left(\frac{e^{\bar{x}_l}}{K_{mlj}}\right)^{-m_{lj}^+} + \prod_{l=1}^{q}\left(\frac{e^{\bar{x}_l}}{K_{mlj}}\right)^{n_{lj}}\right)\left(1 + \hat{\theta_n}(\alpha^+(e^{\bar{x}_l}))\right) \leqslant 1, j \in \{1,\ldots,m\}$$

$$(\tilde{C}_4^-): \quad \frac{1}{e^{s_j^-} e^{e_j} k_j^-}\left(1 + \prod_{l=1}^{q}\left(\frac{e^{\bar{x}_l}}{K'_{mlj}}\right)^{-m_{lj}^-} + \prod_{l=1}^{q}\left(\frac{e^{\bar{x}_l}}{K'_{mlj}}\right)^{n'_{lj}}\right)\left(1 + \hat{\theta_n}(\alpha^-(e^{\bar{x}_l}))\right) \leqslant 1, j \in \{1,\ldots,m\}$$

$$(C_{\mu\,1}^+): \quad m_j^+ \nu_j^+ + n_j^+ e^{s_j^+} - m_j^+ n_j^+ \leqslant 1, \quad j \in \{1,\ldots,m\},$$

$$(C_{\mu\,2}^+): \quad e^{-s_j^+} \leqslant \nu_j^+ \quad j \in \{1,\ldots,m\},$$

$$(C_{\mu\,1}^-): \quad m_j^- \nu_j^- + n_j^- e^{s_j^-} - m_j^- n_j^- \leqslant 1, \quad j \in \{1,\ldots,m\},$$

$$(C_{\mu\,1}^-): \quad e^{-s_j^-} \leqslant \nu_j^-. \quad j \in \{1,\ldots,m\}$$

$$--- \quad n_j^+ \leqslant \nu_j^+ \leqslant N_j^+, \quad j \in \{1,\ldots,m\},$$

$$--- \quad n_j^- \leqslant \nu_j^- \leqslant N_j^-, \quad j \in \{1,\ldots,m\},$$

$$--- \quad n_j^+ \leqslant e^{s_j^+} \leqslant N_j^+, \quad j \in \{1,\ldots,m\},$$

$$--- \quad n_j^- \leqslant e^{s_j^-} \leqslant N_j^-, \quad j \in \{1,\ldots,m\}.$$

$$--- \quad e^{-\bar{x}_i}.\bar{X}_{lmin} \leqslant 1 \quad i = 1,\ldots,q,$$

$$--- \quad e^{\bar{x}_i}.\bar{X}_{lmax}^{-1} \leqslant 1 \quad i = 1,\ldots,q,$$

(3.27)

3.4 Exemple numérique

Dans cette section nous allons présenter l'application de notre approche de relaxation convexe au réseau métabolique utilisé dans la section où nous avons montré que les RBA étendu n'était pas convexe. Nous comparerons alors les solutions du problème relaxé avec celles de l'approche « brute force » qui consiste à résoudre le problème RBA non convexifié (2.45) page 24.

Nous modifions légerement le problème d'optimisation, en considérant qu'on cherche à maximiser le flux ν avec des contraintes non plus sur les métabolites internes uniquement mais sur l'ensemble des concentrations des métabolites en utilisant les contraintes suivantes :

$$0 \leqslant X_i \leqslant 2.$$

En se basant sur le chapitre 2, le sous problème RBA étendu restreint au réseau métbolique s'écrit comme :

$$
\begin{aligned}
&\max_{E_i \geqslant 0, X_i \geqslant 0, \nu_i^+ \geqslant 0, \nu_i^- \geqslant 0, \nu, \nu_T} \quad \nu \\
&\text{tel que :} \\
&(C_1) \quad \nu_1^+ \leqslant E_1 \frac{\max\{0, 1-\frac{X_2}{X_1}\}}{1+\frac{1}{X_1}+\frac{X_2}{X_1}}, \\
&(C_2) \quad \nu_1^- \leqslant E_1 \frac{\max\{0, 1-\frac{X_1}{X_2}\}}{1+\frac{1}{X_2}+\frac{X_1}{X_2}}, \\
&(C_3) \quad \nu_2^+ \leqslant E_2 \frac{\max\{0, 1-\frac{X_3}{X_2}\}}{1+\frac{1}{X_2}+\frac{X_3}{X_2}}, \\
&(C_4) \quad \nu_2^- \leqslant E_2 \frac{\max\{0, 1-\frac{X_2}{X_3}\}}{1+\frac{1}{X_3}+\frac{X_2}{X_3}}, \\
&(C_5) \quad \nu_3^+ \leqslant E_3 \frac{\max\{0, 1-\frac{X_3}{X_4}\}}{1+\frac{1}{X_4}+\frac{X_3}{X_4}}, \\
&(C_6) \quad \nu_3^- \leqslant E_3 \frac{\max\{0, 1-\frac{X_4}{X_3}\}}{1+\frac{1}{X_3}+\frac{X_4}{X_3}}, \\
&(C_7) \quad \nu_i = \nu_i^+ - \nu_i^-, i = \{1, 2, 3\} \\
&(C_8) \quad \nu_T = \nu_1 - \nu_2, \\
&(C_9) \quad \nu = \nu_2 + \nu_3, \\
&(C_{10}) \quad \nu = 5\nu_T, \\
&(C_{11}) \quad E_1 + 0,1E_2 + E_3 \leqslant 1, \\
&(C_{12}) \quad X_i \leqslant 2, i = \{1, 2, 3, 4\}.
\end{aligned}
\tag{3.28}
$$

L'approche de relaxation convexe présentée dans ce chapitre et appliquée au problème ci-

α_i	β_i
2.0027	2.4262
511.5106	64.4793
12.8178	13.5330
4.7544	5.6036
1.0663	1.0124

TABLE 3.4 – Coefficients de la partie interpolation

dessus résulte en un problème d'optimisation convexe s'écrivant sous la forme suivante :

$$
\begin{aligned}
&\max_{\nu_i^+\geqslant 0,\nu_i^-\geqslant 0,e_i,x_i,\nu,\nu_T,s_i^+,s_i^-} \nu \\
&\text{tel que :} \\
&e^{-s_1^+-e_1}(1+e^{-x_1}+e^{x_2-x_1})(1+\sum_{i=1}^{5}\alpha_i(e^{x_2-x_1})^{\beta_i})\leqslant 1,\\
&e^{-s_1^--e-1}(1+e^{-x_2}+e^{x_1-x_2})(1+\sum_{i=1}^{5}\alpha_i(e^{x_1-x_2})^{\beta_i})\leqslant 1,\\
&e^{-s_2^+-e_2}(1+e^{-x_2}+e^{x_3-x_2})(1+\sum_{i=1}^{5}\alpha_i(e^{x_3-x_2})^{\beta_i})\leqslant 1,\\
&e^{-s_2^--e_2}(1+e^{-x_3}+e^{x_2-x_3})(1+\sum_{i=1}^{5}\alpha_i(e^{x_2-x_3})^{\beta_i})\leqslant 1,\\
&e^{-s_3^+-e_3}(1+e^{-x_4}+e^{x_3-x_4})(1+\sum_{i=1}^{5}\alpha_i(e^{x_3-x_4})^{\beta_i})\leqslant 1,\\
&e^{s_3^--e_3}(1+e^{-x_3}+e^{x_4-x_3})(1+\sum_{i=1}^{5}\alpha_i(e^{x_4-x_3})^{\beta_i})\leqslant 1,\\
&a\nu_i^+ + ae^{s_i^+} - a^2 \leqslant 1, \quad i\in\{1,2,3\},\\
&a\nu_i^- + ae^{s_i^-} - a^2 \leqslant 1, \quad i\in\{1,2,3\},\\
&e^{-s_i^+}\leqslant \nu_i^+, \quad i\in\{1,2,3\},\\
&e^{-s_i^-}\leqslant \nu_i^-, \quad i\in\{1,2,3\},\\
&a\leqslant e^{s_i^+}\leqslant 1/a, \quad i\in\{1,2,3\},\\
&a\leqslant e^{s_i^-}\leqslant 1/a, \quad i\in\{1,2,3\},\\
&e^{e_1}+0,1e^{e_2}+e^{e_3}\leqslant 1,\\
&e^{x_i}\leqslant 2, \quad i\in\{1,2,3,4\},\\
&\nu_i=\nu_i^+-\nu_i^-, \quad i\in\{1,2,3\},\\
&\nu_T=\nu_1-\nu_2,\\
&\nu=\nu_2+\nu_3,\\
&\nu=5\nu_T.
\end{aligned}
\tag{3.29}
$$

Les cœfficients α_i et β_i sont calculés conformément à la stratégie détaillée dans le paragraphe 3.3.1 a) page 33 et sont donnés dans le tableau 3.4.

Les nouvelles variables x_i, e_i, s_i^+ et s_i^- sont introduites conformément au tableau 3.3. La constante a a été fixée dans notre exemple à $a = 10^{-5}$. L'objectif c'est de choisir le pa-

$e^{x_i^*}$	$1,9992$	$1,8905$	$1,8825$	$1,9907$
$e^{e_i^*}$	$0,4952$	$0,0794$	$0,4963$	-
$e^{s_i^{+*}}$	$135,3728$	$2,9447.10^4$	$135,0731$	-
$e^{s_i^{-*}}$	$9,9815.10^4$	$8,1478.10^4$	$9,9817.10^4$	-
ν_i^{+*}	$9,9864.10^4$	$6,7758.10^4$	$9,9865.10^4$	-
ν_i^{-*}	$0,0322$	$1,1817.10^3$	$0,0323$	-
ν_i^*	$9,9864.10^4$	$6,6576.10^4$	$9,9865.10^4$	-
ν^*	$1,6644.10^5$	-	-	-
ν_T^*	$3,3288.10^4$	-	-	-

TABLE 3.5 – Solution numérique du problème convexifié (3.29)

ramètre a le plus petit possible afin de balayer un domaine de flux le plus large possible.

La résolution du problème (3.29) sous Matlab via l'interface `CVX` et à l'aide du solver `SeDuMi` donne les solutions données dans le tableau 3.5.

La valeur optimale du flux dans ce cas est $\nu^* = 1,6644.10^5$. Ceci représente une borne supérieure du vrai flux optimal car la solution que nous venons de calculer est une solution d'un problème (problème (3.29)) dont l'ensemble faisable contient l'ensemble faisable du problème brut non convexe (3.28).

Si on effectue un test avec les valeurs de ν_i^{+*}, ν_i^{-*}, $e^{x_i^*}$ et $e^{e_i^*}$, on remarque que les contraintes $(C_1), \ldots, (C_6)$ ne sont pas satisfaites. Ceci est dû au fait qu'avec la relaxation convexe

$$\begin{aligned}
&a\nu_i^+ + ae^{s_i^+} - a^2 \leqslant 1, \quad i \in \{1,2,3\},\\
&a\nu_i^- + ae^{s_i^-} - a^2 \leqslant 1, \quad i \in \{1,2,3\},\\
&e^{-s_i^+} \leqslant \nu_i^+, \quad i \in \{1,2,3\},\\
&e^{-s_i^-} \leqslant \nu_i^-, \quad i \in \{1,2,3\},\\
&a \leqslant e^{s_i^+} \leqslant 1/a, \quad i \in \{1,2,3\},\\
&a \leqslant e^{s_i^-} \leqslant 1/a, \quad i \in \{1,2,3\},
\end{aligned}$$

remplaçant la contrainte originale :

$$\begin{aligned}
e^{s_i^+}\nu_i^+ &= 1,\\
e^{s_i^-}\nu_i^- &= 1,
\end{aligned} \tag{3.30}$$

on obtient des flux ν_i^+ et ν_i^- qui sont supérieurs forcément à $e^{-s_i^+}$ et $e^{-s_i^-}$ respectivement :

$$\begin{aligned}
\nu_i^+ &\geqslant e^{-s_i^+},\\
\nu_i^- &\geqslant e^{-s_i^-}.
\end{aligned} \tag{3.31}$$

Ceci peut être illustré par la figure 3.7 : l'ensemble des points $(e^{s_i^+}, \nu_i)$ avec $\nu_i \in I$ constitue le segment qui relie les deux extrémités du domaine de la relaxation convexe au niveau de

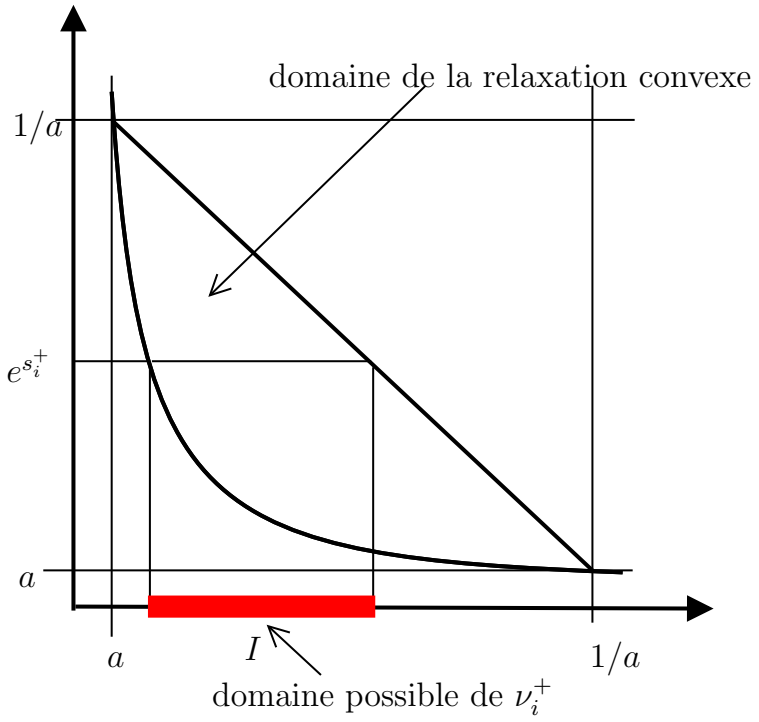

FIGURE 3.7 – Domaine admissible de ν_j^+

$e^{s_i^+}$, et ce segment est situé en dessus de la courbe de l'hyperbole qui à chaque ν_i^+ associe $1/\nu_i^+$.

Or les contraintes $(C_1), \ldots, (C_6)$ sont vérifiées avec les solutions $e^{-s_i^{+*}}$ $e^{-s_i^{-*}}$ $e^{e_i^*}$ et $e^{x_i^*}$:

$$
\begin{aligned}
e^{-s_1^{+*}} &\leqslant e^{e_1^*} \frac{\max\{0, 1-\frac{e^{x_2^*}}{e^{x_1^*}}\}}{1+\frac{1}{e^{x_1^*}}+\frac{e^{x_2^*}}{e^{x_1^*}}}, \\
e^{-s_1^{-*}} &\leqslant e^{e_1^*} \frac{\max\{0, 1-\frac{e^{x_1^*}}{e^{x_2^*}}\}}{1+\frac{1}{e^{x_2^*}}+\frac{e^{x_1^*}}{e^{x_2^*}}}, \\
e^{-s_2^{+*}} &\leqslant e^{e_2^*} \frac{\max\{0, 1-\frac{e^{x_3^*}}{e^{x_2^*}}\}}{1+\frac{1}{e^{x_2^*}}+\frac{e^{x_3^*}}{e^{x_2^*}}}, \\
e^{-s_2^{-*}} &\leqslant e^{e_2^*} \frac{\max\{0, 1-\frac{e^{x_2^*}}{e^{x_3^*}}\}}{1+\frac{1}{e^{x_3^*}}+\frac{e^{x_2^*}}{e^{x_3^*}}}, \\
e^{-s_3^{+*}} &\leqslant e^{e_3^*} \frac{\max\{0, 1-\frac{e^{x_3^*}}{e^{x_4^*}}\}}{1+\frac{1}{e^{x_4^*}}+\frac{e^{x_3^*}}{e^{x_4^*}}}, \\
e^{-s_3^{-*}} &\leqslant e^{e_3^*} \frac{\max\{0, 1-\frac{e^{x_4^*}}{e^{x_3^*}}\}}{1+\frac{1}{e^{x_3^*}}+\frac{e^{x_4^*}}{e^{x_3^*}}}.
\end{aligned}
\tag{3.32}
$$

Par conséquent, en prenant en compte des inégalités (3.31) et (3.32), on ne garantit pas forcément aux flux ν_i^{+*} et ν_i^{-*} (voir tableau 3.5) de vérifier les contraintes $(C_1), \ldots, (C_6)$ à leur tour, c'est-à-dire, de vérifier l'ensemble des inégalités suivantes :

$$
\begin{aligned}
\nu_1^+ &\leqslant e^{e_1} \frac{\max\{0, 1-\frac{e^{x_2}}{e^{x_1}}\}}{1+\frac{1}{e^{x_1}}+\frac{e^{x_2}}{e^{x_1}}}, \\
\nu_1^- &\leqslant e^{e_1} \frac{\max\{0, 1-\frac{e^{x_1}}{e^{x_2}}\}}{1+\frac{1}{e^{x_2}}+\frac{e^{x_1}}{e^{x_2}}}, \\
\nu_2^+ &\leqslant e^{e_2} \frac{\max\{0, 1-\frac{e^{x_3}}{e^{x_2}}\}}{1+\frac{1}{e^{x_2}}+\frac{e^{x_3}}{e^{x_2}}}, \\
\nu_3^+ &\leqslant e^{e_2} \frac{\max\{0, 1-\frac{e^{x_2}}{e^{x_3}}\}}{1+\frac{1}{e^{x_3}}+\frac{e^{x_2}}{e^{x_3}}}, \\
\nu_3^+ &\leqslant e^{e_3} \frac{\max\{0, 1-\frac{e^{x_3}}{e^{x_4}}\}}{1+\frac{1}{e^{x_4}}+\frac{e^{x_3}}{e^{x_4}}}, \\
\nu_3^- &\leqslant e^{e_3} \frac{\max\{0, 1-\frac{e^{x_4}}{e^{x_3}}\}}{1+\frac{1}{e^{x_3}}+\frac{e^{x_4}}{e^{x_3}}}.
\end{aligned}
\tag{3.33}
$$

Ainsi en résolvant le problème convexe (3.29), nous garantissons une borne supérieure sur le flux optimal ν^* de (3.28) mais par contre, nous n'obtenons pas des valeurs de flux métaboliques vérifiant à la fois les contraintes stœchiométriques et les contraintes thermodynamiques, c'est-à-dire (3.33).

Néanmoins, nous pouvons essayer d'obtenir des flux admissibles c'est-à-dire uniquement faisables par rapport aux contraintes stœchiométriques et aux contraintes thermodynamiques sans être optimales, solutions du problème brut (3.28). Ceci est possible en fixant les variables e^{e_i} et e^{x_i} aux valeurs du tableau 3.5 et en les injectant dans les contraintes du problème brut (3.28). Cela résulte en un problème d'optimisation linéaire en les variables ν_i^+, ν_i^-, ν, ν_T (3.34) dont les solutions sont données dans la tableau 3.6.

$$
\begin{aligned}
&\max_{\nu_i^+ \geqslant 0, \nu_i^- \geqslant 0, \nu, \nu_T} \quad \nu \\
&\text{tel que :} \\
&\quad \nu_1^+ \leqslant A_1, \\
&\quad \nu_1^- \leqslant B_1, \\
&\quad \nu_2^+ \leqslant A_2, \\
&\quad \nu_2^- \leqslant B_2, \\
&\quad \nu_3^+ \leqslant A_3, \\
&\quad \nu_3^- \leqslant B_3, \\
&\quad \nu_i = \nu_i^+ - \nu_i^-, i = \{1, 2, 3\} \\
&\quad \nu_T = \nu_1 - \nu_2, \\
&\quad \nu = \nu_2 + \nu_3, \\
&\quad \nu = 5\nu_T,
\end{aligned}
\tag{3.34}
$$

où

ν_i^+	$0,0023$	$0,0001$	$0,0110$
ν_i^-	0	0	0
ν_i	$0,0023$	10^{-4}	$0,0110$
ν	$0,0111$	-	-
ν_T	$0,0022$	-	-

TABLE 3.6 – Solution numérique du problème d'optimisation linéaire (3.34)

$$
\begin{aligned}
A_1 &= e^{e_1^*} \frac{\max\{0,1-\frac{e^{x_2^*}}{e^{x_1^*}}\}}{1+\frac{1}{e^{x_1^*}}+\frac{e^{x_2^*}}{e^{x_1^*}}} = 0,0110,\\
B_1 &= e^{e_1^*} \frac{\max\{0,1-\frac{e^{x_1^*}}{e^{x_2^*}}\}}{1+\frac{1}{e^{x_2^*}}+\frac{e^{x_1^*}}{e^{x_2^*}}} = 0,\\
A_2 &= e^{e_2^*} \frac{\max\{0,1-\frac{e^{x_3^*}}{e^{x_2^*}}\}}{1+\frac{1}{e^{x_2^*}}+\frac{e^{x_3^*}}{e^{x_2^*}}} = 0,0001,\\
B_2 &= e^{e_2^*} \frac{\max\{0,1-\frac{e^{x_2^*}}{e^{x_3^*}}\}}{1+\frac{1}{e^{x_3^*}}+\frac{e^{x_2^*}}{e^{x_3^*}}} = 0,\\
A_3 &= e^{e_3^*} \frac{\max\{0,1-\frac{e^{x_3^*}}{e^{x_4^*}}\}}{1+\frac{1}{e^{x_4^*}}+\frac{e^{x_3^*}}{e^{x_4^*}}} = 0,0110,\\
B_3 &= e^{e_3^*} \frac{\max\{0,1-\frac{e^{x_4^*}}{e^{x_3^*}}\}}{1+\frac{1}{e^{x_3^*}}+\frac{e^{x_4^*}}{e^{x_3^*}}} = 0.
\end{aligned}
\tag{3.35}
$$

Ainsi et contrairement au tableau 3.5 où on obtient des valeurs des flux exorbitantes, le tableau 3.6 nous donne des valeurs de flux métaboliques, certes pas optimales, mais réalisables. De plus, avec la résolution du problème (3.34) on obtient une borne inférieure sur la vraie valeur optimale du flux ν^* (valeur optimale du problème (3.28)).

La résolution du problème (3.28) présente la difficulté qu'on ne peut garantir un optimum global. Néanmoins, pour comparer les solutions obtenues, nous avons résolu le problème en utilisant la macro de Matlab `fmincon` combinée avec l'algorithme `sqp` (méthode de la programmation quadratique séquentielle) en initialisant l'algorithme avec les solutions du tableau 3.5. Les résultats obtenus sont donnés dans le tableau 3.7.

Nous remarquons que les valeurs des flux métaboliques sont plus proches des solutions du problème linéaire (3.34) que des solutions du problème convexe (3.29). La relaxation convexe du problème RBA étendu, représentée par le problème convexe (3.29), ne garantit pas des valeurs de flux admissibles du point de vue thermodynamique. Pour pallier cette difficulté, notre perspective est basée sur les remarques relevées dans ce paragraphe. En effet, pour que les flux ν_i^+ et ν_i^- soient admissibles thermodynamiquement, nous devons garantir les inégalités (3.36).

ν_i^+	$0,1111$	$3,1684.10^{-19}$	$0,5556$
ν_i^-	$4,4736.10^{-16}$	$4,5139.10^{-36}$	$2,2079.10^{-16}$
ν_i	$0,1111$	$3,1684.10^{-19}$	$0,5556$
ν	$0,5556$	-	-
ν_T	$0,1111$	-	-

Table 3.7 – Solution numérique du problème brut (3.28)

$$\begin{aligned} \nu_i^+ &\leqslant e^{-s_i^+}, \\ \nu_i^- &\leqslant e^{-s_i^-}. \end{aligned} \tag{3.36}$$

Nous rappelons que dans notre approche de convexification, nous avons divisé les contraintes originales (3.30) en deux contraintes : (3.31) et (3.36). Les inégalités (3.31) ont pour effet de produire des flux ν_i^+ et ν_i^- très grands et donc pas forcément admissibles thermodynamiquement. Ainsi, il ne faut garder que les contraintes (3.36) qui combinées avec les contraintes (3.32), garantissent aux flux ν_i^+ et ν_i^- d'être thermodynamiquement admissibles. La seule difficulté est que ces contraintes (3.36) sont non convexes. D'où la nécessité de proposer une stratégie permettant de les satisfaire au sein du problème (3.29) tout en le gardant convexe.

3.5 Conclusion

Au regard de la structure des contraintes RBA non convexes, on s'est ramené à reformuler le problème RBA proche d'un problème d'optimisation géométrique-linéaire mixte : la diffèrence fondamentale est que deux contraintes égalités (3.20) non convexes faisant intervenir le produit d'une variable « posynomiale » et d'une variable « linéaire » sont introduites :

$$\begin{aligned} \mu_j^+ \nu_j^+ &= 1 \\ \mu_j^- \nu_j^- &= 1. \end{aligned} \tag{3.37}$$

Une relaxation convexe de ces contraintes a été proposée sous la forme d'un problème d'optimisation géométrique-linéaire mixte dont la résolution est efficace. L'avantage est qu'on a gardé la structure fondamentale des contraintes tout en les ramenant vers la convexité, ce qui garantit l'efficacité numérique de la méthodologie. Néanmoins, nous avons vu à travers l'exemple numérique que l'approche présente des limitations, à savoir la difficulté de garantir aux différents flux métaboliques ν_i^+ et ν_i^- d'être admissibles thermodynamiquement car ces derniers flux garantissent uniquement les contraintes stœchiométriques. A côté de cela, les quantités $e^{-s_i^+}$ et $e^{-s_i^-}$ peuvent être assimilées à des flux admissibles thermodynamiquement.

La perspective que l'on propose est de garantir une cohérence en terme d'ordre de grandeur entre les ν_i^+ et $e^{-s_i^+}$ d'un coté et ν_i^- et $e^{-s_i^-}$ de l'autre. Cela se fait via la prise en compte des contraintes égalités (3.20) (ou leur forme équivalente (3.36)) qui sont non convexes,

ce qui implique de passer par des stratégies de l'optimisation globale [67, 34] adaptées à ce type de contraintes non convexes. Dans ce contexte, la relaxation convexe proposée dans ce chapitre constitue souvent un ingrédient de leur résolution.

Deuxième partie
Gestion des ressources et stochasticité de l'expression des gènes

Chapitre 4

Modélisation stochastique

Dans ce chapitre et plus largement dans la seconde partie de la thèse, nous allons aborder l'impact de la stochasticité dans l'allocation des ressources de la bactérie et plus particulièrement au niveau du réseau métabolique. En effet, il s'agit ici de prendre en compte explicitement la nature stochastique de la production des protéines (appelée aussi l'expression des gènes) dans la production par le réseau métabolique des constituants de la biomasse. Les études théoriques ont suggéré depuis longtemps que l'expression des gènes revêt un caractère stochastique, sa validation expérimentale est beaucoup plus récente et a fait des progrès incroyables ces dernières années puisque nous avons maintenant accès à des mesures du bruit d'expression d'à peu près un millier de protéines sur une bactérie [66]. Ce sujet comme nous l'avons déjà dit a donné lieu à des travaux théoriques dans les années 70 qui ont été résumés dans une publication qui fait référence dans le champ biologique [56].
Dans ce chapitre et cette seconde partie de la thèse, nous nous concentrons sur les effets de ce bruit sur la capacité de production du réseau métabolique. Rappelons ici que le réseau métabolique produit les précurseurs nécessaires pour la croissance de la bactérie et le taux de croissance est directement lié à la capacité de production de ce réseau. Si ce réseau est comparé à un réseau connecté de tuyaux dont la capacité maximale est associée au tuyaux le plus petit (de concentration minimale) alors il est raisonnable de penser que le caractère stochastique de la capacité de chaque tuyau va avoir un impact fort sur la capacité du réseau : à ressources équivalentes, le taux de croissance prédit sera plus faible avec un modèle stochastique qu'avec un modèle déterminsite. Il est aussi intuitif de penser qu'il est plus difficile d'optimiser un réseau modélisé de façon stochastique que de façon déterministe. On déduit donc que les cellules d'une population bactérienne d'une même espèce (et qui donc possèdent le même réseau métabolique) ont du fait de la nature stochastique de l'expression des gènes des concentrations d'enzyme différentes d'une cellule à l'autre et donc des capacités maximales de production de biomasse différentes, c'est-à-dire des taux de croissance différents. Ainsi, en supposant que chaque cellule de la population optimise sa propre capacité de production (taux de croissance), cela conduit à définir le problème de la maximisation de la croissance de la population comme un problème de maximisation du taux moyen de croissance qui correspond un problème d'optimisation stochastique appelé « *two stage* » avec recours [73] :

1. la première étape (« *first stage* ») correspond à l'allocation optimale des enzymes, encodée dans le code génétique, maximisant le taux de croissance moyen ;
2. la seconde étape (« *second stage* ») correspond à l'adaptation du flux métabolique d'une cellule à la production réalisée des enzymes de façon à maximiser son propre

taux de croissance.

Ce taux moyen de croissance est l'espérance mathématique du taux de croissance d'une cellule et rend le problème d'optimisation correspondant difficile. En effet, cette espérance, une intégrale multi-dimensionnelle, s'interprète comme un calcul de volume si on considère des lois uniformes pour la production des protéines [26]. Même dans le cas d'un ensemble convexe, il n'existe pas d'algorithme polynomial (en le nombre de dimensions) pour le calcul exact de son volume [27] : la simple évaluation de la fonction objectif est déjà un problème numériquement difficile (en le nombre de variables de décision). En revanche, si nous recherchons une solution approchée (concept de solution que nous définirons formellement dans la suite du document), il existe des algorithmes en temps polynomial [46, 52] pour les problèmes d'optimisation stochastique convexes, une classe de problèmes englobant celle des problèmes « *two stage* » convexes.

Dans ce chapitre, une modélisation de la nature stochastique de l'expression des gènes est proposée. L'impact de la modélisation stochastique est discutée. Elle prend la forme de problèmes d'optimisation stochastique. La résolution de ces problèmes fait appel à la résolution de problèmes d'optimisation convexes de grande dimension qui sera traitée dans le chapitre 5 où les méthodes de résolution sont présentées et leur convergence est analysée puis dans le chapitre 6 où sera présenté leur extension pour la résolution de problèmes d'optimisation stochastique dont la performance est étudiée sur la modélisation stochastique de réseaux métaboliques.

4.1 Modélisation

4.1.1 Eléments de modélisation

Nous rappelons ici les éléments de modélisation nécessaires à ce chapitre. Nous considérons un réseau métabolique de matrice de stœchiométrie S. En notant ν le vecteur des flux dans ce réseau, la conservation de la masse en régime équilibré s'écrit

$$S\nu = 0 \tag{4.1}$$

où la matrice $S \in \mathbb{R}^{m\times n}$ est, sans perte de généralité, supposée être de rang plein avec $m \leqslant n$. Cette égalité traduit le fait que le réseau métabolique est équilibré en interne, c'est-à-dire que tout flux de production d'un métabolite interne s'équilibre avec un flux de consommation de ce même métabolite interne.

Chaque réaction du réseau est catalysée par une enzyme. En suivant la modélisation usuelle [30], un flux métabolique est borné en valeur absolue par la concentration et par l'efficacité de l'enzyme correspondante :

$$|\nu| \leqslant k.E \tag{4.2}$$

où $k \in \mathbb{R}^n_{+*}$ est le vecteur des efficacités enzymatiques et $E \in \mathbb{R}^n_+$ est le vecteur des concentrations des enzymes. La notation $x.y$ pour deux vecteurs x et y de même taille correspond au produit terme à terme ou produit d'Hadamard. De même, l'inégalité doit se comprendre terme à terme.

En considérant que chaque protéine a un coût de production et que la quantité totale de ressources allouées à la production des enzymes est limitée, nous obtenons l'inégalité

$$w^T E \leqslant E_T \tag{4.3}$$

où $w \in \mathbb{R}_+^n$ est le vecteur de coût des enzymes et $E_T > 0$ est la quantité totale de ressources allouée.

4.1.2 Modèle déterministe

Dans le cas déterministe, nous supposons que la concentration des enzymes est une quantité déterministe, c'est-à-dire que dans les mêmes conditions, chaque cellule de la population a le même vecteur de concentrations enzymatiques. Nous modélisons alors le cas déterministe de la façon suivante.

Formulation du problème 2. *Le problème d'allocation optimale des ressources avec une expression des gènes déterministe est défini comme le problème de programmation linéaire suivant :*

$$\max_{E \in \mathbb{R}_+^n,\, \nu \in \mathbb{R}^n} \quad c^T \nu$$
$$tel\ que \quad \begin{cases} S\nu = 0 \\ |\nu| \leqslant k.E \\ w^T E \leqslant E_T \end{cases}$$

où $c \in \mathbb{R}^n$ est le vecteur connu de composition de la biomasse.

Le fait que le vecteur c soit connu est une hypothèse commune dans les approches de modélisation sous contraintes telle que l'approche FBA (Flux Balance Analysis) [55]. Cette modélisation déterministe est bien une modélisation de type populationnelle comme indiqué dans l'introduction. En effet, dans le cas déterministe, chaque cellule d'une population possède la même concentration d'enzymes et toutes les cellules sont ainsi identiques et ont toutes la même croissance. Considérer la croissance de la population entière se réduit donc à considérer la croissance d'une seule cellule. La modélisation déterministe peut être vue comme une modélisation « dégénérée » : dans la modélisation déterministe, on considère une cellule unique ou plus exactement une population constituée de cellules identiques ; dans la modélisation stochastique, au sein d'une population, les cellules sont considérées de façon individuelle car différentes les unes des autres en terme de concentrations enzymatiques.

4.1.3 Modèle stochastique

Nous considérons maintenant le cas où les concentrations enzymatiques E sont des quantités stochastiques et nous supposons dans un premier temps que ces concentrations suivent une loi de probabilité (noté *loi* dans la suite) dont nous discutons le choix dans la section 4.1.4. Notons $\widetilde{E}$ le vecteur des concentrations enzymatiques moyennes et $\widehat{E}$ une réalisation de la variable aléatoire E suivant la loi *loi* de moyenne $\widetilde{E}$. Nous obtenons alors la modélisation suivante.

Formulation du problème 3. *Le problème d'allocation optimale des ressources avec une expression des gènes stochastique est défini comme le problème d'optimisation suivant*

$$\begin{aligned} \max_{\widetilde{E} \in \mathbb{R}_+^n} \quad & f(\widetilde{E}) = \mathbb{E}_{E \sim loi[\widetilde{E}]} F(E) \\ tel\ que \quad & w^T \widetilde{E} \leqslant E_T \end{aligned} \tag{4.4}$$

où

$$F(\widehat{E}) = \max_{\nu \in \mathbb{R}^n} \quad c^T \nu \qquad \textit{tel que} \quad \begin{cases} S\nu = 0 \\ |\nu| \leqslant k.\widehat{E} \end{cases} . \tag{4.5}$$

avec $\hat{E}$ est une ralisation de la variable alatoire E. Ce problème est la traduction du problème intuitif suivant : le taux moyen de croissance d'une population étant défini comme la moyenne de la croissance sur la population, l'allocation optimale est celle qui maximise le taux moyen de croissance en prenant en compte le fait que la réalisation de la concentration d'enzymes suit une loi donnée.

Si E est de la forme

$$E = H(\widetilde{E}, \xi)$$

avec H une fonction de $\widetilde{E}$ et d'une variable aléatoire ξ dont la loi de probabilité est indépendante de $\widetilde{E}$, le problème 3 est alors un problème « *two stage* » [73]. Dans ce cas, (4.4) et $\widetilde{E}$ définissent la première étape (« *first stage* ») du problème et les variables de décision associées ; (4.5) et ν définissent la seconde étape (« *second stage* ») et les variables de décision (dit aussi de recours) associées. Notons que, dans ce cas, la seconde étape est définie comme

$$F(H(\widetilde{E}, \hat{\xi})) = \max_{\nu \in \mathbb{R}^n} \quad c^T \nu \qquad \text{tel que} \quad \begin{cases} S\nu = 0 \\ |\nu| \leqslant k.H(\widetilde{E}, \hat{\xi}) \end{cases}$$

où $\hat{\xi}$ est une réalisation de la variable aléatoire ξ de loi indépendante de $\widetilde{E}$. La seconde étape est supposée toujours avoir un point faisable (en supposant par exemple que $\nu = 0$ est une solution) et le problème « *two stage* » est ainsi bien défini [71].

Si de plus H est linéaire en $\widetilde{E}$, le problème 3 est alors un problème « *two stage* » *linéaire* [73, 71, 72]. Ce problème de minimisation s'écrit alors de façon équivalente comme un problème de maximisation de la forme :

$$\begin{aligned} \max_{x \in \mathbb{R}^n_+} \quad & \mathbb{E}_\xi \left[F(x, \xi) \right] \\ \text{tel que} \quad & Ax = b \end{aligned}$$

où

$$\begin{aligned} F(x, \hat{\xi}) = \quad & c(\hat{\xi})^T x + \max_{y \in \mathbb{R}^n_+} q(\hat{\xi})^T y \\ \text{tel que} \quad & T(\hat{\xi})x + W(\hat{\xi})y = p(\hat{\xi}) \end{aligned}$$

où ξ est une variable aléatoire indépendante de x et où $\hat{\xi}$ est une réalisation de ξ, A, b, c, q, T, W sont les données du problème et elles sont de dimensions appropriées. Les variables de décision de la première étape x correspondent à la concentration E des enzymes ainsi qu'à une variable supplémentaire permettant de transformer l'inégalité en égalité. Les variables de recours y correspondent aux flux métaboliques ν ainsi qu'à des variables supplémentaires pour transformer les inégalités en égalités.

Même si les problèmes « *two stage* » linéaires sont difficiles à résoudre [26, 63], ils sont concaves [72], ce qui est une propriété essentielle pour le calcul numérique d'une solution.

4.1.4 Modèle stochastique « exponentiel »

La loi de distribution associée à l'expression des gènes est le résultat de l'intégration de plusieurs processus stochastiques élémentaires et dépend alors d'une manière complexe de plusieurs paramètres (voir par exemple [56]). Une caractéristique commune est que la variance est généralement de grande taille et qu'en première approximation, elle peut être reliée à la moyenne par la relation suivante : $\mathbb{V}(E) = a\mathbb{E}^2[E] = a\widetilde{E}^2$ où a est une constante proche de 1. Cette approximation est cohérente avec l'observation expérimentale présentée dans [66]. Suite à cette remarque préliminaire, nous définissons alors un problème spécifique qui permet d'approcher dans une certaine mesure le « problème réel » tout en présentant l'intérêt d'admettre dans certains cas bien déterminés une solution analytique. Cette spécificité est essentielle dans la suite de la thèse lorsque nous devrons évaluer les propriétés de convergence d'algorithmes de résolution.

Formulation du problème 4. *Le problème d'allocation optimale des ressources avec une expression génique de type « loi exponentielle » est défini comme le problème de programmation « two stage » linéaire suivant*

$$\begin{array}{ll} \max\limits_{\widetilde{E}\in\mathbb{R}^n_+} & f(\widetilde{E}) = \mathbb{E}_{\xi\sim exp(1)}\left[F(\widetilde{E},\xi)\right] \\ tel\ que & w^T\widetilde{E} \leqslant E_T \end{array} \tag{4.6}$$

où

$$\begin{array}{rl} F(\widetilde{E},\hat{\xi}) = \max\limits_{\nu\in\mathbb{R}^n} & c^T\nu \\ tel\ que & \left\{ \begin{array}{l} S\nu = 0 \\ |\nu| \leqslant k.(\sqrt{a}\hat{\xi} + 1 - \sqrt{a}).\widetilde{E} \end{array} \right. \end{array} \tag{4.7}$$

avec $0 < a \leqslant 1$.

Remarque 1. *Notons les guillemets autour de loi exponentielle. En effet, la loi réellement exponentielle correspond au cas* $a = 1$ *que nous utilisons maintenant pour obtenir un problème dont nous pouvons calculer la solution analytique.*

4.2 Réseau à solution analytique

4.2.1 Description du réseau

Nous considérons un réseau métabolique qui produit un composé chimique final B. Ce composé final est obtenu à partir de N_p composés intermédiaires produits par N_p voies métaboliques spécifiques. On suppose que les voies métaboliques ont la propriété d'être découplées : voies parallèles indépendantes et produisant un unique composé intermédiaire. Le produit final B est alors construit si chaque voie est capable de fournir une quantité minimale de produits intermédiaires. Par définition et afin de construire une unité de B, il faut utiliser c_i unités de produit intermédiaire de la i^{eme} voie. Nous supposons que chaque réaction enzymatique transforme un unique composé chimique en un autre unique composé chimique avec une stoichiométrie unitaire. Ce processus peut être illustré par la figure 4.1. En notant ν_T le flux de composé final B, les flux enzymatiques $\nu_{i,j}$ sont tels que :

$$\nu_{i,j} = c_i\nu_T, \quad \forall i \in \{1,\ldots,N_p\}, \forall j \in \{1,\ldots,N_v\}, \tag{4.8}$$

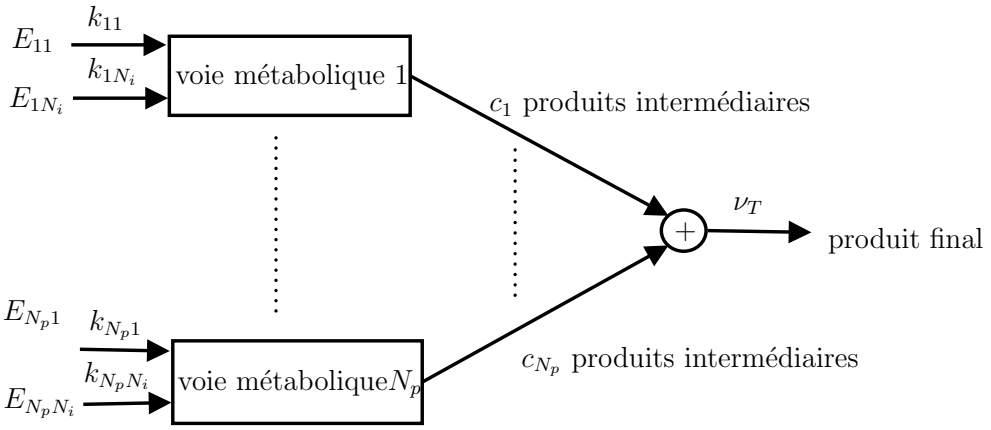

FIGURE 4.1 – Production d'un réseau métabolique découplé

ou de façon équivalente

$$\nu_T = \frac{\nu_{i,j}}{c_i}, \quad \forall i \in \{1, \dots, N_p\}, \forall j \in \{1, \dots, N_v\}. \tag{4.9}$$

Il s'agit de l'instanciation de l'égalité (4.1) au réseau considéré dans cette section où nous notons par ailleurs que $c^T\nu$ vaut ν_T.
La i^{eme} voie fait intervenir N_v (indépendant de i pour alléger les notations) enzymes $\mathbb{E}_{i,j}$, de concentration $E_{i,j}$ où $j = \{1, \dots, N_v\}$ et d'efficacité $k_{i,j}$, c'est-à-dire que le flux enzymatique $\nu_{i,j}$ est tel que :

$$\nu_{i,j} \leqslant k_{i,j} E_{i,j}. \tag{4.10}$$

Il s'agit de l'instanciation de l'inégalité (4.2) au réseau considéré où les flux enzymatiques sont tous positifs.
Ainsi, à partir de (4.9) et (4.10), nous obtenons l'inégalité structurelle suivante sur le flux de production de B :

$$\nu_T \leqslant \frac{k_{i,j}}{c_i} E_{i,j}, \quad \forall i \in \{1, \dots, N_p\}, \forall j \in \{1, \dots, N_v\}. \tag{4.11}$$

Le flux maximal de production ν_T^{max} est alors défini par :

$$\nu_T^{max} = \min_{i,j} \left\{ \frac{k_{i,j}}{c_i} E_{i,j} \right\}.$$

Il est à noter que cette égalité prend en compte les (in)équations (4.1) et (4.2).
Il ne reste alors plus que l'inégalité (4.3) à spécifier. Nous considérons que toutes les enzymes sont de coût unitaire, ce qui nous donne l'instanciation

$$\sum_{i=1}^{N_P} \sum_{j=1}^{N_v} E_{i,j} \leqslant E_T. \tag{4.12}$$

Enfin, nous considérons que dans le cas stochastique, l'expression des gènes est donnée par une loi exponentielle : la concentration $E_{i,j}$ est une variable aléatoire de densité de probabilité exponentielle de paramètre $\widetilde{E}_{ij}^{-1}$:

$$E_{i,j} \sim exp\left(\widetilde{E}_{ij}^{-1}\right)$$

ou de façon équivalente

$$E_{i,j} = \widetilde{E}_{i,j}\xi_{i,j}, \text{ où } \xi_{i,j} \sim exp(1),$$

c'est-à-dire que $\xi_{i,j}$ suit une loi de probabilité exponentielle de paramètre 1. Il s'agit de l'expression proposée dans le problème 4 avec $a = 1$.

Nous allons maintenant présenter les solutions analytiques associées aux modélisations déterministe et stochastique pour pouvoir les comparer et présenter l'intérêt et les implications de la modélisation stochastique.

4.2.2 Modélisation déterministe et résolution

Avec les hypothèses de la section 4.2.1, le problème 2 (problème déterministe) se simplifie en le problème suivant, ce qui permet de déterminer analytiquement la solution optimale.

Formulation du problème 5. *Le problème d'allocation optimale des ressources avec une expression des gènes déterministe et les hypothèses de la section 4.2.1 est défini comme le problème d'optimisation suivant :*

$$\tilde{\nu}_T^* = \max_{E_{ij}} \left(\min_{i,j} \left\{ \frac{k_{i,j}}{c_i} E_{i,j} \right\} \right)$$
$$tel\ que \sum_{i=1}^{N_P} \sum_{j=1}^{N_v} E_{i,j} \leqslant E_T.$$

Pour résoudre ce problème, nous remplaçons, pour simplifier les notations, l'ensemble des enzymes $\{E_{i,j}\}$ par l'ensemble $\{x_r\}$ avec $r \in \{1, \ldots, n = N_p N_v\}$. Les deux ensembles sont en bijection. Nous posons $a_r = \frac{c_i}{k_{i,j}}$ tel que $r = 1, \ldots, n$ et $i = 1, \ldots, N_p$ et $j = 1, \ldots, N_v$. Le problème se réécrit alors sous la forme équivalente :

$$\begin{aligned} -\tilde{\nu}_T^* &= \min_{x_r, \gamma} \gamma \\ -x_r/a_r &\leqslant \gamma, \quad r = 1, \ldots, n \\ \text{tel que : } \sum_{r=1}^{n} x_r &\leqslant E_T, \\ x_r &\geqslant 0 \quad r = 1, \ldots, n. \end{aligned} \tag{4.13}$$

Ce problème s'écrit sous la forme compacte suivante :

$$\begin{aligned} &\min_{\tilde{x}} \tilde{c}^T \tilde{x} \\ &\text{tel que : } \tilde{A}\tilde{x} \geqslant \tilde{b}, \end{aligned} \tag{4.14}$$

où $\tilde{A} \in \mathbb{R}^{2n+1 \times n+1}$, $\tilde{b} \in \mathbb{R}^{2n+1}$, $\tilde{c} \in \mathbb{R}^{n+1}$ sont tels que :

$$\tilde{A} := \left(\begin{array}{cccc} 1/a_1 & & & 1 \\ & \ddots & 0 & \vdots \\ & 0 & 1/a_n & 1 \\ \hline -1 & \ldots & -1 & 0 \\ \hline 1 & & & 0 \\ & \ddots & 0 & \vdots \\ & 0 & 1 & 0 \end{array}\right) \quad \tilde{b} = \left(\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline -E_T \\ \hline 0 \\ \vdots \\ 0 \end{array}\right) \quad \tilde{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad \tilde{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \gamma \end{pmatrix}.$$

Sachant que le problème dual de (4.14) est défini par

$$\begin{aligned} &\max_{y} \tilde{b}' y \\ \text{tel que :} \quad &\tilde{A}^T y = \tilde{c}, \\ &y \geqslant 0, \end{aligned} \tag{4.15}$$

on applique les conditions d'optimalité de la programmation linéaire rappelées par le théorème suivant :

Théorème 5. *[9] $\tilde{x}$ est une solution optimale de (4.14) et y est une solution optimale de (4.15) si et seulement si*

$$y^T \left(\tilde{A}\tilde{x} - \tilde{b} \right) = 0 \tag{4.16}$$

de même que si et seulement si

$$\tilde{c}^T \tilde{x} - \tilde{b}^T y = 0. \tag{4.17}$$

L'application directe de la condition (4.16) donne :

$$\begin{cases} y_r(x_r/a_r + \gamma) = 0, \quad r = 1, \ldots, n \\ y_{n+1}(\sum_{r=1}^{n} x_r - E_T) = 0, \\ y_r x_r = 0, \quad r = n+2, \ldots, 2n+1. \end{cases} \tag{4.18}$$

On obtient alors,

$$\begin{cases} x_r = -\gamma a_r, \quad r = 1 \ldots, n \\ \gamma = -\frac{E_T}{n}, \\ \qquad \sum_{r=1} a_r \\ y_r = -\frac{\gamma}{E_T} \quad \text{si} \quad r = n+1, \\ y_r = 0 \quad \text{sinon.} \end{cases}$$

On en déduit la solution optimale au problème 5.

Corollaire 1. *La solution optimale au problème 5 est donnée par :*

$$E^*_{i,j} = \frac{c_i/k_{i,j} E_T}{\sum_{i=1}^{N_p} \sum_{i=1}^{N_v} c_i/k_{i,j}} \quad et \quad \tilde{\nu}^*_T = \frac{E_T}{\sum_{i=1}^{N_p} \sum_{i=1}^{N_v} c_i/k_{i,j}}. \tag{4.19}$$

4.2.3 Modélisation stochastique exponentielle et résolution

Avec les hypothèses de la section 4.2.1, le problème 4 (problème stochastique) se simplifie en le problème suivant, ce qui permet de déterminer analytiquement la solution optimale.

Formulation du problème 6. *Le problème d'allocation optimale des ressources avec une expression des gènes stochastique et les hypothèses de la section 4.2.1 est défini comme le problème d'optimisation suivant :*

$$\begin{aligned} \nu_T^* = \max_{\widetilde{E}_{ij} \geqslant 0} \quad & \mathbb{E}_\xi \left[\min_{i,j} \left\{ \frac{k_{i,j}}{c_i} \widetilde{E}_{i,j} \xi_{i,j} \right\} \right] \\ \textit{tel que} \quad & \sum_{i=1}^{N_P} \sum_{j=1}^{N_v} \widetilde{E}_{i,j} \leqslant E_T. \end{aligned}$$

Notons que si le problème 6 admet une solution optimale alors elle sature la contrainte inégalité. En effet, on peut raisonner par l'absurde et supposer le contraire. Dans ce cas, il est toujours possible de perturber la solution optimale en multipliant chaque concentration d'enzyme par un même coefficient positif (> 1) afin de saturer la contrainte inégalité. Cela augmentera la concentration de chaque enzyme et par conséquent améliorera la valeur de la fonction objectif, ce qui est absurde.

Formulation du problème 7. *Le problème 6 est équivalent au problème d'optimisation suivant :*

$$\begin{aligned} \nu_T^* = \max_{\widetilde{E}_{ij} \geqslant 0} \quad & \mathbb{E}_\xi \left[\min_{i,j} \left\{ \frac{k_{i,j}}{c_i} \widetilde{E}_{i,j} \xi_{i,j} \right\} \right] \\ \textit{tel que} \quad & \sum_{i=1}^{N_P} \sum_{j=1}^{N_v} \widetilde{E}_{i,j} = E_T. \end{aligned}$$

Remarque 2. *Ce dernier problème a un ensemble faisable qui est le simplexe unité. Cela représente un avantage numérique important pour la résolution du problème*[1]*.*

Dans le problème 7, il est possible de restreindre l'espace de recherche à l'ensemble des $\widetilde{E}_{i,j} > 0$. En effet, comme E_T est strictement positif, il est possible de trouver une allocation des enzymes telle que le flux est strictement positif. A contrario (par l'absurde), si une solution optimale contient une concentration nulle, le flux est nul. Par ailleurs, comme nous l'avons mentionné précédemment, nous avons des variables aléatoires à densité de probabilité exponentielle. Dans ce cas, nous avons le résultat suivant.

Proposition 3. *[3, 2] Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes et à densité de probabilité exponentielle de paramètres $\lambda_1, \ldots, \lambda_n$. Alors $\min\{X_1, \ldots, X_n\}$ est une variable aléatoire à densité de probabilité exponentielle de paramètre $\lambda = \lambda_1 + \ldots + \lambda_n$. La variable aléatoire $\min\{X_1, \ldots, X_n\}$ est dite premier ordre statistique de X.*

Ainsi, le flux défini par $\min_{i,j} \frac{k_{i,j}}{c_i} E_{i,j}$ est une variable aléatoire à densité de probabilité exponentielle et de paramètre

$$\lambda = \sum_{i=1}^{N_p} \sum_{j=1}^{N_v} \frac{c_i}{k_{i,j} \widetilde{E}_{i,j}},$$

ou encore

$$\mathbb{E}_\xi \left[\min_{i,j} \left\{ \frac{k_{i,j}}{c_i} \widetilde{E}_{i,j} \xi_{i,j} \right\} \right] = \frac{1}{\sum_{i=1}^{N_p} \sum_{j=1}^{N_v} \frac{c_i}{k_{i,j} \widetilde{E}_{i,j}}} = \frac{1}{\lambda}. \tag{4.20}$$

Par conséquent, nous avons l'équivalence suivante.

1. La projection nonlinéaire (6.15) de l'algorithme MDSA de ce problème peut être alors calculée efficacement, voir l'algorithme 8 page 107.

Formulation du problème 8. *Les problèmes 6 et 7 sont équivalents au problème d'optimisation (déterministe) suivant :*

$$\begin{aligned} \nu_T^* = \max_{\widetilde{E}_{ij}>0} \quad & \frac{1}{\sum_{i=1}^{N_p}\sum_{j=1}^{N_v}\frac{c_i}{k_{i,j}\widetilde{E}_{i,j}}} \\ \textit{tel que} \quad & \sum_{i=1}^{N_P}\sum_{j=1}^{N_v}\widetilde{E}_{i,j} = E_T. \end{aligned}$$

Nous avons le résultat suivant.

Proposition 4. *La solution optimale des problèmes 6, 7 et 8 est donnée par :*

$$\widetilde{E}_{i,j}^* = \frac{E_T\sqrt{\frac{c_i}{k_{i,j}}}}{\sum_{i=1}^{N_p}\sum_{j=1}^{N_v}\sqrt{\frac{c_i}{k_{i,j}}}} \quad et \quad \nu_T^* = \frac{E_T}{\left(\sum_{i=1}^{N_p}\sum_{j=1}^{N_v}\sqrt{\frac{c_i}{k_{i,j}}}\right)^2}. \tag{4.21}$$

Preuve. Comme les problèmes 6, 7 et 8 sont équivalents, il suffit de montrer que ce dernier a la solution optimale donnée en (4.21). C'est ce que nous allons démontrer.
Pour simplifier les notations, on remplace l'ensemble des enzymes $\{\widetilde{E}_{i,j}\}$ par l'ensemble $\{x_r\}$ avec $r \in \{1,\dots,n = N_pN_v\}$. Les deux ensembles sont en bijection. On pose $a_r = \frac{c_i}{k_{i,j}}$ tel que $r = 1,\dots,n$ et $i = 1,\dots,N_p$ et $j = 1,\dots,N_v$. On considère la fonction h définie par :

$$h(x) = -\frac{1}{\sum_{i=1}^{n}\frac{a_i}{x_i}}. \tag{4.22}$$

Le problème 8 est équivalent au problème suivant :

$$\begin{aligned} & -\min_{x\in(\mathbb{R}_+^*)^n} h(x) \\ \text{tel que : } & \sum_{i=1}^{n} x_i = E_T. \end{aligned} \tag{4.23}$$

On commence d'abord par prouver la convexité de la fonction h. En effet, on peut voir la fonction h comme la composée des deux fonctions f_1 et f_2 telles que $f_1(y) = \frac{1}{y}$ et $f_2(x) = -\sum_{i=1}^{n}\frac{a_i}{x_i}$, avec les domaines de définition $domf_1 = \mathbb{R}_+^*$ et $domf_2 = X$. Puisque f_1 est convexe et que la fonction étendue $\tilde{f}_1$ définie par $\tilde{f}_1(y) = f_1(y)$ si $y \in domf_1$, et $\tilde{f}_1(y) = \infty$ sinon, est décroissante et puisque la fonction f_2 est concave ($-f_2$ étant la somme positive de fonctions convexes a_i/x_i, sur X, est donc convexe) alors on peut conclure que la composée $f_1 \circ f_2$ est convexe sur X. Il s'agit de l'application des résultats standard sur la convexité de la composée de fonctions (pour plus de détails voir [15] page 83).
En appliquant les conditions de KKT du premier ordre au problème (4.23) (et c'est possible car le problème n'a qu'une seule contrainte ; la contrainte de qualification dite d'indépendance linéaire [54] est forcément vérifiée), on obtient : x^* est un minimiseur global de (4.23) si et seulement s'il existe $\mu \in \mathbb{R}$ tel que :

$$h'(x^*) + \mu g'(x^*) = 0$$

où $g(x) = \sum_{i=1}^n x_i - E_T$, h' et g' sont les gradients de h et g respectivement. Ainsi

$$\frac{a_i}{D^2}\frac{1}{x_i^{*2}} + \mu = 0.$$

Donc $x_i^* = \sqrt{\frac{a_i}{-\mu}}\frac{1}{D}$ avec $D = \frac{1}{(\sum_1^n \frac{a_i}{x_i^*})^2}$. Or $\sum_{i=1}^n x_i^* = E_T$ donc μ doit vérifier $\mu = -\frac{\sum_{i=1}^n a_i}{E_T D}$. Avec cette valeur (unique) de μ le point x^* qui vérifie les conditions de KKT est unique et tel que $x_i^* = \frac{E_T\sqrt{a_i}}{\sum_{i=1}^n \sqrt{a_i}}$. On obtient la valeur optimale du flux en remplaçant dans $-h(x)$ les concentrations des enzymes par leur expression optimale, avec h donnée par (4.22). □

4.3 De l'intérêt de la modélisation stochastique

4.3.1 Comparaison entre les solutions déterministe et stochastique

La différence entre les solutions stochastiques (4.21) et déterministes (4.19) peut être facilement évaluée.

(i) Le rapport des flux métaboliques optimaux vaut :

$$\frac{\text{flux optimal déterministe}}{\text{flux optimal stochastique}} = \frac{\left(\sum_{i=1}^{N_p}\sum_{j=1}^{N_v}\sqrt{c_i/k_{i,j}}\right)^2}{\sum_{i=1}^{N_p}\sum_{j=1}^{N_v} c_i/k_{i,j}}.$$

Ce qui donne en revenant aux notations simplifiées :

$$\frac{\text{flux optimal déterministe}}{\text{flux optimal stochastique}} = 1 + 2\frac{\sum_{1\leqslant i<j\leqslant n}\sqrt{a_i a_j}}{\sum_{i=1}^n a_i} \geqslant 1.$$

Par conséquent, afin d'obtenir le même niveau de production en moyenne, il est nécessaire d'engager plus de ressources dans le cas stochastique.

(ii) La concentration optimale de chaque enzyme, dans le cas déterministe et stochastique, est proportionnelle au flux de production maximal ν_T^*. Dans le cas déterministe, nous avons :

$$\text{concentration optimale déterministe} = \frac{c_i}{k_{i,j}}\tilde{\nu}_T^*,$$

qui dépend uniquement de la i^{eme} voie métabolique. Par contre, dans le cas stochastique, nous avons :

$$\text{concentration optimale stochastique} = \sqrt{\frac{c_i}{k_{i,j}}}\sum_{i=1}^{N_p}\sum_{j=1}^{N_v}\sqrt{\frac{c_i}{k_{i,j}}}\nu_T^*,$$

qui dépend de la configuration entière du réseau métabolique.

Ainsi l'integration de l'aspect stochastique au sein du réseau métabolique a pour effet de rendre la production du flux métabolique moins performante et plus impactée par la configuration globale du réseau métabolique.

4.3.2 Sur l'efficacité des enzymes

Jusqu'à présent, nous nous sommes concentrés sur l'impact de la nature stochastique de l'expression des gènes, à ressources et à efficacités données. Nous nous posons maintenant la question « inverse » : quel doit être le lien entre les efficacités des enzymes dans le modèle déterministe, que nous notons $k^d_{i,j}$, et celles dans le modèle stochastique, que nous notons $k^s_{i,j}$, pour que les solutions soient les mêmes ? En effet, les mesures d'efficacité se font via des mesures moyennes de flux et de concentration d'enzymes, l'efficacité étant obtenue comme le ratio du flux sur la concentration [31, 24]. L'efficacité obtenue est donc une efficacité moyenne (ou apparente) et correspond à $k^d_{i,j}$. Intuitivement et vis-à-vis des solutions obtenues précédemment, l'efficacité $k^s_{i,j}$ doit être supérieure à l'efficacité $k^d_{i,j}$. Nous voulons quantifier de combien.

Afin que la résolution des deux problèmes d'optimisation donne le même flux optimal total ν^*_T, les efficacités des enzymes doivent vérifier la relation suivante :

$$k^s_{i,j} = \frac{\left(k^d_{i,j}\right)^2}{c_i} \sum_{k=1}^{N_p} \sum_{l=1}^{N_v} \frac{c_k}{k^d_{k,l}}.$$

On en déduit la relation suivante :

$$\frac{k^s_{i,j}}{k^s_{k,l}} = \left(\frac{k^d_{i,j}}{k^d_{k,l}}\right)^2 \frac{c_k}{c_i}.$$

Remarque : Ce rsultat s'obtient d'une façon immédiate en mettant à égalité les concentrations d'enzymes optimales issues des solutions analytiques obtenues par (4.21) et (4.19).

Dans les deux expressions, un carré apparaît. Les efficacités apparentes pouvant atteindre 10^7, ces formules nous montrent que l'efficacité réelle des enzymes peut être bien supérieure à leur efficacité apparente et que le rapport des efficacités réelles peut aussi être bien supérieur au rapport des efficacités apparentes.

4.4 Conclusion

Dans ce chapitre, nous avons considéré la nature stochastique de l'expression des gènes et proposé une modélisation de son impact sur la croissance d'une population de cellules bactériennes ainsi que sur l'allocation optimale de ses ressources en protéines. Cette modélisation fait émerger un problème d'optimisation « *two stage* » linéaire avec des lois de probabilité relativement représentatives de la variabilité mesurée. Elle résulte d'un compromis entre cette représentativité et les objectifs poursuivis dans cette thèse.
Pour un type de réseau métabolique, nous avons donné une expression analytique de la solution du problème « *two stage* » correspondant, ce qui nous a permis de quantifier, au delà des résultats qualitatifs attendus, l'impact de cette nature stochastique. De par la taille des réseaux métaboliques étudiés (et donc de leur modèle), la question est de déterminer la capacité des algorithmes numériques à résoudre les problèmes d'optimisation associés. Si dans le cas général, le problème 2 (modèle déterministe) peut être résolu efficacement par programmation linéaire en utilisant un solveur comme cplex y compris pour des dimensions élevées, la question de la résolution numérique efficace du problème 3 ou du problème 4 se pose. Les algorithmes de résolution de ces problèmes seront abordés dans le

chapitre 6 et leur performance sera évaluée pour le réseau métabolique présenté dans ce chapitre. Comme ces algorithmes sont dérivés de certains algorithmes de problèmes d'optimisation convexe de grande dimension, ceux-ci sont présentés dans le chapitre suivant.

Chapitre 5

Problèmes d'optimisation déterministe de grande dimension et leur résolution numérique

Ce chapitre est composé de deux parties principales, d'abord les méthodes dites du premier ordre (MPO) pour la résolution numérique de problèmes d'optimisation convexe de grande dimension dans le cas déterministe. Les motivations, le cadre théorique, ainsi que les discussions sur la complexité algorithmique y seront donnés. Dans la deuxième partie, nous nous proposons de revisiter l'analyse de convergence de ces MPO sous le point de vue de l'automatique afin d'aller vers un nouveau cadre d'analyse de convergence des algorithmes numériques, basé sur la théorie de dissipativité [74]. Cette deuxième partie se veut comme un travail d'organisation et de synthèse offrant à l'analyse de convergence des MPO un cadre systématique.

Ces algorithmes d'optimisation convexe de grande dimension seront exploités dans le chapitre 6 pour dériver des algorithmes de résolution de problèmes d'optimisation stochastique recouvrant le problème 3 et le problème 4.

5.1 Méthodes du premier ordre pour l'optimisation convexe

5.1.1 Motivation

Il est bien connu que les méthodes polynomiales (garantissant un temps de résolution polynomial en le nombre de variables de décision et en le nombre de contraintes : les méthodes du type points intérieurs par exemple), sont efficaces pour résoudre les problèmes d'optimisation convexes avec une grande précision et un bas nombre d'itérations [9]. En effet si on considère une classe $\mathcal{C}$ de problèmes d'optimisation convexe :

$$\begin{aligned} (c) : & \min_{x \in X \subset \mathbb{R}^n} f(x) \\ \text{tel que} : & \ f_i(x) \leqslant 0, \quad i = 1, \ldots, m \end{aligned} \tag{5.1}$$

avec une structure donnée (LP, QP, SDP, etc [1]), alors chaque instance (c) de la classe peut être identifiée par un vecteur de données $D(c)$ (l'ensemble des cœfficients de l'instance). On désire calculer une solution avec une précision numérique ϵ pour l'instance (c). Un algorithme qui, en un nombre fini d'opérations arithmétiques, renvoie une telle solution à partir de l'entrée $(D(c), \epsilon)$ est dit *polynomial*, si le nombre total de ses opérations pour toute instance $(c) \in \mathcal{C}$ et pour tout $\epsilon > 0$ est borné supérieurement par $p(m(c), n(c), dim(D(c))) \ln(V(c)/\epsilon)$ [9], où p est un polynôme et $V(c)$ est un cœfficient dépendant de $D(c)$. Le terme $\ln(V(c)/\epsilon)$ peut être interprété comme le nombre de digits de précision souhaité à une solution numérique calculée. Avec cette interprétation, on constate qu'une méthode polynomiale bénéficie de la propriété qui fait que le coût arithmétique par digit de précision est borné par un polynôme de la taille $(m(c), n(c), dim(D(c)))$ du problème. D'un point de vue pratique, cette propriété sous entend une convergence rapide en terme de nombre d'itérations et par conséquent la possibilité de calculer des solutions avec une grande précision, voir [57, 59, 75, 15, 9, 51] pour plus de détails.

Néanmoins, les méthodes polynomiales partagent un inconvénient : le coût calculatoire par itération croît non linéairement en fonction de la taille du problème d'optimisation en question [9, 51]. Par exemple si on choisit la méthode ellipsoïdale [9] pour résoudre une instance du problème (5.2) avec $m = 0$, on a un coût arithmétique par itération de l'ordre de $O(n^2)$ et la méthode a besoin de $2(n+1)^2 \ln(cst/\epsilon)$ itérations pour calculer une solution avec une précision numérique ϵ, où cst est une constante dépendant des paramètres du problème (5.2) [48]. Dans certains cas de problèmes avec une structure favorable (LP), le coût arithmétique d'une itération est cubique en le nombre de variables de décision [9]. Par conséquent, plus le nombre de variables de décision augmente, plus ces méthodes perdent de leur côté efficace : pour un nombre de variables de décision de l'ordre de 10^5 ce qui est typique aux applications biologiques qui nous intéressent, une seule itération peut être très couteuse en temps de calcul. Ceci dit, ces méthodes polynomiales peuvent résoudre des problèmes de dimension similaire, typiquement dans le cas de problème d'optimisation linéaire avec des matrices de contraintes très creuses, ce qui reste un cas très particulier dans la pratique. Ainsi, dans le cas où le nombre maximal d'opérations arithmétiques autorisé (possible sur un ordinateur) devient de l'ordre du nombre de variables de décision n, c'est le cas des problèmes d'optimisation dits de grande dimension, les méthodes polynômiales deviennent limitatives. Actuellement, dans la littérature, les méthodes choisies pour résoudre les problèmes d'optimisation convexes de tailles dépassant le potentiel des méthodes du point intérieur sont les méthodes dites du premier ordre, pour plus de détails voir [35, 36]. En effet, de nombreux travaux de recherche ont été engagés afin de proposer une alternative aux méthodes polynomiales dans le cas de problèmes de grande dimension, on peut citer par exemple [35, 10, 47, 8, 4]. L'idée est de privilégier les méthodes dites « du premier ordre » du fait que leur coût calculatoire (nombre total d'itérations), contrairement aux méthodes polynomiales, est indépendant ou presque indépendant du nombre de variables de décision comme nous le verrons dans la suite. Néanmoins, ce coût calculatoire pour une solution de précision ϵ est de $O(1/\epsilon^2)$ contre $O(\ln(1/\epsilon))$ dans le cas des méthodes polynomiales. Toutefois, les méthodes du premier ordre restent les plus appropriées dans le cas des applications de grande dimension où une très grande précision n'est pas requise.

1. Programmation linéaire, Quadratique, Semi Définie, etc [9]

5.1.2 Cadre général des méthodes basées sur un oracle du premier ordre

Les méthodes numériques présentées dans cette section visent à résoudre le problème d'optimisation suivant :

$$(c) : \min_{x \in X \subset \mathbb{R}^n} f(x) \tag{5.2}$$
$$\text{tel que : } f_i(x) \leqslant 0, \quad i = 1, \ldots, m$$

Elles sont basées sur les hypothèses suivantes.

Hypothèse 2. — *(i) $X \subset \mathbb{R}^n$ est un sous-ensemble convexe borné et fermé de $\mathbb{R}^n$ muni du produit interne $\langle ., . \rangle$.*
— *(ii) La fonction objectif f de (5.2) est convexe, continue et L-Lipschitz, représentée par un oracle du premier ordre.*
— *(iii) Les sous-gradients de f en $x \in X$ $f'(x)$ sont bornés sur X.*

Par rapport à (i), notons que si X est borné et fermé alors il est compact en dimension finie et donc si f est continue, le minimum (5.2) est fini et atteint sur X. Sinon, il suffit de redéfinir la fonction de telle sorte que si $x \notin X$ alors $f(x) = \infty$. Par conséquent la fonction objectif devient coercive sur X ($\lim_{|x| \to \infty} f(x) = \infty$) et donc le minimum (5.2) sera atteint sur X. Ce résultat standard est connu sous le nom du théorème de Weierstrass (voir les théorèmes 1 et 2, pages 11 et 12).

Par rapport à (ii), l'hypothèse de Lipschitz continuité n'est pas sévère. En effet d'après l'analyse convexe [58], f est Lipschitz continue sur tout ensemble fermé et borné contenu dans le domaine de f du moment où f est convexe. On peut considérer l'extension de f à $\mathbb{R}^n$ c'est-à-dire $f(x) = \infty$ pour $x \notin X$. Dans ce cas, si X est borné alors f est Lipschitz continue.

L'hypothèse (iii) est toujours vérifiée car d'après l'analyse convexe on a $\|f'(x)\|_* \leqslant L \quad \forall x \in X$, où L est la constante de Lipschitz de f par rapport à $\|.\|$. On a $\sup_{x \in X} \|f'(x)\|_* = L$ (pour plus de détails sur ces résultats d'analyse convexe voir [58]). Les propriétés de convexité et de Lipschitz continuité de f, comme nous le verrons plus tard, permettent d'assurer la convergence des méthodes que l'on va présenter ici.

Ces méthodes sont représentées par un oracle du premier ordre, c'est-à-dire que l'information disponible sur f étant donné, un point faisable x est uniquement $f(x)$ et $f'(x)$. Cette restriction est due au fait que le nombre des variables de décision est très grand (millions par exemple) et donc, pour éviter des itérations qui durent longtemps, on ne peut utiliser que les méthodes de résolution ayant un coût calculatoire (par itération) linéaire en fonction du nombre des variables de décision. Cette restriction empêche malheureusement d'exploiter notre connaissance analytique *a priori* sur la structure du problème d'optimisation (5.2) : dans le cas où l'on sait, par exemple, que f est $\mathcal{C}^2$ et fortement convexe et que donc son hessien est défini-positif, ces informations sont inexploitables car l'usage d'une méthode basée sur un oracle du deuxième ordre (Newton, point intérieur ... et qui sont, jusqu'à un certain nombre de variables de décisions, plus rapides que ceux basées sur un oracle du premier ordre) a un coût arithmétique, par itération, au moins quadratique en le nombre des variables de décision comme cela était mentionné plus haut. L'avantage des méthodes du premier ordre est que ce coût arithmétique par itération est de fait linéaire

et qu'en plus ces méthodes présentent un taux de convergence quasiment indépendant de la taille du vecteur des variables de décision contrairement à toute autre méthode basée sur un oracle d'ordre 2 ; l'inconvénient c'est qu'elles sont relativement lentes.

La résolution du problème (5.2) consiste à trouver une solution $\epsilon-$précise c'est-à-dire [48] :

$$\text{Trouver} \quad x^{\epsilon} \quad \text{tel que} : f(x^{\epsilon}) - f^{*} \leqslant \epsilon. \tag{5.3}$$

Pour cela, étant donnée une méthode $\mathcal{M}$, on applique un schéma itératif au sein duquel la méthode sera exécutée. Formellement, un tel schéma engendrera la séquence $\{x_k\}_{\{k=1,\ldots,N\}}$. A l'itération k, le point x_k est construit en se basant sur l'ensemble $\mathcal{I}_{k-1}$ des informations acquises lors des itérations précédentes envoyées par l'oracle de premier ordre $\mathcal{O}$ en question. Si on note l'ensemble des informations envoyées par un oracle en un point x, $\mathcal{O}(f(.), x)$, le schéma général de ces méthodes peut être résumé par le processus déterministe détaillé dans l'algorithme 2.

Algorithme 2 Processus déterministe

Initialisation : $\mathcal{I}_0 = \emptyset$
Tant que $0 \leqslant k \leqslant N-1$:
prendre $k = k+1$,
prendre $\mathcal{I}_k = \mathcal{I}_{k-1} \cup \mathcal{O}(f(.), x_{k-1})$
calculer x_k sur la base de $\mathcal{I}_k$.
Fin de tant que
Renvoyer $\hat{x}_N$ solution $\epsilon-$ précise sur la base de $\mathcal{I}_N$.

Dans le cadre de l'algorithme 2, on choisit un nombre d'itérations prédéfini, fonction de la précision désirée et du taux de convergence de la méthode numérique. On peut également se baser sur un critère d'arrêt tel qu'une fois vérifié le schéma itératif s'arrête et renvoie une solution $\epsilon-$ précise mais cela ne rentre pas dans notre cadre d'étude. Ces conditions d'arrêt sont basées sur les conditions d'optimalité caractérisant les solutions optimales. Il existe plusieurs critères d'arrêt (KKT, règle de Fermat, etc.) qui dépendent de la classe de problème d'optimisation considérée et qui sont plus ou moins adaptés pour un problème donné, pour plus de détails sur ce point, voir [11]. Notre choix de présenter un schéma itératif avec un nombre d'itérations prédéfini est basé sur le fait que l'on va intégrer dans le chapitre 6 ce schéma déterministe au sein d'un autre schéma dit « stochastique » afin de résoudre des problèmes d'optimisation stochastiques. Dans ce domaine d'optimisation, on ne dispose malheureusement pas de critère d'arrêt contrairement au cas déterministe [64].

Afin d'illustrer cela, nous passons à la présentation des méthodes numériques pouvant résoudre le problème (5.2) sous les hypothèses 2 page 71.

Il faut noter que le lecteur pourra se reporter à l'annexe du chapitre (section 5.5) où se trouvent quelques définitions de base sur lesquelles la suite de ce chapitre est basée.

5.2 Quelques méthodes principales du premier ordre

Dans cette section, on présente le cadre général de ces méthodes du point de vue de la complexité algorithmique (pour plus de détails voir [11, 48, 47] et les références incluses).

5.2.1 Méthode du sous-gradient projeté

Cadre théorique On considère le produit scalaire euclidien et la norme euclidienne $\|.\|_2$. Sous les hypothèses 2 page 71, le problème (5.2) admet au moins une solution. Notons que du moment que f est convexe, l'ensemble des sous-différentiels de f, en tout point de X, est non vide et la norme des sous-gradients est inférieure à L [11].

La méthode consiste à générer une séquence de points $\{x_k\}_{\{k=0,\dots,N-1\}}$ à partir d'un point $x_0 \in X$ quelconque conformément à l'algorithme 3, où $f'(x)$ est un sous-gradient de f en x, $\gamma_k > 0$ est le pas de la méthode (qui peut être variable ou fixe), $\pi_X(.)$ est la projection euclidienne sur l'ensemble X (voir définition 18 dans l'annexe du chapitre, section 5.5).

Algorithme 3 Algorithme du sous-gradient projeté

Entrées :
- un point initial $x_0 \in X$;
- une constante $R > 0$ telle que $\|x - x_0\|_2 \leqslant R$, $\forall x \in X$
- une constante $L > 0$ telle que $\|y\|_2 \leqslant L$, $\forall y \in \partial f(x)$, $\forall x \in X$;
- un nombre total d'itérations N ;
- un pas γ_k pour la méthode.

Sorties : une solution $\hat{x}_N$ $\epsilon-$ précise du problème (5.2).
Tant que $(0 \leqslant k \leqslant N-1)$:
prendre $k = k + 1$;
prendre $y_{k-1} = f'(x_{k-1}) \in \partial f(x_{k-1})$;
prendre $x_k = \pi_X(x_{k-1} - \gamma_k y_{k-1}/||y_{k-1}||_2)$.
Fin de tant que
Renvoyer $\hat{x}_N := \text{argmin}_{x \in \{x_0,\dots,x_{N-1}\}} f(x_k)$.

Il faut noter que par rapport à un problème d'optimisation sans contrainte avec de bonnes propriétés telles que sa fonction objectif est différentiable (donc ses sous-gradients coïncident avec son gradient), la propriété de Fermat [48] c'est-à-dire $f'(x) = 0$ correspond à x solution optimale est vérifiée. Dans le problème (5.2), ce n'est pas nécessairement le cas. Par suite, on ne peut pas exploiter cette propriété pour l'analyse de convergence.

Le résultat standard sur la convergence de l'algorithme 3 est résumé dans le théorème suivant.

Théorème 6. *[48] Sous les hypothèses 2 page 71, le point $\hat{x}_N$ calculé par l'algorithme 3 vérifie, pour un pas γ_k non fixe, l'inégalité suivante*

$$f(\hat{x}_N) - f^* \leqslant \frac{L}{2} \frac{R^2 + \sum_{k=1}^{N} \gamma_k^2}{\sum_{k=1}^{N} \gamma_k} \tag{5.4}$$

Pour assurer la convergence du membre de droite de (5.4), on doit faire certaines hypothèses sur le pas γ_k. L'hypothèse standard est la suivante [11, 48] :

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \gamma_k \to 0, k \to \infty.$$

Ici, étant donné N, le meilleur choix du pas γ_k qui minimise le membre de droite de (5.4) par rapport à γ_k est le pas fixe :

$$\gamma_k = \gamma^* = \frac{R}{\sqrt{N}},$$

ce qui justifie bien la pertinence de considérer un pas fixe optimal dans le cas d'un algorithme de sous-gradient projeté avec un nombre prédéfini d'itérations. Ainsi si on exploite l'hypothèse de la bornitude de X, X est incluse dans la boule $B(x_0, R)$ de centre x_0 et de rayon R c'est-à-dire $\exists R : \forall x \in X,\ ||x - x_0||_2 \leqslant R$, alors avec une stratégie de pas constant, c'est-à-dire $\gamma_k = R/\sqrt{N}, k = 0, \ldots, N-1$, l'inégalité (5.4) devient :

$$f(\hat{x}_N) - f^* \leqslant \frac{LR}{\sqrt{N}}.$$

Ainsi le taux de convergence de la méthode du sous-gradient est en $O(1/\sqrt{N})$. Par conséquent, pour calculer une solution $\epsilon-$précise, on a besoin d'un nombre d'itérations de l'ordre de $O(1/\epsilon^2)$ (il suffit de prendre $\epsilon \leqslant \frac{LR}{\sqrt{N}}$). Certes la méthode, dans ce sens, est beaucoup plus lente que les méthodes polynomiales qui proposent un nombre d'itérations de $O(\ln(1/\epsilon))$. Néanmoins, le taux de convergence de la méthode ne dépend qu'implicitement, à travers R et L, de la taille du problème d'optimisation contrairement aux méthodes polynomiales qui dépendent lourdement de la taille du problème.

5.2.1.1 Discussion sur l'optimalité de la méthode

A ce stade, la question qui se pose naturellement est de voir si on peut trouver une méthode avec un meilleur taux de convergence en utilisant les mêmes outils (sous-gradients et projection euclidienne).

La mauvaise nouvelle est que sous les hypothèses 2 (dans le cas d'une fonction objectif convexe et $L-$Lipschitz continue non différentiable), la réponse est non. En effet si on considère l'ensemble $\mathcal{M}$ des méthodes de minimisation du premier ordre du problème (5.2) dans le cas non contraint ($X = \mathbb{R}^n$) sous les hypothèses 2, une méthode $m \in \mathcal{M}$ peut s'écrire sous la forme suivante [48] :

$$x_{k+1} \in x_0 + Lin(\{\eta_1, \ldots, \eta_k\}),$$

où $\eta_i \in \partial f(x_i)$. On a le résultat suivant :

Théorème 7. *[48, page 138] Pour tout $k \leqslant n-1$, il existe une fonction f convexe et $L-$Lipschitz continue sur la boule $B(x_0, R)$ de centre x_0 et de rayon R telle que pour toute méthode $m \in \mathcal{M}$, on a :*

$$f(x_k) - f^* \geqslant \frac{LR}{2(1+\sqrt{k+1})}. \tag{5.5}$$

Remarque : Ce théorème est énoncé pour les problèmes d'optimisation sans contrainte : $X = \mathbb{R}^n$. D'une part, l'optimisation sans contrainte est un cas particulier d'optimisation avec contrainte. D'autre part, on peut envisager d'étendre ce résultat à l'optimisation sous contrainte ($X \subset \mathbb{R}^n$). En effet, il suffit de prendre x_k tel que $x_{k-1} - \gamma y_{k-1}/||y_{k-1}||_2 \in X$

ce qui fait que $\pi_X(x_{k-1} - \gamma y_{k-1}/||y_{k-1}||_2) = x_{k-1} - \gamma y_{k-1}/||y_{k-1}||_2$ et par conséquent l'algorithme 3 fera partie de $\mathcal{M}$.

Ce théorème est plein de conséquences. En particulier celle qui nous intéresse, fait que sous l'hypothèse que le nombre d'itérations de l'algorithme 3 n'est pas très grand par rapport au nombre de variables de décision n ($N \leqslant n$), ou, autrement dit, n est très grand ($n \geqslant 1/\epsilon^2$), il existe des instances du problème (5.2) qui requièrent, au moins, un nombre d'itérations de $O(1/\epsilon^2)$ pour atteindre une solution $\epsilon-$précise. Or la méthode du (sous)-gradient projeté requiert, au plus, un nombre d'itérations de $O(1/\epsilon^2)$. Dans ce sens la méthode est déjà optimale et nous ne pouvons pas faire mieux.

Cela dit, il est important de remarquer que si on pouvait modifier le choix de la norme dans l'algorithme 3, cela permettrait de modifier les valeurs de L et R. Ainsi faisant, on arriverait à améliorer la rapidité de l'algorithme en jouant sur les paramètres cachés dans le O même si le taux de convergence $O(1/\sqrt{N})$ est déjà optimal. En effet, on a défini les constantes R et L par rapport à la norme 2. Si on considère leurs homologues en norme 1, L_1, R_1, on sait que le ratio $\frac{LR}{L_1R_1}$ peut être très grand comme il peut être très petit. Dans le premier cas, le choix de la norme 1 est judicieux. Par contre si le ratio est très grand, il est préférable de rester sur le choix de la norme 2 au sein de l'algorithme 3. Malheureusement, on ne peut pas modifier le choix de la norme 2 pour cet algorithme. Par contre, il a été possible de généraliser la méthode du sous-gradient projeté afin de mettre en œuvre des méthodes numériques qui rendent possible d'adapter le choix de la norme à la géométrie du problème en question afin de garantir une meilleure rapidité de convergence (efficacité d'estimation) que la méthode du sous-gradient projeté. On va présenter dans la suite le cadre général des quelques méthodes les plus répandues.

5.2.2 General Mirror Descent Algorithm

Cadre théorique La méthode Mirror Descent Algorithm (MDA) se propose de généraliser la méthode vue précédemment afin de bénéficier d'une meilleure rapidité de convergence (voir [47] pour plus de détails). Pour cela, l'espace $\mathbb{R}^n$ sera muni du produit scalaire $\langle ., . \rangle$ et de la norme $\|.\|$ et on aura besoin d'une fonction de classe $\mathcal{C}^1$ génératrice de distance [8] $\omega(.) : X \to \mathbb{R}$ fortement convexe de paramètre α, c'est-à-dire :

$$\forall x, y \in X : \langle \omega'(x) - \omega'(y), x - y \rangle \geqslant \alpha \|x - y\|^2, \tag{5.6}$$

On introduit la fonction dite « distance de Bregman » [16] ou encore « prox -function » définie par :

$$\omega_x(y) = \omega(y) - \omega(x) - \langle y - x, \omega'(x) \rangle. \tag{5.7}$$

Notons que (5.6) est équivalent au fait que la distance de Bregman (5.7) vérifie :

$$\forall x, y \in X : \omega_x(y) \geqslant \frac{\alpha}{2} \|x - y\|^2. \tag{5.8}$$

La méthode MDA génère une séquence de points $x_k \in X$ selon la règle :

$$x_{k+1} = \underset{y \in X}{\operatorname{argmin}} \left\{ \omega_{x_k}(y) + \gamma_k \left\langle f'(x_k), y - x_k \right\rangle \right\}, \tag{5.9}$$

où $\gamma_k > 0$ est le pas de la méthode. On remarque que, pendant un pas de la méthode, on minimise sur X le modèle du premier ordre au voisinage de x_k de la fonction f augmentée par le terme $w_{x_k}(y)$ qui mesure une espèce de distance (distance de Bregman) entre le point courant x_k et le point suivant x_{k+1} et qui selon (5.8) pénalise toute déviation du point x_k et garantit que l'itération suivante est proche de l'itération précédente (en terme des x_k). Ceci permet d'éviter les pas trop longs résultant éventuellement en un point où le « modèle » linéaire de f est très loin de f. La figure 5.1 illustre la distance de Bregman (initialement introduite dans [16]).

Remarque : Ici on va rebondir sur la discussion de la section précédente et expliquer en quoi la méthode MDA est une généralisation de la méthode du gradient projeté. En effet, dans l'algorithme 3, la projection euclidienne peut être écrite sous la même forme que celle du processus itératif de la méthode MDA (5.9) :

$$\begin{aligned} x_{k+1} &= \pi_X(x_k - \gamma_k f'(x_k)/||f'(x_k)||) \\ &:= \operatorname*{argmin}_{x\in X}\{||x - x_k + \gamma_k f'(x_k)/||f'(x_k)||||_2^2\}. \\ &= \operatorname*{argmin}_{x\in X}\{\tfrac{1}{2}||x - x_k||_2^2 + \gamma_k\langle x - x_k, f'(x_k)/||f'(x_k)||\rangle\}. \end{aligned}$$

Or si on choisit la fonction génératrice de distance $\omega(x) = \frac{1}{2}\|x\|_2^2$, on obtient $\omega_x(y) = \|y - x\|_2^2$.

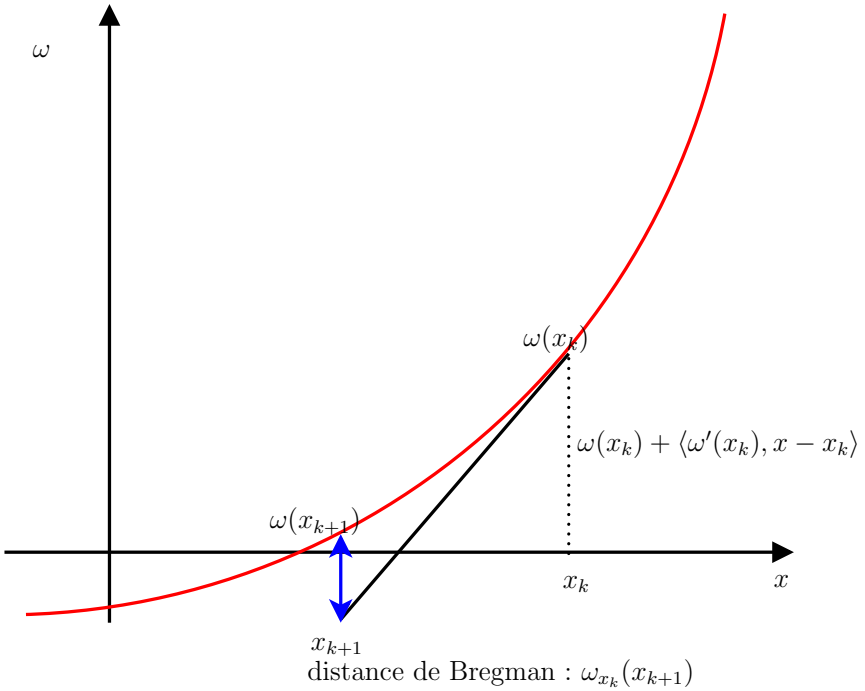

FIGURE 5.1 – Distance de Bregman.

La méthode MDA consiste à générer une séquence de points $\{x_k\}_{\{k=0,\ldots,N-1\}}$ à partir d'un point $x_0 \in int(X)$ quelconque conformément au schéma itératif de l'algorithme 4.

Les propriétés de convergence de la méthode MDA sont résumées dans le théorème suivant.

Théorème 8. *[4, 8] Sous l'hypothèse 2 page 71, et pour un pas γ_k quelconque le point $\hat{x}_N$ calculé par l'algorithme 4 vérifie l'inégalité suivante :*

Algorithme 4 Mirror Descent Algorithm (MDA)

Entrées :

- une fonction génératrice de distance ω de classe $\mathcal{C}^1$ et fortement convexe de cœfficient α ;
- un point initial $x_0 \in int(X)$;
- un nombre total d'itérations N.

Sorties : une solution $\hat{x}_N$ $\epsilon-$ précise du problème (5.2).

Tant que $(0 \leqslant k \leqslant N-1)$:

prendre $k = k+1$;

prendre γ_k pas de la méthode tel que $\gamma_k \to 0$, quand $k \to \infty$ et la série $\sum\limits_k \gamma_k$ diverge ;

prendre $y_{k-1} = f'(x_{k-1}) \in \partial f(x_{k-1})$;

prendre $x_k = \underset{y \in X}{\operatorname{argmin}} \left\{ \omega_{x_{k-1}}(y) + \gamma_{k-1} \left\langle y_{k-1}, y - x_{k-1} \right\rangle \right\}$.

Fin de tant que

Renvoyer $\hat{x}_N := \operatorname{argmin}_{x \in \{x_0, \ldots, x_{N-1}\}} f(x_k)$.

$$f(\hat{x}_N) - f^* \leqslant \frac{\Omega + 2\alpha^{-1} \sum\limits_{k=0}^{N-1} \gamma_k^2 \|f(x_k)\|_*^2}{\sum\limits_{k=1}^{N-1} \gamma_k}, \tag{5.10}$$

avec $\Omega = \max\limits_{x,y \in X} \{\omega(y) - \omega(x) - \langle y - x, \omega'(x) \rangle\}$.

En particulier,

- *si* $\gamma_k \to 0$, *quand* $k \to \infty$ *et la série* $\sum\limits_k \gamma_k$ *diverge, alors* $f(x^N) \to f^*$;
- *si on définit en particulier le pas [4] :*

$$\gamma_k = \frac{\sqrt{2\Omega\alpha}}{L\sqrt{k}},$$

alors la sortie $\hat{x}_N$ *de l'algorithme 4 vérifie*

$$f(\hat{x}_N) - f^* \leqslant \epsilon = L\sqrt{\frac{2\Omega}{\alpha N}}.$$

On note que $\Omega < \infty$ car $\omega(.)$ est $\mathcal{C}^1$ sur X et X est compacte selon l'hypothèse 2 page 71.

5.2.2.1 Adaptabilité

L'avantage de la méthode MDA est le degré de liberté (la norme $\|.\|$ et la fonction génératrice de distance $\omega(.)$) qui permet d'ajuster la méthode à la géométrie spécifique du problème (X est supposé à géométrie simple : simplexe, boule euclidienne, cube, etc.). Nous donnons ici quelques configurations de la méthode MDA ainsi que leurs ensembles de faisabilité les plus appropriés (voir [35]) et nous détaillons la comparaison entre les deux configurations du point de vue de la rapidité de convergence :

- configuration euclidienne : X est un compact convexe inclus dans la boule euclidienne :

$$\Delta_2 = \{x \in \mathbb{R}^n : \|x\|_2 \leqslant 1\}.$$

La fonction génératrice de distance est donnée par : $\omega(x) = 1/2\langle x, x\rangle$. De plus on a $\|.\| = \|.\|_2$ et donc, $\|.\|_* = \|.\|_2$. Ceci implique que $\alpha = 1$ et que $\Omega = D_X := 1/2 \max_{x,y} \|x - y\|_2$ est le diamètre de l'ensemble faisable X. Le théorème 8 devient [8] :

$$\gamma_k = \frac{D_X}{\|f'(x_k)\|_2\sqrt{2k}} \Rightarrow f(\hat{x}_N) - f^* \leqslant \underbrace{O(1)\frac{L_2}{\sqrt{2N}}}_{\epsilon_{l_2}}, \tag{5.11}$$

où L_2 est la constante de Lipschitz de f par rapport à la norme 2, c'est-à-dire $\sup_{x\in X} \|f'(x)\|_2$, ϵ_{l_2} est l'efficacité d'estimation de l'algorithme 4 dans le cas de la configuration euclidienne. Notons que dans cette configuration la méthode MDA est réduite exactement à la méthode du gradient projeté.

- configuration l_1 : X est un compact convexe inclus dans le simplexe standard suivant :

$$\Delta_1 = \{x \in \mathbb{R}^n_+ : \sum_i x_i \leqslant (\text{ou } =)1\}.$$

La fonction génératrice de distance est donnée par : $\omega(x) = \sum_{i=1}^{n} x_i \ln(x_i)$. De plus on a $\|.\| = \|.\|_1$ et donc, $\|.\|_* = \|.\|_\infty$. Ici $\Omega \leqslant O(1)\ln(n)$ et $\alpha = O(1)$ (voir [10]). Le théorème 8 devient [8] :

$$\gamma_k = \frac{\sqrt{\ln(n)}}{\|f'(x_k)\|_\infty\sqrt{k}} \Rightarrow f(\hat{x}_N) - f^* \leqslant \underbrace{O(1)\frac{L_1\sqrt{\ln(n)}}{\sqrt{N}}}_{\epsilon_{l_1}}, \tag{5.12}$$

où L_1 est la constante de Lipschitz de f par rapport à la norme 1, c'est-à-dire $\sup_{x\in X} \|f'(x)\|_\infty$, ϵ_{l_1} est l'efficacité d'estimation de l'algorithme 4 dans le cas de la configuration l_1.

Pour illustrer la capacité de la méthode MDA à s'adapter à la géométrie spécifique du problème en question, on va considérer le problème (5.2) dans le cas où X est la boule unité, donnée par la norme $\|.\|_p$, dans $\mathbb{R}^n$ où $p = 1$ ou $p = 2$, et on va comparer la performance de la méthode dans le cas d'une configuration euclidienne ainsi que dans le cas d'une configuration l_1.

Configuration euclidienne : si $p = 1$ et comme $\|x\|_2 \leqslant \|x\|_1$, $\forall x$, alors X est inclus dans Δ_2. Ainsi on est placé dans la configuration euclidienne. Si $p = 2$ alors on est également placé dans la configuration euclidienne car $X = \Delta_2$. Dans ces deux cas, et si on suppose que la solution optimale du problème (5.2) appartient à $X \cap \Delta_2$, alors l'efficacité d'estimation de l'algorithme ϵ_{l_2} est donnée par le membre de droite de l'inégalité (5.11).

Configuration l_1 : si $p = 1$ alors on est directement placé dans la configuration l_1 car $\Delta_1 = X$. L'efficacité d'estimation de l'algorithme ϵ_{l_1} est donnée par le membre de droite de (5.12). Si $p = 2$ et pour se placer dans la configuration l_1 pour notre problème, il suffit de faire le changement de variable $x = \sqrt{n}u$ ainsi $X = \{x \in \mathbb{R}^n_+ : \exists u \in \mathbb{R}^n_+ : x = \sqrt{n}u, \|u\|_2 \leqslant 1/\sqrt{n}\}$. Du fait que $\|u\|_1 \leqslant \sqrt{n}\|u\|_2 \forall u$, on a $X \subset \Delta_1$. Si on suppose que la solution optimale du problème (5.2) (par rapport à la variable u) appartient à $X \cap \Delta_1$, alors l'efficacité d'estimation de l'algorithme ϵ_{l_1} est donnée par le membre de droite de l'inégalité (5.12) par rapport à la variable u et par suite par rapport à la variable x on a :

$$\epsilon_{l_1} = O(1)\sqrt{n}\sqrt{\ln(n)}\frac{\sup_{x\in X}\|f'(x)\|_\infty}{\sqrt{N}}.$$

Pour résumer, dans le cas d'une configuration, l_1 on a $\epsilon_{l_1} = O(1)n^{1-1/p}\sqrt{\ln(n)}\frac{\sup_{x\in X}\|f'(x)\|_\infty}{\sqrt{N}}$. Dans le cas d'une configuration euclidienne on a : $\epsilon_{l_2} = \frac{L_2 O(1)}{\sqrt{2N}}$. Considérons maintenant le ratio des deux efficacités $\Theta := \epsilon_{l_2}/\epsilon_{l_1}$, on a :

$$\Theta = \underbrace{\frac{O(1)}{n^{1-1/p}\sqrt{\ln(n)}}}_{A}\underbrace{\frac{\sup_{x\in X}\|f'(x)\|_2}{\sup_{x\in X}\|f'(x)\|_\infty}}_{B}. \tag{5.13}$$

Notons que $\Theta << 1$ veut dire que la configuration euclidienne est significativement plus efficace que la configuration l_1. Et que $\Theta >> 1$ signifie exactement le contraire. Le terme A est $\leqslant 1$ et par conséquent est toujours favorable pour la configuration euclidienne. Le facteur B est par contre favorable pour la configuration l_1 et est $\geqslant 1$ ou $\leqslant \sqrt{n}$: il peut être de l'ordre de $\sqrt{n}$ (dans le cas où tous les coefficients de $f'(x)$ sont du même ordre de grandeur). Le facteur A dépend de la géométrie et de la taille du problème d'optimisation et le terme B dépend de la nature de la fonction objectif f. Ainsi pour une meilleure performance de l'algorithme, il faut faire un choix raisonnable entre les deux configurations. Pour illustrer cela, on prend $p = 2$, on a $A = \frac{1}{n^{1/2}\sqrt{\ln(n)}}$, dans ce cas le produit AB est $\leqslant 1$ et la situation est favorable pour la configuration euclidienne. Maintenant quand $p = 1$, $A = \frac{1}{\sqrt{\ln(n)}}$ tandis que B peut être aussi grand que $\sqrt{n}$. Cette situation est en faveur de la configuration l_1 . La performance de la configuration euclidienne peut dépasser (marginalement) celle de la configuration l_1 par un facteur $\leqslant \sqrt{\ln(n)}$ dans le cas où B est de l'ordre de 1. Par contre, il y a de forte chance que la configuration l_1 dépasse celle euclidienne par un facteur $\frac{\sqrt{n}}{\sqrt{\ln(n)}}$ ce qui peut faire une grande différence quand n est grand : il est avantageux de choisir la configuration l_1 quand il s'agit de minimiser sur un simplexe de très grande dimension ! Mais globalement, il est pertinent de choisir la configuration euclidienne quand $p = 2$ et une configuration l_1 quand $p = 1$.

5.2.3 Primal Dual Subgradient Algorithm

Cadre théorique La méthode Primal Dual Subgradient Algorithm (PDA) propose une solution $\epsilon-$optimal à (5.2) sous les hypothèses 2 page 71. Le principe et la base théorique de la méthode ont été initialement introduits dans [49, 50]. Bien qu'on ne puisse pas,

comme on l'a expliqué plus haut (voir théorème 7 et discussions page 74), améliorer le taux de convergence, vers une solution $\epsilon-$optimale, des méthodes basées sur un oracle du premier ordre [48] (il faut un nombre d'itérations $O(1/\epsilon^2)$), la méthode PDA (comme la méthode MDA) présente une amélioration considérable par rapport à la méthode standard du gradient. Essentiellement, la méthode PDA n'a pas besoin de fixer par avance son nombre total d'itérations garantissant au moins une précision ϵ : à la place, un écart « gap »de dualité est évalué pendant chaque itération et est utilisé dans un critère d'arrêt de la méthode.

En effet, si on étend le domaine de f à $\mathbb{R}^n$ en posant $f(x) = \infty$ si $x \notin X$, la fonction conjuguée f_* de f est donnée par :

$$f_*(\zeta) := \sup_{x\in\mathbb{R}^n} \{\langle \zeta, x\rangle - f(x)\},$$

et le dual du problème (5.2) est [25] :

$$\max_{\zeta\in dom f_*} \phi(\zeta), \tag{5.14}$$

où

$$\phi(\zeta) := -f_*(\zeta) + \min_{x\in X} \langle \zeta, x\rangle. \tag{5.15}$$

L'idée de la méthode est d'utiliser les sous-gradients calculés de f pour déterminer une approximation de la solution du problème primal ainsi que celle du problème dual. La méthode est dotée d'une mémoire dans le sens où elle retient les gradients calculés au fur et à mesure des itérations (contrairement à la méthode du gradient classique dont toute la mémoire est « concentrée » dans l'itération courante) pour créer un modèle sous la forme d'une fonction lisse et fortement convexe (du type $x \to \langle y, x\rangle + \beta\omega(x)$ avec $\beta > 0$ et ω la fonction génératrice de distance pour la méthode) dont le minimiseur tend vers la solution optimale du problème (5.2).

Ingrédients de la méthodes :

- une norme $\|.\|$ associée à l'espace vectoriel E contenant X, sa norme duale $\|.\|_*$ sur l'espace dual E_* de E : $\|\xi\|_* = \max_{x\in X}\{\langle \xi, x\rangle : \|x\| \leqslant 1\}$;
- une fonction génératrice de distance $\omega(x) : X \to \mathbb{R}$ $\mathcal{C}^1$ et fortement convexe de paramètre α. Notons que ω n'a pas besoin d'être différentiable sur l'intérieur de X contrairement à la méthode MDA ;
- un point initial

$$x^0 := \operatorname{argmin}_{x\in X} \omega(x). \tag{5.16}$$

L'algorithme 5 correspond au schéma général de la méthode PDA. Son idée principale est illustrée figure 5.2 : l'accumulation des sous-gradients calculés et le calcul d'un modèle approché de la fonction objective à l'aide d'une fonction fortement convexe génératrice de distance ω. La minimisation de ce modèle sur l'ensemble faisable X donne le point suivant auquel un sous-gradient de f doit se calculer et ainsi de suite jusqu'à ce que le gap de dualité soit $\leqslant \epsilon$.
Les propriétés de convergence de la méthode sont résumées dans le théorème suivant.

Algorithme 5 Primal-Dual Subgradient Algorithm (PDA)

Entrées :

- une fonction génératrice de distance ω fortement convexe de cœfficient α ;
- un point initial $x^0 := \operatorname{argmin}_{x\in X} \omega(x)$;
- un nombre total d'itération N ;
- une suite numérique β_k ;
- une suite numérique λ_k.

Sortie : une solution $\hat{x}$ $\epsilon-$précise pour le problème (5.2).

Initialisation :

choisir $\beta_0 > 0$;
choisir $\lambda_0 > 0$;
mettre $\Lambda_0 = \lambda_0$;
mettre $x_0 = x^0$;
calculer $\xi_0 \in \partial f(x_0)$;
mettre $\zeta_0 = \lambda_0 \xi_0$;
mettre $k = 0$.

Tant que $(0 \leqslant k \leqslant N-1)$:

mettre $k = k + 1$;
choisir $\beta_k \geqslant \beta_{k-1}$;
calculer $x_k = \operatorname{argmin}_{x\in X}\{\langle \zeta_{k-1}, x\rangle + \beta_k \omega(x)\}$;
choisir $\lambda_k > 0$;
mettre $\Lambda_k = \Lambda_{k-1} + \lambda_k$;
calculer $\xi_k \in \partial f(x_k)$ et mettre $\zeta_k = \zeta_{k-1} + \lambda_k \xi_k$.

Fin de tant que

Renvoyer : $\hat{x}_N = \frac{1}{\Lambda_{N-1}} \sum_{i=0}^{N-1} \lambda_i x_i$.

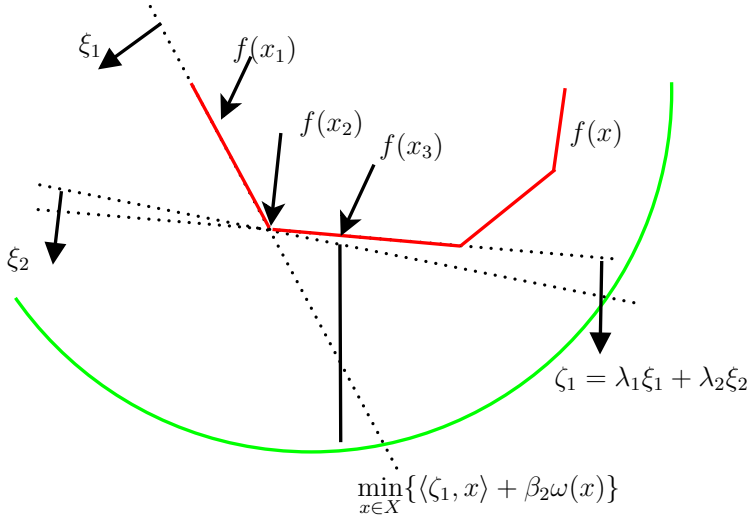

FIGURE 5.2 – Primal Dual Subgradient Method [25].

Théorème 9. *[50] Sous les hypothèses 2 page 71, l'efficacité d'estimation de l'algorithme 5 au bout de N itérations est donnée par la borne supérieure suivante :*

$$f(\hat{x}_N) - f^* \leqslant \frac{1}{\sum_{i=0}^{N}\lambda_i}\left(\beta_{N+1}D_X + \frac{1}{2\alpha}\sum_{i=0}^{N}\frac{\lambda_i^2}{\beta_i}\|\xi_i\|_*^2\right) \tag{5.17}$$

où $D_X = \max_{x\in X}\omega(x)$.

Au regard de l'inégalité (5.17), la rapidité de convergence de la méthode PDA dépend du choix des suites λ_k et β_k. En particulier les deux cas de figures suivants ont été considérés dans [50] :

Moyennage Simple (Simple averaging) :

$$\lambda_k = 1, \quad \beta_k = \frac{L}{\sqrt{2\alpha D_X}}\hat{\beta}_k, \quad \forall k \geqslant 0 \tag{5.18}$$

Moyennage Pondéré (Weighted averaging) :

$$\lambda_k = \frac{1}{\|\xi_k\|_*}, \quad \beta_k = \frac{1}{\sqrt{2\alpha D_X}}\hat{\beta}_k, \quad \forall k \geqslant 0 \tag{5.19}$$

où $\hat{\beta}_k$ est telle que :

$$\hat{\beta}_0 = \hat{\beta}_1 = 1 \quad et \quad \hat{\beta}_{k+1} = \hat{\beta}_k + \frac{1}{\hat{\beta}_k} \quad \forall k \geqslant 1. \tag{5.20}$$

On note ici que les β_k représentent des pas optimaux dans le sens où ils ont été choisis de telle sorte à ce que le membre de droite de (5.17) soit minimal. Par ailleurs, avec le choix (5.20) on peut démontrer que [50] :

(a) $\hat{\beta}_{k+1} = \sum_{i=0}^{k}\frac{1}{\hat{\beta}_i}, \quad k \geqslant 0,$

(b) $\sqrt{2k-1} \leqslant \hat{\beta}_k \leqslant \frac{1}{1+\sqrt{3}} + \sqrt{2k-1} \quad k \geqslant 1.$

Dans ce cas le théorème 9 entraîne le théorème 10.

Théorème 10. *[50] Sous les hypothèses 2 page 71, l'efficacité d'estimation de l'algorithme 5 au bout de N itérations par simple averaging (5.18) ou bien par weighted averaging (5.19) est donnée par la borne supérieure suivante :*

$$f(\hat{x}_N) - f^* \leqslant \frac{2L}{\sqrt{N}}\sqrt{\frac{2D_X}{\alpha}}, \tag{5.21}$$

où $D_X = \max_{x \in X} \omega(x)$.

Notons que la méthode PDA a la même capacité de s'adapter à la géométrie spécifique du problème en question du fait de sa configuration « proximale ». La même discussion de la section précédente vaut pour la méthode PDA avec dans ce cas une fonction génératrice de distance $\omega(x) = \ln(n) + \sum_{i=1}^{n} x_i \ln(x_i)$ dans le cas d'une configuration l_1 (voir la discussion page 77).

5.3 Analyse de la convergence des méthodes du premier ordre par la dissipativité

Les différents algorithmes présentés dans la section précédente sont des algorithmes itératifs. Par suite, ils peuvent être interprétés comme des systèmes dynamiques en temps discret. L'intérêt de cet angle de vue est d'étudier leur convergence par application des approches d'Automatique développées pour l'étude de la stabilité des systèmes dynamiques et de comprendre par exemple les propriétés que peut apporter la structure plus complexe de l'algorithme « Primal Dual Averaging » par rapport à l'algorithme du sous-gradient projeté. Pour cela, on va ré-écrire les équations de récurrence des algorithmes itératifs comme le modèle d'un système bouclé et appliquer une démarche inspirée par la théorie de la dissipativité, voir [74, 17] ou plus récemment [32]. On peut rapprocher notre démarche de la démarche adoptée dans [39]. Dans ce papier, il s'agit d'étudier la convergence d'algorithmes de résolution de problèmes d'optimisation convexe *sans contrainte* dont la fonction de coût est une fonction fortement convexe avec un gradient Lipschitz continu. Dans ce cas-là, l'algorithme s'interprète comme le rebouclage d'un système dynamique linéaire stationnaire et d'une non-linéarité statique, ce qui permet aux auteurs de réaliser l'étude en appliquant l'analyse par contraintes intégrales quadratiques (IQC) [44] qui a des liens très forts avec la dissipativité [2]. Dans ce chapitre, nous considérons une forme plus générale de problèmes d'optimisation convexe puisqu'ils sont avec contraintes et avec moins d'hypothèses sur la fonction de coût. Comme nous allons voir dans la suite, les algorithmes considérés dans ce cas s'interprètent comme le rebouclage d'un système dynamique *non-linéaire* et d'une non-linéarité statique, ce qui ne permet pas l'application de l'analyse par IQC.

A travers cette étude, nous présentons un cadre systématique et commun à l'analyse de convergence des MPO. En effet, nous montrons que l'analyse de convergence des 3 MPO

2. Les contraintes intégrales quadratiques peuvent se reformuler comme des inégalités de dissipativité sur des signaux bien précis.

passe par trois étapes clés à savoir : *i*) l'identification des sous-systèmes dynamiques constituant la MPO, *ii*) l'identification de leur interconnexion, et *iii*) la recherche d'une fonction de dissipativité globale en se basant sur certaines propriétés locales de chaque sous-système ainsi que leur interconnexion.

5.3.1 Notions sur l'analyse de la stabilité des systèmes bouclés par la dissipativité

Un des grands intérêts de la théorie de la dissipativité est de relier les propriétés de stabilité d'un système bouclé (ou d'une interconnexion de sous-systèmes) aux propriétés de ses éléments constitutifs (sous-systèmes). Dans cette sous-section, cette idée va être illustrée de façon informelle par l'étude de la stabilité asymptotique d'un système bouclé constitué par l'interconnexion du sous-système 1, en temps discret, défini par

$$\left\{\begin{array}{lcl} x^1_{k+1} & = & f_1(x^1_k, w_k) \\ z_k & = & g_1(x^1_k) \\ x^1_0 & = & x^1_0 \end{array}\right. \tag{5.22}$$

avec le sous-système 2, en temps discret, défini par

$$\left\{\begin{array}{lcl} x^2_{k+1} & = & f_2(x^2_k, z_k) \\ w_k & = & g_2(x^2_k) \\ x^2_0 & = & x^2_0. \end{array}\right. \tag{5.23}$$

On cherche ici à montrer que pour tout x^1_0 et x^2_0, le vecteur

$$x_k = \left[\begin{array}{c} x^1_k \\ x^2_k \end{array}\right]$$

tend vers 0 quand k tend vers l'infini. Pour cela, on peut chercher une fonction $V(x)$ équivalente à une norme euclidienne ($V(x) = \|x\|^2$ par exemple), c'est-à-dire qu'il existe $\alpha > 0$ et $\beta > 0$ tels que

$$\forall x \quad \alpha\|x\|^2 \leqslant V(x) \leqslant \beta\|x\|^2$$

et pour x_{k+1} et x_k vérifiant (5.22) et (5.23), on a

$$V(x_{k+1}) - V(x_k) < 0. \tag{5.24}$$

Supposons qu'il existe deux fonctions $V^1(x^1)$ et $V^2(x^2)$ équivalentes à une norme euclidienne et une forme quadratique $q(z, w)$ telles que les inégalités de dissipativité ci-dessous soient satisfaites :

$$V^1(x^1_{k+1}) - V^1(x^1_k) + q(z_k, w_k) < 0 \tag{5.25}$$

et

$$V^2(x^2_{k+1}) - V^2(x^2_k) - q(z_k, w_k) < 0. \tag{5.26}$$

Alors en sommant les inégalités (5.25) et (5.26) et en choisissant $V(x) = V^1(x^1) + V^2(x^2)$, on obtient une fonction V équivalente à une norme euclidienne qui satisfait (5.24), ce qui démontre la stabilité. Le procédé peut être étendu pour démontrer des propriétés autres que la stabilité, qui peuvent se traduire par l'existence d'une fonction V vérifiant des

inégalités plus complexes que (5.24), voir par exemple [13].

Dans les sous-sections qui suivent, les algorithmes de résolution vont être réécrits comme l'interconnexion de deux sous-systèmes pour lesquels il est possible de proposer des fonctions V^1, V^2 et q.

5.3.2 Algorithme du sous-gradient projeté

L'algorithme du sous-gradient projeté s'exprime comme un système en temps discret défini par la représentation d'état, avec $\pi_X(x) = \text{argmin}_{y\in X} \|x-y\|_2$ la projection euclidienne de x sur X :

$$\left\{ \begin{array}{lcl} x_{k+1} & = & \pi_X\left(x_k - \gamma_k \frac{g(x_k)}{\|g(x_k)\|_2}\right) \\ x_0 & = & x_0 \end{array} \right.$$

avec x_k le vecteur d'état et $g(x_k) \in \partial f(x_k)$. Ce système peut se récrire comme l'interconnexion de deux sous-systèmes :

$$\left\{ \begin{array}{lcl} x_{k+1} & = & \pi_X(x_k - w_k) \\ x_k & = & x_k \\ x_0 & = & x_0 \end{array} \right. \tag{5.27}$$

et

$$w_k = \gamma_k \frac{g(x_k)}{\|g(x_k)\|_2} \tag{5.28}$$

avec $\gamma_k \to 0$ quand $k \to +\infty$. Le premier sous-système correspond à une non-linéarité rebouclée sur un retard d'une période d'échantillonnage et a pour sortie x_k ; le second sous-système est une non-linéarité statique ($\gamma_k \frac{g(x_k)}{\|g(x_k)\|_2}$). Nous allons maintenant dériver une inégalité de dissipativité pour chaque sous-système.

Inégalité de dissipativité associée à (5.27)

Lemme 1. *Soit x^* tel que $x^* = \pi_X(x^*)$. Pour tout x_k solution de :*

$$\left\{ \begin{array}{lcl} x_{k+1} & = & \pi_X(x_k - w_k) \\ x_k & = & x_k \\ x_0 & = & x_0 \end{array} \right.$$

on a

$$V(x_{k+1}) - V(x_k) + \langle w_k, x_k - x^* \rangle - \tfrac{1}{2}\|w_k\|_2^2 \ \leqslant \ 0 \tag{5.29}$$

où V est définie par :

$$V(x) = \frac{1}{2}\|x - x^*\|_2^2. \tag{5.30}$$

Preuve. Par définition de la fonction V et d'après (5.27) :

$$\begin{array}{rcl} V(x_{k+1}) & = & \frac{1}{2}\|x_{k+1} - x^*\|_2^2 \\ & = & \frac{1}{2}\|\pi_X(x_k - w_k) - \pi_X(x^*)\|_2^2 \end{array}$$

Puisque π_X est une projection, on a [11, Proposition 2.1.3] :

$$\begin{aligned} V(x_{k+1}) &\leqslant \tfrac{1}{2}\|x_k - w_k - x^*\|_2^2 \\ &\leqslant \tfrac{1}{2}\|x_k - x^*\|_2^2 - \langle w_k, x_k - x^*\rangle + \tfrac{1}{2}\|w_k\|_2^2 \end{aligned}$$

soit l'inégalité (5.29). □

Inégalité de dissipativité associée à (5.28)

Lemme 2. *Soit f une fonction convexe. Pour $\gamma_k > 0$,*

$$\frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) - \frac{\gamma_k}{\|g(x_k)\|_2} f^* - \langle w_k, x_k - x^*\rangle \leqslant 0. \tag{5.31}$$

avec $f^* = f(x^*)$ et γ_k défini par (5.28) avec $g(x_k) \in \partial f(x_k)$.

Preuve. La fonction f étant convexe :

$$f(x^*) \geqslant f(x_k) + \langle g(x_k), x^* - x_k\rangle$$

Par suite, puisque $\gamma_k > 0$, avec $f^* = f(x^*)$:

$$\frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) - \frac{\gamma_k}{\|g(x_k)\|_2} f^* - \left\langle \gamma_k \frac{g(x_k)}{\|g(x_k)\|_2}, x_k - x^* \right\rangle \leqslant 0$$

soit l'inégalité de dissipativité (5.31). □

Des lemmes 1 et 2, on en déduit le théorème de convergence ci-dessous.

Théorème 11. *Soit*

- *N un entier strictement positif ;*
- *$\{x_k\}_{k\in\{0,\cdots,N\}}$ solution de :*

$$\begin{cases} x_{k+1} &= \pi_X\left(x_k - \gamma_k \frac{g(x_k)}{\|g(x_k)\|_2}\right) \\ x_0 &= x_0 \end{cases} ;$$

- *$R > 0$ telle que $\|x - x_0\|_2 \leqslant R$, $\forall x \in X$;*
- *$L > 0$ une borne sur la constante de Lipschitz de f.*

Alors la borne

$$f(\hat{x}_N) - f^* \leqslant \frac{L}{2} \frac{\left(R^2 + \frac{1}{2}\sum_{k=0}^{N} \gamma_k^2\right)}{\sum_{k=0}^{N} \gamma_k} \tag{5.32}$$

est satisfaite pour

$$\hat{x}_N := \underset{\{x_k\}_{k\in\{0,\cdots,N\}}}{\operatorname{argmin}} \; f(x_k)$$

et pour

$$\hat{x}_N = \frac{1}{\sum_{k=0}^{N} \gamma_k} \sum_{k=0}^{N} \gamma_k x_k.$$

Preuve. En additionnant (5.29) et (5.31), on obtient : pour tout entier naturel k

$$V(x_{k+1}) - V(x_k) + \frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) - \frac{\gamma_k}{\|g(x_k)\|_2} f^* - \frac{1}{2}\|w_k\|_2^2 \leqslant 0 \tag{5.33}$$

Or $\|w_k\|_*^2 = \gamma_k^2$. En sommant pour k allant de 0 à N les inégalités (5.33) et en simplifiant, on obtient :

$$V(x_{N+1}) - V(x_0) + \sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) - \sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2} f^* - \frac{1}{2}\sum_{k=0}^{N} \gamma_k^2 \leqslant 0 \tag{5.34}$$

Puisque par définition de V, voir (5.30), $V(x_{N+1}) \geqslant 0$, l'inégalité (5.34) implique :

$$\sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) - \sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2} f^* \leqslant V(x_0) + \frac{1}{2}\sum_{k=0}^{N} \gamma_k^2 \tag{5.35}$$

D'une part, avec

$$\hat{x}_N := \underset{\{x_k\}_{k\in\{0,\cdots,N\}}}{\operatorname{argmin}} f(x_k),$$

on a :

$$\sum_{k=1}^{k=N} \frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) \geqslant \left(\sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2}\right) f(\hat{x}_N).$$

Puisque $\|g(x_k)\|_2 \leqslant L$,

$$\sum_{k=0}^{k=N} \frac{\gamma_k}{\|g(x_k)\|_2} \geqslant \tfrac{1}{L}\left(\sum_{k=0}^{N} \gamma_k\right)$$

on obtient :

$$f(\hat{x}_N) - f^* \leqslant \frac{L}{\sum_{k=0}^{N} \gamma_k}\left(V(x_0) + \frac{1}{2}\sum_{k=0}^{N} \gamma_k^2\right) \tag{5.36}$$

D'autre part, avec

$$\hat{x}_N = \frac{1}{\sum_{k=0}^{N} \gamma_k} \sum_{k=0}^{N} \gamma_k x_k$$

on a dans le terme de gauche de l'inégalité (5.35) qui donne :

$$\begin{aligned}\sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2} f(x_k) - \sum_{k=0}^{N} \frac{\gamma_k}{\|g(x_k)\|_2} f^* &\geqslant \tfrac{1}{L}\sum_{k=0}^{N} \gamma_k f(x_k) - \sum_{k=0}^{N} \gamma_k f^* \\ &\geqslant \frac{\sum_{k=0}^{N} \gamma_k}{L}\left(\frac{\sum_{k=0}^{N} \gamma_k f(x_k)}{\sum_{k=0}^{N} \gamma_k} - f^*\right)\end{aligned}$$

Comme par convexité de f :

$$\frac{\sum_{k=0}^{N}\gamma_k f(x_k)}{\sum_{k=0}^{N}\gamma_k} \geqslant f\left(\frac{1}{\sum_{k=0}^{N}\gamma_k}\sum_{k=0}^{N}\gamma_k x_k\right)$$

on obtient aussi (5.36).
En notant que, d'après l'algorithme du sous-gradient projeté, $V(x_0) \leqslant \frac{R^2}{2}$, on obtient à partir de (5.36) l'inégalité (5.32).

□

5.3.3 Primal-Dual Subgradient Algorithm

Soit la fonction

$$\Pi_\beta(\zeta) = \underset{x\in X}{\operatorname{argmin}}\{-\langle\zeta, x\rangle + \beta\omega(x)\}.$$

Alors l'algorithme PDA (Primal-Dual subgradient Algorithm) peut s'exprimer comme un système en temps discret défini par la représentation d'état :

$$\begin{cases} \zeta_{k+1} &= \zeta_k + \lambda_k g(x_k) \\ x_k &= \Pi_{\beta_k}(-\zeta_k) \\ \zeta_0 &= \lambda_0 g(x^0) \end{cases}$$

avec ζ_k le vecteur d'état, x_k la sortie du système et $g(x_k) \in \partial f(x_k)$.

Ce système peut se récrire comme l'interconnexion de deux sous-systèmes :

$$\begin{cases} \zeta_{k+1} &= \zeta_k + w_k \\ x_k &= \Pi_{\beta_k}(-\zeta_k) \end{cases} \tag{5.37}$$

et

$$w_k = \lambda_k g(x_k). \tag{5.38}$$

Le premier sous-système correspond à un intégrateur ($\zeta_{k+1} = \zeta_k + w_k$) suivi d'une non-linéarité en sortie ($\Pi_\beta(-\zeta_k)$) ; le second sous-système est une non-linéarité statique ($\lambda_k g(x_k)$).

Remarque : Notons que pour $\omega(x) = \frac{1}{2}\|x\|_2^2$ et $\beta = 1$, $\Pi_\beta = \frac{1}{2}\pi_X$ avec π_X défini sous-section 5.3.2. Par suite, dans ce cas-là, la différence fondamentale entre cet algorithme et l'algorithme du gradient projeté est la présence du terme intégrateur en amont de l'opération de projection. Il est bien connu en automatique classique que la présence d'un intégrateur dans une interconnexion permet de réaliser le rejet de perturbations « basses fréquences »se produisant en aval de l'intégrateur et en amont du signal d'intérêt. Ici le signal d'intérêt est x_k : l'introduction de l'intégrateur permet donc de limiter les effets d'erreurs numériques lors du calcul de la projection (perturbations « basses fréquences »).

Nous allons maintenant dériver une inégalité de dissipativité pour chaque sous-système.

Inégalité de dissipativité associée à (5.37)

Lemme 3. *Soit x^* tel que $f(x^*) = f^*$. Pour tout ζ_k solution de :*

$$\left\{ \begin{array}{lcl} \zeta_{k+1} & = & \zeta_k + w_k \\ x_k & = & \Pi_{\beta_k}(-\zeta_k) \end{array} \right.$$

on a

$$V_{\beta_{k+1}}(-\zeta_{k+1}) - V_{\beta_k}(-\zeta_k) + \langle w_k, x_k - x^* \rangle - \tfrac{1}{2\sigma\beta_k}\|w_k\|_*^2 \;\leqslant\; 0 \tag{5.39}$$

où V_β est définie par :

$$V_\beta(\zeta) = \max_{x \in X}\{\langle \zeta, x - x^* \rangle - \beta\omega(x)\}. \tag{5.40}$$

Preuve. La fonction V_β a les propriétés suivantes [50] :

Propriété 1

$$\forall \beta_2 \geqslant \beta_1 > 0, \;\; \forall \zeta, \;\; V_{\beta_2}(\zeta) \leqslant V_{\beta_1}(\zeta) \tag{5.41}$$

Propriété 2

$$\forall \zeta, \nabla V_\beta(\zeta) = \Pi_\beta(\zeta) - x^* \tag{5.42}$$

Propriété 3

$$\forall \zeta, \forall \delta, V_\beta(\zeta + \delta) \leqslant V_\beta(\zeta) + \langle \delta, \nabla V_\beta(\zeta) \rangle + \frac{1}{2\sigma\beta}\|\delta\|_*^2 \tag{5.43}$$

Propriété 4

$$\forall \zeta, V_\beta(\zeta) \leqslant \frac{1}{2\sigma\beta}\|\zeta\|_*^2 \tag{5.44}$$

Propriété 5

$$\forall \zeta, V_\beta(\zeta) \geqslant -\beta\omega(x^*) \tag{5.45}$$

D'après l'équation (5.37) :

$$V_{\beta_{k+1}}(-\zeta_{k+1}) \;=\; V_{\beta_{k+1}}(-(\zeta_k + w_k)).$$

Puisque $\beta_{k+1} \geqslant \beta_k$, d'après (5.41) :

$$V_{\beta_{k+1}}(-\zeta_{k+1}) \;\leqslant\; V_{\beta_k}(-(\zeta_k + w_k)).$$

D'après (5.43) :

$$V_{\beta_{k+1}}(-\zeta_{k+1}) \;\leqslant\; V_{\beta_k}(-\zeta_k) + \langle -w_k, \nabla V_{\beta_k}(-\zeta_k) \rangle + \tfrac{1}{2\sigma\beta_k}\|w_k\|_*^2$$

ce qui donne, d'après (5.42) :

$$V_{\beta_{k+1}}(-\zeta_{k+1}) \;\leqslant\; V_{\beta_k}(-\zeta_k) - \langle w_k, x_k - x^* \rangle + \tfrac{1}{2\sigma\beta_k}\|w_k\|_*^2$$

soit l'inéquation de dissipativité (5.39). □

Inégalité de dissipativité associée à (5.38)

Lemme 4. *Soit f une fonction convexe. Pour $\lambda_k \geqslant 0$,*

$$\lambda_k f(x_k) - \lambda_k f^* - \langle w_k, x_k - x^* \rangle \leqslant 0 \tag{5.46}$$

avec $f^* = f(x^*)$ et $g(x_k) \in \partial f(x_k)$.

Preuve. La fonction f étant convexe :

$$f(x^*) \geqslant f(x_k) + \langle g(x_k), x^* - x_k \rangle$$

Par suite, puisque $\lambda_k \geqslant 0$, avec $f^* = f(x^*)$

$$\lambda_k f(x_k) - \lambda_k f^* - \langle \lambda_k g(x_k), x_k - x^* \rangle \leqslant 0$$

soit l'inéquation de dissipativité (5.46).

□

Des lemmes 3 et 4, on en déduit le théorème de convergence ci-dessous.

Théorème 12. *Soit*

- *N un entier strictement positif ;*
- *$\{x_k\}_{k\in\{0,\cdots,N\}}$ solution de :*

$$\left\{ \begin{array}{lcl} \zeta_{k+1} & = & \zeta_k + \lambda_k g(x_k) \\ x_k & = & \Pi_{\beta_k}(-\zeta_k) \\ \zeta_0 & = & \lambda_0 g(x^0) \end{array} \right.$$

- *$R > 0$ telle que $\omega(x) \leqslant R$, $\forall x \in X$;*
- *$L > 0$ une borne sur la constante de Lipschitz de f*

Alors la borne

$$f(\hat{x}_N) - f^* \leqslant \frac{1}{\sum_{k=1}^{N} \lambda_k} \left(\beta_{N+1} R + \frac{L^2}{2\sigma} \sum_{k=0}^{N} \frac{\lambda_k^2}{\beta_k} \right) \tag{5.47}$$

est satisfaite pour

$$\hat{x}_N := \operatorname{argmin} \{x_k\}_{k\in\{0,\cdots,N\}} f(x_k)$$

et pour

$$\hat{x}_N = \frac{1}{\sum_{k=0}^{N} \lambda_k} \sum_{k=0}^{N} \lambda_k x_k.$$

Preuve. En additionnant (5.39) et (5.46), on obtient : pour tout entier naturel k

$$V_{\beta_{k+1}}(-\zeta_{k+1}) - V_{\beta_k}(-\zeta_k) + \lambda_k f(x_k) - \lambda_k f^* - \tfrac{1}{2\sigma\beta_k} \|w_k\|_*^2 \leqslant 0 \tag{5.48}$$

En sommant pour k allant de 0 à N les inégalités (5.53) et en simplifiant, on obtient :

$$V_{\beta_{k+1}}(-\zeta_{k+1}) - V_{\beta_0}(-\zeta_0) + \sum_{k=1}^{N} \lambda_k f(x_k) - \sum_{k=1}^{N} \lambda_k f^* - \sum_{k=1}^{N} \frac{\lambda_k^2}{2\sigma\beta_k} \|g(x_k)\|_*^2 \leqslant 0 \tag{5.49}$$

D'après (5.37) et (5.38) et avec $\zeta_0 = 0$, $\zeta_1 = \lambda_0 g(x_0)$. De plus, d'après l'inégalité (5.44)

$$V_{\beta_0}(-\zeta_0) \leqslant 0.$$

Ainsi puisqu'on a l'inégalité (5.45), (5.49) implique :

$$\frac{\sum_{k=0}^{N} \lambda_k f(x_k)}{\sum_{k=0}^{N} \lambda_k} - f^* \leqslant \frac{1}{\sum_{k=0}^{N} \lambda_k} \left(\beta_{N+1}\omega(x^*) + \sum_{k=0}^{N} \frac{\lambda_k^2}{2\sigma\beta_k} \|g(x_k)\|_*^2 \right)$$

D'une part, dans le cas où :

$$\hat{x}_N = \frac{1}{\sum_{k=0}^{N} \lambda_k} \sum_{k=0}^{N} \lambda_k x_k.$$

comme par convexité de f :

$$\frac{\sum_{k=0}^{N} \lambda_k f(x_k)}{\sum_{k=0}^{N} \lambda_k} \geqslant f\left(\frac{1}{\sum_{k=0}^{N} \lambda_k} \sum_{k=0}^{N} \lambda_k x_k \right)$$

on obtient :

$$f(\hat{x}_N) - f^* \leqslant \frac{1}{\sum_{k=0}^{N} \lambda_k} \left(\beta_{N+1}\omega(x^*) + \sum_{k=0}^{N} \frac{\lambda_k^2}{2\sigma\beta_k} \|g(x_k)\|_*^2 \right) \tag{5.50}$$

D'autre part, dans le cas où :

$$\hat{x}_N := \operatorname{argmin} \{x_k\}_{k\in\{0,\cdots,N\}} f(x_k)$$

en notant que :

$$\frac{\sum_{k=0}^{N} \lambda_k f(x_k)}{\sum_{k=0}^{N} \lambda_k} \geqslant f(\hat{x}_N)$$

on obtient l'inégalité (5.50). Comme $L > 0$ est une borne sur la constante de Lipschitz de f, on a pour tout k $\|g(x_k)\|_*^2 \leqslant L^2$. Comme $\omega(x^*) \leqslant R$, on obtient l'inégalité (5.58). □

5.3.4 Mirror Descent Algorithm

Soit la fonction [3] ω_x définie par :

$$\omega_x(y) = \omega(y) - \omega(x) - \langle y - x, \omega'(x) \rangle$$

3. Cette fonction est appelée « distance de Bregman » [16] ou encore « prox-function ».

où ω est une fonction fortement convexe de coefficient α et continument différentiable sur l'intérieur de X. Alors l'algorithme MDA (Mirror Descent Algorithm) peut s'exprimer comme un système en temps discret défini par la représentation d'état :

$$\left\{ \begin{array}{lcl} x_{k+1} & = & \operatorname{argmin}_{x\in X}\{\omega_{x_k}(x) + \langle \lambda_k g(x_k), x - x_k\rangle\} \\ x_0 & = & x^0 \end{array} \right.$$

avec x_k le vecteur d'état et $g(x_k) \in \partial f(x_k)$. Ce système peut se récrire comme l'interconnexion de deux sous-systèmes :

$$x_{k+1} = \operatorname{argmin}_{x\in X}\{\omega_{x_k}(x) + \langle w_k, x - x_k\rangle\} \tag{5.51}$$

et

$$w_k = \lambda_k g(x_k). \tag{5.52}$$

Inégalité de dissipativité associée à (5.51)

Lemme 5. *Soit x^* le point de convergence. Pour tout x_k solution de :*

$$\left\{ \begin{array}{lcl} x_{k+1} & = & \operatorname{argmin}_{x\in X}\{\omega_{x_k}(x) + \langle \lambda_k g(x_k), x - x_k\rangle\} \\ x_0 & = & x^0 \end{array} \right.$$

on a

$$V(x_{k+1}) - V(x_k) + \langle w_k, x_k - x^*\rangle - \tfrac{1}{2\alpha}\|w_k\|_*^2 \leqslant 0 \tag{5.53}$$

où V est définie par :

$$V(x) = \omega_x(x^*). \tag{5.54}$$

Remarque : Nous avons noté que pour la fonction génératrice de distance $\omega(x) = \frac{1}{2}\|x\|_2^2$, on obtient $\omega_x(y) = \|y - x\|_2^2$. Dans ce cas là, le choix de V défini par (5.54) correspond au V effectué pour le gradient projeté (5.30), page 85. Ce lemme est donc une généralisation du lemme 1, page 85.

Preuve. Pour établir l'inégalité, le terme ci-dessous est décomposé en trois :

$$\langle x_k - x^*, w_k\rangle = s_1 + s_2 + s_3$$

avec

$$\begin{array}{lcl} s_1 & = & \langle x^* - x_{k+1}, \nabla\omega(x_k) - \nabla\omega(x_{k+1}) - w_k\rangle \\ s_2 & = & -\langle x^* - x_{k+1}, \nabla\omega(x_k) - \nabla\omega(x_{k+1})\rangle \\ s_3 & = & -\langle x_k - x_{k+1}, -w_k\rangle. \end{array}$$

Etudions chacun des termes s_1, s_2 et s_3.

Etude de s_1 Puisque $x_{k+1} = \operatorname{argmin}_{x\in X}\{\omega_{x_k}(x) + \langle w_k, x - x_k\rangle\}$,

$$\forall x, \quad \langle x - x_{k+1}, -\nabla\omega(x_k) + \nabla\omega(x_{k+1}) + w_k\rangle \geqslant 0.$$

Ce qui est vrai en particulier pour $x = x^*$:

$$\langle x - x_{k+1}, -\nabla\omega(x_k) + \nabla\omega(x_{k+1}) + w_k\rangle \geqslant 0.$$

Par suite

$$s_1 \leqslant 0. \tag{5.55}$$

Etude de s_2 Elle se base sur le lemme ci-dessous.

Lemme 6 ([19])**.** *Soit X un sous ensemble ouvert de $\mathbb{R}^n$, d'intérieur $\mathring{X}$ et ω une fonction de $\mathring{X}$ dans $\mathbb{R}$ continuement différentiable sur X. Alors pour tout a, $b \in X$ et $c \in \mathring{X}$, on a :*

$$\omega_a(c) + \omega_b(a) - \omega_b(c) = \langle \nabla\omega(b) - \nabla\omega(a), c - a\rangle$$

Par application de ce lemme, on obtient :

$$s_2 = \omega_{x_k}(x^*) - \omega_{x_{k+1}}(x^*) - \omega_{x_k}(x_{k+1}). \tag{5.56}$$

Etude de s_3 En appliquant à s_3 la propriété élémentaire :

$$\langle \nabla a, b\rangle \leqslant \frac{\alpha}{2}\|a\|^2 + \frac{1}{2\alpha}\|b\|^2$$

on obtient :

$$s_3 \leqslant \frac{\alpha}{2}\|x_k - x_{k+1}\|^2 + \frac{1}{2\alpha}\|w_k\|^2$$

Comme ω est fortement convexe de coefficient α :

$$\omega_{x_k}(x_{k+1}) \geqslant \frac{\alpha}{2}\|x_k - x_{k+1}\|^2$$

ce qui donne :

$$s_3 \leqslant \omega_{x_k}(x_{k+1}) + \frac{1}{2\alpha}\|w_k\|^2. \tag{5.57}$$

D'après (5.55), (5.56) et(5.57) :

$$\begin{aligned}\langle x_k - x^*, w_k\rangle &= s_2 + s_3 \\ &\leqslant \omega_{x_k}(x^*) - \omega_{x_{k+1}}(x^*) + \tfrac{1}{2\alpha}\|w_k\|^2,\end{aligned}$$

ce qui correspond à l'inégalité (5.53). □

Le cas traité dans cette section est très similaire au cas traité dans la section 5.3.3 :
- les inégalités de dissipativité satisfaites par les premiers sous-systèmes sont similaires avec dans l'inégalité (5.39) le terme $\frac{1}{\sigma\beta_k}$ qui devient $\frac{1}{\alpha}$ dans l'inégalité (5.53) ;
- les seconds sous-systèmes sont identiques.

Par suite, le théorème suivant découle directement du théorème 12.

Théorème 13. *Soit*
- *N un entier strictement positif ;*
- *$\{x_k\}_{k\in\{0,\cdots,N\}}$ solution de :*

$$\begin{cases} x_{k+1} &= \underset{x\in X}{\operatorname{argmin}}\{\omega_{x_k}(x) + \langle \lambda_k g(x_k), x - x_k\rangle\} \\ x_0 &= x^0 \end{cases}$$

- *$R > 0$ telle que $\omega_x(y) \leqslant R$, $\forall x \in X$, $\forall y \in X$;*

— $L > 0$ *une borne sur la constante de Lipschitz de* f

Alors la borne

$$f(\hat{x}_N) - f^* \leqslant \frac{1}{\sum_{k=0}^{N} \lambda_k} \left(R + \frac{L^2}{2\alpha} \sum_{k=0}^{N} \lambda_k^2 \right) \tag{5.58}$$

est satisfaite pour

$$\hat{x}_N := \underset{\{x_k\}_{k\in\{0,\cdots,N\}}}{\operatorname{argmin}} \quad f(x_k)$$

et pour

$$\hat{x}_N = \frac{1}{\sum_{k=0}^{N} \lambda_k} \sum_{k=0}^{N} \lambda_k x_k.$$

Remarque 1 : le fait de choisir une sortie

$$\hat{x}_N := \underset{\{x_k\}_{k\in\{0,\cdots,N\}}}{\operatorname{argmin}} \quad f(x_k)$$

ou bien

$$\hat{x}_N = \frac{1}{\sum_{k=0}^{N} \gamma_k} \sum_{k=0}^{N} \gamma_k x_k,$$

n'a aucune influence sur la complexité algorithmique des trois méthodes car pour les deux sorties, nous avons obtenu la même borne de convergence pour chaque méthode. Néanmoins, du point de vue pratique la première sortie permet de choisir le point garantissant la meilleure précision obtenue sur la totalité des itérations. Nous allons voir dans le cas stochastique que cette sortie devient malheureusement inaccessible.

Remarque 2 : concernant la deuxième sortie, l'idée est de calculer une somme pondérée sur l'ensemble ou un sous-ensemble seulement des points engendrés par les trois algorithmes. Nous avons choisi dans notre présentation de prendre la somme pondérée sur l'ensemble des points. Rien n'empêche de considérer uniquement un sous-ensemble c'est-à-dire de calculer la sortie

$$\hat{x}_N = \frac{1}{\sum_{k=i}^{N} \gamma_k} \sum_{k=i}^{N} \gamma_k x_k,$$

avec $1 \leqslant i \leqslant N$. Le choix de i implique un certain niveau de précision de la solution numérique. Le choix de i est purement empirique et relève du niveau d'expertise et de compréhension que l'on a accumulé sur le problème d'optimisation à résoudre.

5.4 Conclusion

Nous avons présenté les hypothèses et le cadre général des méthodes du premier ordre (MPO) [47]. Nous avons présenté la méthode du sous-gradient projeté, ainsi que son efficacité d'estimation et nous avons expliqué en quoi cette efficacité est optimale. On a présenté également des généralisations de cette méthode, on a souligné leurs principales motivations et on a comparé leur efficacité par rapport à la méthode du sous-gradient projeté. On a vu comment ces méthodes du premier ordre donnent des degrés de liberté dans le sens où elles permettent de prendre en compte la nature géométrique de l'ensemble faisable X afin d'améliorer leur efficacité d'estimation (voir paragraphe 5.2.2.1). On a conclu qu'un bon choix de la configuration (norme et fonction génératrice de la distance) de ces méthodes peut nettement améliorer leur efficacité d'estimation. On a aussi vu que les méthodes présentées ici sont optimales au sens du théorème (7) : le taux de convergence des MPO est de l'ordre de $O(1/\sqrt{N})$ et ne peut pas être amélioré dans le cas des hypothèses 2 page 71. Nous avons pu également étudier l'analyse de convergence en se basant sur une interprétation systémique des MPO : en effet, le fait d'interpréter les MPO comme étant des systèmes dynamiques non linéaires nous a permis d'exploiter les outils d'automatique pour l'analyse de la stabilité afin de redémontrer la convergence des ces MPO. Ceci a l'avantage d'organiser l'analyse de convergence des MPO diversifiée et variée au sein d'un cadre unique. Il est clair qu'il n'existe pas de méthodes numériques universelles capables de résoudre efficacement tout problème d'optimisation (voir [48] pour plus de discussions). Selon la nature du problème en question, on a vu comment il est possible d'adapter les MPO afin d'accélérer le calcul des solutions et en particulier dans notre cas où n est très grand. La littérature sur les méthodes MPO est assez large et bénéficie de développements récents massifs grâce à la popularité montante de l'optimisation convexe grande échelle et vu leur application intensive dans les applications ayant un aspect « big data ». On cite à titre d'exemple [6, 5, 7, 69].

Dans le prochain chapitre, on va expliquer comment exploiter ces méthodes au sein d'un processus itératif « stochastique » afin de résoudre des problèmes d'optimisation, dits stochastiques, avec une bonne precision et une grande marge de confiance.

5.5 Annexe du chapitre

Rappels de quelques notions en optimisation convexe

Dans cette partie, on va rappeler brièvement les notions standard d'optimisation convexe nécessaires à la compréhension de ce chapitre. On considère l'espace vectoriel $\mathbb{R}^n$ muni du produit scalaire $\langle .,. \rangle$ et d'une norme $\|.\|$. La norme duale de $\|.\|$ est notée $\|.\|_*$ et définie par $\|y\|_* := \max_{\|x\| \leqslant 1} \langle y, x \rangle$.

Définition 7. *(Ensemble convexe) X est un ensemble convexe si pour tout $x, y \in X$ et pour tout $\alpha \in [0,1]$, $\alpha x + (1-\alpha)y \in X$.*

Définition 8. *(Domaine d'une fonction) Soit $f : X \to \mathbb{R}$, le domaine de f noté $domf$ est défini par $domf = \{x \in X | f(x) \neq \pm\infty\}$.*

Définition 9. *(Fonction convexe) Soit $f : X \to \mathbb{R}$ une fonction, f est dite convexe si $domf$ est convexe et si pour tout $\alpha \in [0,1]$ et pour tout $x, y \in domf$: $f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y)$.*

Définition 10. *(Fonction Lipschitz) Soit $f : X \to \mathbb{R}$ une fonction, f est $L-$Lipschitz sur X par rapport à la norme $\|.\|$, si*

$$|f(x) - f(y)| \leqslant L\|x - y\| \quad \forall x, y \in X.$$

Définition 11. *(Sous-différentiel et sous-gradient) Soit $f : X \to \mathbb{R}$ une fonction convexe. Le sous-différentiel de f en x, s'il existe, est défini par :*

$$\partial f(x) := \{\eta \in X, f(y) \geqslant f(x) + \langle \eta, y - x \rangle, \forall y \in X\}.$$

Les éléments de $\partial f(x)$ sont appelés les sous-gradients de f en x.

Définition 12. *(Problème d'optimisation convexe) Le problème d'optimisation*

$$f^* := \min_{x \in X} f(x) \tag{5.59}$$

est convexe si la fonction f est convexe sur X et X est convexe. Un point x est un point faisable pour ce problème d'optimisation si $x \in X$. $f^* = f(x^*)$ est la valeur optimale, avec x^* est une solution optimale de (5.59).

Sauf mention contraire, on se limite au cas où la fonction objectif f est non différentiable car cela est typiquement le cas pour nos applications basées sur la méthode RBA qui seront présentées au fur et à mesure. Il faut noter que les définitions qui suivent sont posées par rapport au problème d'optimisation (5.59).

Définition 13. *(solution $\epsilon-$ précise) Une solution x est dite $\epsilon-$ précise du problème (5.59), si $x \in X$ et $f(x) - f^* \leqslant \epsilon$, où f^* est la valeur optimale de (5.59).*

Définition 14. *(Oracle) Un oracle est une procédure numérique qui en un point donné, renvoie une réponse (une sortie).*
Un oracle du premier ordre est un processus numérique qui en un point x, renvoie les quantités $f(x)$ et/ou $f'(x)$, où $f'(x)$ est un gradient ou un sous-gradient de f en x.

Définition 15. *(Méthode numérique) Une méthode de résolution numérique pour ce problème, est un processus itératif qui à partir d'un point initial $x_0 \in X$, engendre une suite de points $(x_k)_k$ en collectant des informations spécifiques sur le problème d'optimisation via un oracle et en appliquant un ensemble de règles spécifique à la méthode.*
Une méthode de premier ordre est une méthode numérique basée sur un oracle de premier ordre.

Définition 16. *(Méthode numérique convergente) Une méthode numérique est dite convergente si* $\lim_k f(x_k) - f^* = 0$.

Définition 17. *(Taux de convergence d'une méthode numérique) Une méthode numérique convergente à un taux de convergence $r(k)$ si*

$$f(x_k) - f^* \leqslant r(k),$$

et $\lim_k r(k) = 0$. *On peut dire aussi que l'efficacité d'estimation de la méthode numérique est de $r(k)$.*

Définition 18. *(Projection euclidienne) Soit $x \in \mathbb{R}^n$, la projection euclidienne de x sur X est définie par*

$$\pi_X(x) = \operatorname*{argmin}_{y \in X} \|x - y\|_2.$$

Chapitre 6

Méthodes pour l'optimisation stochastique et mise en œuvre

Dans ce chapitre, nous allons considérer la classe des problèmes d'optimisation dite stochastique définie par le problème

$$\min_{x\in X}\{f(x) := \mathbb{E}_\xi[F(x,\xi)]\}, \tag{6.1}$$

où ξ est un vecteur aléatoire à valeur dans $\Xi \subset \mathbb{R}^d$, $d \in \mathcal{N}$ et Ξ est le support du vecteur aléatoire ξ (c'est-à-dire le plus petit ensemble vérifiant $\mathbb{P}_\xi(\Xi) = 1$) où $\mathbb{P}_\xi$ est la densité de probabilité, continue, de ξ, F est une fonction convexe de $X \times \Xi$ dans $\mathbb{R}$ et $X \subset \mathbb{R}^n$ est un ensemble convexe de $\mathbb{R}^n$. La fonction objectif est donnée par

$$f(x) := \int_\Xi F(x,\xi)\mathbb{P}_\xi(\xi)\mathrm{d}\xi. \tag{6.2}$$

Cette classe recouvre les problèmes d'optimisation stochatiques présentés dans le chapitre 4. L'objectif de ce chapitre est de présenter des algorithmes de résolution, de les évaluer sur la classe de réseau définie dans le chapitre 4 et si nécessaire de proposer des améliorations pour résoudre les problèmes qui nous intéressent.

6.1 Méthodes pour l'optimisation stochastique

Nous présentons deux approches permettant de résoudre le problème (6.1).

i) La méthode dite *Stochastic Approximation* (SA) (voir [46, 64, 38]). C'est une méthode numérique basée sur des méthodes déterministes du premier ordre au sein desquelles on introduit à chaque itération des réalisations de la variable aléatoire ξ au niveau de l'oracle du premier ordre. Cette version stochastique doit ensuite vérifier une certaine propriété de convergence moyenne. Nous allons voir dans la suite que toutes les méthodes présentées dans le chapitre 5 admettent des versions stochastiques possédant une propriété de convergence en moyenne que nous expliciterons dans la suite dans le cas de chaque méthode. L'intérêt principal de cette méthode est qu'elle permet d'intégrer des méthodes de résolution déterministes

dans un cadre algorithmique stochastique permettant de résoudre (6.1).

ii) L'approche dite *Sample Average Approximation* (SAA) (voir [64, 63, 62, 61]). C'est une approche visant à approcher la fonction coût $\mathbb{E}_\xi[F(x,\xi)]$ du problème (6.1) par la moyenne empirique

$$\hat{f}_M(x) = \frac{1}{M}\sum_{i=1}^{M} F(x,\xi^i)$$

de $F(x,\xi)$ en considérant un certain nombre M de réalisations ξ^i du vecteur aléatoire ξ. Ainsi on se ramène à un problème d'optimisation déterministe. L'approche propose une méthode pour calculer le nombre M tel que la solution du problème SAA soit une solution du problème initial avec une bonne précision et une grande marge de confiance. L'intérêt de cette approche est qu'elle permet de faire passer la résolution du problème stochastique (6.1) par la résolution d'un problème d'optimisation déterministe que nous présentons explicitement dans la suite.

Nous précisons que contrairement à la méthode SA, la méthode SAA n'est pas une méthode numérique : elle permet de construire un problème d'optimisation (modèle) « approché » du problème (6.1). Encore faut-il alors choisir une méthode numérique permettant la résolution efficace du problème SAA. Nous allons expliquer dans la suite ce que « approché » signifie.

Finalement, après avoir présenté les deux méthodes SA et SAA, nous finirons par comparer leur complexité algorithmique afin de ne retenir que la plus efficace.

6.1.1 *Stochastic Approximation* (SA)

6.1.1.1 Cadre général

Essentiellement, la méthode SA est un processus stochastique basé sur le processus déterministe associé à l'algorithme 2, présenté page 72. Il s'agit de générer une séquence finie de points suivant une certaine règle tout en collectant des informations sur le problème et sur sa nature stochastique en faisant des tirages sur le vecteur aléatoire ξ au fur et à mesure des itérations. Ce processus stochastique est présenté dans l'algorithme 6. On note ici que

Algorithme 6 Processus stochastique

Initialisation : $\mathcal{I}_0 = \emptyset$.
Tant que $(0 \leqslant k \leqslant N-1)$:
prendre $k = k+1$;
tirer ξ^{k-1} ;
mettre $\mathcal{I}_k = \mathcal{I}_{k-1} \cup \mathcal{O}(F(.,\xi^{k-1}), x_{k-1})$;
calculer x_k sur la base de $\mathcal{I}_k$.
Fin de tant que
Renvoyer $\hat{x}_N$ solution $\epsilon-$ précise sur la base de $\mathcal{I}_N$.

l'oracle $\mathcal{O}$ est un oracle « stochastique » du premier ordre qui, après avoir effectué une réalisation de ξ, calcule en un point donné x un sous-gradient de $F(x,\xi)$.

Remarque : Etant donnée une réalisation de ξ et un point x, la sortie de l'oracle $\mathcal{O}$ est certes déterministe. Toutefois, cet oracle est appelé stochastique car, contrairement à un oracle déterministe, le calcul d'une sortie tient compte, implicitement, de la densité de probabilité de ξ car la réalisation de ξ est issue d'un tirage au sort.

Certes, ce processus n'a de sens que s'il est capable de nous garantir au bout d'un certain nombre d'itérations N, une « solution » au problème (6.1) avec au moins la précision désirée. Mais la question est : comment une solution au problème (6.1) peut être définie ? Nous pouvons immédiatement penser à définir cette solution d'une manière similaire à la solution d'un problème d'optimisation déterministe c'est-à-dire :

Définition 19. *$\hat{x}_N$ est solution de (6.1) avec au moins une précision ϵ si*

$$f(\hat{x}_N) - f^* \leqslant \epsilon. \tag{6.3}$$

Cette définition présente deux difficultés. La première difficulté est calculatoire car il s'agit d'évaluer, pour calculer l'espérance, l'intégrale multidimensionnelle (6.2) et cela rend l'évaluation de $f(\hat{x}_N)$ difficile voire pratiquement impossible à partir de la dimension 4 ou 5. La deuxième difficulté est que $\hat{x}_N$ est issue du processus stochastique associé à l'algorithme 6, dépend donc des réalisations du vecteur aléatoire ξ et, par conséquent, la sortie de l'algorithme 6 est également un vecteur aléatoire dont $\hat{x}_N$ n'est qu'une réalisation particulière. Ainsi, l'inégalité (6.3) devient un événement avec une probabilité d'occurrence.

Ainsi, vu la nature stochastique de $\hat{x}_N$, nous devons poser des définitions basées sur des critères probabilistes. Nous pouvons exiger de $\hat{x}_N$ qu'elle soit précise en moyenne c'est-à-dire

Définition 20. *$\hat{x}_N$ est une solution avec au moins une précision moyenne ϵ ou $\epsilon-$précise en moyenne si*

$$\mathbb{E}[f(\hat{x}_N) - f^*] \leqslant \epsilon. \tag{6.4}$$

Cette définition présente la même difficulté calculatoire que la définition 19 car il faut calculer deux fois une espérance mathématique (une première fois par rapport au vecteur aléatoire ξ et une deuxième fois par rapport au vecteur aléatoire $\hat{x}_N$). Néanmoins la difficulté peut être contournée. En effet si l'algorithme 6 vérifie l'hypothèse ci-dessous alors la suite $\mathbb{E}[f(\hat{x}_N) - f^*]$ converge vers zéro.

Hypothèse 3. *[convergence moyenne] il existe une fonction numérique ϕ telle que*

$$\mathbb{E}[f(\hat{x}_N) - f^*] \leqslant \phi(N), \tag{6.5}$$

où $\phi(N) \to 0$ quand $N \to \infty$.

Ainsi pour tout $\epsilon > 0$ il existe un N_ϵ à partir duquel $\hat{x}_N$ est $\epsilon-$précise. Pour spécifier N_ϵ, il suffit de le prendre tel que $N_\epsilon = \phi^{-1}(\epsilon)$. Comme nous allons le voir, les méthodes déterministes vues au chapitres 3, intégrées au sein du processus SA associé à l'algorithme 6 vérifient l'hypothèse 3 de convergence moyenne.

Une autre définition basée sur des critères probabilistes est la définition suivante.

Définition 21. *[53] $\hat{x}_N$ est une solution avec au moins une précision moyenne ϵ avec une marge de confiance d'au moins β ou solution $(\epsilon, \beta)-$précise si*

$$Prob\{f(\hat{x}_N) - f^* \geqslant \epsilon\} \leqslant 1 - \beta. \tag{6.6}$$

L'intérêt de la définition 21 est qu'elle permet de prendre en compte la nature de la solution $\hat{x}_N$ qui ne satisfait l'inégalité (6.3) qu'avec une certaine probabilité. La vérification de cette définition est très difficile même si la distribution de ξ est connue car il s'agit d'évaluer en un point une fonction définie par une intégrale multidimensionnelle. Toutefois cette difficulté peut être palliée en vérifiant le critère (6.6) indirectement. Il existe différentes façons de le vérifier indirectement suivant les conditions qu'on prend. Nous allons donner ici un exemple. Il s'agit de l'application directe de l'inégalité de Markov [68]. En effet, sous l'hypothèse 3, nous avons :

$$\text{Prob}\{f(\hat{x}_N) - f^* \geqslant \epsilon\} \leqslant \epsilon^{-1}\mathbb{E}[f(\hat{x}_N) - f^*] \leqslant \epsilon^{-1}\phi(N). \tag{6.7}$$

Ainsi pour en déduire le nombre d'itérations minimal pour garantir une solution $(\epsilon, \beta)-$précise, il suffit de prendre $1 - \beta \geqslant \epsilon^{-1}\phi(N)$ et de déduire à partir de là ce nombre d'itérations minimal garantissant (6.7).

Ainsi la pertinence de l'hypothèse 3 réside dans le fait qu'elle permet à n'importe quelle méthode déterministe, intégrée au sein du processus stochastique associé à l'algorithme 6, de fournir des solutions au problème (6.1). Il s'agit en effet de la condition commune que vérifient les méthodes déterministes, vues au chapitre 5, pour faire la transition vers le cadre de l'optimisation stochastique.

Dans la suite de ce chapitre, nous allons présenter l'intégration des méthodes déterministes, présentées au chapitre 5, au sein du processus général associé à l'algorithme 6. Ceci va nous spécifier différentes variantes de la SA. Ensuite, nous présentons l'efficacité d'estimation moyenne spécifique à chaque méthode.

L'intégration des méthodes du chapitre 5 au sein du processus stochastique associé à l'algorithme 6 est basée sur les hypothèses suivantes.

Hypothèse 4.
- *Il est possible d'engendrer des réalisations $\xi^1, \xi^2, \ldots$ du vecteur aléatoire ξ de telle sorte qu'elles soient i.i.d. (indépendantes et identiquement distribuées).*
- *Il existe un oracle tel qu'étant donné l'entrée $(x, \xi) \in X \times \Xi$, il est capable de renvoyer un vecteur $G(x, \xi) \in \partial F(x, \xi)$ tel que $\mathbb{E}[G(x, \xi)]$ est bien définie et est un sous-gradient de f.*
- *On suppose que l'ampleur de l'incertitude au sein de l'oracle stochastique est bornée c'est-à-dire*

$$\sup_{x \in X} \mathbb{E}[||G(x, \xi)||_*^2]^{1/2} \leqslant L < \infty. \tag{6.8}$$

- *$F(., \xi)$ est convexe sur X pour tout $\xi \in \Xi$.*
- *X est un ensemble convexe borné et fermé de $\mathbb{R}^n$.*

Notons que si $F(.,\xi)$, $\xi \in \Xi$ est convexe et f est à valeur finie au voisinage de x alors :

$$\partial f(x) = \mathbb{E}[\partial F(x,\xi)], \tag{6.9}$$

pour plus de détails on peut voir [46]. Dans ce cas, $G(x,\xi)$ sera un estimateur non biaisé des sous gradients de f en x. La méthode SA est itérative : à chaque étape k, elle construit un point x_k et avec une nouvelle simulation (réalisation) du vecteur aléatoire ξ, l'oracle stochastique calcule un estimateur non-biaisé d'un sous-gradient de la fonction objectif en x_k : $G(x_k,\xi)$. A partir de cette nouvelle information, la méthode construit un nouveau point x_{k+1} qui sera présenté à l'oracle qui, par sa réponse, augmentera l'information accumulée sur le problème, et ainsi de suite.

6.1.1.2 Méthode du sous-gradient projeté stochastique

La méthode est similaire à son homologue déterministe, l'algorithme 3 page 73. En effet, elle génère une séquence finie de points $\{x_k\}, k = 1, 2, \ldots, N$ afin de calculer une solution $\epsilon-$précise en moyenne et cela conformément à l'algorithme 7 page 103. On précise que la norme considérée dans le cas de cet algorithme est uniquement la norme euclidienne notée $\|.\|_2$.

Dans cette section, nous allons voir en détails la complexité algorithmique de cette méthode afin d'illustrer la discussion et les définitions de la section précédente.

Algorithme 7 Algorithme du sous-gradient projeté stochastique

Entrées :

- un point initial $x_0 \in X$;
- une constante $R > 0$ telle que $\|x - x_0\|_2 \leqslant R$, $\forall x \in X$;
- une constante $L > 0$ vérifiant (6.8) ;
- un nombre total d'itérations N ;
- un pas γ_k pour la méthode ;
- une séquence de réalisation $\xi[N-1] = (\xi^1, \ldots, \xi^{N-1})$ i.i.d. de la variable aléatoire ξ.

Sorties : une solution $\hat{x}_N^i$ $\epsilon-$ précise en moyenne du problème (6.1).
Tant que $(1 \leqslant k \leqslant N)$:
prendre $k = k + 1$;
prendre $G(x_{k-1}, \xi^{k-1}) = F'(x_{k-1}, \xi^{k-1}) \in \partial F(x_{k-1}, \xi^{k-1})$;
prendre $x_k = \pi_X(x_{k-1} - \gamma_k G(x_{k-1}, \xi^{k-1}))$.
Fin de tant que
Renvoyer $\hat{x}_N^i = \hat{x}_N^i(\xi[N-1]) = \frac{\sum\limits_{i \leqslant k \leqslant N} \gamma_k x_k}{\sum\limits_{i \leqslant k \leqslant N} \gamma_k}$ avec $1 \leqslant i \leqslant N$.

On note ici que chaque point x_k dépend des $k-1$ réalisations du vecteur aléatoire $\xi[k-1] = (\xi^1, \ldots, \xi^{k-1})$. On rappelle que dans le cas déterministe, la solution proposée par la méthode du sous-gradient projeté, algorithme 3 page 73, est le point correspondant à la plus petite valeur, de la fonction objectif, de toutes les valeurs des points de la séquence générée par la méthode c'est-à-dire $\text{argmin}_{x \in \{x_1, \ldots, x_N\}} f(x)$. Dans le cas du problème (6.1), on n'a pas accès en général à la valeur de la fonction objectif donc on ne peut pas savoir quel point choisir parmi les points de la séquence engendrée par la SA. L'alternative est de définir, à partir d'une séquence de $K = N - i + 1$ points, la solution moyenne suivante :

$$\hat{x}_N^i = \hat{x}_N^i(\xi[N-1]) = \frac{\sum\limits_{i\leqslant k\leqslant N} \gamma_k x_k}{\sum\limits_{i\leqslant k\leqslant N} \gamma_k}, \tag{6.10}$$

avec $1 \leqslant i \leqslant N$, on peut prendre $i = 1$, $i = N/2$, etc. Maintenant on va regarder en détails l'efficacité de la méthode.

Théorème 14. *[46] Sous les hypothèses 4 page 102, le point $\hat{x}_N^i$ calculé par l'algorithme 7 vérifie l'inégalité*

$$\mathbb{E}[f(\hat{x}_N^i) - f^*] \leqslant \frac{R^2 + L^2 \sum\limits_{k=i}^{N} \gamma_k^2}{2\sum\limits_{k=i}^{N} \gamma_k} = \phi(N). \tag{6.11}$$

Preuve. En effet,

$$\begin{aligned}
\delta_{k+1}^2 &:= ||x_{k+1} - x^*||_2^2 = ||\pi_X(x_k(\xi[k-1]) - \gamma_k G(x_k(\xi[k-1]), \xi^k)) - x^*||_2^2 \\
&\leqslant ||x_k(\xi[k-1]) - \gamma_k G(x_k(\xi[k-1]), \xi^k) - x^*||_2^2 \text{ (voir [48, lemme 3.1.5])} \\
&\leqslant \delta_k^2 - 2\gamma_k \langle x_k(\xi[k-1]) - x^*, G(x_k(\xi[k-1]), \xi^k)\rangle + \gamma_k^2 ||G(x_k(\xi[k-1]), \xi^k)||_2^2
\end{aligned}$$

En passant à l'espérance on obtient avec $d_{k+1}^2 := \mathbb{E}[\delta_{k+1}^2]$

$$d_{k+1}^2 \leqslant d_k^2 - 2\gamma_k \mathbb{E}[\langle x_k(\xi[k-1]) - x^*, G(x_k(\xi[k-1]), \xi^k)\rangle] + \gamma_k^2 \mathbb{E}[||G(x_k(\xi[k-1]), \xi^k)||_2^2].$$

Puisque les $k-1$ réalisations de ξ sont indépendantes de la k^{eme} réalisation ξ^k par hypothèse, alors pour calculer le second terme du membre de droite de cette inégalité on applique la loi des espérances itérées [12]. En effet, on a :

$$\begin{aligned}
\mathbb{E}[\langle x_k(\xi[k-1]) - x^*, G(x_k(\xi[k-1]), \xi^k)\rangle] &= \mathbb{E}_{\xi[k-1]}[\mathbb{E}_{\xi^k}[\langle x_k(\xi[k-1]) - x^*, G(x_k(\xi[k-1]), \xi^k)\rangle | \xi[k-1]]], \\
&= \mathbb{E}_{\xi[k-1]}[\langle x_k(\xi[k-1]) - x^*, \mathbb{E}_{\xi^k}[G(x_k(\xi[k-1]), \xi^k) | \xi[k-1]]\rangle], \\
\text{(avec } f' \text{ est un sous-gradient de } f) &= \mathbb{E}_{\xi[k-1]}[\langle x_k(\xi[k-1]) - x^*, f'(x_k)\rangle], \\
\text{(par convexité de la fonction } f) &\geqslant \mathbb{E}[f(x_k) - f(x^*)].
\end{aligned}$$

Le même raisonnement s'applique sur le dernier terme :

$$\begin{aligned}
\mathbb{E}[||G(x_k(\xi[k-1]), \xi^k)||_2^2] &= \mathbb{E}_{\xi[k-1]}[\mathbb{E}_{\xi^k}[||G(x_k(\xi[k-1]), \xi^k)||_2^2 | \xi[k-1]]], \\
&\leqslant \mathbb{E}[L^2], \\
&= L^2.
\end{aligned}$$

par (6.8) car dans ce cas $\|.\|_* = \|.\|_2$ (la norme 2 est auto duale [9]). Ainsi :

$$d_{k+1}^2 \leqslant d_k^2 - 2\gamma_k \mathbb{E}[f(x_k) - f(x^*)] + \gamma_k^2 L^2,$$

par suite,

$$2\gamma_k \mathbb{E}[f(x_k) - f(x^*)] \leqslant d_k^2 - d_{k+1}^2 + \gamma_k^2 L^2.$$

On suppose que l'ensemble faisable est borné et qu'en particulier il est inclus dans une boule de diamètre R, c'est-à-dire pour tout $x, y \in X, ||x - y||_2 \leqslant R$. En sommant terme à

terme les deux membres de cette inégalité de $k = i$ à $k = N$, N étant le nombre de points engendrés par la méthode et $1 \leqslant i \leqslant N$ on obtient :

$$\mathbb{E}[2\sum_{k=i}^{N}\gamma_k(f(x_k) - f(x^*))] = 2\sum_{k=i}^{N}\gamma_k\mathbb{E}[f(x_k) - f(x^*)] \leqslant R^2 + L^2\sum_{k=i}^{N}\gamma_k^2.$$

Or par convexité de f on a $f(\hat{x}_N^i) \leqslant (\sum_{i\leqslant k\leqslant N}\gamma_k)^{-1}\sum_{i\leqslant k\leqslant N}\gamma_k f(x_k)$ voir (6.10). Par linéarité de l'espérance on obtient : $(\sum_{i\leqslant k\leqslant N}\gamma_k)\mathbb{E}[f(\hat{x}_N^i) - f^*] \leqslant \sum_{i\leqslant k\leqslant N}\gamma_k\mathbb{E}[f(x_k) - f^*]$. Ainsi, on arrive à l'inégalité (6.11). □

La borne (6.11) renseigne sur l'efficacité d'estimation moyenne de la méthode du sous-gradient projeté stochastique. Il s'agit de voir comment cette borne varie en fonction du nombre total d'itérations N. Comme nous l'avons montré dans la section précédente, cela renseigne aussi sur la complexité de l'algorithme (le nombre d'itérations minimal avant d'atteindre une solution ϵ^N-précise en moyenne). Pour illustrer cela, dans le cas d'un pas fixe $\gamma_k = \gamma$ et à partir de (6.11), on déduit immédiatement que :

$$\mathbb{E}[f(\hat{x}_N^1) - f^*] \leqslant \frac{R^2 + L^2N\gamma^2}{2N\gamma},$$

en minimisant le terme de droite par rapport à γ, on obtient le pas fixe optimal γ^* assurant la meilleure rapidité de convergence de la borne :

$$\gamma^* = \frac{R}{L\sqrt{N}}. \tag{6.12}$$

L'efficacité d'estimation moyenne de l'algorithme 7, devient

$$\mathbb{E}[f(\hat{x}_N^1) - f^*] \leqslant \frac{RL}{\sqrt{N}}. \tag{6.13}$$

Ainsi, un nombre d'itérations de

$$N_\epsilon = \frac{L^2R^2}{\epsilon_N^2} \tag{6.14}$$

garantit une solution ϵ_N-précise en moyenne.

Par conséquent, on retrouve la complexité de la méthode du gradient dans le cas déterministe. Il s'agit d'une méthode qui présente les mêmes performances dans le cas déterministe que stochastique. De plus, on ne peut pas améliorer le taux de convergence de $O(1/\sqrt{N})$ tant que l'on n'a pas rajouté des hypothèses supplémentaires sur la fonction objectif de (6.1) autres que la convexité, seule hypothèse dans ce cadre (voir [47] pour plus de détails). Il s'agit du même résultat concernant les méthodes du premier ordre déterministes (voir théorème 7 page 74 et la discussion qui le suit).

Par ailleurs, dans le cas du pas fixe γ^*, pour obtenir une solution $\hat{x}_N^1$ $(\epsilon, \beta)-$précise, on a le résultat suivant :

Proposition 5. *Un nombre d'itérations*

$$N = \frac{R^2 L^2}{(1-\beta)^2 \epsilon^2}$$

garantit une solution $(\epsilon, \beta)-$*précise dans le cas du pas fixe* γ^* *(6.12).*

Preuve. Le résultat est immédiat, il suffit de considérer l'inégalité (6.7) pour $\phi(N) = \frac{RL}{\sqrt{N}}$ (voir (6.13)) et ensuite prendre $1 - \beta \geqslant \frac{RL}{\sqrt{N}}$. □

Nous constatons que l'algorithme 7 permet d'obtenir une solution $\epsilon-$précise ou $(\epsilon, \beta)-$précise, au bout d'un nombre d'itérations de l'ordre de $O(1/\epsilon^2)$. C'est le même niveau de complexité qui a été mis en évidence dans le cas de son homologue déterministe (voir algorithme 3, page 73) ainsi que les autres méthodes du premier ordre présentées dans le chapitre 5. Ainsi, on remarque que ce taux de convergence a été hérité dans le cadre de la SA. Dans la suite, nous allons présenter deux autres variantes de la SA basées sur les méthodes déterministes associées à l'algorithme 4, page 77 et à l'algorithme 5, page 81. Nous allons montrer également que le nombre total d'itérations pour une solution $\epsilon-$précise ($O(1/\epsilon^2)$) est hérité dans les deux cas.

Par ailleurs, on note que l'intérêt de choisir un pas fixe permet d'obtenir a priori un nombre d'itérations maximal garantissant un certain niveau de précision préfixé. Dans la suite de ce chapitre, nous nous contenterons de cette stratégie du pas fixe pour la présentation des autres variantes de la SA.

6.1.1.3 Méthode Mirror Descent Stochastique (MDSA)

L'intégration de la méthode MDA de l'algorithme 4, page pagerefschemaMDA, au sein du schéma itératif aléatoire associé à l'algorithme 6 permet d'obtenir explicitement l'algorithme 8 qu'on appellera MDA stochastique (MDSA). Nous avons le théorème suivant sur la complexité algorithmique de l'algorithme 8.

Théorème 15. *[46] L'algorithme 8 a la propriété suivante :*

$$\mathbb{E}[f(\hat{x}_N) - f^*] \leqslant D_{\omega,X} L \sqrt{\frac{2}{\alpha N}} = \phi_{MDSA}(N), \tag{6.16}$$

où

$$D_{\omega,X} := \left[\max_{x \in X} \omega(x) - \min_{x \in X} \omega(x)\right]^{1/2}. \tag{6.17}$$

$D_{\omega,X}$ est le « rayon » de l'ensemble faisable X du point de vue de la fonction génératrice de distance ω, c'est-à-dire un scalaire vérifiant

$$D^2_{\omega,X} \geqslant \max_{z \in X} \omega_{x_0}(z) \tag{6.18}$$

où $\omega_{x_0}(z)$ représente la distance de Bregman (voir (5.7) page 75).

Algorithme 8 Algorithme MDSA

Entrées :

- une fonction génératrice de distance ω de classe $\mathcal{C}^1$ et fortement convexe de cœfficient α ;
- un point initial $x_0 \in int(X)$;
- une constante $L > 0$ vérifiant (6.8) ;
- un nombre total d'itérations N ;
- pas de la méthode $\gamma = \frac{\sqrt{2\alpha}D_{\omega,X}}{L\sqrt{N}}$;
- une séquence de réalisations $\xi[N-1] = (\xi^1, \ldots, \xi^{N-1})$ i.i.d. de la variable aléatoire ξ.

Sorties : une solution $\hat{x}_N$ $\epsilon-$précise en moyenne du problème (6.1)

Tant que $(1 \leqslant k \leqslant N)$:

prendre $k = k + 1$;

prendre $G(x_{k-1}, \xi^{k-1}) = F'(x_{k-1}, \xi^{k-1}) \in \partial F(x_{k-1}, \xi^{k-1})$;

prendre

$$x_k = \underset{y \in X}{\operatorname{argmin}} \left\{ \omega_{x_{k-1}}(y) + \gamma \left\langle G(x_{k-1}, \xi^{k-1}), y - x_{k-1} \right\rangle \right\}. \tag{6.15}$$

Fin de tant que

Renvoyer $\hat{x}_N = \hat{x}_N(\xi[N-1]) = \frac{1}{N} \sum_{i=0}^{N-1} x_k$.

Ainsi, un nombre d'itérations de

$$N_\epsilon = \frac{2}{\alpha} \frac{D_{\omega,X}^2 L^2}{\epsilon_N^2} \tag{6.19}$$

garantit une solution ϵ_N-précise en moyenne. Nous avons obtenu un nombre d'itérations similaire à (6.14) pour l'algorithme 7 du sous-gradient projeté stochastique. La différence par rapport à cette dernière est le paramètre de forte convexité α qui permet de moduler la rapidité de convergence. Plus le coefficient de forte convexité de la fonction génératrice de distance ω est grand, plus petit est N_ϵ. Sinon, les deux méthodes partagent la propriété intrinsèque aux méthodes du première ordre à savoir une efficacité d'estimation de l'ordre $O(1/\epsilon^2)$.

Par application directe de l'inégalité de Markov (6.7), on obtient un résultat similaire à celui de la proposition 5.

Proposition 6. *L'algorithme 8 calcule une solution $(\epsilon, \beta)-$précise au moins au bout d'un nombre d'itérations*

$$N = \frac{2L^2 D_{\omega,X}^2}{\epsilon^2 (1-\beta)^2 \alpha}. \tag{6.20}$$

On peut obtenir des bornes de complexité plus fines en considérant des conditions plus restrictives sur la distribution de probabilité de $G(x, \xi)$, que la condition (6.8). En effet, sous la condition suivante [46, 64, 38] :

$$\mathbb{E}[\exp\left(\|G(x,\xi)\|_*^2 / L^2\right)] \leqslant exp(1), \forall x \in X \tag{6.21}$$

on a le théorème suivant.

Théorème 16. *[46, 64, 38] Sous la condition (6.21), la solution renvoyée par la MDSA de l'algorithme 8 page 107 vérifie :*

$$Prob\left\{ f(\hat{x}_N) - f^* \geqslant \frac{\sqrt{2}LD_{\omega,X}(12+2\Theta)}{\sqrt{\alpha N}} \right\} \leqslant 2\exp(-\Theta), \quad \forall \Theta \geqslant 1. \tag{6.22}$$

Le théorème 16 permet d'obtenir la borne de complexité suivante :

Proposition 7. *Avec $1-\beta \leqslant \frac{1}{2}$, la méthode MDSA de l'algorithme 8 page 107 garantit une solution $(\epsilon, \beta)-$précise pour le problème (6.1) au moins au bout d'un nombre d'itérations*

$$N = \left\lceil 8\left[6 - \ln\left(\frac{1-\beta}{2}\right)\right]^2 \frac{L^2 D_{\omega,X}^2}{\alpha\epsilon^2} \right\rceil. \tag{6.23}$$

Preuve. Le résultat est immédiat : soit $\Theta \geqslant 1$ et soit une précision $\epsilon > 0$. On pose $\epsilon = \frac{\sqrt{2}LD_{\omega,X}(12+2\Theta)}{\sqrt{\alpha N}}$. Un tel entier N vérifiant cette égalité existe. En effet, on a $\Theta = \frac{\epsilon\sqrt{\alpha N}}{2\sqrt{2}LD_{\omega,X}} - 6$. Il suffit de choisir N tel que $\Theta \geqslant 1$ c'est-à-dire

$$N \geqslant \frac{8\text{x}7^2 L^2 D_{\omega,X}^2}{\alpha\epsilon^2}.$$

Ainsi on a par le théorème 16 :

$$\text{Proba}\left\{ f(\hat{x}_N) - f^* \geqslant \epsilon \right\} \leqslant 2\exp\left\{ 6 - \frac{\epsilon\sqrt{\alpha N}}{2\sqrt{2}LD_{\omega,X}} \right\}.$$

Il suffit ensuite de prendre $1-\beta \geqslant 2\exp\left\{6 - \frac{\epsilon\sqrt{\alpha N}}{2\sqrt{2}LD_{\omega,X}}\right\}$, c'est-à-dire $N \geqslant 8\left[6 - \ln\left(\frac{1-\beta}{2}\right)\right]^2 \frac{L^2 D_{\omega,X}^2}{\alpha\epsilon^2}$. Par suite N doit vérifier :

$$N \geqslant \max\left\{ 8\left[6 - \ln\left(\frac{1-\beta}{2}\right)\right]^2 \frac{L^2 D_{\omega,X}^2}{\alpha\epsilon^2}, \frac{8\text{x}7^2 L^2 D_{\omega,X}^2}{\alpha\epsilon^2} \right\}.$$

Pour $1-\beta \leqslant \frac{1}{2}$, nous avons $6 - \ln\left(\frac{1-\beta}{2}\right) \geqslant 7$ (car $\ln(4) \geqslant 1$). Ainsi le nombre d'itérations minimal pour obtenir une solution $(\epsilon, \beta)-$ précise, correspond à l'entier juste au dessus de $8\left[6 - \ln\left(\frac{1-\beta}{2}\right)\right]^2 \frac{L^2 D_{\omega,X}^2}{\alpha\epsilon^2}$, ce qui donne le résultat.

□

Ainsi on obtient une amélioration nette par rapport au nombre d'itérations (6.20) : le niveau de complexité présenté par le nombre d'itérations (6.23) obtenu sous la condition (6.21), présente un terme en $\ln(1/1-\beta)^2$ au lieu $1/(1-\beta)^2$ ce qui peut faire une très grande différence pour $1-\beta$ assez petit.

Algorithme 9 Méthode General Primal Dual Subgradient Stochastique PDSA

Entrées :

- une fonction génératrice de distance ω fortement convexe de cœfficient α ;
- un point initial $x^0 := \text{argmin}_{x\in X}\,\omega(x)$;
- une constante $R > 0$ telle que $\|x - x^0\|_2 \leqslant R$, $\forall x \in X$;
- un pas de l'algorithme $\gamma = \frac{L}{\sqrt{2\alpha D_X}}$;
- un nombre total d'itérations N ;
- une séquence de réalisations $\xi[N-1] = (\xi^0, \ldots, \xi^{N-1})$ i.i.d. de la variable aléatoire ξ.

Sortie : une solution $\hat{x}_N$ $\epsilon-$précise pour le problème (6.1).

Initialisation :

choisir $\beta_0 > 0$;
prendre $x_0 = x^0$;
calculer $G(x_0, \xi^0) \in \partial F(x_0, \xi^0)$;
prendre $\zeta_0 = G(x_0, \xi^0)$.

Tant que $0 \leqslant k \leqslant N-1$:

prendre $k = k + 1$;
choisir $\beta_k \geqslant \beta_{k-1}$;
calculer

$$x_k = \underset{x\in X}{\text{argmin}}\{\langle \zeta_{k-1}, x\rangle + \gamma\beta_k\omega(x)\}\,; \tag{6.25}$$

calculer $G(x_k, \xi^k) \in \partial F(x_k, \xi^k)$ et prendre $\zeta_k = \zeta_{k-1} + \lambda_k G(x_k, \xi^k)$.

Fin de tant que

Renvoyer $\hat{x}_N = \frac{1}{N}\sum\limits_{i=0}^{N-1} x_i$.

6.1.1.4 Méthode General Primal Dual Subgradient Stochastique (PDSA)

L'intégration de la méthode PDA au sein du processus stochastique associé à l'algorithme 6 permet d'obtenir une variante de la SA donnée explicitement dans l'algorithme 9 qu'on appellera PDSA.

Dans le cadre de la PDSA, il y a une légère différence par rapport aux hypothèses 4 page 102 : la condition (6.8) sera remplacée par la suivante

$$\|G(x,\xi)\|_* \leqslant L, \quad \forall x \in X, \quad \forall \xi \in \Xi. \tag{6.24}$$

Dans le cas « simple averaging » (5.18) page 82 (qui correspond à une stratégie pas fixe), l'algorithme 9 a la propriété suivante :

Théorème 17. *[50, 25]*

$$\mathbb{E}[f(\hat{x}_N) - f^*] \leqslant \frac{2L}{\sqrt{N}}\sqrt{\frac{2D_X}{\alpha}} = \phi_{PDSA}(N), \tag{6.26}$$

où $D_X := \max_{x\in X} \omega(x)$.

Ainsi pour obtenir une solution ϵ_N-précise, il faut au moins un nombre d'itérations

$$N_\epsilon = \frac{8D_X L^2}{\alpha \epsilon_N^2}. \tag{6.27}$$

Il s'agit d'un niveau de complexité de la méthode similaire à celui de la MDSA (voir (6.19)).

Par application directe de l'inégalité de Markov (6.7), on obtient un résultat similaire à celui des propositions 5 et 6.

Proposition 8. *L'algorithme 9 calcule une solution* $(\epsilon, \beta)-$*précise au moins au bout d'un nombre d'itérations :*

$$N = \frac{8L^2 D_X}{(1-\beta)^2 \epsilon^2 \alpha}. \tag{6.28}$$

On peut obtenir un niveau de complexité plus fin que celui garanti par la proposition 8 en considérant la condition (6.21) qui est vérifiée grâce à (6.24). Ainsi nous pouvons obtenir le résultat suivant :

Théorème 18. *Sous la condition (6.24), la solution renvoyée par la méthode PDSA associée à l'algorithme 9 page 109 vérifie pour* $\Theta \geqslant 0$:

$$Prob\left\{ f(\hat{x}_N) - f^* \geqslant \frac{C(1+\Theta)}{\sqrt{N+1}} \right\} \leqslant 4\exp(-\Theta), \tag{6.29}$$

avec $\gamma = \frac{L}{\sqrt{2\alpha D_X}}$, $C = 2\max\{\gamma, \gamma^{-1}\}\kappa + L\sqrt{\frac{96 D_X}{\alpha}}$ *et* $\kappa := \max\{D_X, \frac{L^2}{2\alpha}\}$.

Preuve. Notons d'abord qu'on trouve dans [50] le résultat suivant (pour ξ_i fixés) :

$$\forall x_i \in X, \quad \max_{x\in X}\left\{ \sum_{i=0}^{N} \langle G(x_i,\xi_i), x_i - x\rangle \right\} \leqslant \beta_{N+1} D_X + \frac{1}{2\sigma}\sum_{i=0}^{N}\frac{1}{\beta_i}\|G(x_i,\xi_i)\|_*^2.$$

Alors on a en particulier pour $x = x^*$:

$$\forall x_i \in X, \quad \sum_{i=0}^{N} \langle G(x_i, \xi_i), x_i - x^* \rangle \leqslant \beta_{N+1} D_X + \frac{1}{2\sigma} \sum_{i=0}^{N} \frac{1}{\beta_i} \|G(x_i, \xi_i)\|_*^2,$$

On introduit la quantité suivante : $\Delta_i = G(x_i, \xi_i) - g(x_i)$, où $g(x_i) \in \partial f(x_i)$.
Alors,

$$\sum_{i=0}^{N} \langle g(x_i), x_i - x^* \rangle \leqslant \beta_{N+1} D_X + \sum_{i=0}^{N} \langle \Delta_i, x^* - x_i \rangle + \frac{1}{2\sigma} \sum_{i=0}^{N} \frac{1}{\beta_i} \|G(x_i, \xi_i)\|_*^2.$$

Ainsi par convexité de f $(f(x_i) - f^* \leqslant \langle g(x_i), x_i - x^* \rangle)$, on a :

$$\frac{1}{N+1} \sum_{i=0}^{N} f(x_i) - f^* \leqslant \frac{1}{N+1} \beta_{N+1} D_X + \frac{1}{N+1} \sum_{i=0}^{N} \langle \Delta_i, x^* - x_i \rangle + \frac{1}{N+1} \frac{1}{2\sigma} \sum_{i=0}^{N} \frac{1}{\beta_i} \|G(x_i, \xi_i)\|_*^2,$$

$$f(\hat{x}_N) - f^* \leqslant \underbrace{\frac{1}{N+1} \beta_{N+1} D_X + \frac{1}{N+1} \frac{1}{2\sigma} \sum_{i=0}^{N} \frac{1}{\beta_i} \|G(x_i, \xi_i)\|_*^2}_{A_1} + \underbrace{\frac{1}{N+1} \sum_{i=0}^{N} \langle \Delta_i, x^* - x_i \rangle}_{A_2}. \tag{6.30}$$

On pose $Y_i = \frac{\|G(x_i, \xi_i)\|_*^2}{\beta_i}$ et $c_i = \frac{L^2}{\beta_i}$.
Notons que par (6.24) on a

$$\mathbb{E}[\exp(Y_i / c_i)] \leqslant \exp(1), i = 0, \ldots, N. \tag{6.31}$$

Par convexité de la fonction exp(.) on a $\exp\left(\frac{\sum_{i=0}^{N} Y_i}{\sum_{i=0}^{N} c_i} \right) = \exp\left(\sum_{i=0}^{N} \frac{c_i (Y_i / c_i)}{\sum_{i=0}^{N} c_i} \right) \leqslant \sum_{i=0}^{N} \frac{c_i}{\sum_{i=0}^{N} c_i} \exp(Y_i / c_i)$.
Ainsi par passage à l'espérance des deux côtés de l'inégalité précédente et en utilisant (6.31) on obtient :

$$\mathbb{E}\left[\exp\left(\frac{\sum_{i=0}^{N} Y_i}{\sum_{i=0}^{N} c_i} \right) \right] \leqslant \exp(1).$$

Ainsi et par l'inégalité de Markov on a pour tout $\Theta \in \mathbb{R}$:

$$\text{Prob}\left\{ \exp\left(\frac{\sum_{i=0}^{N} Y_i}{\sum_{i=0}^{N} c_i} \right) \geqslant \exp(\Theta) \right\} \leqslant \exp(1) / \exp(\Theta) = \exp(1 - \Theta),$$

Ainsi (par application de la fonction logarithme qui est monotone)

$$\text{Prob}\left\{ \sum_{i=0}^{N} Y_i \geqslant \Theta \sum_{i=0}^{N} c_i \right\} \leqslant \exp(1 - \Theta) \leqslant 3 \exp(-\Theta). \tag{6.32}$$

Ce qui veut dire que pour tout Θ, on a :

$$\text{Prob}\left\{\frac{1}{N+1}\frac{1}{2\sigma}\sum_{i=0}^{N}\frac{1}{\beta_i}\|G(x_i,\xi_i)\|_*^2 \geqslant \Theta\frac{1}{2\sigma}\frac{1}{N+1}L^2\sum_{i=0}^{N}\frac{1}{\beta_i}\right\} \leqslant 3\exp(-\Theta). \tag{6.33}$$

Ainsi on obtient :

$$\text{Prob}\left\{A_1 \geqslant \hat{\beta}_{N+1}\frac{1}{N+1}\left(\gamma D_X + \Theta\frac{1}{2\sigma}L^2\frac{1}{\gamma}\right)\right\} \leqslant 3\exp(-\Theta),$$

où $\beta_k := \gamma\hat{\beta}_k$. Par le lemme 3 de [50], on a : $\hat{\beta}_k \leqslant \frac{1}{1+\sqrt{3}} + \sqrt{2k-1}$, pour $k \geqslant 0$, ainsi $\hat{\beta}_k \leqslant 2\sqrt{k}$ par suite $\frac{\hat{\beta}_N}{N} \leqslant \frac{2}{\sqrt{N}}$. Ainsi on obtient :

$$\text{Prob}\left\{A_1 \geqslant 2\frac{\max\{\gamma,\gamma^{-1}\}}{\sqrt{N+1}}\left(D_X + \Theta\frac{1}{2\sigma}L^2\right)\right\} \leqslant 3\exp(-\Theta). \tag{6.34}$$

On considère maintenant la variable aléatoire A_2. Le lemme 6 de [50] nous donne $\omega(x) \geqslant \frac{1}{2}\alpha\|x-x_0\|^2, \quad \forall x \in X$. Par conséquent on a :

$$\|x^*-x_i\| \leqslant \|x^*-x_0\| + \|x_0-x_i\| \leqslant \sqrt{\frac{2}{\alpha}}(\sqrt{\omega(x^*)}+\sqrt{\omega(x_i)}) \leqslant \sqrt{\frac{8}{\alpha}D_X},$$

par suite,

$$|\Delta_i^T(x^*-x_i)|^2 \leqslant \|\Delta_i\|_*^2\|x^*-x_i\|^2 \leqslant \|\Delta_i\|_*^2\frac{8}{\alpha}D_X.$$

Or $\|\Delta_i\|_* \leqslant 2L$, donc,

$$|\Delta_i^T(x^*-x_i)|^2 \leqslant \frac{32}{\alpha}DL^2.$$

On a également :

$$\mathbb{E}_{\xi_i}\left[\Delta_i^T(x^*-x_i)\middle|\,\xi[i-1]\right] = (x^*-x_i)'\mathbb{E}_{\xi_i}\left[\Delta_i|\xi[i-1]\right] = 0.$$

De plus et par (6.24) on a aussi

$$\mathbb{E}_{\xi_i}\left[\exp\{\|\Delta_i\|_*^2/(4L^2)\}|\xi[i-1]\right] \leqslant \exp(1).$$

Ainsi et en appliquant l'inégalité (7.194) de la proposition (7.64) de [64, page 391] qu'on rappelle ici :

Proposition 9. *[64] Soit $\xi^0,\xi^1,\ldots$ une suite de vecteurs aléatoires i.i.d., et soit $\sigma_i > 0$, $i = 1,\ldots,$ une suite numérique déterministe et soit $\phi_i = \phi_i(\xi[i])$ une fonction de $\xi[i] = (\xi^0,\ldots,\xi^i)$ telle que*

$$\mathbb{E}\left[\phi_i|\xi[i-1]\right] = 0 \quad et \quad \mathbb{E}\left[exp\left(\phi_i^2/\sigma_i^2\right)|\xi[i-1]\right] \leqslant exp(1),$$

alors pour tout $\Theta \geqslant 0$,

$$Prob\left\{\sum_{i=0}^{N}\phi_i \geqslant \Theta\sqrt{\sum_{i=0}^{N}\sigma_i^2}\right\} \leqslant exp(-\Theta^2/3).$$

avec $\phi_i = \frac{1}{N+1}\Delta_i^T(x^* - x_i)$ et $\sigma_i^2 = \frac{32D_X L^2}{(N+1)^2\sigma}$, on obtient pour tout $\Theta \geqslant 0$

$$\text{Prob}\left\{A_2 \geqslant \Theta\frac{L}{\sqrt{N+1}}\sqrt{\frac{32D_X}{\alpha}}\right\} \leqslant \exp(-\Theta^2/3). \tag{6.35}$$

En changeant $\Theta^2/3$ par Θ et en remarquant que $\Theta^{1/2} \leqslant 1 + \Theta$ pour $\Theta \geqslant 0$, on obtient

$$\text{Prob}\left\{A_2 \geqslant \sqrt{3}(\Theta+1)\frac{L}{\sqrt{N+1}}\sqrt{\frac{32D_X}{\alpha}}\right\} \leqslant \exp(-\Theta). \tag{6.36}$$

Alors, de (6.30), (6.34) et (6.36) on obtient :

$$\text{Prob}\left\{f(\hat{x}_N) - f^* \geqslant \frac{C_1 + C_2\Theta}{\sqrt{N+1}}\right\} \leqslant 4\exp(-\Theta),$$

avec $C_1 := 2\max\{\gamma, \gamma^{-1}\}D_X + L\sqrt{\frac{96D}{\alpha}}$, $C_2 := 2\max\{\gamma, \gamma^{-1}\}\frac{L^2}{2\sigma} + L\sqrt{\frac{96D}{\alpha}}$.
En posant $\kappa := \max\{D_X, \frac{L^2}{2\sigma}\}$, et $C = 2\max\{\gamma, \gamma^{-1}\}\kappa + L\sqrt{\frac{96D}{\alpha}}$, on a : $C(1+\Theta) \geqslant C_1 + C_2\Theta$, et ainsi on obtient le résultat (6.29).

□

Ainsi, par l'application du théorème 18, on obtient le niveau de complexité de la PDSA suivant :

Proposition 10. *Avec $1-\beta \leqslant \frac{1}{2}$, la méthode PDSA définie par l'algorithme 9, page 109, garantit une solution $(\epsilon, \beta)-$précise pour le problème (6.1) au moins au bout d'un nombre d'itérations N :*

$$N + 1 = \left\lceil \frac{C^2}{\epsilon^2}\ln^2\left(\frac{11}{1-\beta}\right)\right\rceil. \tag{6.37}$$

Preuve. Le résultat est immédiat : soit $\Theta \geqslant 0$ et soit une précision $\epsilon > 0$. On pose $\epsilon = \frac{C(1+\Theta)}{\sqrt{N+1}}$. Un tel entier N vérifiant cette inégalité existe. En effet, on a $\Theta = \frac{\epsilon}{C}\sqrt{N+1} - 1$. Il suffit de choisir N tel que $\Theta \geqslant 0$ c'est-à-dire $N+1 \geqslant \frac{C^2}{\epsilon^2}$. Ainsi on a par le théorème 18 :

$$\text{Prob}\left\{f(\hat{x}_N) - f^* \geqslant \epsilon\right\} \leqslant 4\exp\{-\epsilon\sqrt{N+1}C^{-1}\}\exp\{1\} \approx 11\exp\{-\epsilon\sqrt{N+1}C^{-1}\}.$$

Il suffit maintenant de prendre $1-\beta \geqslant 11\exp\{-\epsilon\sqrt{N+1}C^{-1}\}$, c'est-à-dire $N+1 \geqslant \frac{C^2}{\epsilon^2}\ln\left(\frac{11}{1-\beta}\right)^2$. Par suite N doit vérifier :

$$N + 1 \geqslant \max\left\{\frac{C^2}{\epsilon^2}\ln\left(\frac{11}{1-\beta}\right)^2, \frac{C^2}{\epsilon^2}\right\}.$$

Pour $1-\beta \leqslant \frac{1}{2}$, nous avons $\ln\left(\frac{11}{1-\beta}\right) \geqslant 1$ (car $\ln\left(\frac{11}{1-\beta}\right) \geqslant \ln\left(\frac{11}{2}\right) \geqslant \ln(4) \geqslant 1$). Ainsi on obtient le résultat. □

Les propriétés de la distribution de probabilité des sous-gradients $G(x, \xi)$ contribuent directement à l'amélioration de la PDSA. Le nouveau niveau de complexité donné dans (6.37) présente un terme en $\ln(1/1-\beta)^2$ au lieu $1/(1-\beta)^2$ pour la borne (6.28) ce qui peut faire une très grande différence pour $1-\beta$ assez petit.

6.1.1.5 Discussion et premières conclusions

Nous avons vu qu'avec l'hypothèse 3 (convergence moyenne), page 101, une méthode du première ordre déterministe intégrée dans le processus stochastique associé à l'algorithme 6 permet de calculer pour le problème (6.1) des solutions numériques au sens des définitions 20 et 21. Ceci a induit un cadre théorique de l'analyse de complexité de la méthode SA basé sur l'analyse de complexité dans le cas des méthodes de résolution déterministes. En effet, l'analyse de convergence (les démonstrations de convergences) de la méthode SA est similaire à celle des méthodes du premier ordre vues au chapitre 5. Ceci a été illustré avec la démonstration de la borne 6.11 du théorème 14 de l'analyse de convergence moyenne de la méthode du sous-gradient projeté stochastique. Cette analyse combinée avec des inégalités de probabilité nous a permis d'évaluer la complexité algorithmique de chaque variante de la méthode SA. Nous avons également constaté que ce niveau de complexité a été hérité de celui des versions déterministes de chaque variante de la méthode SA. Il s'agit d'un nombre d'itérations de l'ordre de $O(1/\epsilon^2)$ pour une précision moyenne de ϵ. Il s'agit sous les hypothèses 4 du taux de convergence optimal que nous ne pouvons pas améliorer en se basant juste sur la simple hypothèse de convexité de la fonction objectif de (6.1) sans qu'elle présente d'autres propriétés supplémentaires telle la forte convexité ou la différentiabilité (voir [47] pour plus de détails). Cette propriété a été également héritée du cadre déterministe présenté dans le chapitre 5.

Par ailleurs, la méthode SA présente le privilège de ne pas passer par l'évaluation de la fonction objectif qui est définie par une intégrale multidimentionnelle. Car, s'agissant d'une fonction objectif définie par une intégrale multiple, l'évaluation passera forcément par une simulation Monte Carlo (car le calcul de sa valeur exacte est impossible pratiquement à partir d'une dimension 4 ou 5). Ceci coûtera un nombre de simulations (tirage sur la variable aléatoire ξ) de $O(1/\epsilon^2)$ pour une précision ϵ (voir [18] par exemple) et cela que pour une seule évaluation en un seul point pour une seule itération alors que la méthode SA propose $O(1/\epsilon^2)$ appels de l'oracle du premier ordre pour calculer une solution $\epsilon-$précise en moyenne.

Après avoir passé en revue les différentes variantes de la méthode SA et leurs complexités algorithmiques, nous allons dans la suite de ce chapitre présenter une deuxième alternative pour la résolution de (6.1) : il s'agit de la méthodologie Sample Average Approximation (SAA).

6.1.2 Méthodologie Sample Average Approximation (SAA)

6.1.2.1 Principe et propriétés principales

Le principe de la SAA est le suivant : nous supposons que l'on puisse engendrer M réalisations i.i.d. (indépendants identiquement distribués) $\xi^1, \ldots, \xi^M$ du vecteur aléatoire ξ afin de construire le problème d'optimisation suivant :

$$\min_{x \in X} \left\{ \hat{f}_M(x) := \frac{1}{M} \sum_{j=1}^{M} F(x, \xi^j) \right\}. \tag{6.38}$$

La méthodologie SAA est basée sur la propriété que les solutions du problème (6.38) convergent vers celles du problème (6.1) avec une probabilité 1 quand M tend vers l'infini.

En effet, la fonction $\hat{f}_M$ dépend des échantillons ξ^j et est, par conséquent, une variable aléatoire. Pour tout $x \in X$, cette fonction est un estimateur non biaisé de $f(x)$, c'est-à-dire, $\mathbb{E}[\hat{f}_M(x)] = f(x)$ et par la loi des grands nombres on démontre que $\hat{f}_M(x)$ tend vers $f(x)$ avec une probabilité un quand $M \to \infty$ (en fait, cette convergence est uniforme sur tout sous-ensemble compact C de X, voir conditions [64, chapitre 5]). De plus, par le théorème de la limite centrale, on démontre que pour tout $x \in X$, $M^{1/2}(\hat{f}_M(x) - f(x))$ converge en loi de probabilité vers la loi normale $\mathcal{N}(0, \sigma^2(x))$, avec $\sigma^2(x) = Var[F(x, \xi)]$ (par contre, la vitesse de convergence de $\hat{f}_M(x)$ vers $f(x)$ est notoirement lente : de l'ordre de $O(M^{-1/2})$ [63]). Compte tenu de toutes ces observations, pour tout $x \in X$, on peut estimer la valeur de $f(x)$ en calculant la moyenne des valeurs $F(x, \xi^j)$, $j = 1, \ldots, M$. De plus, la valeur optimale ainsi que l'ensemble des solutions optimales de (6.38) sont des estimateurs consistants [1] de leurs homologues du problème (6.1).

En effet, on a le résultat suivant :

Théorème 19 (Shapiro *et al.* 2007 [63])**.** *S'il existe un compact $C \in \mathbb{R}^n$ tel que :*
(i) l'ensemble S des solutions optimales de (6.1) est non vide et inclus dans C ;
(ii) la fonction $f(x)$ est continue et à valeurs finies sur C ;
(iii) $\hat{f}_M(x)$ converge, uniformément en $x \in C$, vers $f(x)$ quand $M \to \infty$;
(iv) avec probabilité 1 et pour M suffisamment grand, $\hat{S}_M$ est non vide et $\hat{S}_M \subset C$.
Alors, $\hat{f}^*_M \to f^*$ et $\mathbb{D}(\hat{S}_M, S) \to 0$ [2] avec probabilité 1 quand $N \to \infty$. $\hat{S}_M$ et $\hat{f}^*_M$ sont respectivement l'ensemble des solutions optimales et la valeur optimale de (6.38), et f^* est la valeur optimale de (6.1).

L'assertion $\mathbb{D}(\hat{S}_M, S) \to 0$ avec probabilité 1 veut dire que pour toute sélection $\hat{x}_M \in \hat{S}_M$, on a $dist(\hat{x}_M, S) \to 0$ avec probabilité 1, où $dist(\hat{x}_M, S)$ est la distance du point $\hat{x}_M$ à l'ensemble S. Si de plus on a $S = \{x^*\}$ un singleton, c'est-à-dire, le problème (6.1) a une solution unique x^*, alors cela revient à dire $\hat{x}_M \to x^*$ avec probabilité 1. En plus, cette convergence se fait avec un taux de convergence de l'ordre de $O(M^{-1/2})$ [62]. Par ailleurs, on a [61] :

$$\hat{f}^*_M = \min_{x \in S} \hat{f}_M(x) + o(M^{-1/2})$$

avec probabilité 1. Ce qui veut dire, dans le cas où $S = \{\bar{x}\}$ est un singleton, $\hat{f}^*_M$ converge vers f^* avec le même taux de convergence que celui de $\hat{f}_M(x)$ vers $f(x)$ c'est-à-dire $O(M^{-1/2})$.

D'après la discussion ci-dessus, il est clair que la méthode SAA a des propriétés de convergence lentes : la vitesse de convergence de $\hat{f}_M(x)$ vers $f(x)$ est de l'ordre de $O(M^{-1/2})$. Par conséquent, pour améliorer la précision de l'estimateur $\hat{f}_M(x)$ d'un digit, on a besoin d'un M 100 fois plus grand. Cette convergence lente a été héritée par l'estimateur $\hat{f}^*_M$ aussi (pour plus de détails et discussions, voir [64]).

Par ailleurs, à l'instar de l'approche SA, la méthodologie SAA permet de garantir une solution au sens de la définition 21. En effet, on trouve dans [64, 65] un résultat permettant

1. $\hat{\theta}_N$ est un estimateur consistant de θ si $\hat{\theta}_N$ converge vers θ avec probabilité 1 quand $N \to \infty$.
2. $\mathbb{D}(A, B) := \sup_{x \in A} dist(x, B)$ est appelé déviation de A dans B.

de calculer le nombre M pour qu'une solution $\delta-$précise de (6.38) soit $(\epsilon,\beta)-$précise pour (6.1).

Théorème 20. *[65] Si l'ensemble X a un diamètre fini R (c'est-à-dire $R := \sup_{x,x'\in X} \|x - x'\|$) et si $F(.,\xi)$ est continue et $L-$Lipschitz sur X pour tout $\xi \in \Xi$ c'est-à-dire*

$$|F(x',\xi) - F(x,\xi)| \leqslant L\|x - x'\|, \forall x, x' \in X, \forall \xi \in \Xi,$$

alors avec au moins un nombre M de tirages sur la variable aléatoire ξ tel que

$$M = O(1)\left(\frac{RL}{\epsilon}\right)^2 \left[n \ln \frac{RL}{\epsilon} + \ln \frac{O(1)}{1-\beta}\right], \tag{6.39}$$

toute solution $\epsilon/2-$précise de (6.38) est une solution $(\epsilon,\beta)-$précise de (6.1).

La méthode SAA propose de ramener la résolution d'un problème stochastique (6.1) à celle d'un problème déterministe convexe (6.38). Il s'agit d'une sorte d'équivalence avec une grande marge de confiance $1-\beta$ entre les deux problèmes. L'avantage est de réduire la difficulté de résolution de (6.1) liée à sa nature stochastique en se ramenant à la résolution d'un problème d'optimisation déterministe mais qui dépend toujours des tirages sur ξ.

6.1.2.2 Comparaison entre la SAA et la SA (à travers ses deux variantes MDSA et PDSA)

Pour comparer la complexité algorithmique des deux approches, on va considérer les deux méthodes MDSA et PDSA. On a vu que sous la condition (6.21), un nombre d'itérations de

$$N_{MDSA} = \left\lceil 8\left[6 - \ln\left(\frac{1-\beta}{2}\right)\right]^2 \frac{L^2 D^2_{\omega,X}}{\alpha\epsilon^2}\right\rceil \tag{6.40}$$

garantit le calcul d'une solution $(\epsilon,\beta)-$précise pour la méthode MDSA (voir proposition 7) alors que dans le cas de la méthode PDSA ce nombre est de (voir proposition 10)

$$N_{PDSA} = \left\lceil \frac{C^2}{\epsilon^2} \ln^2\left(\frac{11}{1-\beta}\right)\right\rceil + 1. \tag{6.41}$$

Dans le cas de l'approche SAA, la résoulution du problème (6.38) avec un nombre de réalisations sur ξ (6.39) de

$$M_{SAA} = O(1)\left(\frac{RL}{\epsilon}\right)^2 \left[n \ln \frac{RL}{\epsilon} + \ln \frac{1}{1-\beta}\right]$$

garantit une solution $(\epsilon,\beta)-$précise pour le problème (6.1).

Nous notons la même dépendance, en β et en ϵ : une dépendance logarithmique en $1-\beta$ et quadratique (ou à peu près quadratique) en $1/\epsilon$. Par ailleurs, pour la MDSA et la PDSA, ce nombre d'itérations ne dépend pas explicitement du nombre de variables de décision contrairement à la SAA, où cette dépendance du nombre de réalisations est explicite et linéaire. Nous avons, pour un jeu donné de constantes, tracé les allures correspondantes avec $1 - \beta = 95\%$ et pour différents ϵ. Le résultat est donné dans la figure 6.1. On voit

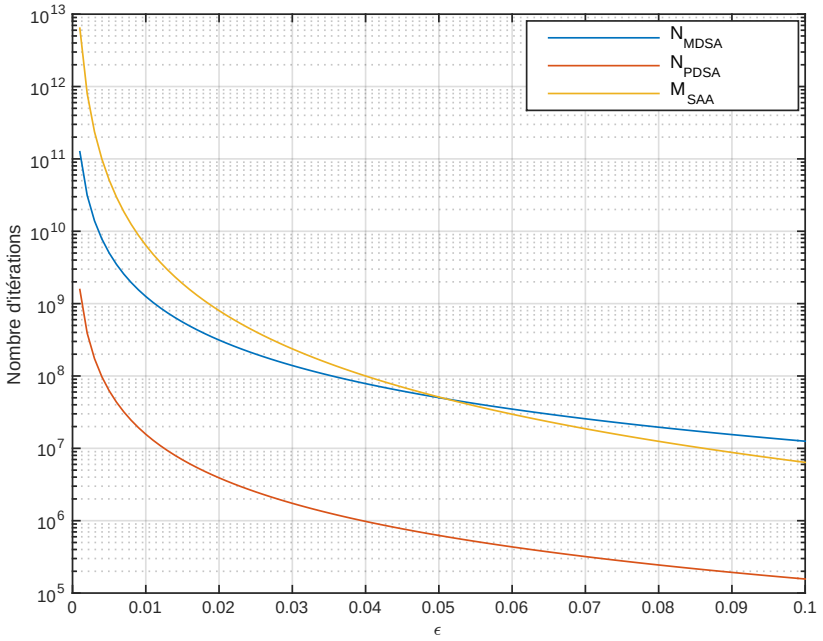

FIGURE 6.1 – Allures des nombres d'itérations en fonction de ϵ

que les allures varient d'une façon pratiquement similaire vis-à-vis de la précision ϵ et ont une allure quadratique en $1/\epsilon$.

Dans le cas des deux méthodes MDSA et PDSA, pour aboutir à une solution $(\epsilon, \beta)-$précise pour le problème (6.1), cela coûtera au total $O(1/\epsilon^2)$ appels à l'oracle du premier ordre calculant F ainsi que ses sous-gradients G (soit un seul appel par itération). A côté de cela, l'approche SAA va résoudre le problème (6.38) avec M_{SAA} appels à l'oracle du premier ordre évaluant la fonction F ainsi qu'un de ses sous-gradients et cela à chaque itération ! Maintenant si on choisit de résoudre (6.38) avec une méthode du première ordre (cela est typiquement le cas quand il s'agit d'un problème de grande dimension), alors pour obtenir une solution $\epsilon/2-$précise (de (6.38)) nous aurons besoin d'un nombre d'itérations de $O(1/\epsilon^2)$. Cela fait au total $O(1/\epsilon^2)M_{SAA}$ appels à l'oracle du premier ordre calculant la fonction objectif F ainsi que ses sous-gradients. Ainsi, en terme de dépendance en n et en terme du nombre d'appels à l'oracle évaluant F et ses sous-gradients, la MDSA et la PDSA se montrent largement plus efficaces que la SAA.

Finalement, pour plus de détails concernant la comparaison entre la MDSA (uniquement) et SAA, on peut se reporter à [65].

6.2 Cas d'étude : Réseau à solution analytique

Nous allons maintenant aborder la résolution du problème 7, défini dans le chapitre 4, page 63, avec la méthode MDSA. Rappelons que pour ce problème précis, la solution a été établie analytiquement, voir la proposition 4, page 64.

Pour cela, nous passerons par les étapes suivantes :

1) intégrer le problème 7 au sein du cadre de la MDSA :
   - mettre le problème 7 sous la forme d'un problème de minimisation c'est-à-dire sous la forme du problème (6.1),
   - vérifier que le problème ainsi reformulé satisfait les hypothèses 4 page 102 ;
2) spécifier l'algorithme MDSA qui correspond à notre cas d'étude.

Finalement, nous passerons à la résolution numérique. A travers cette étape, nous illustrerons les propriétés de convergence de la MDSA ainsi que ses performances pratiques.

6.2.1 Intégration du problème 7 au sein du cadre de la MDSA

6.2.1.1 Mise sous la forme du problème d'optimisation (6.1)

Pour intégrer le problème 7 au sein du cadre MDSA, nous allons tout d'abord mettre notre problème sous la forme d'un problème de minimisation c'est-à-dire sous la forme (6.1). Pour cela, le problème 7 peut être mis sous la forme équivalente suivante :

$$\nu_T^* = -\min_{\tilde{E}_{ij}>0} \mathbb{E}\left[\max_{i,j}\left\{-\frac{k_{i,j}}{c_i}\tilde{E}_{i,j}\xi_{i,j}\right\}\right] \tag{6.42}$$
$$\text{tel que : } \sum_{i=1}^{N_P}\sum_{j=1}^{N_v}\tilde{E}_{i,j} = E_T.$$

Pour simplifier les notations, on remplace l'ensemble des enzymes $\{\tilde{E}_{i,j}\}$ par l'ensemble $\{x_i\}$ avec $r \in \{1,\dots,n = N_pN_v\}$, les deux ensembles étant en bijection. On remplace également l'ensemble $\{\xi_{i,j}\}$ par l'ensemble $\{\xi_r\}$ avec $r \in \{1,\dots,n = N_pN_v\}$, les deux ensembles étant également en bijection ; les ξ_r sont des variables aléatoires exponentielles de paramètre 1. On pose $a_r = \frac{c_i}{k_{i,j}}$ tel que $r = 1,\dots,n = N_pN_v$ et $i = 1,\dots,N_p$ et $j = 1,\dots,N_v$. Ainsi le problème (6.42) devient :

$$\nu_T^* = -\min_{x} \mathbb{E}\left[\max_{r\in\{1,\dots,N_vN_p\}} -x_r\xi_r/a_r\right] \tag{6.43}$$
$$\text{tel que : } \sum_{r=1}^{n} x_r = E_T, x_r > 0.$$

6.2.1.2 Vérification des hypothèses 4 page 102

Il est pertinent, pour l'efficacité numérique de l'algorithme MDSA , comme nous l'avons vu au chapitre précédent chapitre 5 et dans ce chapitre de considérer la configuration l_1. Par conséquent, dans la suite de ce chapitre, l'espace $\mathbb{R}^n$ sera muni de la norme l_1, et donc la norme duale dans ce cas sera $\|.\|_* = \|.\|_\infty$. Le deuxième avantage réside dans le fait que le sous problème d'optimisation (6.15) de la MDSA (voir l'algorithme 8, page 107), c'est-à-dire la projection non linéaire, peut se résoudre analytiquement comme nous allons le montrer dans la suite. Pour compléter la vérification des hypothèses 4 page 102. Nous avons la propriété suivante sur la fonction qui, à tout $x \in X$, associe $\max_{r\in\{1,\dots,n\}}\{-x_r\xi_r/a_r\}$.

Proposition 11. *(i) La fonction F définie sur X par $F(x,\xi) = \max_r\{-x_r\xi_r/a_r\}$ est une fonction convexe, par rapport à $x \in \mathbb{R}_+^n$ pour tout $\xi \in \Xi \subset \mathbb{R}_+^n$.*

(ii) F est $(\max_r\{|\xi_r/a_r|\})-$Lipschitz.

(iii) La constante

$$L := \sqrt{2+(1+\ln(n))^2}/\min_r a_r \tag{6.44}$$

vérifie la condition (6.8) page 102.

Preuve. (i) En effet, la fonction $F(.,\xi)$ est définie comme étant le maximum de plusieurs fonctions affines donc convexes. Le maximum de fonctions convexes étant convexe [15] on en déduit que $F(.,\xi)$ est convexe.

(ii) Nous introduisons la fonction $F_r(x,\xi) := -x_r\xi_r/a_r$. On a pour tout $x, y \in X$, pour tout $r \in \{1,\dots,n\}$ et pour un $\xi \in \Xi$:

$$\begin{aligned}|F_r(x,\xi) - F_r(y,\xi)| &= |-\xi_r x_r/a_r + \xi_r y_r/a_r| \\ &\leqslant |\xi_r/a_r||x_r - y_r| \text{ (par inégalité de Cauchy-Schwarz)} \\ &\leqslant |\xi_r/a_r| \max_r\{|x_r - y_r|\} \\ &= |\xi_r/a_r|\|x-y\|_*.\end{aligned}$$

Par suite, on a pour tout $i \in \{1,\dots,n\}$:

$$\begin{aligned}F_r(x,\xi) &\leqslant F_r(y,\xi) + |\xi_r/a_r|\|x-y\|_* \\ &\leqslant F(y,\xi) + \max_r\{|\xi_r/a_r|\}\|x-y\|_*\end{aligned}$$

par suite $F(x,\xi) \leqslant F(y,\xi) + \max_r\{|\xi_r/a_r|\}\|x-y\|_*$. Par symétrie entre x et y, on en déduit

$$|F(x,\xi) - F(y,\xi)| \leqslant \max_r\{|\xi_r/a_r|\}\|x-y\|_*, \quad \forall x, y > 0.$$

Donc $F(.,\xi)$ est $(\max_r\{|\xi_r/a_r|\})-$Lipschitz par rapport à la norme $\|.\|_*$.

(iii) On en déduit par conséquent que :

$$\|s\|_* \leqslant \max_r\{|\xi_r/a_r|\} \leqslant \frac{\max_r \xi_r}{\min_r a_r} \quad \forall s \in \partial F(x,\xi), x > 0 \quad \text{(voir [58] pour la première inégalité).}$$

Par passage à l'espérance nous avons :

$$\mathbb{E}[\|s\|_*^2] \leqslant \mathbb{E}[(\max_r \xi_r)^2]/(\min_r a_r)^2, \quad \forall s \in \partial F(x,\xi), x > 0.$$

L'une des propriétés typiques de la distribution exponentielle est la propriété d'absence de mémoire. Cela fait que le maximum des variables exponentielles de paramètre λ n'est pas une variable exponentielle (voir par exemple [2]). La variable aléatoire $\max_{r=1,\dots,n} \xi_r$ est dite n^{eme} ordre statistique de ξ. On trouve en particulier les relations récursives suivantes sur le n^{eme} ordre statistique dans [3] :

$$\begin{aligned}\mathbb{E}[\xi_{n:n}] &= \mathbb{E}[\xi_{n-1:n-1}] + \tfrac{1}{n}, \\ \mathbb{E}[\xi_{n:n}^2] &= \mathbb{E}[\xi_{n-1:n-1}^2] + \tfrac{2}{n}\mathbb{E}[\xi_{n:n}]\end{aligned}$$

où $\xi_{n:n} := \max\{\xi_1,\dots,\xi_n\}$. Ainsi on obtient :

$$\mathbb{E}[\xi_{n:n}^2] = \frac{2}{n}\mathbb{E}[\xi_{n:n}] + \cdots + \frac{2}{2}\mathbb{E}[\xi_{2:2}] + \mathbb{E}[\xi_{1:1}^2],$$

mais aussi

$$\mathbb{E}[\xi_{n:n}] = 1 + \frac{1}{2} + \ldots + \frac{1}{n}.$$

Ainsi, si on note $H_r = \sum_{k=1}^{r} \frac{1}{k}$ la i^{eme} somme partielle de la série harmonique, on obtient le résultat final :

$$\mathbb{E}[\xi_{n:n}^2] = 2\sum_{r=2}^{n} \frac{H_r}{r} + \mathbb{E}[\xi_{1:1}^2] = 2\sum_{r=2}^{n} \frac{H_r}{r} + 2 = \sum_{r=1}^{n} \frac{1}{r^2} + (\sum_{r=1}^{n} \frac{1}{r})^2 < 2 + (1 + \ln(n))^2,$$

du fait que $\mathbb{E}[\xi_{1:1}^2] = Var[\xi_{1:1}^2] + \mathbb{E}[\xi_{1:1}]^2 = 2$, car $\xi_{1:1}$ est une variable exponentielle de paramètre 1 et du fait que l'on a :

$$2\sum_{r=2}^{n} \frac{H_r}{r} + 2 = \sum_{r=1}^{n} \frac{1}{r^2} + (\sum_{r=1}^{n} \frac{1}{r})^2, \tag{6.45}$$

que nous pouvons démontrer facilement par récurrence. En effet, pour $n = 2$ c'est évident. Supposons maintenant que (6.45) est vraie pour un entier $n \in \mathbb{N}$ et démontrons la pour $n + 1$:

$$\begin{aligned}
2\sum_{r=2}^{n+1} \frac{H_r}{r} + 2 &= 2\frac{H_{n+1}}{n+1} + 2\sum_{i=2}^{n} \frac{H_r}{r} + 2 \\
&= 2\frac{H_{n+1}}{n+1} + \sum_{r=1}^{n} \frac{1}{r^2} + (\sum_{r=1}^{n} \frac{1}{r})^2 \text{ (hypothèse de récurrence)} \\
&= \frac{2}{(n+1)^2} + 2\frac{1}{n+1}\sum_{r=1}^{n} \frac{1}{r} + 2\sum_{r=1}^{n} \frac{1}{r^2} + 2\sum_{1 \leqslant i < j \leqslant n} \frac{1}{i}\frac{1}{j} \\
&= \frac{2}{(n+1)^2} + 2\sum_{r=1}^{n} \frac{1}{r^2} + 2\sum_{1 \leqslant i < j \leqslant n+1} \frac{1}{i}\frac{1}{j} \\
&= \sum_{r=1}^{n+1} \frac{1}{r^2} + (\sum_{r=1}^{n+1} \frac{1}{r})^2,
\end{aligned}$$

ce qui termine la démonstration de l'égalité (6.45). Par ailleurs, on peut remarquer que :

$$\sum_{r=1}^{n} \frac{1}{r^2} \leqslant 1 + \sum_{r=2}^{n} \frac{1}{r(r-1)} = 1 + \sum_{r=2}^{n} \frac{1}{r-1} - \frac{1}{r} = 1 + 1 - \frac{1}{n} < 2.$$

Finalement, H_n étant la somme partielle de la série harmonique, on a

$$H_n < \ln(n) + 1.$$

Par suite on a :

$$\mathbb{E}[\|s\|_*^2] \leqslant L^2, \quad \forall s \in \partial F(x, \xi), x > 0,$$

avec

$$L := \sqrt{2 + (1 + \ln(n))^2} / \min_r a_r.$$

Ainsi L vérifie de fait la condition (6.8).

□

En remarquant que $F(x, \xi) = -\min_r \{x_r \xi_r / a_r\}$ et par application directe de la proposition 3, on en déduit que la fonction objectif du problème (6.43) est donnée par $f(x) = -\frac{1}{\sum_{r=1}^{n} a_r / x_r}$, ainsi elle prend des valeurs finies au voisinage de tout $x > 0$, par conséquent les sous-gradients de $F(., \xi)$ vérifient (6.9) (voir [46]) pour tout x non nul.

Il reste maintenant à construire un oracle du premier ordre qui pour un point faisable, x, donné du problème et une réalisation ξ^k de ξ, renvoie un sous-gradient de la fonction $F(x, \xi^k)$. Pour cela on a le résultat suivant :

Proposition 12. *[48] Soit $\xi \in \Xi$, le sous-différentiel de la fonction F définie par $F(x, \xi) = \max_r\{-x_r \xi_r / a_r\}$ est donné par*

$$\partial F(x, \xi) = Conv\{F'_r(x, \xi) | r \in I(x)\} \tag{6.46}$$

où $I(y) = \{r : F_r(x, \xi) := -x_r \xi_r / a_r = F(x, \xi)\}$, $F'_r(x, \xi) = (0, \ldots, 0, -\xi_r / a_r, 0, \ldots, 0)$, et Conv désigne l'enveloppe convexe.

Preuve. Le résultat est obtenu par application directe du [48, lemme 3.1.10]. □

Cette proposition caractérise le sous-différentiel de $F(., \xi)$, c'est-à-dire l'ensemble des sous-gradients de la fonction $F(., \xi)$. Pour la MDSA et la SA en général, un seul élément de cet ensemble est requis étant donné une entrée x. En se basant sur cette proposition on a

$$G(x, \xi) = (0, \ldots, 0, -\xi_{r^*} / a_{r^*}, 0, \ldots, 0) \in \partial F(x, \xi), \tag{6.47}$$

où $r^* = \text{argmax}_r\{-\xi_r x_r / a_r\}$. Ainsi la construction de notre oracle du premier ordre est donnée dans l'algorithme 10.

Algorithme 10 Oracle du premier ordre pour le problème (6.43)

Entrée : un point x et une réalisation $\hat{\xi}$ du vecteur aléatoire ξ
Sortie : $G(x, \hat{\xi}) \in \partial F(x, \hat{\xi})$
Faire : calculer $r^* = \text{argmax}_{r \in \{1, \ldots, n\}}\{-\xi_r x_r / a_r\}$,
Renvoyer : $G(x, \hat{\xi}) = (0, \ldots, 0, -\hat{\xi}_{r^*} / a_{r^*}, 0, \ldots, 0)$.

Cet oracle requiert un nombre d'opérations arithmétiques de $O(n)$.

Ainsi on arrive à conclure que le problème (6.43) rentre dans le cadre des hypothèses 4 et par suite il peut être résolu par la méthode MDSA.

6.2.2 Algorithme MDSA spécifique au problème (6.43)

Pour déterminer l'algorithme MDSA spécifique au problème (6.43), il nous reste maintenant à spécifier la projection non linéaire (6.15) déjà donnée sous sa forme générale dans l'algorithme 8.

Dans le cas de notre application et comme on l'a mentionné plus haut, on va considérer la configuration l_1 car cela représente un choix judicieux du point de vue de la vitesse de convergence algorithmique. La fonction génératrice considérée est la fonction entropie définie par

$$\omega(x) = \sum_{r=1}^{n} x_r \ln(x_r). \tag{6.48}$$

Ce choix a l'avantage de rendre le calcul de la projection non linéaire définie par (6.15) facile comme le montre le lemme 7.

Lemme 7. *On considère la projection non linéaire (6.15) de l'algorithme 8, définie par :*

$$P_x(y) = \operatorname*{argmin}_{z \in X} \{\langle y, (z-x)\rangle + \omega(z) - \omega(x) - \langle \omega'(x), (z-x)\rangle\}, \tag{6.49}$$

où $X = \{x \in \mathbb{R}_+^n | \sum_{r=1}^{n} x_r = E_T\}$, *et* ω *est définie par (6.48). La* r^{eme} *composante de la projection peut être mise sous la forme explicite suivante :*

$$[P_x(y)]_r = E_T \frac{x_r e^{-y_r}}{\sum_{r=1}^{n} x_r e^{-y_r}}. \tag{6.50}$$

Preuve. On note $\alpha(z) = \langle y, (z-x)\rangle + \omega(z) - \omega(x) - \langle \omega'(x), (z-x)\rangle$ et $\gamma(z) = \sum_{r=1}^{n} x_r - E_T$. Comme le problème (6.49) est convexe en z (la fonction entropie ω est convexe) alors l'application des conditions KKT donne que z^* est un optimum global (c'est-à-dire $z^* = P_x(y)$) si et seulement s'il existe un multiplicateur $\lambda \in \mathbb{R}$ tel que

$$\alpha'(z^*) + \lambda \gamma'(z^*) = 0$$

c'est-à-dire

$$y_r + \ln(z_r^*) - \ln(x_r) + \lambda = 0, \quad \forall r \in \{1, \ldots, n\}.$$

En combinant cela avec le fait que $\sum_{r=1}^{n} z_r^* = 1$ on obtient le résultat (6.50). □

Maintenant pour expliciter l'adaptation de l'algorithme MDSA, il reste à calculer les paramètre $D_{\omega,X}$ et L (algorithme 8 page 107). La constante L est déjà calculée (voir équation (6.44)). On rappelle que

$$D_{\omega,X} := \left[\max_{x \in X} \omega(x) - \min_{x \in X} \omega(x)\right]^{1/2}.$$

On peut montrer que

$$D_{\omega,X} = \sqrt{E_T \ln(n)}. \tag{6.51}$$

En effet, la fonction ω est convexe donc atteint son maximum au sommet de $\bar{X}$ (l'adhérence de X). Les sommets ont la forme suivante $(0,..,E_T,0,..,0)$. Ce sont des points avec tous les cœfficients nuls sauf un qui vaut E_T. La valeur de la fonction ω aux sommets de $\bar{X}$ vaut $E_T \ln(E_T)$ (on sait que par continuité on a $x_r \ln(x_r)$ est égale à 0 pour $x_r = 0$). Par ailleurs la valeur minimale de la fonction ω sur X (ou $\bar{X}$) est $E_T \ln(E_T) - E_T \ln(n)$ obtenue en $x_r = \frac{E_T}{n}$ (une application directe des conditions nécessaires d'optimalité du premier ordre KKT permet de le voir).

Remarque : On note que nous pouvons choisir n'importe quel point faisable dans X pour démarrer l'algorithme. Nous avons choisi le point $x_0 = \text{argmin}_{x \in X}\, \omega(x) = (\frac{E_T}{n}, \ldots, \frac{E_T}{n})$ car il représente le point « central » de notre simplexe.

Maintenant nous avons tout pour expliciter l'adaptation de l'algorithme MDSA 8 page 107 à notre cas d'étude. Il s'agit de l'algorithme 11.

Algorithme 11 Algorithme MDSA pour le problème (6.43)

Initialisation :
- prendre $x_0 = E_T(1/n, \ldots, 1/n)$;
- prendre une séquence de $N-1$ réalisations $\hat{\xi}[N-1] = (\hat{\xi}^0, \ldots, \hat{\xi}^{N-1})$ i.i.d du vecteur de variables aléatoires ξ ;
- prendre $\gamma = \frac{\sqrt{2E_T}\min\limits_r\{a_r\}}{\sqrt{N}\sqrt{2+(1+\ln(n))^2}}$.

Tant que $(1 \leqslant k \leqslant N-1)$:
prendre $k = k+1$;
appeler l'oracle 10 page 121 pour calculer un sous-gradient

$$G(x_{k-1}, \xi^{k-1}) \in \partial F(x_{k-1}, \xi^{k-1})\,;$$

prendre

$$(x_k)_r = \frac{(x_{k-1})_r e^{-\gamma(G(x_{k-1},\xi^{k-1}))_r}}{\sum\limits_{r=1}^{n}(x_{k-1})_r e^{-\gamma(G(x_{k-1},\xi^{k-1}))_r}}. \tag{6.52}$$

Fin de tant que

Renvoyer $\hat{x}_N = \hat{x}_N(\xi[N-1]) = \frac{1}{N}\sum\limits_{k=0}^{N-1} x_k$.

6.2.3 Retour sur les aspects stochastique et la grande dimension

Nous allons rebondir sur la discussion engagée dans l'introduction du chapitre concernant la nature de la difficulté de résoudre les problèmes d'optimisation stochastique. Nous allons illustrer grâce au problème (6.43), le fait que les problèmes d'optimisation stochastique portent en eux, du fait de leur nature, l'aspect de la grande dimensionnalité discutée dans le chapitre 5.

Pour ce faire, nous allons nous placer dans le cas où le vecteur aléatoire ξ a une densité de probabilité discrète avec un nombre K fini de valeurs possibles (ou scénarios) $\xi^1, \ldots, \xi^K$

ayant les probabilités respectives $p_1, \ldots, p_K$ avec $\sum_{k=1}^{K} p_k = 1$. Dans ce cas, nous avons :

$$\mathbb{E}[F(x,\xi)] = \sum_{k=1}^{K} p_k F(x,\xi^k).$$

La fonction $F(x,\xi) = \max_{r\in\{1,\ldots,N_v N_p\}} \{-x_r\xi_r/a_r\}$ peut être définie par le problème d'optimisation (6.53).

$$\begin{aligned} F(x,\xi) &= \min_{y} y \\ \text{tel que : } -x_r\xi_r/a_r &\leqslant y, \quad r = 1,\ldots,n \end{aligned} \tag{6.53}$$

Par conséquent, le problème (6.43) s'écrit sous la forme d'un grand problème d'optimisation linéaire déterministe :

$$\begin{aligned} \nu_T^* &= -\min_{x_r, y_k} \sum_{k=1}^{K} p_k y_k \\ \text{tel que : } -x_r\xi_r^k/a_r &\leqslant y_k, \quad r = 1,\ldots,n; k = 1,\ldots,K \end{aligned} \tag{6.54}$$

On rappelle que le vecteur aléatoire ξ appartient à $\mathbb{R}^n$ avec $n = N_v N_p$. Dans le cas où chaque coefficient peut prendre un nombre fini m de valeurs possibles, le nombre de scénarios K à prendre en compte devient $K = n^m$. Ainsi nous avons un problème d'optimisation avec $n + n^m$ variables de décision. Dans le cas de lois de probabilité continues, nous pouvons se ramener à approcher finement la loi de probabilité de ξ par une loi discrète. Cela nécessitera un nombre m important de valeurs possibles pour les ξ_r. Par suite, on se ramènera à un problème du type (6.54) avec un très grand nombre de variables de decision même pour un n assez faible. Par suite, l'utilisation des méthodes du première ordre et de la méthode SA ne perd pas de sa pertinence, ne serait ce que du point de vue théorique, et cela même si le nombre de variables de décision du problème (6.43), c'est-à-dire n, est assez petit.

6.2.4 Exemple de résolution numérique

Dans cette section, nous allons expérimenter les performances pratiques de la méthode MDSA présentée par l'algorithme 11 en considérant une instance du problème (6.43). En effet, nous allons passer à la résolution d'une instance du problème (6.43) en ayant pour objectif :

1 la $\epsilon-$précision : tester le pouvoir de l'algorithme pour calculer une solution numérique admissible du point de vue précision numérique moyenne ;
2 l'illustration de la rapidité de la convergence en $O(1/\sqrt{N})$ présentée dans le chapitre 5 ;
3 la $(\epsilon,\beta)-$précision : évaluer du point de vue pratique le niveau de marge de confiance que présente l'algorithme et le comparer avec celui théorique garanti par les équations sur le nombre d'itérations (6.20) et (6.23).

a) Description de l'expérience Nous avons effectué l'expérimentation numérique suivante : nous nous sommes fixé une instance du problème qui correspond à $N_p = 5$ et $N_v = 2$ soit un nombre de variables de décision $n = N_v N_p = 10$. Les paramètres a_i ont été générés aléatoirement entre 0 et 1. Nous avons choisi E_T de manière à ce que la valeur optimale du problème $\nu_T^* = \frac{E_T}{\left(\sum_{r=1}^n \sqrt{a_r}\right)^2}$ soit égale à 1, soit $E_T = \left(\sum_{r=1}^n \sqrt{a_r}\right)^2$. Ensuite, pour résoudre cette instance, nous avons fait tourner l'algorithme MDSA 11 avec un nombre d'itérations N allant de 1 à 1000. A l'issue de cette expérience nous obtenons une trajectoire, $\{f(\hat{x}_N)\}_{\{N=1,\dots,10^3\}}$, de l'algorithme sur N points. Nous avons ensuite répété cette expérience $N_e = 1000$ fois et nous avons calculé la moyenne empirique sur les N_e trajectoires obtenues afin d'estimer le terme $\mathbb{E}[f(\hat{x}_N)]$ qui intervient dans l'inégalité (6.16), caractérisant la rapidité de convergence de l'algorithme, ainsi que dans l'évaluation de la précision moyenne d'une solution numérique calculée par cet algorithme.

b) Premières observations Comme il s'agit initialement d'un problème de maximisation du flux et non pas d'un problème du minimum, nous avons tracé la convergence de l'estimation empirique de la trajectoire donnée par l'ensemble de points $\{-\mathbb{E}[f(\hat{x}_N)]\}_{\{N=1,\dots,10^3\}}$ vers la valeur optimale du problème 7 qui n'est que l'opposée de celle du problème (6.42). Nous notons qu'en se basant sur l'identité $-\min f = \max -f$, le problème (6.42) est équivalent au problème :

$$\nu_T^* = \max_x -f(x)$$
$$\text{tel que : } \sum_{r=1}^n x_r = E_T, x_r > 0,$$

avec

$$f(x) := \mathbb{E}\left[\max_{r\in\{1,\dots,N_v N_p\}} -x_r \xi_r / a_r\right].$$

Le résultat de cette simulation est donné sur la figure 6.2 et représente donc la trajectoire moyenne de la fonction objectif du problème 7 ou encore la trajectoire de l'espérance de cette fonction objectif.

Nous remarquons que la convergence vers la valeur optimale est lente : au bout de $N = 1000$ itérations, l'algorithme a amélioré la fonction objectif de moins de $0,05$. Ceci signifie que l'algorithme déplace lentement les x_k vers la solution optimale x^*. Ce déplacement est contrôlé par les sous-gradients. En effet, dans le cadre de l'optimisation déterministe, si on considère le problème (5.2) dans le cas où la fonction objectif n'est pas différentiable, alors ses sous-gradients $g(x)$ vérifient pour tout point x faisable, la propriété fondamentale suivante [48] :

$$\langle g(x), x - x^* \rangle \geqslant 0,$$

ce qui signifie que la distance entre x et x^* décroît dans la direction de $-g(x)$. Dans notre cas, ce principe s'applique sur l'équivalent déterministe de (6.43), c'est-à-dire le problème donné par :

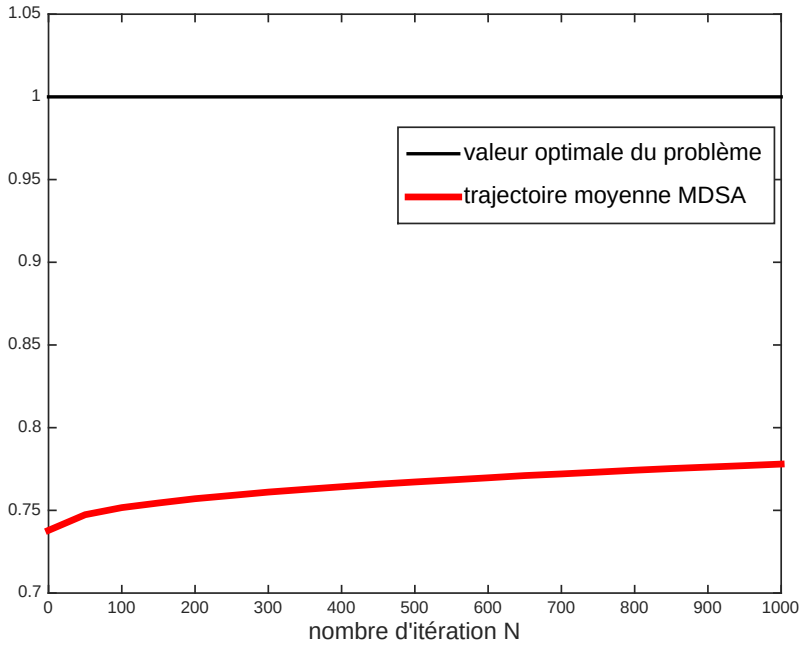

FIGURE 6.2 – Trajectoire moyenne $-\mathbb{E}[f(\hat{x}_N)]$ de la fonction objectif en fonction du nombre d'itérations N

$$\nu_T^* = -\min_x f(x) \tag{6.55}$$
$$\text{tel que : } \sum_{r=1}^{n} x_r = E_T, x_r > 0,$$

avec

$$f(x) := \mathbb{E}\left[\max_{r\in\{1,\dots,N_v N_p\}} -x_r\xi_r/a_r\right].$$

Dans ce cas, on a grâce à la condition (6.9) (qui est vérifiée pour nous) :

$$\mathbb{E}[G(x,\xi)] \in \partial f(x).$$

Ainsi la distance entre x^* (la solution de (6.43)) et x décroît dans la direction de $-\mathbb{E}[G(x,\xi)]$.

Afin de pouvoir illustrer ce phénomène, nous avons calculé la moyenne empirique sur les N_e sous-gradients obtenus à chaque itération, c'est-à-dire :

$$\frac{1}{N_e}\sum_{j=1}^{N_e}\left(G(x_{N,j},\xi^{N,j})\right)_{r=1,\dots,n} \tag{6.56}$$

avec $x_{N,j}$ un point calculé par l'algorithme 11 à l'itération N pendant l'expérience j, $\xi^{N,j}$ la réalisation du vecteur aléatoire ξ à l'itération N pendant l'expérience j et i l'indice pointant sur les coefficients du sous-gradient avec $n = 10$. La moyenne empirique (6.56) est calculée afin d'approcher l'espérance mathématique des cœfficients des sous-gradients car on a :

$$\mathbb{E}[(G(x_N, \xi^N))_r] \approx \frac{1}{N_e} \sum_{j=1}^{N_e} \left(G(x_{N,j}, \xi^{N,j})\right)_r \quad r = 1, \ldots, n,$$

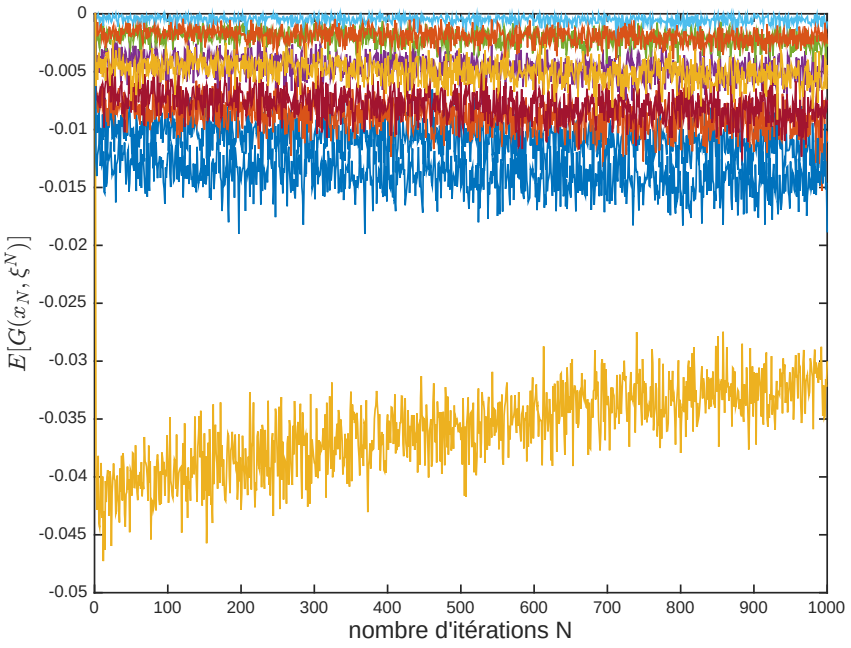

FIGURE 6.3 – Trajectoire moyenne $\mathbb{E}[(G(x_N, \xi^N))_r]$ des sous-gradients en fonction du nombre d'itérations N

Le résultat est tracé dans la figure 6.3. Nous remarquons sur la figure 6.3 que les sous-gradients ont des amplitudes faibles : de l'ordre de 10^{-3}, ce qui est négligeable devant les valeurs des cœfficients de la solution optimale x^* qui sont de l'ordre de 10 dans notre exemple. En plus au niveau des valeurs des cœfficients des sous-gradients, on remarque que chaque cœfficient varie autour d'une composante moyenne qui garde presque toujours la même valeur et il y a une composante avec une amplitude beaucoup plus importante (en valeur absolue) que celle des autres et sa variation est plus importante que les autres. Le fait d'avoir ce phénomène signifie que la décroissance de la distance entre les x_N et la solution optimale x^* est très lente du fait de la faiblesse des amplitudes des sous-gradients et du fait qu'il n'y a qu'une seule direction favorisée : celle qui correspond au cœfficient le plus prépondérant (en valeur absolue) des sous-gradients. Cette direction dans notre exemple est $e_3 = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$. Cette direction est favorisée par $\min_r\{a_r\}$ car les a_r interviennent dans les sous-gradients en $1/a_r$ (voir (6.47)). Donc, sur les N_e expériences, la moyenne des sous-gradients en chaque x_N aura toujours un cœfficient prépondérant en valeur absolue. L'indice de ce cœfficient est constant et égal à l'indice du plus petit a_r. Dans notre exemple $\min_r\{a_r\} = a_3$. Ainsi le déplacement des x_N vers x^* sera plus important pour $(x_N)_3$. Ce phénomène peut être affirmé également par la formule de la projection non linéaire de la MDSA (6.52) qui nous dit que plus les cœfficients des sous-gradients sont faibles, plus faible sera la variation au niveau des x_N et par suite l'évolution vers x^*. Ainsi, comme les sorties $\hat{x}_N$ ne sont que des moyennes sur les x_k précédents ($k = 1, \ldots, N$), alors le même phénomène est valide en moyenne et se répercutera sur les $\hat{x}_N$: la décroissance la plus rapide se fera dans la direction $((x_k)_3)$ et donc en moyenne, cette décroissance

sera toujours plus rapide dans la direction de $(\hat{x}_N)_3$. Cette conclusion est confirmée par la figure 6.4. En effet, on voit qu'il y a une seule composante de $\hat{x}_N$ qui prend des valeurs plus importantes que les autres composantes. Il s'agit de la 3^{eme} composante dans notre cas.

L'idéal serait que les cœfficients des sous-gradients évoluent vers des valeurs de plus en plus importantes tout en ayant des valeurs proches les unes des autres. Ceci permettrait aux x_N d'avoir des variations importantes mais aussi une évolution qui s'effectue à peu près dans toutes les directions et non pas dans une seule, ce qui permettrait d'aller plus vite vers x^*. Cette constatation peut être confirmée au regard des deux figures 6.2 et 6.3. En effet la tendance des cœfficients des sous-gradients, sur la figure 6.3, de se rejoindre au fil des itérations se traduit simultanément par une montée, sur la figure 6.2, de la trajectoire moyenne de l'algorithme vers la valeur optimale du problème.

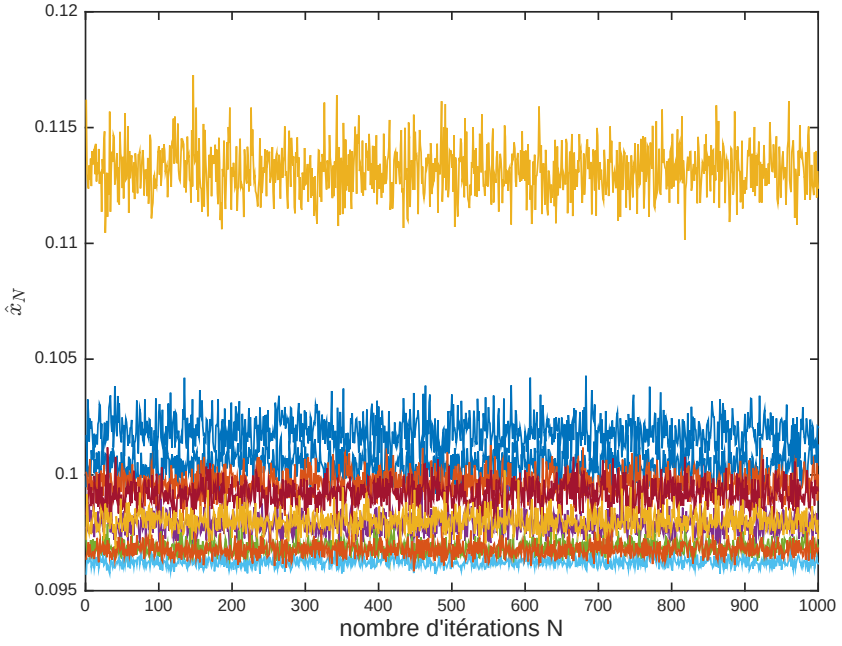

FIGURE 6.4 – Solutions numériques $\hat{x}_N$ en fonction du nombre d'itérations N

c) Le taux de convergence $O(1/\sqrt{N})$ et $\epsilon-$précision Pour conclure sur notre premier objectif de cette exploration numérique, la MDSA nous a calculé une solution avec une précision moyenne de $0,25$. Cette précision peut être améliorée bien sûr mais très lentement en fonction des itérations.

Par rapport à notre deuxième objectif de cette simulation, nous rappelons que le taux de convergence de $O(1/\sqrt{N})$ que présente la méthode est donné par la borne de convergence (6.16). Comme il s'agit à la base d'un problème de maximisation, nous allons adapter cette inégalité de la façon suivante

$$-\mathbb{E}[f(\hat{x}_N)] + f^* \geqslant -D_{\omega,X} L \sqrt{\frac{2}{\alpha N}}. \tag{6.57}$$

Nous avons tracé la trajectoire moyenne de l'algorithme ainsi que le terme δ_N minorant cette trajectoire :

$$\delta_N = -f^* - D_{\omega,X} L \sqrt{\frac{2}{\alpha N}}. \tag{6.58}$$

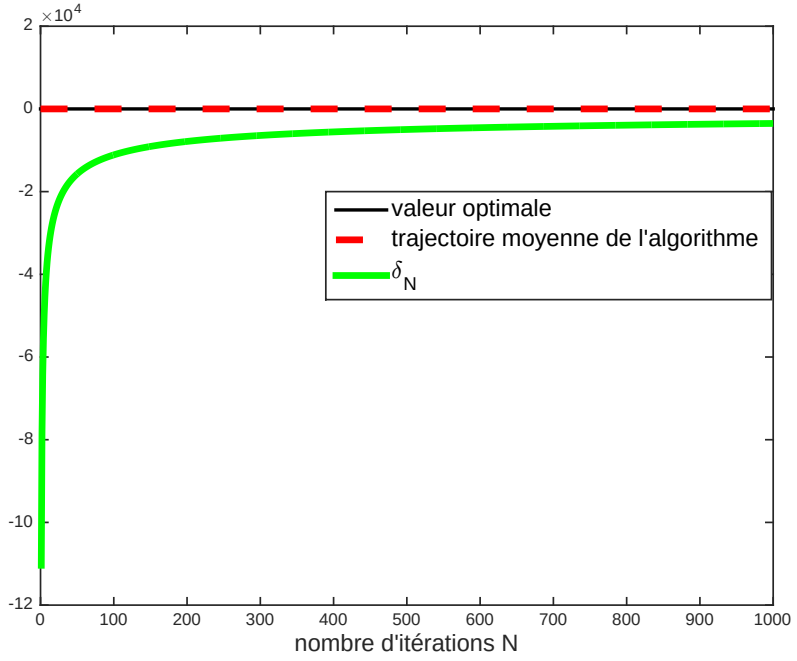

FIGURE 6.5 – Borne de convergence et trajectoire moyenne de l'algorithme en fonction du nombre d'itérations N

Le résultat donné dans la figure 6.5, montre que la propriété de convergence (6.16) ou, d'une façon équivalente, (6.57) est bien vérifiée dans notre cas. Toutefois du point de vue pratique, elle reste pessimiste et grossière. En effet d'après cette figure, nous constatons que la borne (6.58) nous indique un biais d'estimation ayant un ordre de grandeur de 10^4 pour N compris entre 1 et 1000 (la trajectoire moyenne de l'algorithme se trouve au dessus de la courbe δ_N avec un biais de l'ordre de 10^4).

d) La précision (ϵ, β) Comme cela a été présenté au chapitre 5, un nombre d'itérations garantissant à une solution, calculée par l'algorithme 11, d'être (ϵ, β) précise (voir définition 21) dépend essentiellement de la borne de convergence moyenne (6.16). Le fait que cette borne soit pessimiste dans la pratique entraîne que la précision (ϵ, β) théorique est également pessimiste. Pour examiner ce point, nous allons calculer la précision de la solution $\hat{x}_{1000}$ correspondant à une marge de confiance $1 - \beta = 95\%$. Nous calculons d'abord cette précision du point de vue théorique. Nous pouvons calculer deux niveaux de précision qu'on notera ϵ_{Markov} et $\epsilon_{L.D}$ la première est calculée en exploitant la borne sur le nombre d'itérations (6.20) issue de l'inégalité de Markov, la deuxième est calculée en exploitant la borne sur le nombre d'itérations (6.23) issue de la théorie des larges déviations. Ainsi on a :

$$\begin{aligned} \epsilon_{Markov} &= \frac{LD_{\omega,X}}{1-\beta}\sqrt{\frac{2}{\alpha N}}, \\ \epsilon_{LD} &= \frac{2\sqrt{2}(6-\ln\frac{1-\beta}{2})LD}{\sqrt{N}\sqrt{\alpha}}. \end{aligned} \tag{6.59}$$

Remarque : Afin d'exploiter le nombre d'itérations (6.23), il faut s'assurer que la condition (6.21) est satisfaite au moins sur les N_e trajectoires considérées dans notre expérience. Pour cela, nous avons calculé la quantité

$$g_{N_e} = \frac{1}{N_e}\sum_{r=1}^{N_e} \exp\left\{\left[\max_{i=1,\ldots,10}\left(G(x_{N,j},\xi^{N,j})_r\right)\right]^2/L^2\right\},$$

ce qui nous permet d'approcher le membre de gauche de l'inégalité (6.21). Nous avons calculé cette moyenne empirique en fonction du nombre d'itérations N. Le résultat est donné dans la figure 6.6.

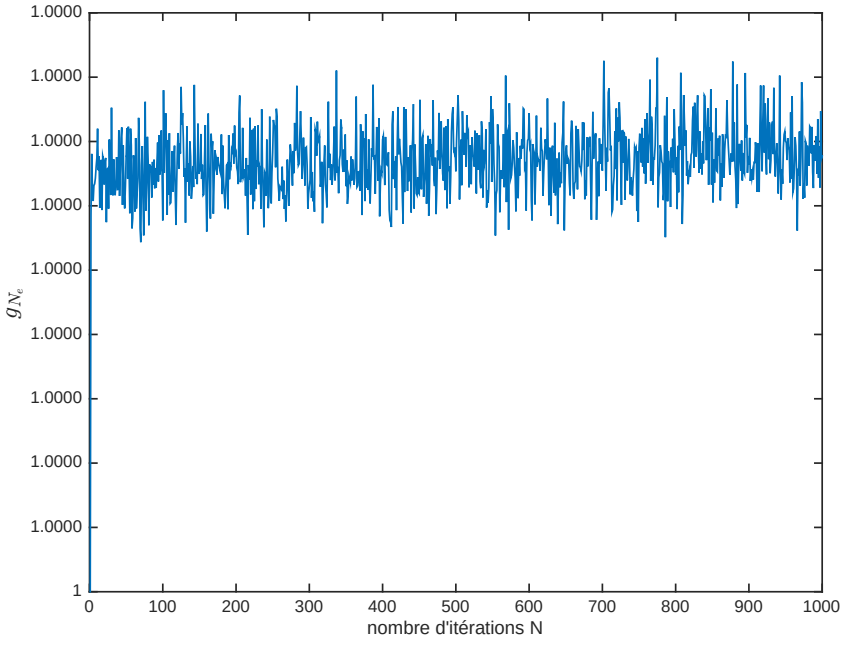

FIGURE 6.6 – Vérification de la condition (6.21) dans le cas de notre problème

On va bien sur la figure 6.6, que le terme g_{N_e} est bien inférieur à $\exp(1) \approx 2,7183$.

Du point de vue pratique, nous avons pris la valeur de $-f(\hat{x}_{1000})$ la plus petite parmi les 950 plus grandes valeurs. Cette valeur correspond à 0.7739 alors que $\nu_T^* = 1$. Ainsi au bout de 1000 itérations, l'algorithme a calculé une solution avec une précision expérimentale $\epsilon_{exp} = 0,2261$ et une marge de confiance de 95%. Nous avons appliqué la même procédure pour estimer des précisions expérimentales pour différentes valeurs de N. Les résultats comparant les précisions théoriques et expérimentales sont donnés dans le tableau 6.1. Ce tableau montre que les bornes (6.20) et (6.23) sur le nombre d'itérations permettant de

N	10^3	5.10^2	10^2
ϵ_{Markov}	45.4676	64.3009	143.7811
ϵ_{LD}	42.5581	60.1863	113, 08
ϵ_{exp}	0, 2261	0, 2389	0, 2531

TABLE 6.1 – Précision théorique vs précision expérimentale

garantir une solution (ϵ, β) précise ne sont pas utiles du point de vue pratique.

Inversement, nous avons obtenu une précision expérimentale $\epsilon_{exp} = 0, 2261$ avec une marge de confiance $\beta = 95\%$ pour un nombre d'itérations $N = 10^3$. Pour garantir la même précision et la même marge de confiance nous avons besoin, en appliquant l'inégalité (6.20), de $N_{Markov} = 1, 4599.10^{10}$ itérations au moins. Si on utilise l'inégalité (6.23), on a besoin de $N_{LD} = 1, 3704.10^{10}$ itérations au moins, ce qui est pratiquement prohibitif.

Le pessimisme de la borne de convergence moyenne illustré a rendu l'estimation théorique du nombre d'itérations nécessaire pour obtenir une solution (ϵ, β) précise, sans grande utilité pratique.

Dans la section suivante, nous allons expliquer la cause de ce niveau de conservatisme élevé de la méthode MDSA à travers l'analyse de sa borne de convergence (6.16).

6.2.5 Analyse de la borne de convergence

La comparaison de la borne de convergence (6.16), ou de façon équivalente (6.57), avec les résultats obtenus en simulation (voir figure 6.5) a montré un très haut niveau de conservatisme excluant son intérêt pratique. L'objectif de notre étude est d'identifier l'origine de ce conservatisme, puis de réfléchir à des modifications permettant d'obtenir un pouvoir prédictif en adéquation avec une utilisation pratique de l'algorithme.

6.2.5.1 Borne initiale

Nous nous plaçons dans le cadre du problème (6.1) et des hypothèses 4. D'après (2.44) de [46], l'algorithme MDSA garantit une erreur moyenne d'estimation ϵ vérifiant

$$\epsilon := \mathbb{E}[f(\hat{x}_N) - f(x^*)] \leqslant \frac{D_{\omega,X}^2 + \frac{1}{2\alpha} L^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N \gamma_k}. \tag{6.60}$$

Pour une stratégie pas fixe, c'est-à-dire $\gamma_k = \gamma$ pour tout k, on minimise le terme de droite de (6.60) par rapport à γ, en prenant comme pas optimal

$$\gamma = \frac{\sqrt{2\alpha} D_{\omega,X}}{L\sqrt{N}}, \tag{6.61}$$

ce qui amène à la borne finale proposée dans [46] (voir (2.46))

$$\epsilon \leqslant \frac{\sqrt{2/\alpha} D_{\omega,X} L}{\sqrt{N}}. \tag{6.62}$$

La figure 6.5 donne une idée du conservatisme de cette borne. Tandis que la valeur optimale du problème est de 1, la borne nous indique que le biais d'estimation sera de l'ordre de 10^4 sur les 1000 premières itérations (les trajectoires de l'algorithme se trouvent entre les deux courbes). A cette échelle, les résultats obtenus avec l'algorithme sont confondus avec la valeur optimale. Il y a ainsi un rapport de l'ordre de 10^4 entre la prédiction et l'observation.

Afin de comprendre l'origine de ce conservatisme, on a passé en revue les différentes étapes de la démonstration aboutissant à la borne, en commençant par les paramètres L et $D_{\omega,X}$.

6.2.5.2 Étude du L

a) Influence du paramètre Le paramètre L a une double influence, à la fois pratique et théorique. En pratique, c'est-à-dire au niveau de l'algorithme, il intervient dans la taille du pas réalisé à chaque itération (voir (6.61)). Une faible valeur de L engendre un pas important, donc une convergence plus rapide vers un état stationnaire possiblement plus bruité. En théorie, c'est-à-dire au niveau de l'erreur ϵ, un faible L rend la méthode plus performante. Toutefois L^2 est par définition un majorant du moment d'ordre 2 de la norme des sous-gradients de la fonction objectif sur l'ensemble faisable (voir (6.8)). L est ainsi une caractéristique imposée par la définition du problème.

Dans [46], le développement menant à la borne ϵ utilise la propriété

$$\sum_{k=1}^{N} \mathbb{E}\left[\|G(x_k, \xi_k)\|_*^2\right] \leqslant NL^2, \tag{6.63}$$

qui découle de la définition (6.8). C'est donc ici une approximation de type « pire-cas ». Le cas d'égalité correspond à une trajectoire immobile sur le point de X pour lequel l'espérance du gradient est maximale.

Notons L_* le scalaire associé à une trajectoire de l'algorithme vérifiant

$$\sum_{t=1}^{N} \mathbb{E}\left[\|G(x_t, \xi_t)\|_*^2\right] = NL_*^2. \tag{6.64}$$

Sur l'ensemble des simulations réalisées pour le problème de production des protéines, nous observons que le rapport L/L_* est de l'ordre de 10^2. Ce rapport donne l'ordre de grandeur du conservatisme introduit par l'utilisation de L.

b) Borne a posteriori L'idéal serait de remplacer directement L par L_* à la fois dans le calcul de la borne ϵ et dans le calcul du pas γ de l'algorithme. Cependant la valeur de L_* est liée à une trajectoire de l'algorithme et ne peut donc être connue qu'a posteriori. L'idée est alors de découpler l'aspect théorique (la borne) de l'aspect pratique (le pas de l'algorithme).

Concernant la borne, le calcul a posteriori de L_* nous permet d'exprimer une borne a posteriori, notée ϵ_{post}. Si cette borne ne nous permet pas de prédire le nombre d'itérations N nécessaire pour atteindre une précision donnée, elle nous permettra a posteriori de

dire la précision obtenue sur des simulations réalisées avec une nombre d'itérations N. La borne a posteriori qui nous intéressera dorénavant s'écrit ainsi

$$\epsilon_{\text{post}} \triangleq \mathbb{E}[f(\hat{x}_N) - f(x^*)|x_1, ..., x_N] \leqslant \frac{D_{\omega,X}^2 + \frac{1}{2\alpha} L_*^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N \gamma_k}, \tag{6.65}$$

Concernant l'aspect pratique, L_* n'étant connu qu'a posteriori, il ne peut pas être utilisé pour définir le pas de l'algorithme. Il faut donc avoir recours à un choix a priori, noté $\tilde{L}$, permettant de définir le pas de l'algorithme selon

$$\gamma = \frac{\sqrt{2\alpha} D_{\omega,X}}{\tilde{L}\sqrt{N}}. \tag{6.66}$$

En introduisant ce nouveau pas dans le calcul de la borne donnée en (6.65), on obtient la borne a posteriori

$$\epsilon_{\text{post}} \leqslant \frac{D_{\omega,X}\tilde{L}}{\sqrt{2\alpha}\sqrt{N}} \left[1 + \left(\frac{L_*}{\tilde{L}}\right)^2\right]. \tag{6.67}$$

Cette borne atteint son minimum lorsque $\tilde{L} = L_*$ et vaut alors

$$\epsilon_{\text{post}}^* = \frac{\sqrt{2/\alpha} D_{\omega,X} L_*}{\sqrt{N}}. \tag{6.68}$$

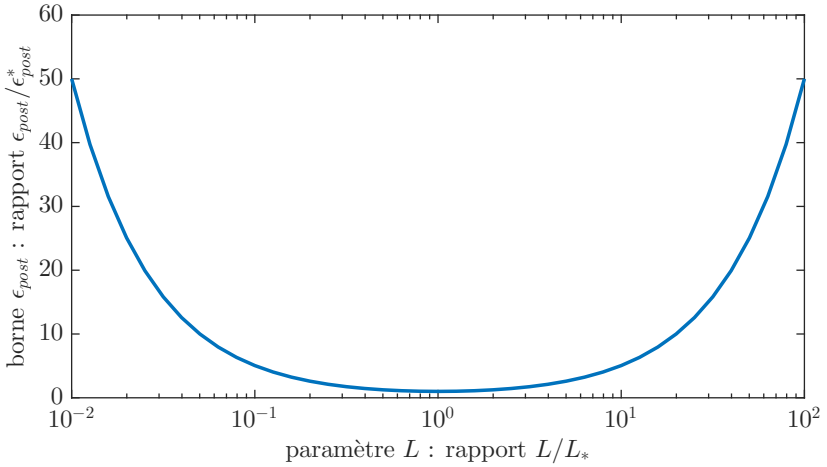

FIGURE 6.7 – Impact sur la borne ϵ_{post} d'une erreur sur le choix (l'estimation) de $\tilde{L}$

La figure 6.7 illustre la dégradation sur la borne engendrée par un mauvais choix de $\tilde{L}$. Le rapport $\epsilon_{\text{post}}/\epsilon_{\text{post}}^*$ est tracé en fonction du rapport $\tilde{L}/L_*$. On note un maintient raisonnable des performances pour des rapports $\tilde{L}/L_*$ compris entre 0.1 et 10. Au delà, une forte dégradation est observée. Il semble donc pertinent de mettre en place une stratégie permettant de donner à $\tilde{L}$ une valeur proche de L_*.

c) Choix de $\tilde{L}$ Nous n'avons pas mis en place de stratégie proprement fondée pour le choix de ce paramètre $\tilde{L}$. Cependant, pour l'ensemble des simulations réalisées sur le problème de production de protéines, il s'avère que le choix de

$$\tilde{L} = \frac{2}{\left(\sum_{r=1}^n \frac{c_r}{k_r}\right)^2} \tag{6.69}$$

est satisfaisant. Pour illustration, la figure 6.8 indique la borne a posteriori obtenue en utilisant ce choix de $\tilde{L}$ ainsi que la borne a posteriori optimale utilisant $\tilde{L} = L_*$. En comparaison avec la figure 6.5, un rapport 100 pour ϵ^*_{post} et un rapport 10 pour ϵ_{post} ont été gagnés.

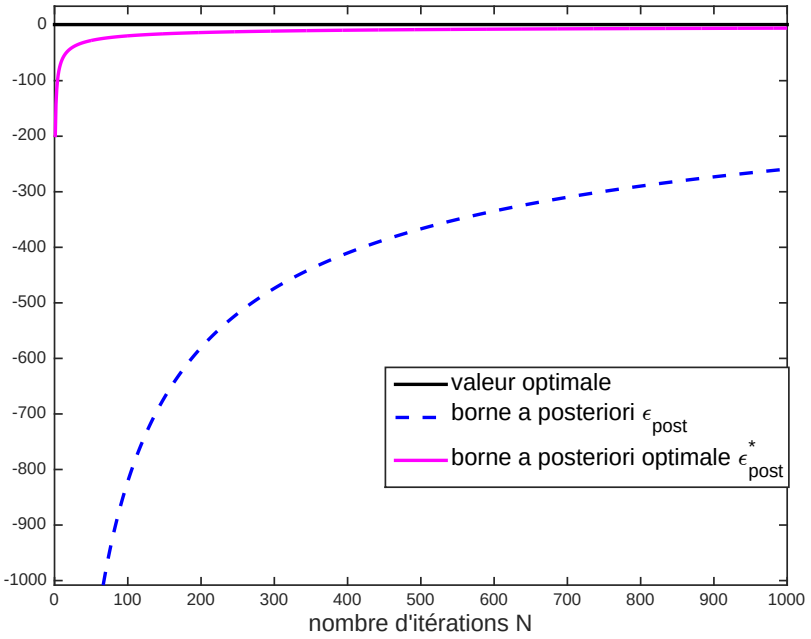

FIGURE 6.8 – Solution du problème et borne a posteriori pour $N = 1000$

6.2.5.3 Étude du $D_{\omega,X}$

a) Influence du paramètre Le paramètre $D_{\omega,X}$, défini en (6.17), intervient de façon proportionnelle dans l'expression de la borne (voir (6.67)) ainsi que dans le pas de l'algorithme (voir (6.66)). Ce paramètre est utilisé pour majorer la distance (de Bregman) entre le point initial de l'algorithme x_0 et la solution x^*.

Dans la démonstration de Nemirovski [46], cette majoration est réalisée en deux temps. Une première majoration intervient entre l'équation (2.36) et (2.37), à savoir

$$\sum_{k=0}^{N-1} \omega_{x_k}(x^*) - \omega_{x_{k+1}}(x^*) = \omega_{x_0}(x^*) - \omega_{x_N}(x^*) \leqslant \omega_{x_0}(x^*), \tag{6.70}$$

qui se justifie du fait que la distance de Bregman $\omega_\cdot(.)$ est toujours positive ou nulle. Cette première majoration a un impact limité car lorsque l'algorithme converge, $\hat{x}_N$ se rapproche de la solution x^* et $\omega_{x_N}(x^*)$ tend à s'annuler. L'impact pourrait éventuellement devenir non négligeable pour des petits N. Pour $N = 10^4$, on observe un rapport $\omega_{x_0}(x^*)/\omega_{x_N}(x^*)$ de l'ordre de 100, soit un conservatisme de 1%. Dans un second temps, $D^2_{\omega,X}$ est choisi de façon à majorer $\omega_{x_0}(x^*)$ quelle que soit la solution x^* du problème (cf définition (6.18)), ce qui permet d'obtenir la double inégalité finale

$$\sum_{k=0}^{N-1} \omega_{x_k}(x^*) - \omega_{x_{k+1}}(x^*) \leqslant \omega_{x_0}(x^*) \leqslant D^2_{\omega,X}. \tag{6.71}$$

L'impact de cette seconde majoration est possiblement plus important. En effet, pour $x_0 = E_T[1/n, ..., 1/n]$, les solutions majorant la distance $\omega_{x_0}(x^*)$ sont les points admettant une unique composante non-nulle, soit par exemple $x^* = E_T[1, 0, ..., 0]$. On obtient alors $D^2_{\omega,X} = E_T \ln(n)$. Dans notre cadre biologique, cela signifie que seule une voie métabolique est utilisée. Sur l'ensemble des problèmes testés pour lesquels les paramètres de coût et d'efficacité enzymatique (c_i et k_i respectivement) sont tirés de façon aléatoire, on observe que le rapport $D^2_{\omega,X}/\omega_{x_0}(x^*)$ est de l'ordre de 5.

b) Amélioration possible ? Le paramètre $D_{\omega,X}$ est profondément dépendant de la solution du problème, aussi il n'est pas simple d'en construire un estimateur a priori permettant de réduire le conservatisme qu'il génère.

Pour simplifier notre présentation, on se ramène au simplexe unitaire, c'est-à-dire on considère le cas où $E_T = 1$ (cela se fait sans perte de généralité via un simple changement de variable normalisant sur E_T). Une première possibilité serait néanmoins d'utiliser un critère de régularité sur la solution nous permettant de dire que l'on ira jamais chercher les points extrêmes : sauf cas trivial, on n'aura jamais $x^* = [1, 0, ..., 0]$. Une idée de cette régularité peut possiblement être a priori obtenue en observant le vecteur $\{a_r = c_r/k_r\}_{r=1,...,n}$. On définit l'entropie de a (vue comme une loi de probabilité) par :

$$H(a) = -\sum_{i=1}^{n} a_i \ln a_i. \tag{6.72}$$

Sur le simplexe unitaire, on note que la distance de Bregman $\omega_x(y)$ est par définition la divergence de Kullback-Liebler $\mathcal{D}(x\|y)$ entre les lois de probabilité x et y, soit

$$\omega_x(y) = \sum_{r=1}^{n} x_r \ln \frac{x_r}{y_r} = \mathcal{D}(x\|y). \tag{6.73}$$

Or, pour x une loi quelconque et $y = u$ la loi uniforme à n états, on a l'égalité

$$\mathcal{D}(x\|u) = \ln n - H(x). \tag{6.74}$$

Si l'on montre maintenant que l'entropie de la solution x^* vérifie

$$H(x^*) \geqslant H(a), \tag{6.75}$$

ce qui signifie que le vecteur solution x^* est au moins aussi régulier que le vecteur a, on obtient l'inégalité

$$\omega_{x_0}(x^*) = \mathcal{D}(x^*\|x_0) = \ln n - H(x^*) \leqslant \ln n - H(a), \tag{6.76}$$

dans laquelle on utilise le fait que le point d'initialisation $x_0 = [1/n, ..., 1/n]$ correspond à la loi uniforme. L'intérêt de ce développement est bien entendu le fait que la quantité $\ln n - H(a)$ est connue a priori. Par contre, tout repose sur l'inégalité (6.75) qui reste à montrer en général mais qui est vérifiée dans le cas de notre exemple numérique. Il s'agit alors d'une approche a posteriori permettant d'avoir une estimation empirique du terme $D_{\omega,X}$ en tant que majorant du terme $\omega_{x_0}(x^*)$ (voir équation (6.71)).

6.2.5.4 Étude de la convexité

a) Majorations utilisées Le troisième élément de cette étude est l'analyse d'une double majoration utilisée dans la démonstration de la borne se basant sur la convexité de la fonction objectif. Cette double majoration est celle permettant de passer de l'équation (2.37) à l'équation (2.39) de [46]. La première inégalité implique que f est toujours au dessus de sa tangente en x_k, soit

$$A = \sum_{k=1}^{N} \gamma_k (x_k - x^*)^T g(x_k) \geqslant \sum_{k=1}^{N} \gamma_k \left[f(x_k) - f(x^*) \right] = B \tag{6.77}$$

avec g le sous-gradient de f. La seconde inégalité implique la définition de la convexité de f :

$$B = \sum_{k=1}^{N} \gamma_k \left[f(x_k) - f(x^*) \right] \geqslant \left(\sum_{k=1}^{N} \gamma_k \right) \left[f(\hat{x}_N) - f(x^*) \right] = C, \tag{6.78}$$

dans laquelle on introduit le moyennage final sur les x_k, à savoir

$$\hat{x}_N = \frac{1}{\sum_{k=1}^{N} \gamma_k} \sum_{k=1}^{N} \gamma_k x_k. \tag{6.79}$$

La majoration utilisée pour établir la borne est finalement

$$C \leqslant A. \tag{6.80}$$

b) Impact de ces majorations Nous avons évalué le degré de conservatisme introduit par cette majoration en calculant le rapport A/C. Sur l'ensemble des tests réalisés, nous observons un rapport de l'ordre de 40.

Afin de corriger ce conservatisme au niveau de la borne, nous introduisons la *borne a posteriori après correction ABC* définie selon

$$\epsilon_{\text{post}}^{ABC} = \frac{C}{A} \epsilon_{\text{post}} \tag{6.81}$$

Pour illustration, la figure 6.9 indique le résultat moyen de l'algorithme pour N allant de 100 à 1000, ainsi que la borne a posteriori après correction ABC. En observant le niveau de similitude entre la « prédiction » et l'observation, on déduit que les principaux facteurs de conservatisme ont été identifiés.

c) Amélioration possible ? Le conservatisme introduit par cette majoration est entièrement déterminé par la nature de la fonction objectif. Il n'est donc plus possible d'agir dessus une fois le problème fixé. On peut tout de même faire deux remarques :

On notera premièrement que les « meilleures » fonctions objectifs, c'est-à-dire celles amenant des rapports A/C proches de 1 sont les fonctions linéaires.

On ajoute ensuite que dans la démonstration de la borne, la double majoration ci-dessus est essentielle car elle permet de passer d'un terme impliquant les points x_k (terme A) à

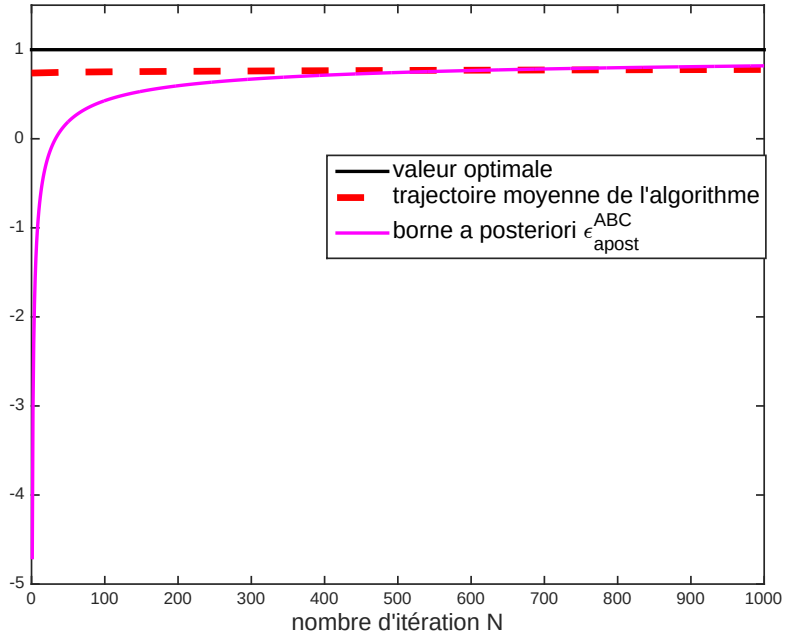

FIGURE 6.9 – Solution d'une instance du problème, convergence de l'algorithme MDSA et borne a posteriori après correction ABC pour $N = 1000$

un terme impliquant la valeur finale $f(\hat{x}_N)$ (terme C) pour laquelle la borne est exprimée. Cependant, cette étape n'est peut-être pas obligatoire si nous souhaitons exprimer une borne non plus sur la valeur finale $f(\hat{x}_N)$ mais sur le point solution $\hat{x}_N$ comme cela est le cas pour le problème de production des protéines.

6.2.5.5 Bilan

Nous avons étudié trois facteurs de conservatisme dans la borne théorique proposée dans [46] pour l'algorithme MDSA. Le premier concerne l'utilisation de L. Il introduit un conservatisme de l'ordre de 10^2. Pour y remédier, le calcul d'une borne a posteriori a été proposé ainsi qu'une possibilité pour définir à priori le pas de l'algorithme. Le second concerne le paramètre $D_{\omega,X}$. Il introduit un conservatisme de l'ordre de 5. Une piste a été soulevée pour améliorer ce facteur. Le troisième facteur concerne la convexité de la fonction objectif. Il introduit un conservatisme de l'ordre de 40. Concernant ce facteur, aucune solution n'est envisageable une fois le problème fixé.

6.2.6 Discussion sur l'impact de la dimension

Nous rappelons que l'un des avantages des méthodes du premier ordre est que le nombre total d'itérations nécessaire pour atteindre une solution $\epsilon-$précise est quasi-indépendant voir indépendant du nombre de variables de décision. Nous avons montré que cette propriété a été héritée dans le cadre de la méthode SA et en particulier pour la MDSA. En effet, si on reprend la borne de convergence (6.62), on remarque que la dépendance par rapport au nombre de variables de décision n, ne peut intervenir qu'au niveau des paramètres $D_{\omega,X}$ et L. En fonction des problèmes, ces paramètres peuvent même être

indépendants de n. Dans notre cas d'étude, ces paramètres ne dépendent que légèrement de n : nous avons une dépendance en $\ln(n)$ (voir (6.51) et (6.44)). Cela fait que la borne de convergence (6.62) spécifique à notre problème ne dépend que légèrement de n. En effet la borne (6.62) est donnée explicitement par

$$\epsilon \leqslant \frac{\sqrt{2E_T}}{\min_r\{a_r\}} \frac{1}{\sqrt{N}} \sqrt{\ln(n)\left(2 + (1 + \ln(n))^2\right)}. \tag{6.82}$$

Du point de vue de la complexité algorithmique, c'est ici un avantage par rapport aux méthodes polynomiales qui présentent une dépendance polynomiale en n. Cela peut faire une très grande différence notamment pour n grand.

Du point de vue pratique et par rapport à la discussion engagée dans le paragraphe b) de la section 6.2.4 sur la nature du problème (6.43), nous avons mis en évidence le phénomène suivant :
le fait d'avoir un vecteur a avec des écarts importants entre ses cœfficients fait que la moyenne des sous gradients a toujours un seul cœfficient prépondérant en valeur absolue. L'indice de ce cœfficient correspondant à $\min a_i$. Cela a pour effet de favoriser l'évolution des x_N, vers x^*, dans une seule direction. Ce phénomène est d'autant plus important quand la dimension du problème est importante car le fait de déplacer les x_N suivant une seule dimension (direction) rend la décroissance de la distance entre les x_N et x^* encore plus lente que dans un espace de faible dimension. Il s'agit d'une difficulté liée aux sous-gradients et donc à la nature même du problème. Malheureusement, il n'y a aucune remède ce stade car le problème est fixé.
Nous avons résolu le problème dans le cas de plusieurs valeurs de n. Les paramètres a_i ont été générés aléatoirement entre 0 et 1. Nous avons choisi E_T de manière à ce que la valeur optimale du problème $\nu_T^* = \frac{E_T}{\left(\sum_{r=1}^n \sqrt{a_r}\right)^2}$ soit égale à 1, soit $E_T = \left(\sum_{r=1}^n \sqrt{a_r}\right)^2$ et nous avons pris $N = 1000$. Le résultat est donné dans la figure 6.10.
La dégradation de la précision numérique illustrée dans la figure 6.10 est l'effet de la combinaison de deux causes. La première cause est intrinsèque à la méthode. Il s'agit de son efficacité d'estimation qui se dégrade en $O(\ln(n)\sqrt{\ln(n)})$. Mais aussi la nature « pathologique » du problème détecté au niveau de ses sous-gradients.

6.2.7 Amélioration de la vitesse de convergence

La nature de notre problème ainsi que les observations issues de l'exemple numérique motivent une modification de la MDSA afin d'améliorer la vitesse de convergence, si ce n'est à un niveau garanti par une borne théorique, au niveau pratique. Cette modification nous permettra d'aborder des problèmes d'allocation pour un réseau métabolique de dimension plus élevée (10^3 pour l'exemple traité dans le cadre de cette thèse). Cette modification porte sur deux aspects :

Pas γ_k cette modification se base sur la remarque de la section 6.2.5.2. Au lieu d'être utilisée pour obtenir une borne *a posteriori*, elle sera utilisée afin de proposer un réglage du pas de l'algorithme MDSA ;

Sortie $\hat{x}_N$ en fonction de la trajectoire cette modification est mentionnée dans l'article initial [46] et consiste à moyenner la trajectoire de l'algorithme non pas à partir de la première itération mais à partir d'une itération plus tardive. La raison intuitive derrière cette modification est qu'*a priori* la première partie de la

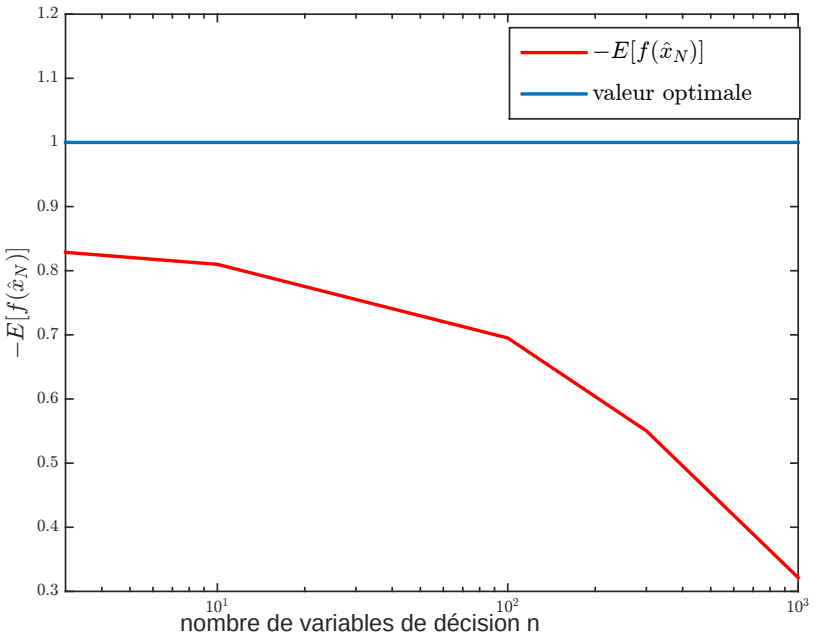

FIGURE 6.10 – Dégradation de la $\epsilon-$précision en fonction de n

trajectoire est plus éloignée de la solution optimale que la seconde ; la solution obtenue en moyennant la seconde partie de la trajectoire uniquement serait donc plus intéressante que celle obtenue en moyennant sur toute la trajectoire.

Si sur le calcul du pas γ_k, nous proposons une raison rationnelle à la modification proposée, le choix de l'itération à partir de laquelle le moyennage se fera est arbitraire (nous avons choisi de couper la trajectoire en deux parties égales). Nous présentons maintenant la modification sur le calcul du pas.

Comme évoqué précédemment dans la section 6.2.5.2, la borne de convergence obtenue dans [46] correspond à un pire cas, celui où la trajectoire de l'algorithme serait telle qu'à chaque itération la norme de l'espérance du gradient est égale à la constante de Lipschitz. Cette borne est utilisée pour obtenir le pas constant « optimal » qui la minimise. Pour un nombre d'itérations N donné, elle est donnée par (6.66), rappelé ici :

$$\gamma = \frac{\sqrt{2\alpha} D_{\omega,X}}{\tilde{L}\sqrt{N}}$$

avec $\tilde{L} = L/\theta$ égal à la constante de Lipschitz L, *i.e.* $\theta = 1$.

Elle garantit par ailleurs la convergence en moyenne vers la solution optimale, ce qui est contradictoire avec la trajectore pire cas utilisée pour l'obtenir. En effet, en pratique, le pas constant utilisé est obtenu à partir d'une valeur de $\tilde{L}$ différente de L [46, 38], c'est-à-dire $\theta \neq 1$.
L'analyse développée dans la section 6.2.5.2 nous a permis de proposer une stratégie d'amélioration de la MDSA basée sur un choix judicieux de θ dans l'article présenté dans le chapitre Annexe du document de thèse. Elle a permis de calculer la solution du problème d'optimisation dans le cas d'un réseau de grande dimension (1000 enzymes),

avec une bonne précision (erreur relative inférieure à un pourcent) et un temps de calcul raisonnable (de l'ordre de 10h).

6.3 Conclusion

Dans ce chapitre nous avons présenté des définitions d'une solution pour un problème d'optimisation stochastique et avons discuté leur pertinence du point de vue pratique ainsi que du point de vue théorique. Nous avons également présenté deux approches pour la résolution de ce problème : la SA et la SAA. Nous avons montré que dans le cadre de la SA nous pouvons intégré, avec l'hypothèse 3 (convergence moyenne), des méthodes du premier ordre déterministes tout en héritant leur efficacité d'estimation. Nous avons montré par la même occasion que le fait d'intégrer ces méthodes du premier ordre au sein de la SA a permis de faire émerger un cadre d'analyse de complexité algorithmique, pour le problème d'optimisation stochastique (6.1), basé sur celui du cas déterministe ainsi que sur les inégalités de probabilité que nous avons présentées pour la MDSA et avons démontrées pour la PDSA. Nous avons présenté la méthodologie SAA qui, contrairement à la méthode SA, n'est pas une méthode de résolution numérique, mais plutôt une approche qui permet de ramener la résolution de (6.1) à celle de (6.38). Encore faut-il choisir une méthode numérique déterministe pour le résoudre. Nous avons comparé la SAA et la SA en terme de dépendance en le nombre de variables de décision n et en le nombre d'appels total à l'oracle du premier ordre. Nous avons montré que selon ces deux critères la méthode SA à travers ses variantes MDSA et PDSA s'est avérée la plus efficace.

Fort de cette comparaison, nous avons exploré la résolution numérique par la MDSA du problème d'optimisation correspondant au modèle stochastique du réseau introduit dans le chapitre 4. A travers ce problème, nous avons illustré le degré de conservatisme que présente la borne de convergence de l'algorithme. Nous avons montré que la précision numérique est impactée par la dimension du problème. Ici il y a un enchainement de deux difficultés qui ont pour effet la dégradation de la précision en fonction de n. La première difficulté est de nature algorithmique, la deuxième est liée à la nature « pathologique » du problème qui réduit la puissance des sous-gradients à améliorer la fonction objectif. La conséquence est que les résultats obtenus lors de l'évaluation du MDSA sur un problème de taille 10, c'est-à-dire faible par rapport à une taille réaliste pour un réseau métabolique, sont très décevants. Cependant l'analyse des résultats obtenus a permis de proposer des améliorations pratiques de cet algorithme. Les résultats montrent une amélioration importante des performances de l'algorithme ainsi que la possibilité de s'attaquer à des réseaux de taille réaliste (1000 enzymes). Ces résultats se trouvent dans un article en soumission au prochain Conference on Decision and Control qui a été joint en annexe de cette thèse.

Chapitre 7

Conclusion générale et perspectives

Cette thèse s'inscrit dans le cadre de l'extension de la méthodologie RBA afin d'améliorer son pouvoir prédictif. Cet objectif est inséparable du fait que le modèle RBA doit garantir un bon compromis entre la complexité des phénomènes physiques intégrés et l'efficacité de sa résolution numérique. Un modèle complexe et représentatif de la réalité reste inexploitable tant qu'il ne peut pas être traité et résolu efficacement. Un modèle simple est certes efficacement résoluble mais peu représentatif de la réalité. Nous avons, durant cette thèse, exploré la possibilité d'atteindre ce compromis.

La première partie a été consacrée à l'extension du cadre RBA du point de vue de la modélisation et l'intégration des contraintes thermodynamiques et cinétiques régissant les réactions biochimiques au sein d'un réseau métabolique. Dans le chapitre 1 une extension du cadre RBA a été proposée en se basant sur la loi empirique (2.26), page 17, ce qui a permis d'intégrer l'aspect thermodynamique et cinétique dans les RBA à travers la nouvelle contrainte structurelle (2.40) page 21. Cette contrainte représente le couplage existant entre flux métabolique, concentration des métabolites et concentration des enzymes. Ce couplage fixe des contraintes sur les concentrations des métabolites et des enzymes rendant possible un flux métabolique du point de vue thermodynamique et cinétique. Certes cela présente une amélioration du modèle RBA classique car les contraintes thermodynamiques et cinétiques sont fortement présentes dans les interactions biochimiques des réseaux métaboliques, mais cela conduit un problème d'optimisation non convexe. Par conséquent à ce stade, l'intégration de cette nouvelle contrainte structurelle au sein du cadre RBA a conduit en une perte de l'efficacité numérique garantie pour les RBA classiques. Comme notre extension doit maintenir l'avantage de l'efficacité de résolution numérique, le deuxième axe de notre contribution de cette première partie de la thèse était de proposer une stratégie permettant d'approcher le modèle RBA non convexe par un modèle approché convexe. Le problème résultant de cette stratégie appartient à la classe d'optimisation géométrique. Cette stratégie nous a permis de garder la structure fondamentale des nouvelles contraintes structurelles non convexes tout en se plaçant dans la convexité et en bénéficiant des propriétés sympatiques de la classe de l'optimisation géométrique que nous avons présentée au chapitre 2. A travers l'exemple numérique traité dans le même chapitre nous nous sommes arrêtés sur les limitations d'une telle approximation. Nous avons montré à travers ce cas d'étude que la relaxation que nous avons proposée pour les contraintes (3.20) a pour effet de ne pas garantir l'admissibilité thermodynamique aux flux métaboliques ν_i^+ et ν_i^-. Pour pallier cette difficulté nous proposons de garder une cohérence dans les ordres de grandeur entre ν_i^+ et $e^{-s_i^+}$ d'un coté et ν_i^- et $e^{-s_i^-}$

de l'autre. Nous pouvons assimiler les variables ($e^{-s_i^+} = 1/\mu_i^+$) et ($e^{-s_i^-} = 1/\mu_i^-$) à des flux métaboliques admissibles. Ainsi nous proposons de revenir sur la relaxation formelle des contraintes (2.40) en les remplaçant uniquement par les contraintes inégalité du type :

$$\mu_i^+ \nu_i^+ \leqslant 1,$$

ce qui revient à dire que les flux métaboliques sont toujours limités par les flux admissibles cinétiquement et thermodynamiquement. Une fois que la pertinence de cette contrainte est validée, nous pouvons à partir de là leur construire une relaxation basée sur les stratégies de l'optimisation globales proposées dans [67, 34].

La deuxième partie a été consacrée à la prise en compte de l'aspect stochastique au sein du modèle RBA classique. Le problème ainsi construit est maintenant un problème d'optimisation stochastique appartenant à la classe de problèmes définie par (6.1) page 99. Le problème RBA stochastique est un problème de grande dimension. Nous avons montré dans le chapitre 5, à travers le cas d'étude, le lien étroit entre l'aspect stochastique et la très grande dimension. Dans ce cas les méthodes polynomiales s'avèrent peu efficaces. Nous avons présenté dans les chapitres 3 et 4 la bibliographie sur les méthodes du premier ordre. Nous avons montré qu'elles présentent la propriété importante d'être indépendantes ou quasi-indépendantes du nombre des variables de décision du problème d'optimisation en question. Cette propriété qui n'est pas vérifiée pour les méthodes polynômiales les rend moins efficaces que les méthodes du premier ordre s'agissant d'un problème de très grande dimension. Nous avons également proposé un cadre pour l'analyse systématique de la convergence des MPO à travers une interprétation systémique. En effet, nous pouvons interpréter les MPO comme des systèmes dynamiques composés d'interconnexions de sous-systèmes non-linéaires possédant certaines propriétés dynamiques à savoir la vérification d'inégalités de dissipativité. Ceci permet de conclure sur leur stabilité et par conséquent démontrer la convergence de ces MPO.

Dans le chapitre 4 nous avons présenté deux approches différentes pour la résolution des problèmes d'optimisation stochastique. La première approche consiste en l'intégration des MPO déterministes au sein du processus stochastique associé à l'algorithme 6 page 100 basé sur un oracle stochasticisé du premier ordre. Il s'agit de la méthode « *Stochastic Approximation* » (SA). Nous avons montré que toute MPO admet une version stochastique sous réserve de vérifier une certaine condition de convergence moyenne. Les méthodes de résolution numériques présentées dans ce chapitre étant des versions stochasticisées de MPO déterministes, nous avons montré que leur complexité algorithmique est hérité de leur homologue, les MPO déterministes. Dans ce chapitre nous avons étudié la complexité algorithmique de la méthode PDSA et avons montré que sa complexité est équivalente à la MDSA. La deuxième approche consiste à formuler à partir du problème stochastique un problème déterministe dont les solutions numériques approchent celles du problème stochastique. Il s'agit de l'approche « *Stochastic Average Approximation* » (SAA). Nous avons dressé un comparatif de complexité algorithmique entre les différentes approches permettant de résoudre des problèmes d'optimisation stochastique : l'approche SA s'avère plus efficace que la SAA.

Dans les chapitre 5 nous nous sommes proposé d'étudier et de résoudre un cas d'étude spécifique issu de la méthode RBA. Nous avons résolu le problème analytiquement dans un cadre stochastique. Afin de pousser la compréhension d'une telle solution nous avons aussi résolu le problème dans le cas déterministe et avons comparé les solutions obtenues

dans les deux contextes. Nous avons montré que contrairement au contexte déterministe et afin d'obtenir les mêmes solutions, nous devons engager plus de ressources en protéines dans le cas stochastique. Nous avons également montré que la nature de la solution dans le cas stochastique est plus complexe que dans le cas déterministe dans la mesure où la première dépend de la structure globale du réseau métabolique en question. Nous avons également mis en œuvre la méthode MDSA pour résoudre notre cas d'étude. Nous avons vérifié que notre cas d'étude peut être intégré au sein de la méthode MDSA et avons calculé une estimation de tous les paramètres nécessaires à l'algorithme dans notre cas. Nous nous sommes également arrêtés sur les performances de cette algorithme par rapport à notre cas d'étude et avons montré certaines limitations de l'algorithme. Ces limitations sont de deux natures différentes, il y a des limitation liées à la nature numérique même de notre problème d'optimisation ainsi que d'autres limitations liées à l'algorithme à travers ses paramètres. Nous avons proposé une amélioration de l'algorithme consistant à modifier ces paramètres en question à travers une procedure expérimentale a posteriori. En particulier nous avons montré que le paramètre introduisant le plus de difficulté est la constante de lipschitz L de la fonction objectif de notre problème d'optimisation. La borne de convergence de la MDSA est en fonction du paramètre L. Cela introduit un degré de conservatisme vis-à-vis de l'estimation d'une solution optimale. Nous avons montré qu'une amélioration possible des performance de la MDSA est d'agir sur le paramètre L : une estimation a posteriori nous a permis de réduire considérablement ce degré de conservatisme. Nous proposons ici comme perspective, de pousser notre approche expérimentale plus loin : quelle stratégie de pas de la méthode garantit une meilleure rapidité de convergence ? Quelle stratégie fiable permettant une bonne estimation du paramètre L ? La solution optimale calculée par la MDSA représente une moyenne sur tous les points engendrés par la méthode sur la totalité des itérations. Une moyenne calculée uniquement sur un sous-ensemble des points est pertinente du point de vue de l'analyse de convergence. Comment choisir ce sous-ensemble de point afin de garantir une convergence plus rapide ? Ceci dit, la nature numérique de notre problème d'optimisation pose des limites intrinsèques réduisant considérablement la capacité de résoudre efficacement le problème dans le cas d'un très grand nombre de variables de décision : la nature des sous-gradients fait que leur contribution dans le déplacement vers la solution optimale du problème x^* est faible. Ceci résulte du point de vue pratique en une perte de la rapidité de convergence de la méthode.

Finalement et pour évaluer la précision et la représentativité des modèles RBA établis (déterministes et stochastiques), nous proposons de les résoudre dans le cas de dimension réaliste et de comparer les solutions obtenues avec les données observées via l'expérimentation. Ceci permettra également de juger de la pertinences des modèles ainsi établis (déterministe et stochastique) par rapport au modèle RBA classique.

Chapitre 8

Annexe

Article soumis à Conference on Decision and Control 2016

Bibliographie

[1] F. A. Al-Khayyal. Jointly constrained bilinear programs and related problems : An overview. Computers & Mathematics with Applications, 19(11) :53 – 62, 1990.

[2] K. Balakrishnan. *Exponential Distribution : Theory, Methods and Applications*. Taylor & Francis, 1996.

[3] N. Balakrishnan and C.R. Rao. Preface. In *Order Statistics : Theory & Methods*, volume 16 of *Handbook of Statistics*, pages v – vii. Elsevier, 1998.

[4] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3) :167–175, 2003.

[5] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Trans. Img. Proc.*, 18(11) :2419–2434, 2009.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1) :183–202, 2009.

[7] S. Becker, J. Bobin, and E. J. Candès. Nesta : A fast and accurate first-order method for sparse recovery. *SIAM J. Img. Sci.*, 4(1) :1–39, 2011.

[8] A. Ben-Tal, T. Margalit, and A. Nemirovskii. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1) :79–108, 2001.

[9] A. Ben-Tal and A. Nemirovskii. *Lectures on modern convex optimization : analysis, algorithms, and engineering applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[10] A. Ben-Tal and A. Nemirovskii. Non-euclidean restricted memory level method for large-scale convex optimization. *Math. Program.*, 102(3) :407–456, 2005.

[11] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[12] D. P. Bertsekas and J. Tsitsiklis. *Introduction to probability*. Athena Scientific, Belmont (Mass.), 2008.

[13] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in Systems and Control Theory*, volume 15 of *Studies in Appl. Math.* SIAM, Philadelphia, 1994.

[14] S. Boyd, S-J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optimization and Engineering*, 8(1) :67–127, 2007.

[15] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2010.

[16] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3) :200 – 217, 1967.

[17] Scherer C. and W. Siep. *Linear matrix inequalities in control*. Lecture Notes, Dutch Institute for Systems and Control, Delft, The Netherlands, 2005.

[18] R. E. Caflisch. Monte carlo and quasi-monte carlo methods. *Acta Numerica*, 7 :1–49, 1998.

[19] G. Chen and M. Teboulle. Convergence analysis of a proximal like minimization algorithm using bregman functions. *SIAM J. Optim.*, 3 :538–543, 1993.

[20] A. R. Conn, N. I. M. Gould, and P. Toint. *Trust-region methods*. MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[21] A. Cornish-Bowden. *Fundamentals of Enzyme Kinetics*. PORTLAND PRESS, 3rev ed edition, 2004.

[22] A. Cornish-Bowden, M. Jamin, and V. Saks. *Cinétique enzymatique (Traduction du livre anglais "Fundamentals of enzyme kinetics" de Cornish-Bowden, A.* Collection Grenoble sciences. EDP Sciences, 2005.

[23] M. E. Csete and J. C. Doyle. Reverse engineering of biological complexity. *Science*, 295 :1664, 2002.

[24] Dan Davidi, Elad Noor, Wolfram Liebermeister, Arren Bar-Even, Avi Flamholz, Katja Tummler, Uri Barenholz, Miki Goldenfeld, Tomer Shlomi, and Ron Milo. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. *Proceedings of the National Academy of Sciences*, page 201514240, March 2016.

[25] V. L. Dos Santos Eleutério. *Finding Approximate Solutions for Large Scale Linear Programs*. ETH, 2009.

[26] M. Dyer and L. Stougie. Computational complexity of stochastic programming problems. *Math. Program.*, 106(3) :423–432, 2006.

[27] Z. Furedi and I. Barany. Computing the volume is difficult. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC '86, pages 442–447, New York, NY, USA, 1986. ACM.

[28] A. Goelzer. *Emergence de structures modulaires dans les régulations des systèmes biologiques : théorie et applications à Bacillus subtilis*. Thèses, Ecole Centrale de Lyon, November 2010.

[29] A. Goelzer, V. Fromion, and G. Scorletti. Cell design in bacteria as a convex optimization problem. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 4517–4522, Dec 2009.

[30] A. Goelzer, V. Fromion, and G. Scorletti. Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6) :1210–1218, 2011.

[31] A. Goelzer, J. Muntel, V. Chubukov, M. Jules, E. Prestel, R. Noelker, M. Mariadassou, S. Aymerich, M. Hecker, P. Noirot, D. Becher, and V. Fromion. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic Engineering*, 32 :232–243, November 2015.

[32] G. H. Hines, M. Arcak, and A. K. Packard. Equilibrium-independent passivity : A new definition and numerical certification. *Automatica*, 47(9) :1949–1956, 2011.

[33] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer Verlag, Heidelberg, 1996. Two volumes - 2nd printing.

[34] R. Horst and T. Hoang. *Global optimization : deterministic approaches*. Springer-Verlag, Berlin, New York, 1993.

[35] A. Juditsky and A. Nemirovskii. First Order Methods for Nonsmooth Convex Large-Scale Optimization, I : General Purpose Methods. In *Optimization for Machine Learning*, pages 1–28. MIT Press, 2010.

[36] A. Juditsky and A. Nemirovskii. First Order Methods for Nonsmooth Convex Large-Scale Optimization, II : Utilizing Problem's Structure. In *Optimization for Machine Learning*, pages 29–63. MIT Press, 2010.

[37] H. Kitano. Systems biology : a brief overview. *Science*, 295(5560) :1662–1664, 2002.

[38] G. Lan, A. Nemirovskii, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Math. Program.*, 134(2) :425–458, 2012.

[39] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1) :57–95, 2016.

[40] W. Liebermeister, J. Uhlendorf, and E. Klipp. Modular rate laws for enzymatic reactions : thermodynamics, elasticities and implementation. *Bioinformatics*, 26(12) :1528–1534, 2010.

[41] Jorge J. M. The Levenberg-Marquardt algorithm : Implementation and theory. In G. A. Watson, editor, *Numerical Analysis*, pages 105–116. Springer, Berlin, 1977.

[42] O. Mangasarian. *Nonlinear Programming*. Society for Industrial and Applied Mathematics, 1994.

[43] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11 :431–441, 1963.

[44] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *IEEE Trans. Aut. Control*, pages 819–830, 1997.

[45] S. B. Mokhtar, D. S. Hanif, and C. M. Shetty. *Nonlinear programming - theory and algorithms (2. ed.)*. Wiley, 1993.

[46] A. Nemirovskii, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4) :1574–1609, 2009.

[47] A. Nemirovskii and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester, New York, 1983. A Wiley-Interscience publication.

[48] Y. Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.

[49] Y. Nesterov. Minimizing functions with bounded variation of subgradients. Technical Report 2005079, Catholic University of Louvain (UCL) - Center for Operations Research and Econometrics (CORE), 2005.

[50] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.*, 120(1) :221–259, 2009.

[51] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[52] Y. Nesterov and J.-Ph. Vial. Confidence level solutions for stochastic programming. *CORE Discussion Papers*, 2000 :00–2000, 2000.

[53] Y. Nesterov and J.-Ph. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6) :1559 – 1568, 2008. Stochastic Modelling, Control, and Robust Optimization at the Crossroads of Engineering, Environmental Economics, and Finance.

[54] J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag New York, 2nd edition, 2006.

[55] B. O. Palsson. *Systems Biology : Properties of Reconstructed Networks*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK, January 2006.

[56] J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2) :157 – 175, 2005.

[57] J. Renegar. *A Mathematical View of Interior-point Methods in Convex Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[58] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

[59] C. Roos, T. Terlaky, and J.-Ph. Vial. *Interior Point Approach to Linear Optimization : Theory and Algorithms*. John Wiley & Sons, Chichester, New York, 1997. (second print 1998).

[60] H. M. Sauro. *Enzyme Kinetics for Systems Biology*. Ambrosius Publishing, 2012.

[61] A. Shapiro. Asymptotic analysis of stochastic programs. *Ann. Oper. Res.*, 30(1-4) :169–186, 1991.

[62] A. Shapiro. Monte carlo simulation approach to stochastic programming. In *Proceedings of the 33nd conference on Winter simulation*, WSC '01, pages 428–431, Washington, DC, USA, 2001. IEEE Computer Society.

[63] A. Shapiro. Stochastic programming approach to optimization under uncertainty. *Math. Program.*, 112(1) :183–220, July 2007.

[64] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming : modeling and theory*. MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[65] A. Shapiro and A. Nemirovskii. On complexity of stochastic programming problems, 2004.

[66] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying *e. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991) :533–538, juillet 2010.

[67] M. Tawarmalani and N. V. Sahinidis. *Convexification and global optimization in continuous and mixed-integer nonlinear programming : theory, algorithms, software, and applications*. Nonconvex optimization and its applications. Kluwer Academic Publishers, Dordrecht, Boston, London, 2002.

[68] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized algorithms for analysis and control of uncertain systems*. Springer, Berlin, 2004.

[69] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

[70] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38 :49–95, 1994.

[71] D. W. Walkup and R. J. B. Wets. Stochastic programs with recourse. *SIAM Journal on Applied Mathematics*, 15(5) :1299–1314, septembre 1967.

[72] R. J. B. Wets. Programming under uncertainty : The equivalent convex program. *SIAM Journal on Applied Mathematics*, 14(1) :89–105, janvier 1966.

[73] R. J. B. Wets. Programming under uncertainty : The solution set. *SIAM Journal on Applied Mathematics*, 14(5) :1143–1151, septembre 1966.

[74] J. C. Willems. Dissipative dynamical systems I General theory. II Linear systems with quadratic supply rates. *Archive for Rational Mechanics and Analysis*, 45 :321–343, 1972.

[75] Y. Ye. *Interior Point Algorithms : Theory and Analysis*. Wiley-Interscience series in Discrete Mathematics and Optimization. John Wiley & Sons, New York, 1997.

dernière page de la thèse

AUTORISATION DE SOUTENANCE

Vu les dispositions de l'arrêté du 7 août 2006,

Vu la demande du Directeur de Thèse

Monsieur G. SCORLETTI

et les rapports de

M. D. DUMUR
Professeur - CentraleSupélec - L2S - Pôle Systèmes - Département Automatique
Plateau de Moulon - 3 rue Joliot-Curie - 91192 GIF SUR YVETTE cedex

et de

M. L. DUGARD
Directeur de Recherche CNRS - Bureau B246 - ENSE3 - Domaine Universitaire - BP 46
38402 SAINT-MARTIN-D'HERES

Monsieur AIT EL FAQIR Marouane

est autorisé à soutenir une thèse pour l'obtention du grade de **DOCTEUR**

Ecole doctorale ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE

Fait à Ecully, le 14 novembre 2016

P/Le directeur de l'E.C.L.
La directrice des Etudes

ECOLE CENTRALE DE LYON
Service Scolarité
ECULLY

M-A. GALLAND