Autour De L'Usage des gradients en apprentissage statistique

Pierre-Yves Massé

université
PARIS-SACLAY

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

NNT : 2017SACLS568

THÈSE DE DOCTORAT
DE
L'UNIVERSITÉ PARIS-SACLAY
PREPARÉE À
L'UNIVERSITÉ PARIS-SUD

Laboratoire de Recherche en Informatique

ÉCOLE DOCTORALE N°580
Sciences et Technologies de l'Information et de la Communication

Spécialité

Mathématiques et Informatique

Par

Pierre-Yves MASSÉ

Autour De L'Usage des gradients en apprentissage statistique

Thèse présentée et soutenue à Orsay, le 14 décembre 2017

Composition du Jury :

M. Moulines Eric	Professeur, École Polytechnique	Président
M. Trélat Emmanuel	Professeur, Université Pierre et Marie Curie	Rapporteur
M. Bubeck Sébastien	Chercheur, Theory Group, Microsoft Research	Rapporteur
M. Ollivier Yann	Chercheur, Facebook Artificial Intelligence Research	Directeur de thèse

Remerciements

L'équipe TAO, pour « Apprentissage et Optimisation », du Laboratoire de Recherche en Informatique, m'a accueilli pour la durée de mon doctorat. Yann Ollivier m'a dit un jour que les responsables de l'équipe, Marc Schoenauer et Michèle Sebag, laissaient une totale liberté de travail aux membres de celle-ci. La liberté dont a pu bénéficier Yann a ainsi été également la mienne, ce dont je voudrais les remercier.

Les membres de l'administration ont toujours été d'une grande gentillesse et d'une grande disponibilité pour m'aider à accomplir les nombreuses formalités requises. Je pense en particulier à Stéphanie Druetta, de l'école doctorale, à Glady Bakayoko, du LRI, aux membres de la scolarité de l'Université Paris-Sud, et bien sûr à Olga Mwana Mobulakani, qui a assuré pendant presque toute la durée de mon doctorat le secrétariat de l'équipe TAO.

Olga Mwana Mobulakani faisait également presque partie de l'équipe des doctorantes et des doctorants, dont je ne cite pas les noms mais qui se reconnaîtront, avec qui j'ai passé de nombreux moments qui resteront parmi les très bons souvenirs de mon doctorat.

Jérémy Bensadon, qui était en cours de doctorat sous la direction de Yann Ollivier quand j'ai commencé le mien, m'a accompagné dans mes débuts, avec toujours beaucoup de gentillesse et d'humour !

Gaétan Marceau Caron m'a, à de nombreuses reprises, apporté son aide en informatique, avec patience, calme et toujours un grand souci de précision et d'exhaustivité. Il m'a également souvent fait profiter de sa grande connaissance des publications intéressantes des différents domaines de l'apprentissage. Il a enfin, spontanément, proposé de contribuer à la relecture du manuscrit.

Ma famille m'a soutenu pendant toute la durée de mon doctorat, dans les bons moments et dans les plus difficiles, et m'a aidé notamment lors de la relecture du manuscrit.

Séréna m'a apporté son soutien indéfectible tout au long de celui-ci. Elle a participé à la relecture du manuscrit, et l'apparence définitive de celui-ci doit beaucoup à ses connaissances en LaTeX !

Yann Ollivier a investi un temps et une énergie considérables pour que ma thèse soit la meilleure possible. Il m'a fait travailler sur des sujets intéressants, tant du point de vue du domaine que de mes goûts propres. Enfin, il m'a également laissé du temps, notamment pendant les périodes où je n'arrivais pas à avancer. Sans sa

patience et sa volonté de me laisser progresser et apprendre à mon rythme, je n'aurais pas pu obtenir les résultats présentés dans ce manuscrit.

Enfin, je voudrais remercier tous ceux qui, à un moment ou à un autre, m'ont témoigné leur amitié ou leur soutien pendant mon doctorat, et ont ainsi contribué à m'aider à le mener à terme.

Sommaire

Introduction

Sommaire de l'introduction

Introduction

La thèse que nous présentons s'intitule, sur la suggestion de Yann Ollivier, « Autour de l'usage des gradients en apprentissage statistique ». Son objet est l'étude « théorique et pratique » d'algorithmes d'optimisation de fonctions différentiables, en vue de l'entraînement des modèles d'apprentissage.

Ces modèles, parmi lesquels les réseaux de neurones bénéficient de la plus grande notoriété et représentent une forme d'archétype, constituent une avancée tant qualitative que quantitative dans l'effort historique d'automatisation des tâches remplies par des fonctions biologiques comme, par exemple, la vue, et de contournement des limites de celles-ci, en particulier des limites liées à la vitesse d'accomplissement de ces fonctions, donc à la perception du temps [1]. L'introduction des réseaux de neurones convolutifs a par exemple bouleversé, au tournant des années deux-mille, le traitement des signaux, tant sonores que visuels, pour lequel dominaient naguère les méthodes spectrales. Des situations où la puissance de calcul supérieure des ordinateurs n'avait jusqu'alors pas suffi à inverser le rapport de force avec les meilleurs spécialistes humains ont vu des programmes d'apprentissage prévaloir. Un cas récent très médiatisé est la victoire au jeu de Go du programme AlphaGo face au joueur sud-coréen Lee Sedol [2]. Le programme Libratus a pour sa part fait jeu égal avec certains des meilleurs joueurs mondiaux de Poker, un jeu qui, contrairement au jeu de Go, est à information incomplète [3]. De plus, dans les deux cas, les comportements développés par les programmes d'apprentissage s'éloignaient fortement des stratégies connues.

Si la capacité des modèles d'apprentissage à produire des comportements riches ou, ce qui en est l'analogue en termes mathématiques, à approximer avec une grande précision des fonctions appartenant à une vaste classe, était pressentie, voire connue

1. Remarquons que cet énoncé n'implique nullement l'idée d'une reproduction totale à terme de l'ensemble de ces fonctions biologiques par des automates, ni d'un dépassement de l'humanité. En effet, et *a minima*, rien ne garantit que viendra un moment où la substitution de procédures automatiques aux fonctions biologiques sera concomitante, et pour toujours, d'une absence de développement de nouvelles facultés.

2. Tanguy Chouard. "The Go Files: AI computer wraps up 4-1 victory against human champion". In : *Nature* (mar. 2016).

3. Olivia Solon. "Oh the humanity! Poker computer trounces humans in big step for AI". In : *The Guardian* (jan. 2017).

depuis au moins les années quatre-vingts, qui ont vu l'établissement des propriétés de meilleure approximation des réseaux de neurones [4], leur déploiement avec succès et l'intérêt qu'ils n'ont cessé de susciter depuis résultent de la mise au point au début des années quatre-vingt-dix des méthodes d'entraînement. En effet, la configuration souhaitée des modèles d'apprentissage n'est pas connue *a priori*, et nécessite une recherche heuristique.

L'utilisateur conçoit alors des procédures évaluant, pour chaque entrée soumise au système, la qualité de la sortie qu'il produit. Écrire celle-ci comme une fonction régulière de la configuration du système permet d'exprimer la variation de qualité du système relative à un changement de configuration comme la différentielle de la procédure évaluant la qualité par rapport à cette configuration. L'obtention de procédures permettant en pratique le calcul de cette différentielle, donc le calcul effectif de la différentielle d'une fonction composée, telles que notamment la rétropropagation, ouvrit la voie à la recherche de configurations satisfaisantes, soit à l'optimisation du système, par descente de gradient. La capacité, pour un modèle, d'acquérir un comportement désirable en présence de retours sur sa performance explique le terme d'apprentissage. Le grand nombre de retours requis, qui prennent la forme d'exemples cibles appelés « données » dans le cas de l'apprentissage supervisé est, quant à lui, à l'origine du terme d'apprentissage statistique.

Mais si la puissance de calcul des ordinateurs, conjuguée au pouvoir représentatif des modèles d'apprentissage, a permis les progrès mentionnés ci-dessus, la taille de ces modèles, le nombre des « données » et l'ampleur de l'ensemble des configurations possibles sont telles que le coût de l'entraînement est un facteur limitant les procédures d'optimisation utilisables en pratique. La recherche de procédures de descente de gradient efficaces, mais néanmoins peu coûteuses, connut ainsi un regain d'intérêt à cause de l'apprentissage. Des travaux en ce sens avaient en effet été menés dès les années quarante et cinquante, également en raison des contraintes de puissance de calcul disponibles, par la communauté de l'optimisation stochastique, sur des modèles différents et à l'époque où la puissance de calcul disponible était bien plus faible [5]. Ces travaux avaient notamment porté sur la compréhension « théorique » des procédures d'optimisation stochastique.

Si, actuellement, les modèles d'apprentissage donnent des résultats très intéressants, leur conception et leur entraînement font toujours face à de grandes difficultés. Parallèlement, les représentations disponibles ne rendent pas encore compte de façon satisfaisante de leur efficacité ni de l'efficacité des procédures d'optimisation stochastique. Notre travail s'inscrit donc à la croisée de ces problèmes pratiques et théoriques, et à celle des domaines de l'apprentissage et de l'optimisation stochastique. Une introduction générale à l'apprentissage et aux questions soulevées est fournie par la leçon inaugurale de Yann LeCun au Collège de France [6].

En-dehors de leur régularité, très peu de propriétés des fonctions à optimiser lors de l'entraînement sont connues et peuvent ainsi informer la conception des algorithmes d'optimisation. Nous expliquons dans le premier chapitre de cette introduc-

4. Kurt Hornik, Maxwell Stinchcombe et Halbert White. "Multilayer Feedforward Networks are Universal Approximators". In : *Neural Networks* 2 (1989), p. 359–366.

5. Léon Bottou. "On-line learning and stochastic approximations". In : *On-line learning in neural networks*. Sous la dir. de David Saad. Cambridge University Press New york, NY, USA, 1999. Chap. 2, p. 9–42.

6. Yann LeCun. "L'apprentissage profond : une révolution en intelligence artificielle". Leçon inaugurale au Collège de France, disponible à l'adresse `https://www.college-de-france.fr/site/yann-lecun/inaugural-lecture-2016-02-04-18h00.htm`. 2016.

tion qu'outre le recours à des procédures de conception relativement simples, l'utilisateur se voit contraint d'effectuer des choix de conception de manière arbitraire, qui influencent fortement le comportement de la descente. La fixation numérique des hyper-paramètres des procédures de descente constitue ainsi un des problèmes récurrents de ce type auxquels est confronté le praticien. L'algorithme « LLR », pour « Learning the Learning Rate » ou « Apprentissage du taux d'apprentissage » en français, est une réponse partielle au problème de la détermination des valeurs des hyper-paramètres. Il fut conçu par Yann Ollivier. Nous avons contribué à sa description formalisée, et conduit les expériences. Ce travail, qui a été le premier que nous avons réalisé pour notre thèse, a fait l'objet du dépôt d'une prépublication sur « arxiv », que nous reproduisons dans la troisième partie du manuscrit [7].

Le deuxième chapitre de notre introduction est consacré à la justification de la convergence des algorithmes d'optimisation stochastique. Nous présentons dans un premier temps les résultats majeurs obtenus dont nous avons pris connaissance. Nous décrivons ensuite une étude de ces algorithmes centrée sur la dynamique du système auquel est appliquée la procédure d'optimisation. Ce choix conduit en particulier à distinguer le temps machine, donné par les itérations de l'algorithme, d'un temps intrinsèque du système donné par sa dynamique propre. Nous n'avons pas connaissance de manière similaire de procéder dans les travaux disponibles. Ce point de vue permet, à notre avis, de mieux rendre compte des mécanismes à l'œuvre dans la convergence des procédures d'optimisation que les preuves disponibles jusqu'à présent. Les techniques développées président à l'obtention du résultat principal de notre thèse : l'établissement d'un résulat de convergence locale d'un algorithme célèbre d'entraînement de système dynamique, l'algorithme RTRL [8], appliqué à un système non linéaire.

L'interprétation des modèles d'apprentissage comme des systèmes dynamiques implique alors que nos résultats justifient la convergence de l'entraînement de ces modèles par l'algorithme RTRL, en particulier pour le cas important des réseaux de neurones récurrents. Bien que peu utilisé en pratique, en raison de son coût en mémoire très élevé, l'algorithme RTRL est un algorithme en ligne, qui permet ainsi un traitement en temps réel des données. Cette propriété n'est pas vérifiée par l'algorithme de la rétro-propagation, qui est l'algorithme principal d'entraînement des réseaux de neurones. Celui-ci nécessite en effet un retour sur l'ensemble des données disponibles pour procéder à la mise à jour du paramètre, bien qu'il soit parfois possible de circonvenir cet inconvénient, en partageant le flux de données en sous flux de petite taille, dans le cadre de la rétro-propagation tronquée. Mais cela n'est pas adapté aux cas de données présentant des dépendances temporelles longues, et peut être incompatible avec les contraintes de traitement, de sorte que ce recours ne saurait représenter une solution systématique.

L'algorithme « NoBackTrack », qui a été présenté en 2016 par Yann Ollivier, Guillaume Charpiat et Corentin Tallec [9], propose de résoudre le problème du coût en mémoire très important de l'algorithme RTRL, tout en conservant sa propriété d'être en ligne, en maintenant une approximation de faible taille des informations maintenues par RTRL. Celles-ci sont calculées de manière aléatoire de sorte à être

7. Pierre-Yves Massé et Yann Ollivier. "Speed learning on the fly". In : *preprint* (2015).

8. B.A. Pearlmutter. "Gradient calculations for dynamic recurrent neural networks: a survey". In : *IEEE Transactions on Neural Networks* 6 (5 sept. 1995), p. 1212–1228. doi : 10.1109/72.410363.

9. Yann Ollivier, Guillaume Charpiat et Corentin Tallec. "Training recurrent networks online without backtracking". In : *preprint* (2016).

non biaisées. L'algorithme « UORO », construit sur les mêmes principes que l'algorithme « NoBackTrack », représente une amélioration de celui-ci [10]. Les résultats expérimentaux obtenus sont prometteurs, et surpassent en particulier dans des cas complexes les résultats obtenus par l'usage de la rétro-propagation. Nous prouvons alors, ce qui constitue le deuxième résultat de notre thèse, qu'avec probabilité arbitrairement proche de un, l'algorithme « NoBackTrack » appliqué au même système que l'algorithme RTRL atteindra le même optimum local. Nos résultats sont transposables avec des modifications mineures à l'algorithme « UORO », de sorte qu'ils justifient également la convergence de cet algorithme.

10. Corentin Tallec et Yann Ollivier. "Unbiased Online Recurrent Optimization". In : *preprint* (2017).

Chapitre

1 Descente de gradient et structure du modèle

Nous étudions dans ce chapitre un modèle sur lequel nous pouvons raisonner de manière algébrique : l'optimisation d'une fonction quadratique, sans recourir à des arguments d'analyse. La structure de la fonction est connue, mais nous ne disposons pas d'estimations quantitatives sur celle-ci. La convergence de la descente est alors fortement dépendante du pas de descente choisi par l'utilisateur.

1.1 Descente de gradient sur une fonction quadratique

Considérons la situation d'optimisation classique suivante. Considérons, sur un espace vectoriel réel, une application bilinéaire a, symétrique définie positive, une forme linéaire b et l'application p qui, à tout vecteur θ, associe le réel

$$\frac{1}{2}\,a\,(\theta,\theta)+b\,(\theta)\,.$$

Pour trouver le vecteur qui réalise le minimum de p nous pouvons, en tout point θ de l'espace, calculer la différentielle de p qui, à un vecteur h tangent en θ à l'espace vectoriel, associe le réel

$$a\,(\theta,h)+b\,(h)\,.$$

La différentielle seconde de p associe, à tous vecteurs h et k de l'espace tangent en θ à l'espace vectoriel, le réel

$$a\,(h,k)\,.$$

Comme p est d'ordre 2 en θ, elle coïncide en tout point avec son développement à l'ordre 2. Pour tout θ de l'espace vectoriel, et h du tangent à celui-ci en θ, nous avons alors

$$p\,(\theta+h)-p\,(\theta)=a\,(\theta,h)+b\,(h)+\frac{1}{2}\,a\,(h,h)\,.$$

Le terme $a\,(h,h)$ est d'ordre 2 en h et, pour de petits h, il est ainsi négligeable devant le terme de différentielle, qui est d'ordre 1. Ainsi, pour de petits h, le signe de

$$p\,(\theta+h)-p\,(\theta) \tag{1.1}$$

est celui de

$$\mathrm{d}\,p\,(\theta)\cdot h=a\,(\theta,h)+b\,(h)\,.$$

En tout point θ, $\mathrm{d}\,p\,(\theta)$ est une forme linéaire sur l'espace vectoriel. Elle est donc positive sur un demi-espace, négative sur l'autre, et s'annule sur l'hyperplan qui

les délimite. Nous souhaitons diminuer p, et nous pouvons donc choisir un vecteur tangent h dans le demi-espace

$$\mathrm{d}\, p\,(\theta) < 0,$$

suffisamment petit pour que le signe de la différence de l'Équation (1.1) soit bien donné par $\mathrm{d}\, p\,(\theta) \cdot h$. Malheureusement, un tel algorithme peut échouer.

Algorithme 1.1 (Algorithme d'optimisation pouvant échouer)**.** *Soit la suite de vecteurs (θ_t) définie par la donnée d'un vecteur initial θ_0 et la relation de récurrence, pour $t \geq 0$,*

$$\theta_{t+1} = \theta_t + h_t,$$

où h_t est un vecteur tangent à l'espace vectoriel en θ_t et tel que

$$\mathrm{d}\, p\,(\theta_t) \cdot h_t < 0 \quad \textit{et} \quad \frac{1}{2}\, a\,(h_t,\, h_t) < -\mathrm{d}\, p\,(\theta_t) \cdot h_t.$$

Échec de la démonstration de convergence. À chaque itération, le signe de

$$p\,(\theta_{t+1} = \theta_t + h_t) - p\,(\theta_t)$$

est bien celui de

$$\mathrm{d}\, p\,(\theta_t) \cdot h_t,$$

d'après la deuxième condition sur h_t, donc est négatif, d'après la première condition. Donc, pour tout $t \geq 0$,

$$p\,(\theta_{t+1}) < p\,(\theta_t)\,.$$

Or, p est minorée car a est définie positive. Par conséquent, par convergence monotone, la suite de terme général $p\,(\theta_t)$ converge. Mais nous ne savons pas si cette limite est le minimum de p. □

La manière classique de procéder pour résoudre ce problème d'optimisation consiste alors à le représenter de la manière suivante. Considérons un produit scalaire $\langle \cdot, \cdot \rangle$ sur l'espace vectoriel. Par isomorphisme de l'application

$$y \mapsto \langle y, \cdot \rangle,$$

nous pouvons trouver un vecteur $\tilde{b}$ et, pour tout vecteur θ, un vecteur $\tilde{a}\,(\theta)$ tels que, pour tout vecteur h tangent en θ,

$$a\,(\theta,\, h) = \langle \tilde{a}\,(\theta)\,, h \rangle \quad \text{et} \quad b\,(h) = \langle \tilde{b}, h \rangle.$$

De plus, a est bilinéaire et symétrique, donc $\tilde{a}$ est une application linéaire autoadjointe pour le produit scalaire considéré. De la sorte, pour tout vecteur θ et tout vecteur tangent h,

$$\mathrm{d}\, p\,(\theta) \cdot h = \langle \tilde{a}\,(\theta)\,, h \rangle + \langle \tilde{b}, h \rangle = \langle \tilde{a}\,(\theta) + \tilde{b}, h \rangle. \tag{1.2}$$

Notons alors Λ la plus grande valeur propre de $\tilde{a}$. Nous avons ainsi

$$p\,(\theta + h) - p\,(\theta) \leq \langle \tilde{a}\,(\theta) + \tilde{b}, h \rangle + \frac{1}{2}\, \Lambda\, \langle h, h \rangle.$$

D'après l'Équation (1.2), sur la droite dirigée par le vecteur $\tilde{a}\,(\theta) + \tilde{b}$, la différentielle de p en θ est négative. Nous cherchons donc un vecteur h de la forme :

$$h\,(\theta, \eta) = -\eta\, \left(\tilde{a}\,(\theta) + \tilde{b} \right),$$

où η est un réel. Nous cherchons les η tels que

$$\begin{aligned} p\left(\theta + h\left(\theta,\, \eta\right)\right) - p\left(\theta\right) &\leq -\eta \left\|\tilde{a}\left(\theta\right) + \tilde{b}\right\|^2 + \frac{\eta^2}{2} \Lambda \left\|\tilde{a}\left(\theta\right) + \tilde{b}\right\|^2 \\ &= \eta \left(\frac{\eta \Lambda}{2} - 1\right) \left\|\tilde{a}\left(\theta\right) + \tilde{b}\right\|^2 \end{aligned} \tag{1.3}$$

est strictement négatif. Cette condition est satisfaite pour tout $0 < \eta < 2\,\Lambda^{-1}$. Nous disposons ainsi de l'algorithme suivant.

Algorithme 1.2 (Algorithme d'optimisation classique)**.** *Soit la suite de vecteurs (θ_t) définie par la donnée d'un vecteur initial θ_0 et la relation de récurrence, pour $t \geq 0$,*

$$\theta_{t+1} = \theta_t - \eta \left(\tilde{a}\left(\theta_t\right) + \tilde{b}\right),$$

où $0 < \eta < 2\,\Lambda^{-1}$.

Pour tout vecteur θ, le vecteur $\tilde{a}\left(\theta\right) + \tilde{b}$ est le gradient de l'application $\mathrm{d}\, p\left(x\right)$ (relativement au produit scalaire considéré), et nous le notons

$$\nabla p\left(\theta\right) = \tilde{a}\left(\theta\right) + \tilde{b}.$$

Nous pouvons ainsi réécrire l'algorithme précédent sous la forme suivante.

Algorithme 1.3 (Algorithme de descente de gradient classique)**.** *Soit la suite de vecteurs (θ_t) définie par la donnée d'un vecteur initial θ_0 et la relation de récurrence, pour $t \geq 0$,*

$$\theta_{t+1} = \theta_t - \eta \nabla p\left(\theta\right),$$

où $0 < \eta < 2\,\Lambda^{-1}$.

Convergence de l'algorithme de descente de gradient classique. Pour tout $t \geq 0$,

$$\theta_{t+1} = \left(\mathrm{Id} - \eta\, \tilde{a}\right) \theta_t + \eta\, \tilde{b},$$

où Id est l'identité de l'espace vectoriel. Or, l'opérateur $\mathrm{Id} - \eta\, \tilde{a}$ est auto-adjoint car $\tilde{a}$ l'est. Notons alors λ la plus petite valeur propre de $\tilde{a}$. Alors, les valeurs propres de $\mathrm{Id} - \eta\, \tilde{a}$ sont comprises entre

$$1 - \eta\, \Lambda \quad \text{et} \quad 1 - \eta\, \lambda.$$

D'après la condition sur le pas, $-1 < 1 - \eta\, \Lambda$ et $1 - \eta\, \lambda \leq 1$. Ainsi, les valeurs propres de $\mathrm{Id} - \eta\, \tilde{a}$ sont toutes de module inférieur au sens large à 1. Or, comme a est définie positive, λ est strictement positive et ainsi, leur module est strictement inférieur à 1. Ainsi, $\mathrm{Id} - \eta\, \tilde{a}$ est contractant, et (θ_t) est une suite convergente. D'après l'Équation (1.3), la limite θ_∞ satisfait

$$0 \leq \eta \left(\frac{\eta \Lambda}{2} - 1\right) \left\|\nabla p\left(\theta_\infty\right)\right\|^2.$$

Or, le terme majorant est négatif, et il est nul si, et seulement si,

$$\nabla p\left(\theta_\infty\right) = 0,$$

ce qui implique que θ_∞ réalise bien le minimum de p. □

1.2 Conception d'un algorithme de descente sans information quantitative : le problème du pas de descente

La conception de la descente est ainsi dépendante d'un premier choix de l'utilisateur, celui d'un produit scalaire sur l'espace vectoriel. En l'absence d'un candidat privilégié, le produit retenu est souvent le produit euclidien dans la base de travail. Ce choix détermine alors les valeurs numériques des valeurs propres de l'opérateur $\tilde{a}$ (la hessienne de la perte). Or, nous ne connaissons pas *a priori* leurs valeurs numériques, car nous ne disposons que d'hypothèses analytiques sur la perte p, soit sa différentiabilité, mais pas d'information quantitative. Nous pourrions utiliser des mesures effectuées au cours de l'optimisation afin d'obtenir celles-ci . De manière plus générale, nous pourrions tenter d'obtenir par des mesures des informations supplémentaires sur la fonction. Or, très souvent en apprentissage, le coût des évaluations de la fonction ou de ses dérivées empêche d'en effectuer plus de quelques-unes à chaque itération de l'algorithme d'entraînement. En particulier, il n'est souvent pas souhaitable d'évaluer de manière satisfaisante la dérivée seconde d'une fonction par différence finie de ses gradients dans plusieurs directions. L'algorithme de descente de gradient « classique » peut ainsi sembler excessivement frustre mais, en l'absence d'informations supplémentaires sur la fonction à optimiser, il représente souvent, sous la forme sous laquelle nous l'avons présenté ou sous des formes voisines, un compromis acceptable en terme de vitesse de convergence.

Mais la simplicité de l'algorithme implique que l'obtention de résultats satisfaisants sera en pratique délicate, même pour des fonctions quadratiques. L'algorithme dont nous disposons nous permet en effet d'optimiser toute fonction quadratique p, pourvu que le pas choisi soit assez petit. En l'absence d'information quantitative, le choix d'une valeur numérique pour le pas est à la discrétion du praticien. Or, un choix de pas trop grand entraîne l'échec de l'optimisation, car il ne permet pas de contrôler les grandes valeurs propres. De même, un choix de pas trop petit fournit un algorithme ineffectif, car la convergence sera trop lente pour les besoins de l'utilisateur, même si la procédure aboutira en temps (trop) long.

Pour remédier au problème des pas trop grands, une solution consiste à utiliser une suite de pas (η_t) tendant vers 0. Au bout d'un certain temps, le pas sera ainsi dans la zone admissible. Mais ce temps peut être trop long, de sorte que l'algorithme explose avant l'entrée dans la zone. Symétriquement, le pas peut décroître trop vite, de sorte que la descente soit infiniment lente. S'il est toutefois retenu, ce schéma nécessite que la série de terme général η_t diverge. Dans le cas contraire, la suite (θ_t) satisfaisant l'équation de récurrence

$$\theta_{t+1} = \theta_t - \eta_t \nabla p\left(\theta_t\right), \tag{1.4}$$

converge vers une valeur θ_∞ arbitraire.

Convergence de la descente de gradient pour une série des pas divergente. Si la série des pas diverge, nous remarquons que la différence de deux suites (θ_t) et (θ'_t) satisfaisant l'équation précédente satisfait, au bout d'un temps assez long,

$$\left\|\theta_{t+1} - \theta'_{t+1}\right\| \leq (1 - \eta_t \lambda) \left\|\theta_t - \theta'_t\right\|,$$

et tend ainsi vers 0, comme le produit des termes $1 - \eta_t \lambda$. Or, le vecteur minimisant p,

$$\tilde{a}^{-1}\left(\tilde{b}\right),$$

satisfait identiquement l'Équation (1.4), et toute solution de celle-ci converge ainsi vers ce vecteur. □

Si la suite de pas choisie est $\eta_t = \eta/t$, l'ordre de grandeur du taux de convergence, égal au produit des termes $1 - \eta_t \lambda$, est en

$$\frac{1}{t^{\eta \lambda}},$$

ce qui quantifie notre assertion précédente sur l'influence du choix du pas initial sur la vitesse de convergence de la descente. Le choix du pas est un problème très fréquemment abordé dans les articles d'apprentissage. Des choix de pas plus ou moins heuristiques ont été proposés, mais ils ne représentent au mieux que des solutions ponctuelles au problème [1].

L'algorithme « LLR », ainsi que nous l'avons dit précédemment, représente une tentative de remédier à ce problème [2]. Les expériences numériques furent effectuées sur des données synthétiques. Des tests sur la base de données MNIST avaient été initiés sur la plateforme « Torch », avec des résultats sinon spectaculaires, du moins prometteurs, mais faute de temps nous n'avons pu les mener à terme jusqu'à présent. Nous entendons y remédier dès que possible.

L'algorithme met à jour le pas de descente au fur et à mesure que s'effectue la descente de gradient, pour un coût algorithmique comparable à celui de la descente simple. La mise à jour est effectuée par descente de gradient sur le pas de descente, grâce à la construction d'une perte appropriée. L'algorithme se révèle efficace pour augmenter un pas de descente initial trop faible, mais échoue à diminuer le pas à temps pour prévenir l'explosion numérique dans le cas inverse. Nous reproduisons dans ce manuscrit la prépublication déposée en ligne. Des travaux similaires avaient été effectués dans les années quatre-vingt-dix, notamment par l'équipe de Richard S. Sutton [3].

1.3 Descente de gradient sur des fonctions inconnues

Malgré les difficultées liées au choix du pas, l'utilisation d'une descente de gradient pour les fonctions quadratiques se justifie par les garanties apportées par les résultats de convergence ci-dessus. Des résultats analogues s'obtiennent, moyennant quelques précautions, pour des fonctions régulières au voisinage d'un optimum, et ainsi presque quadratiques dans ce voisinage, car admettant une différentielle seconde bornée. Établir un résultat de convergence suppose d'avoir un contrôle de la fonction à optimiser en tout point susceptible d'être atteint par la procédure. La distinction entre les résultats globaux et locaux provient ainsi de l'admissibilité ou non de l'extension des hypothèses de contrôle de la fonction à tout l'espace considéré ou à un sous-espace.

1. Nous pensons par exemple à l'article suivant. Tom Schaul, Sixin Zhang et Yann LeCun. "No More Pesky Learning Rates". In : *Proceedings of The 30th International Conference on Machine Learning*. Sous la dir. de Sanjoy Dasgupta et David McAllester. JMLR, 2013, p. 343–351.

2. Massé et Ollivier, "Speed learning on the fly", art. cit.

3. Nous indiquons les travaux suivants, dont le premier a été réalisé par un étudiant de Richard S. Sutton. Ashique Mahmood. "Automatic step-size adaptation in incremental supervised learning". Mém.de mast. Edmonton, Alberta, United States of America : University of Alberta, 2010 ; Harold J. Kushner et J. Yang. "Analysis of adaptive step-size SA algorithms for parameter tracking". In : *IEEE Transactions on Automatic Control* 40 (8 1995), p. 1403–1410.

En particulier, en l'absence d'hypothèses globales sur une fonction autres que sa régularité, ce qui est un cas courant en apprentissage statistique, domaine sur lequel nous reviendrons plus tard, le sens de la notion d'optimum global semble incertain. En effet, sous cette seule hypothèse, prouver q'un point particulier est un optimum global est impossible et ainsi, il ne semble pas évident que la notion même d'existence d'un tel point ait un contenu. Mais alors, l'évaluation des procédures d'optimisation à l'aune de leur capacité à atteindre un optimum global est à son tour problématique. Le meilleur résultat d'une descente de gradient ne peut être ici que l'obtention d'un optimum local. L'utilisation de techniques de relance de la descente ne concerne pas l'algorithme, mais les ressources que l'utilisateur peut consacrer à la recherche empirique du meilleur point possible.

Pour une fonction qui est juste supposée régulière, il n'existe pas de garantie de convergence d'une descente de gradient. Le problème n'est en effet plus celui d'atteindre un optimum pourvu qu'il en existe un et que la descente commence en son voisinage, mais de se déplacer sur une surface arbitraire. Sans connaissance sur la vitesse à laquelle cette surface varie, il n'est pas possible à notre connaissance de prévoir le comportement d'une descente. Cette mauvaise connaissance des surfaces est en particulier due, en apprentissage statistique, à la grande dimension des espaces sur lesquels sont définis les fonctions à optimiser. Des travaux récents ont été entrepris qui tentent de mieux comprendre la structure de ces surfaces, même si ceux-ci cherchent principalement à comprendre la structure des optima locaux, en ayant notamment recours à la théorie des matrices aléatoires en grande dimension [4].

4. Levent Sagun et al. "Explorations on high dimensional landscapes". Article accepté pour un atelier à ICLR 2015, disponible sur arxiv à l'adresse `https://arxiv.org/pdf/1412.6615.pdf`. 2015 ; Yann N. Dauphin et al. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In : *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Sous la dir. de Zoubin Ghahramani et al. 2014, p. 2933–2941.

Chapitre

2 Dynamique du modèle et dynamique de la descente

Les procédures d'optimisation décrites dans le chapitre précédent ont été étudiées à notre connaissance depuis les années cinquante, par la communauté de l'optimisation stochastique. L'intérêt de ces procédures pour l'entraînement des algorithmes d'apprentissage a également suscité l'intérêt de la communauté de l'apprentissage [1]. Nous présentons ici certains des résultats principaux de convergence que nous avons lus. Nous exposons ensuite les arguments principaux des résultats de convergence que nous avons obtenus.

2.1 Algorithme du gradient stochastique, résultats de convergence

La situation typique d'utilisation de descente de gradient en apprentissage statistique est un peu plus générale que celle qui a été décrite jusqu'à présent. Souvent en effet, le recours à l'algorithme de descente de gradient est contrarié par l'impossibilité d'obtenir les gradients de la fonction à optimiser, dont seules des évaluations perturbées sont accessibles. Pour une perturbation additive, l'équation de descente de gradient devient alors

$$\theta_{t+1} = \theta_t - \eta_t \left(\nabla p \left(\theta_t \right) + \xi_t \right), \tag{2.1}$$

où ξ_t est un vecteur représentant la perturbation. Nous supposons p exactement quadratique, car la situation étudiée n'est pas liée à une étude locale de la fonction, ni à la présence d'un terme d'ordre 1. Ainsi,

$$p \left(\theta \right) = \frac{1}{2} \left\| \tilde{a} \left(\theta \right) \right\|^2, \quad \text{et} \quad \nabla p \left(\theta \right) = \tilde{a} \left(\theta \right).$$

Considérons une modélisation probabiliste classique de la perturbation, où les perturbations successives sont des réalisations de variables indépendantes, identiquement distribuées, centrées et de variance égale à 1. Notons $\mathcal{F}_t$ la tribu engendrée par $\xi_0, \ldots, \xi_{t-1}$. Alors, pour tout $t \geq 0$,

$$\begin{aligned} \mathrm{E} \left[\left\| \theta_{t+1}^2 \right\|^2 \middle| \mathcal{F}_t \right] &= \left\| \left(\mathrm{Id} - \eta_t \, \tilde{a} \right) \theta_t \right\|^2 + \eta_t^2 \\ &\leq \left(1 - \eta_t \, \lambda \right)^2 \left\| \theta_t \right\|^2 + \eta_t^2. \end{aligned} \tag{2.2}$$

En particulier,

$$\mathrm{E} \left[\left\| \theta_{t+1}^2 \right\|^2 \right] \leq \left(1 - \eta_t \, \lambda \right)^2 \mathrm{E} \left[\left\| \theta_t \right\|^2 \right] + \eta_t^2.$$

1. Bottou, "On-line learning and stochastic approximations", op. cit.

Pourvu que la somme des pas soit divergente, et que la somme des carrés des pas soit finie, la suite de variables aléatoires θ_t converge ainsi en norme 2 vers 0, l'optimum de p.

Convergence vers 0 *en norme* 2 *des variables aléatoires* θ_t. Pour tout $t \geq 0$,

$$\mathrm{E}\left[\left\|\theta_t^2\right\|\right]$$

est inférieur à la norme à l'instant initial multipliée par le produit jusqu'à $t-1$ des $(1-\eta_s\,\lambda)^2$, qui tend vers 0 car la série des pas est divergente, à laquelle s'ajoute la somme

$$\sum_{s\leq t-1}\left(\prod_{p=s}^{t-1}(1-\eta_s\,\lambda)\right)\eta_s^2.$$

Or, pour tous $1 \leq t_0 \leq t$, nous majorons celle-ci par

$$\sum_{s\leq t_0-1}\left(\prod_{p=s}^{t-1}(1-\eta_s\,\lambda)\right)\eta_s^2 + \sum_{s\geq t_0}\eta_s^2.$$

Le deuxième terme est arbitrairement petit dès que t_0 est assez grand, car la somme des carrés des pas est convergente et, à t_0 fixé, le premier terme tend vers 0 avec t, en tant que somme finie de termes tendant vers 0. □

Il est même possible d'obtenir une convergence presque sûre, en considérant la sur-martingale perturbée de l'Équation (2.2).

La justification courante des hypothèses sur les pas de descente veut que la divergence de la somme des pas permette à l'algorithme d'atteindre potentiellement n'importe quel point de l'espace, tandis que la convergence de la somme des carrés assure que la perturbation totale sera finie.

L'algorithme de descente de gradient appliqué dans une telle situation, qui est donc donné par l'Équation (2.1), est appelé algorithme du gradient stochastique. Il a été introduit en 1951 dans un article de Herbert Robbins et Sutton Monro [2], présenté couramment comme à l'origine de l'étude de l'optimisation en présence de bruit, et qui constitue presque une citation obligée de tout travail se rapportant à ce sujet. Les conditions sur la somme des pas et sur la somme des carrés des pas sont appelées conditions de Robbins-Monro. De même, les procédures qui ressemblent à celle de l'Équation (2.1) sont souvent appelées procédures de Robbins-Monro.

Dans leur article, Robbins et Monro présentent une preuve de convergence dans un cadre très proche de celui discuté ci-dessus. En particulier, la fonction n'est pas supposée quadratique, mais strictement convexe, de sorte que la différentielle deuxième reste strictement positive. Les résultats de convergence ultérieurs obtenus par la communauté de l'optimisation stochastique utilisent, à notre connaissance, principalement deux stratégies : l'étude d'une équation différentielle associée à l'équation de récurrence de la descente, ou l'utilisation d'une fonction de Lyapunov. Aucune de ces stratégies ne repose de manière décisive sur la modélisation probabiliste du bruit, même si celle-ci peut être exploitée. Dans chaque cas, des hypothèses sont formulées qui garantissent qu'en moyenne, en un sens à préciser, le bruit est nul, puis l'étude traite de la descente sur la fonction de perte moyennée,

2. Herbert Robbins et Sutton Monro. "A Stochastic Approximation Method". In : *The Annals of Mathematical Statistics* 22.3 (1951), p. 400–407.

p avec nos notations. Ainsi, les pertes sont établies dans un cadre déterministe. De plus, la suite (θ_t) est supposée bornée.

Le premier résultat d'importance sur la stratégie de l'équation différentielle ordinaire est communément attribué à Harold J. Kushner et Dean S. Clark, qui le publient en 1978 dans le livre *Stochastic Approximation Methods for Constrained and Unconstrained Systems*[3]. Il deviendra connu sous le nom de Théorème de Kushner et Clark. Présentons de manière simplifiée l'argument.

« Preuve » simplifiée du théorème de Kushner et Clark . Kushner et Clark définissent le changement de temps $H_0 = 0$ et, pour $t \geq 1$,

$$H_t = \eta_0 + \eta_1 + \ldots + \eta_{t-1}.$$

Ils appellent ensuite θ l'interpolation affine de la suite (θ_t) sur les intervalles $[H_t,\ H_{t+1}[$, et θ_{e} l'interpolation en escalier. Alors, pour tout $s \geq H_t$, ces interpolations satisfont

$$\theta(s) = \theta_t - \int_{H_t}^{s} \nabla p\left(\theta_{\mathrm{e}}(s)\right) ds - \int_{H_t}^{s} \xi_t,$$

Négligeabilité de la perturbation Pour $s \geq H_t$, notons t_s le temps maximal tel que $H_{t_s} \leq s$. Alors, à un terme dominé par $s - H_{t_s}$ près,

$$\int_{H_t}^{s} \xi_t \approx \sum_{k=t}^{t_s} \eta_k \xi_k.$$

Or, Kushner et Clark supposent que la série de terme général $\eta_t\ |\xi_t|$ est convergente, de sorte que le terme de droite est arbitrairement petit, uniformément en $s \geq t$, dès que t est assez grand.

Équation différentielle associée Supposons alors que la suite (θ_t) admet une valeur d'adhérence $\bar{\theta}$. Considérons une sous-suite $\left(\theta_{\phi(t)}\right)$ de la suite (θ_t), qui converge vers $\bar{\theta}$. Alors, θ va être proche, sur des intervalles de la forme $[H_{\phi(t)}, H_{\phi(t)} + T_t[$, d'une solution de l'équation différentielle

$$\dot{y}(s) = -\nabla p\left(y(s)\right),$$

issue de $\bar{\theta}$ à l'instant $H_{\phi(t)}$, qui vérifie bien sûr, pour $H_{\phi(t)} \leq s < H_{\phi(t)} + T_t$,

$$y(s) = \bar{\theta} - \int_{H_{\phi(t)-1}}^{s} \nabla p\left(y(u)\right) du.$$

T_t est le temps maximal sur lequel est définie l'équation différentielle. Or, Kushner et Clark supposent que la suite (θ_t) est bornée. Ainsi, pourvu que la fonction ∇p soit continue, T_t est au moins supérieur à $T > 0$, pour de grandes valeurs de t.

Alors, comme η_t tend vers 0, de plus en plus de valeurs θ_s seront présentes dans les intervalles $[H_{\phi(t)}, H_{\phi(t)} + T_t[$, lorsque t sera grand. Ainsi, les comportements asymptotiques de la suite (θ_t) et des solutions des équations différentielles successives seront proches. *A priori*, seules les valeurs de la suite le long de l'extraction et les solutions des équations différentielles seraient reliées, mais le changement de temps considéré nous permet d'obtenir plus.

3. Harold J. Kushner et Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. T. 26. Applied Mathematical Sciences. Springer-Verlag New York, 1978.

Afin d'établir leur résultat de convergence, Kushner et Clark supposent alors que cette équation [4] admet des solutions asymptotiquement stables au sens de Lyapunov, et que la suite (θ_t) est infiniment souvent dans le bassin d'attraction d'une telle solution. Une étude du comportement des solutions de l'équation différentielle leur permet alors de conclure. □

L'hypothèse formulée, connue sous le nom d'hypothèse de Kushner et Clark, permet certes d'établir le résultat de convergence, mais le consensus parmi les articles que nous avons consultés et qui discutent le théorème de Kushner et Clark est qu'elle est peu utilisable en pratique, car difficilement vérifiable.

La démonstration du résultat est reprise, de manière un peu clarifiée par rapport à l'originale, dans l'ouvrage *Stochastic Approximation and Recursive Algorithms and Applications*, coécrit par Harold J. Kushner et George Yin et publié pour la première fois en 1997 [5]. Celui-ci dresse un panorama des différents domaines qui utilisent des méthodes d'optimisation stochastique, et présente des résultats de convergence pour celles-ci. Si ces résultats prêtent peut-être à des critiques similaires à celle du théorème décrit ci-dessus, l'exposé des domaines qui utilisent l'optimisation stochastique couvre de nombreux champs et est très détaillé. L'article "Convergence of Stochastic Algorithms : From the Kushner-Clark Theorem to the Lyapunov Functional Method" [6], publié en 1996 par Jean-Claude Fort et Gilles Pagès, redémontre de manière concise le théorème de Kushner et Clark, et propose différents résultats qui approfondissent la méthode de l'équation différentielle ordinaire.

Une occurrence de la stratégie qui utilise la fonction de Lyapunov se trouve au théorème deux de l'article "Convergence of a stochastic approximation version of the EM algorithm", publié en 1999 [7]. Celui-ci repart d'un travail précédent du premier auteur, Bernard Delyon [8]. La preuve du premier article cité requiert que la somme des pas diverge. La convergence de la somme des carrés des pas n'est pas requise, mais c'est au prix d'une hypothèse plus forte, c'est-à-dire la convergence de la série de terme général $\eta_t \, \xi_t$, qui sert à contrôler le terme de bruit [9]. De plus, la preuve suppose que la suite (θ_t) est bornée. Examinons rapidement son argument.

Méthode de la fonction de Lyapunov . Notons $\mathcal{L}$ l'ensemble des zéros de ∇p. Notons V la fonction de Lyapunov, et $V(\mathcal{L})$ l'image de $\mathcal{L}$ par celle-ci. Alors, pour un certain $\varepsilon > 0$, pour tout $t \geq 0$,

$$V(\theta_{t+1}) \leq V(\theta_t) - \eta_t \, \varepsilon,$$

dès que $V(\theta_t)$ est loin de $V(\mathcal{L})$. Or, la série des η_t est divergente donc, si cette inégalité était vérifiée asymptotiquement, $V(\theta_t)$ tendrait vers moins l'infini, ce qui est

4. Les équations différentielles successives ont un comportement uniforme en temps, car elles ont toutes pour valeur initiale $\bar{\theta}$. D'ailleurs, Kushner et Clark translatent toutes les fonctions de $-H_t$ pour ramener le comportement asymptotique de la suite θ en 0.

5. Harold J. Kushner et George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag New York, 2003.

6. Jean-Claude Fort et Gilles Pagès. "Convergence of Stochastic Algorithms : From the Kushner-Clark Theorem to the Lyapunov Functional Method". In : *Advances in Applied Probability* 4 (28 1996), p. 1072–1094.

7. Bernard Delyon, Marc Lavielle et Eric Moulines. "Convergence of a stochastic approximation version of the EM algorithm". In : *The Annals of Statistics* 27 (1 1999), p. 94–128.

8. Bernard Delyon. "General results on the convergence of stochastic algorithms". In : *IEEE Transactions on Automatic Control* 41 (9 1996), p. 1245–1255.

9. Dans le cas de où les ξ_t sont des variables aléatoires indépendantes et identiquement distribuées, la convergence presque sûre de la série de terme général $\eta_t \, \xi_t$ est une conséquence de la convergence de la somme des carrés des pas.

impossible car les θ_t sont bornés, et V continue. Ainsi, infiniment souvent, $V(\theta_t)$ est proche de $\mathcal{L}$. Des hypothèses sur ce dernier et des arguments topologiques permettent alors de conclure que $V(\theta_t)$ converge vers un point de $V(\mathcal{L})$.

Un argument du même type que celui de la première partie établit alors que, infiniment souvent, θ_t sera proche de $\mathcal{L}$. Or, le pas de descente tend vers 0 et ainsi, tous les termes de la suite (θ_t) seront proches de $\mathcal{L}$, au bout d'un temps assez long (et non seulement les termes d'une sous-suite qui réaliserait le « infiniment souvent » précédent). □

Nous mentionnons également les travaux suivants. Les deux stratégies figurent dans un article de Lennart Ljung publié en 1997 [10]. L'auteur cite des travaux de Harold J. Kushner de la même période, et lui-même est cité dans le livre publié peu après par celui-ci [11]. L'article "On-line learning and stochastic approximations" étudie les questions de convergence en lien avec les algorithmes d'apprentissage [12].

2.2 Dynamique d'une descente de gradient

Réaliser une descente sur une fonction dont le gradient à l'instant t est

$$\nabla p + \xi_t$$

peut suggérer d'étudier les descentes sur une famille de pertes p_t, plutôt que sur une perte unique et perturbée, ce qui est une situation courante d'apprentissage. Considérons ainsi une famille de fonctions (p_t). Nous cherchons un paramètre optimal pour cette famille, et nous devons donc donner un sens à cette notion. Pour une fonction seule, nous disposons d'un critère d'ordre zéro, soit la minimisation de la fonction. Mais ce critère est souvent difficile à vérifier, et en pratique ce sont le couple formé par un critère d'ordre un, l'annulation du gradient, et d'ordre deux, le caractère défini positif de la hessienne, qui garantissent l'optimalité (locale) d'un paramètre. En optimisation stochastique, souvent, raisonner sur l'espérance des fonctions permet de se ramener au cas d'une fonction seule. Est ainsi appelé paramètre optimal de la famille de fonctions un optimum local de l'espérance des fonctions. Mais cela pose deux difficultés. D'une part, la notion n'a de sens que pour des fonctions indépendantes et identiquement distribuées, ce qui restreint son champ d'application. Pour de nombreuses familles de fonctions, la modélisation probabiliste n'est de plus pas pertinente. D'autre part, même dans les cas où elle le serait, l'utilisateur ne dispose souvent que d'une réalisation de la famille de fonctions, et non de plusieurs réalisations sur lesquelles il pourrait effectuer des descentes avant de les moyenner. Ainsi, prouver la convergence en raisonnant sur l'espérance des fonctions ne permet pas de rendre compte du mécanisme effectivement à l'œuvre dans la descente, quand bien même la preuve déboucherait sur des résultats vrais pour toute trajectoire de descente (soit avec probabilité égale à 1).

Il ne serait pas pertinent de demander qu'un optimum local réalise un optimum de chaque fonction individuellement. Nous souhaitons disposer d'une notion qui quantifie le caractère privilégié d'un optimum. Celle-ci devra exprimer un comportement collectif de la famille de pertes. Nous souhaitons qu'en un optimum,

10. Lennart Ljung. "Analysis of recursive stochastic algorithms". In : *IEEE Transactions on Automatic Control* 22 (4 1977), p. 551–575.

11. Kushner et Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, op. cit.

12. Bottou, "On-line learning and stochastic approximations", op. cit.

les gradients de celles-ci ne présentent pas de direction privilégiée dans laquelle se diriger. Nous souhaitons donc qu'ils soient uniformément répartis dans toutes les directions. Nous choisissons de demander qu'« en moyenne », en un optimum θ^*, les gradients soient nuls, soit vérifient

$$\frac{1}{t}\sum_{s\leq t}\nabla p_s\left(\theta^*\right)\approx 0.$$

Nous remplaçons ainsi la moyenne spatiale à un instant donné, représentée par l'espérance, par une moyenne temporelle. Cependant, la renormalisation par $1/t$ est arbitraire. Il nous faut trouver une manière liée au problème d'optimisation de renormaliser la somme. Or, pour une famille de pas (η_t), l'équation de mise à jour du paramètre s'écrit

$$\theta_{t+1}=\theta_t-\eta_t\nabla p_t\left(\theta_t\right).$$

Considérons alors l'algorithme tel que la mise à jour du paramètre est calculée, mais que l'application de celle-ci au calcul des gradients est différée. L'équation de mise à jour de celui-ci est alors, pour un paramètre initial θ,

$$\theta_{t+1}=\theta_t-\eta_t\nabla p_t\left(\theta\right).$$

Ainsi, au bout d'un temps T,

$$\theta_T=\theta-\sum_{t\leq T-1}\eta_t\nabla p_t\left(\theta\right). \tag{2.3}$$

Nous pouvons réécrire l'équation ci-dessus

$$\theta_T=\theta-\sum_{t\leq T-1}\eta_t\left(\nabla p_t\left(\theta\right)-\nabla p_t\left(\theta^*\right)\right)-\sum_{t\leq T-1}\eta_t\nabla p_t\left(\theta^*\right).$$

Pour des pertes quadratiques $p_t\left(\theta\right)=a_t\left(\theta,\,\theta\right)/2$, nous avons

$$\nabla p_t\left(\theta\right)=\tilde{a}_t\left(\theta\right),$$

avec $\tilde{a}_t$ une application linéaire, et ainsi l'équation l'Équation (2.3) se réécrit

$$\begin{aligned}\theta_T-\theta^*&=\theta-\theta^*-\sum_{t\leq T-1}\eta_t\left(\tilde{a}_t\left(\theta\right)-\tilde{a}_t\left(\theta^*\right)\right)-\sum_{t\leq T-1}\eta_t\nabla p_t\left(\theta^*\right)\\&=\left(\mathrm{Id}-\sum_{t\leq T-1}\eta_t\,\tilde{a}_t\right)\left(\theta-\theta^*\right)-\sum_{t\leq T-1}\eta_t\nabla p_t\left(\theta^*\right).\end{aligned}$$

Nous cherchons alors un temps T tel que d'une part, la moyenne des hessiennes pondérée par les pas de descente soit définie positive sur l'intervalle $[0,T[$. Nous supposons donc que, pour un réel $\lambda>0$, la plus petite valeur propre de la somme des hessiennes est supérieure à

$$\lambda\sum_{s\leq T-1}\eta_t.$$

D'autre part, nous voulons que la somme des gradients évalués en le paramètre θ^* soit négligeable devant ce terme, sur l'intervalle considéré. Plus exactement, nous voulons disposer d'une suite d'instants (T_k) telle que, pour tout $k\geq 0$, la plus petite valeur propre de la somme des hessiennes sur $[T_k,\,T_{k+1}[$ soit supérieure à

$$\lambda\sum_{t=T_k}^{T_{k+1}-1}\eta_t$$

et telle que

$$\sum_{t=T_k}^{T_{k+1}-1} \eta_t \nabla p_t(\theta^*) = o\left(\sum_{t=T_k}^{T_{k+1}-1} \eta_t\right),$$

quand k tend vers l'infini. Nous appelons alors optimal un paramètre θ^* qui vérifie ces deux conditions. Nous cherchons ainsi à nous placer dans l'échelle de temps donnée par la dynamique de la famille de fonctions à optimiser. Nous pouvons optimiser une famille de fonctions (p_t) avec la suite de pas (η_t) s'il existe un changement de temps

$$\sum_{t=T_k}^{T_{k+1}-1} \eta_t$$

tel que la famille satisfait la dynamique ci-dessus dans une échelle de temps. Nous appelons algorithme en boucle ouverte l'algorithme pour lequel la mise à jour du paramètre est effectuée aux instants T_k. C'est l'algorithme qui procède dans l'échelle de temps donnée par la dynamique de la famille de pertes. Sous des hypothèses peu restrictives, la différence entre l'algorithme de départ et l'algorithme en boucle ouverte est d'ordre deux en la somme des pas sur les intervalles $[T_k, T_{k+1}[$. Nous pouvons ainsi établir la convergence de la procédure de descente.

Algorithme 2.1 (Algorithme d'optimisation en boucle ouverte)**.** *Considérons une famille (T_k) d'instants tels que, sur les intervalles $[T_k, T_{k+1}[$, les conditions précédentes sont vérifiées. Soit la suite de vecteurs $\left(\theta^k\right)$ définie par la donnée d'un vecteur initial θ_0 et la relation de récurrence, pour $k \geq 0$,*

$$\theta^{k+1} = \theta^k - \sum_{t=T_k}^{T_{k+1}-1} \eta_t \nabla p_t\left(\theta^k\right),$$

où (η_t) est une suite de pas de descente.

Dans la pratique, nous n'étudions pas directement cet algorithme mais, sur chaque intervalle $[T_k, T_{k+1}[$, nous comparons l'algorithme de descente à l'algorithme en boucle ouverte issu de la valeur à l'instant T_k de l'algorithme de descente.

Preuve de convergence de la descente de gradient sur une famille de fonctions. Notons (θ_t) la suite de paramètres produite par l'algorithme de descente de gradient. Alors, pour tout $k \geq 0$, en comparant celui-ci à l'algorithme d'optimisation en boucle ouverte sur $[T_k, T_{k+1}[$, nous obtenons

$$\left\|\theta_{T_{k+1}} - \theta^*\right\| \leq \left(1 - \lambda \sum_{t=T_k}^{T_{k+1}-1} \eta_t\right) \left\|\theta_{T_k} - \theta^*\right\| + o\left(\sum_{t=T_k}^{T_{k+1}-1} \eta_t\right).$$

Ainsi, pour peu que la somme des pas diverge, et que la somme des pas entre deux instants T_k tende vers 0, θ_{T_k} converge vers θ^*, quand k tend vers l'infini. Mais, entre deux instants T_k consécutifs, nous avons

$$\sup_{T_k \leq t < T_{k+1}} \left\|\theta_t - \theta_{T_k}\right\| = O\left(\sum_{t=T_k}^{T_{k+1}-1} \eta_t\right),$$

de sorte que la suite (θ_t) converge bien vers θ^* [13]. □

13. La preuve de convergence ci-dessus figure en détail dans la preuve de convergence de l'algorithme RTRL.

Considérer l'algorithme en boucle ouverte revient à ne pas assimiler, arbitrairement, l'échelle de temps donnée par les itérations de l'algorithme de descente et l'échelle de temps donnée par la dynamique de la famille de fonctions. La séparation des deux permet, à notre avis, de comprendre le mécanisme de descente sur une famille. Bien sûr, en pratique, nous n'avons pas accès à la dynamique de la famille de fonctions, et nous n'avons pas d'autre choix que de mettre à jour le paramètre à chaque itération. Ce point de vue permet de comprendre les mécanismes à l'œuvre dans la descente, mais non de la mener.

Dans le cas d'une perte bruitée, nous pouvons définir p_t par

$$\nabla p_t(\theta) = \nabla p(\theta) + \xi_t,$$

et nous pouvons alors appliquer ce qui précède pour prouver la convergence de la descente.

2.3 Optimisation d'un système dynamique

L'approche de la section précédente se transpose à l'optimisation d'un système dynamique paramétré, ce qui n'était pas le cas de l'approche avec modélisation probabiliste. Un système dynamique paramétré comporte un état, noté e_t à l'instant t, et un paramètre, que nous notons θ. À l'instant t, l'état est mis à jour grâce à une fonction de transition $\boldsymbol{T}_t$, par l'équation

$$e_{t+1} = \boldsymbol{T}_t(e_t, \theta).$$

À chaque instant, la qualité de l'état courant du système est évaluée par une perte p_t. L'état courant dépend du paramètre, et la composition de la perte, avec la fonction qui au paramètre associe l'état, définit ainsi une perte sur le paramètre. Nous appelons alors optimal un paramètre tel que les gradients et les hessiennes produits par le système utilisant ce paramètre, vérifient les conditions discutées ci-dessus.

Un article récent considère également le comportement temporel des données, dans une preuve de convergence d'un certain système dynamique, plutôt qu'une modélisation probabiliste [14]. Peut-être le recours très fréquent à une modélisation probabiliste est-il discutable dans certains cas, et considérer les comportements temporels des systèmes permettra de mieux comprendre les phénomènes de convergence.

La contribution principale de notre travail consiste en la preuve de convergence de l'algorithme RTRL, un algorithme classique d'optimisation de système dynamique. Une description de l'algorithme et des problèmes liés se trouve dans l'introduction de la preuve. Nous prouvons également la convergence des algorithmes « NoBackTrack » et UORO, comme une conséquence non immédiate de la convergence de l'algorithme RTRL. De même, une présentation détaillée de ceux-ci se trouve dans l'introduction de la preuve correspondante.

L'algorithme « NoBackTrack » propose une solution au coût en mémoire prohibitif de l'algorithme RTRL. Mais, ainsi que l'article "Training recurrent networks online without backtracking" [15] le précise, cette solution peut être utilisée pour d'autres

14. Borja Balle et Odalric-Ambrym Maillard. "Spectral Learning from a Single Trajectory under Finite-State Policies". In : *Proceedings of the 34th International Conference on Machine Learning*. Sous la dir. de Doina Precup et Yee Whye Teh. T. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia : PMLR, juin 2017, p. 361–370.

15. Yann Ollivier, Corentin Tallec et Guillaume Charpiat. "Training recurrent networks online without backtracking". In : *preprint* (2015).

algorithmes, comme le filtre de Kalman. Ce dernier, qui est notamment réputé avoir été utilisé dans le système de navigation du programme Apollo, permet d'estimer un signal à partir d'observations de celui-ci, mais peut également être utilisé à des fins d'entraînement de systèmes dynamiques. Pour un signal suivant une équation d'évolution linéaire en présence d'un bruit gaussien, et des observations perturbées par un autre bruit gaussien, le filtre de Kalman maintient l'espérance conditionnelle de la valeur du signal par rapport aux observations. Cet estimateur est de variance minimale [16]. Il est de plus optimal pour certaines classes d'estimateurs, même pour un bruit non gaussien. Afin d'optimiser un système dynamique avec le filtre de Kalman, il faut considérer que le signal à observer est le paramètre optimal du système, et est donc un signal constant. Ce paramètre est observé au travers des pertes sur le système. Ce point de vue est par exemple présenté par Herbert Jaeger dans le rapport technique *A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach* [17], afin d'entraîner un réseau de neurones récurrents. Enfin, pour des systèmes non linéaires, le filtre de Kalman est calculé à partir de la linéarisation du système étudié.

Un article récent de Yann Ollivier établit une correspondance algébrique entre le filtre de Kalman appliqué au couple état du système–paramètre à entraîner et l'algorithme RTRL utilisé avec des gradients naturels (nous reviendrons sur cette notion plus tard) [18].

16. Brian D. O. Anderson et John B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, 1979.

17. Herbert Jaeger. *A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach*. Rapp. tech. 159. German National Research Center for Information Technology, GMD, 2002.

18. Yann Ollivier. "Online natural gradient as a Kalman filter". In : *preprint* (2017).

Chapitre

3 Entraînement des modèles d'apprentissage statistique

3.1 L'entraînement des modèles d'apprentissage comme optimisation d'un système dynamique

L'étude de l'optimisation de systèmes dynamiques s'applique à la compréhension de l'entraînement des modèles d'apprentissage, car nous pouvons interpréter ceux-ci comme des systèmes dynamiques. Un réseau de neurones est par exemple composé d'états, les neurones, dont les valeurs au cours du temps, les activations, sont modifiées en prenant en compte les poids du réseau, qui font office de paramètres. Éventuellement, cette modification se fait en présence d'entrées soumises au réseau, les données. Nous pouvons ainsi, à la suite de l'article "Unbiased Online Recurrent Optimization" [1], décrire un modèle d'apprentissage comme composé de trois instances : un jeu de paramètres, un ensemble d'états et un graphe de calcul. Le graphe décrit la manière dont les états sont transformés au cours du temps. Pour un réseau de neurones, il consiste en la donnée de l'ensemble des équations de mise à jour des activations.

Cette description algorithmique trouve alors son pendant mathématique dans la notion de système dynamique. Les états du modèle sont les états du système, le paramètre du modèle est celui du système, et le graphe de calcul représente la fonction de transition de celui-ci. Il permet ainsi notamment l'obtention de la dérivée des pertes sur le modèle par rapport au paramètre. Lors de l'entraînement du système d'apprentissage, les pertes dont les gradients sont fournis au système dynamique s'écrivent souvent

$$p_t(\theta) = p(x_t, \theta),$$

où x_t est une donnée. En général, un tel cas ne justifie pas de recourir à une modélisation probabiliste. La situation où p_t est choisie aléatoirement parmi un ensemble de pertes peut en revanche s'y prêter. Comme nous le disions précédemment, la grande dimension des espaces où les états et les paramètres prennent leurs valeurs rend coûteuses les évaluations de gradients, ce qui constitue une justification pour l'utilisation d'algorithmes du type gradient stochastique, économes en ces appels [2], pour l'entraînement des modèles d'apprentissage.

Ainsi, les résultats obtenus sur l'optimisation de systèmes dynamiques rendent compte de l'entraînement des modèles d'apprentissage. Les réseaux de neurones récurrents sont en particulier redevables de nos résultats. Malheureusement, la preuve

1. Tallec et Ollivier, "Unbiased Online Recurrent Optimization", art. cit.
2. Léon Bottou. "Large-scale machine learning with stochastic gradient descent". In : *Proceedings of COMPSTAT'2010*. Springer, 2010, p. 177–186.

de convergence que nous proposons ne permet pas en l'état de justifier la convergence de l'entraînement des modèles en chaîne de Markov cachée, ou « HMM » pour « Hidden Markov Models » en anglais, en particulier utilisés pour la reconnaissance de textes [3]. En effet, des « HMM » n'oublient les états antérieurs qu'en plusieurs itérations machine, alors que notre preuve demande une propriété d'oubli instantané. Nous pensons cependant être près de disposer d'une extension du résultat de convergence qui n'exigera plus cette propriété, et sera en mesure de traiter également le cas « HMM ».

3.2 Trois derniers points sur l'entraînement

Nous concluons en mentionnant trois aspects de l'entraînement des modèles d'apprentissage qui nous semblent intéressants mais ne sont pas traités dans nos travaux.

Le choix de paramétrisation, souvent arbitraire, influence la procédure de descente, ce qui n'est pas désirable. L'utilisation de métriques permet de rendre la descente indépendante de ce choix [4]. Mais le recours à des métriques est souvent coûteux, et les articles cités proposent des approximations de celles-ci préservant leurs propriétés désirables. Parmi les métriques possibles, la métrique naturelle, due à Sun-ichi Amari, d'où provient le gradient naturel, et qui est construite à partir de l'information de Fischer sur les données, implique un choix de distance entre paramètres fonction de l'écart des comportements du réseau en ces paramètres, et semble donc d'un intérêt particulier [5].

La difficulté des modèles d'apprentissage à être sensibles aux dépendances temporelles longues des données est souvent déplorée [6]. Un problème fréquent lors de l'obtention des gradients des réseaux de neurones à plusieurs couches est en effet la très faible valeur des gradients au-delà des premières couches, lors de la rétropropagation. Mais de plus, calculer les gradients sur une longue suite de données est coûteux. Une solution fréquemment retenue en pratique est de tronquer la suite de données en segments de longueur arbitraire. Mais cela empêche d'être sensible à des dépendances temporelles longues. L'article "Unbiasing Truncated Backpropagation Through Time" [7] propose d'effectuer des troncations de taille aléatoire, de sorte qu'en moyenne l'ensemble de la suite de données soit considéré. Ainsi, souvent le segment à utiliser est court, et le calcul du gradient est rapide, mais de temps en temps, l'ensemble de la suite est utilisé.

Mentionnons enfin, parmi les travaux qui ont tenté de rendre raison des performances des réseaux de neurones, la construction de la « scattering transform » [8].

3. Olivier Cappé, Eric Moulines et Tobias Ryden. *Inference in Hidden Markov Models*. Springer-Verlag New York, 2005. DOI : 10.1007/0-387-28982-8.

4. Yann Ollivier. "Riemannian metrics for neural networks I: feedforward networks". In : *Information and Inference: A journal of the IMA* 4 (2 2015), p. 108–153 ; Yann Ollivier. "Riemannian metrics for neural networks II: recurrent networks and learning symbolic data sequences". In : *Information and Inference: A journal of the IMA* 4 (2 2015), p. 153–193.

5. Shun-ichi Amari. "Natural gradient works efficiently in learning". In : *Neural Comput.* 10 (2 fév. 1998), p. 251–276. ISSN : 0899-7667. DOI : 10.1162/089976698300017746. URL : http://portal.acm.org/citation.cfm?id=287476.287477.

6. Yoshua Bengio, Patrice Simard et Paolo Frasconi. "Learning Long-Term Dependencies with Gradient Descent is Difficult". In : *IEEE Transactions on Neural Networks* 5 (2 mar. 1994), p. 157–166.

7. Corentin Tallec et Yann Ollivier. "Unbiasing Truncated Backpropagation Through Time". In : *preprint* (2017).

8. Stéphane Mallat. "Group Invariant Scattering". In : *Communications in Pure and Applied*

Cette transformée destinée au traitement des images utilise une base de décomposition des fonctions spécifiquement construite afin d'encoder des propriétés d'invariance par transformations géométriques, qui sont capturées par les réseaux de neurones sans qu'il soit possible, dans ce cas, de l'expliquer de manière satisfaisante.

Mathematics 65 (10 oct. 2012), p. 1331–1398.

Première partie

Preuve de convergence de l'algorithme RTRL

Sommaire de la Partie I

Introduction

Nous présentons une preuve de convergence d'un algorithme classique d'entraînement d'un système dynamique à paramètre (tel qu'un réseau de neurones récurrent) par descente de gradient en temps réel sur les paramètres : l'algorithme *Real-time recurrent learning* (RTRL) [9], soit « Apprentissage Récurrent en Temps Réel ». Bien que cet algorithme soit connu, d'après Pearlmutter [10], au moins depuis les années 1950, aucune preuve de convergence n'était disponible à notre connaissance.

Considérons un système dynamique de paramètre θ, dont l'état e_t à l'instant $t \geq 0$ satisfait l'équation de récurrence

$$e_{t+1} = \boldsymbol{T}_t \left(e_t, \, \theta \right), \tag{3.1}$$

où $\boldsymbol{T}_t$ est le t-ème opérateur de transition. Ce modèle fournit entre autres une représentation mathématique des réseaux de neurones récurrents. L'objectif de l'entraînement est de trouver une valeur de θ produisant des trajectoires désirables, par exemple minimisant un certain critère de perte le long de la trajectoire.

Dans beaucoup des situations que nous considérerons, l'opérateur de transition $\boldsymbol{T}_t$ sera en fait un opérateur fixe indépendant du temps, mais prenant en plus un argument représentant une éventuelle entrée ou observation à chaque instant :

$$\boldsymbol{T}_t \left(e_t, \, \theta \right) = \boldsymbol{T} \left(e_t, \, u_{t+1}, \, \theta \right). \tag{3.2}$$

Nous considérerons dans ce cas que la suite d'observations u_t est fixée une fois pour toutes. Nos résultats sont valables quelle que soit la suite u_t : en particulier nous ne faisons aucune hypothèse stochastique sur u_t.

Nous renvoyons au tutoriel de Jaeger [11] pour une présentation de différentes méthodes d'entraînement récurrent. L'algoritme habituel, la rétropropagation dans le temps, a l'inconvénient de nécessiter une séquence d'observations complète et de ne pas pouvoir s'utiliser en temps réel avec l'arrivée de nouvelles valeurs du système. Plus coûteux, l'algorithme RTRL, à l'inverse, procède itérativement, comme suit.

9. Jaeger, *A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach*, op. cit. ; Pearlmutter, "Gradient calculations for dynamic recurrent neural networks: a survey", art. cit.

10. Idem, "Gradient calculations for dynamic recurrent neural networks: a survey", art. cit.

11. Jaeger, *A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach*, op. cit.

L'algorithme RTRL procède par descente de gradient sur le paramètre. À chaque instant, le système encourt une perte $p_t(e_t)$ (par exemple, une erreur entre une prédiction calculée à partir de e_t, et une observation à l'instant suivant). L'algorithme présuppose qu'il est possible à chaque instant de calculer la dérivée de cette perte par rapport à l'état instantané e_t (ce qui est le cas pour une perte quadratique par exemple).

La différentielle de cette perte par rapport au paramètre θ est alors calculée par récurrence grâce aux relations

$$\frac{\partial p_t}{\partial \theta}(e_t) = \frac{\partial p_t}{\partial e}(e_t) \cdot \frac{\partial e_t}{\partial \theta} \tag{3.3}$$

et

$$\frac{\partial e_t}{\partial \theta} = \frac{\partial \boldsymbol{T}_{t-1}}{\partial \theta}(e_{t-1}, \theta) + \frac{\partial \boldsymbol{T}_{t-1}}{\partial e_{t-1}}(e_{t-1}, \theta) \cdot \frac{\partial e_{t-1}}{\partial \theta}, \tag{3.4}$$

cette dernière équation résultant de la différentiation de l'équation de récurrence (3.1) définissant le système.

L'algorithme RTRL effectue à chaque instant un pas de gradient sur le paramètre θ, en utilisant l'estimation de gradient calculée par les équations de récurrence ci-dessus. Une des difficultés est qu'en conséquence, le paramètre varie au cours de la trajectoire, de sorte que les relations utilisées pour calculer ce gradient ne sont plus exactement valables (elles le sont seulement dans la limite où la taille de pas de la descente de gradient sur θ tend vers 0).

Convergence vers un optimum local

Nous énonçons ci-dessous un théorème de convergence *local* pour cet algorithme RTRL : si le paramètre θ est initialisé près d'un paramètre θ^* qui est un *optimum local* (en un sens à définir) pour les pertes considérées alors, avec des tailles de pas assez petites, l'algorithme RTRL produira une suite de paramètres convergeant vers θ^*. [12]

Ainsi que nous l'avons dit dans l'introduction, nous définissons un extremum local d'un système dynamique comme un point tel que la moyenne temporelle des gradients des pertes tende vers 0. Nous supposons ainsi disposer d'une fonction ℓ telle que

$$\frac{1}{\ell(t)} \sum_{s=0}^{\ell(t)} \frac{\partial p_s}{\partial \theta}(\theta^*) \to 0,$$

quand t tend vers l'infini. Cette hypothèse disqualifie notamment le cas où de larges plages de gradients consécutifs pointent dans une direction, puis de larges plages dans la direction opposée. Dans ce cas en effet, la moyenne des gradients (calculée le long d'une certaine sous-suite) est nulle, mais un entraînement du paramètre en temps réel oscille autour de θ^*, au lieu de converger. Notre hypothèse demande donc une forme d'homogénéité en temps de la convergence vers 0 des moyennes de gradients.

Si les gradients consécutifs de la perte étaient des variables aléatoires i.i.d., de moyenne nulle et variance finie, la relation ci-dessus serait vérifiée pour toute fonction $\ell(t) = t^a$, avec $a > 1/2$, d'après la loi du logarithme itéré. Nous avons préféré ne

12. De manière générale, pour un système dynamique quelconque avec des pertes non convexes, il est difficile d'espérer une convergence autre que locale au voisinage d'une trajectoire non chaotique.

pas poser d'hypothèse probabiliste sur les données et partir d'une hypothèse plus faible [13].

Afin de nous assurer que l'extremum est un minimum, et d'assurer la convergence de la descente de gradient, nous aurons également besoin d'une condition d'ordre deux. Pour cela, notons H_t^* les hessiennes successives calculées lorsque le système dynamique utilise le paramètre θ^*. Ainsi,

$$H_t^* = \frac{\partial^2 p_t}{\partial \theta^2}(\theta^*).$$

Nous supposons que, en moyenne sur des intervalles de temps suffisamment grands, les valeurs propres de ces hessiennes sont strictement positives soit que, pour tout $t \geq 0$, la plus petite valeur propre de

$$\frac{1}{\ell(t)} \sum_{s=t}^{t+\ell(t)} H_s^*$$

est supérieure à $\lambda > 0$. Cette hypothèse est évidemment vérifiée si chacune des hessiennes H_t^* vérifie déjà cette propriété, mais cela n'est pas toujours le cas. Par exemple en régression linéaire, la contribution de chaque exemple individuel à la hessienne de la perte totale est une matrice de rang 1, et la hessienne n'est définie positive qu'en moyenne sur les exemples [14].

Nous appellerons paramètre optimal du système dynamique un paramètre θ^* satisfaisant ces deux conditions.

L'ensemble des pas de descente admissibles, c'est-à-dire pour lesquels nous garantissons la convergence de la descente, va alors dépendre de la fonction ℓ. En effet, les pas de descente définissent la vitesse à laquelle les gradients successifs sont incorporés. Or, cette incorporation doit être suffisamment lente pour que le phénomène de moyennage temporel puisse se produire. Dans la preuve, cela se traduit par la segmentation de l'axe des temps en des intervalles

$$[T_k, T_{k+1}[.$$

Nous voulons que les intervalles soient suffisamment longs pour que les conditions précédentes sur les gradients et les hessiennes soient vérifiées, relativement à la durée effective des intervalles, qui est la somme des pas sur ceux-ci,

$$\sum_{t=T_k}^{T_{k+1}-1} \eta_t.$$

Grâce à ces intervalles, nous prouvons que la suite extraite (θ_{T_k}) est convergente. Afin de prouver la convergence de la suite, et non seulement de la suite extraite,

13. Dans le cas d'un système dynamique, il est tout à fait irréaliste de supposer que les pertes consécutives dans le temps sont indépendantes et identiquement distribuées (i.i.d.), puisque les états consécutifs du système dynamique sont corrélés. Une modélisation probabiliste *i.i.d.* est donc inadéquate. Nous pourrions la remplacer par des modèles probabilistes plus complexes (par exemple markoviens). Néanmoins, l'un de nos objectifs est d'obtenir des résultats applicables à des modèles plus complexes utilisés sur des données réelles, tels que des réseaux de neurones récurrents appliqués à du texte en langue naturelle. Supposer que le texte est markovien pour ensuite en apprendre un modèle plus complexe que markovien serait étrange. Par la suite, nous abandonnons donc la modélisation probabiliste des données en faveur de cette hypothèse moins forte.

14. Aymeric Dieuleveut, Nicolas Flammarion et Francis Bach. *Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression.* Rapp. tech. 2016.

nous avons alors besoin que

$$\sup_{T_k \leq t < T_{k+1}} d\left(\theta_t,\, \theta_{T_k}\right)$$

tende vers 0 avec k. Or, ce supremum est contrôlé par la somme des pas sur l'intervalle. De manière générale, cette somme contrôle l'évolution de l'algorithme préalablement à la prise en compte de la contractivité [15]. Nous avons donc besoin que les pas soient suffisamment petits pour que cette somme tende vers 0. La taille des intervalles dépend de ℓ, et ainsi la taille des pas en dépend également. La dynamique du système entraîné détermine donc la dynamique des procédures de descente admissibles.

Ensuite, dans un cadre non-i.i.d., les tailles de pas η_t de la descente de gradient doivent satisfaire des contraintes plus strictes que le classique critère de Robbins–Monro [16],

$$\sum \eta_t = \infty \quad \text{et} \quad \sum \eta_t^2 < \infty.$$

En effet, si par exemple η_t est nul pour tous les t pairs, et que le système dynamique affiche des phénomènes de période 2 se manifestant sur les valeurs des gradients, il est clair que la descente de gradient de pas η_t pourrait s'éloigner fortement d'une valeur annulant la moyenne temporelle des gradients. Ce phénomène est évité en ajoutant une hypothèse d'homogénéité temporelle des tailles de pas d'apprentissage, précisée ci-après, et satisfaite pour toutes les tailles de pas du type $\eta_t = 1/t^\gamma$. Cette hypothèse n'est pas nécessaire dans le cas i.i.d. car le système est identique à chaque instant : tous les gradients ont le même rôle, et il n'est donc pas gênant de sélectionner une sous-suite des gradients plutôt que la suite initiale (celles-ci sont en effet indistinguables).

Enfin, l'hypothèse la plus restrictive que nous formulons sur notre modèle est une hypothèse d'oubli. Nous devons nous assurer qu'à paramètre fixé, le système étudié ne sera plus sensible aux conditions initiales au bout d'un certain temps, faute de quoi l'apprentissage serait impossible. Nous supposerons qu'à paramètre et entrées identiques, les opérateurs de transition sur le système dynamique sont uniformément contractants sur les états : il existe une norme sur l'espace des états telle que la différentielle

$$\frac{\partial \boldsymbol{T}_t}{\partial e_t}$$

a une norme d'opérateur inférieure à $1 - \alpha$. Cette hypothèse devra seulement être vérifiée le long de la trajectoire optimale $\theta = \theta^*$, et peut donc être vérifiée *a posteriori* sur le paramètre obtenu lors de l'entraînement. Cette hypothèse est restrictive, et nous tenterons de l'affaiblir par la suite, mais elle permet tout de même de conserver de nombreux modèles, comme l'attestent les exemples suivants.

Exemples

Voici quelques exemples de systèmes dynamiques satisfaisant ces hypothèses. Nous notons u_t une éventuelle entrée du système : dans beaucoup de situations, l'opérateur de transition $\boldsymbol{T}_{t-1}$ au temps t est un opérateur fixe prenant en argument une observation u_t.

15. Nous avons besoin de savoir *a priori* que l'algorithme reste dans des ensembles bornés pour pouvoir ensuite appliquer les hypothèses qui donnent la contractivité.

16. Robbins et Monro, "A Stochastic Approximation Method", art. cit.

Les systèmes non récurrents. Un cas particulier important est celui où le système dynamique est en fait un système non récurrent, c'est-à-dire, où l'état à l'instant t ne dépend que des entrées à l'instant t et d'un paramètre,

$$e_t = F(u_t, \theta). \tag{3.5}$$

Ce cas recouvre les modèles courants d'apprentissage statistique i.i.d., où la tâche est de prédire une quantité y_t à partir d'observations u_t supposées indépendantes et identiquement distribuées, au travers d'un modèle de prédiction paramétré par θ (comme un réseau de neurones « feedforward » de poids θ). Dans ce cas, e_t encode la prédiction sur y_t, et F représente la fonction calculée par le modèle. Dans cette situation, l'hypothèse d'oubli est trivialement satisfaite puisque $\frac{\partial \boldsymbol{T}_t}{\partial e_t} = 0$.

L'algorithme RTRL se réduit dans ce cas à la descente de gradient stochastique ordinaire,

$$\theta_{t+1} = \theta_t - \eta_t \frac{\partial p_t}{\partial \theta}.$$

Le théorème ci-dessous donne déjà un énoncé non trivial dans cet important cas particulier.

Les systèmes dynamiques linéaires dont la matrice de transition est contractante. Ils satisfont une équation de récurrence

$$e_t = A(u_t, \theta)\, e_{t-1} + B(u_t, \theta) \tag{3.6}$$

où A est une matrice dont les valeurs singulières sont inférieures à 1 :

$$\|A(u_t, \theta)\|_{\text{op}} \leq 1 - \alpha$$

pour tous u_t, θ. Cette condition est légèrement plus forte que la stabilité d'un système dynamique linéaire, qui traditionnellement demande que le *rayon spectral* de A soit inférieur à 1.

Les *automates finis probabilistes.* Ici e_t est un vecteur qui contient les probabilités de se trouver dans chacun des états de l'automate à l'instant t ; les probabilités de transition vers le prochain état de l'automate dépendent de la dernière entrée u_t. Ce sont donc des systèmes linéaires comme ci-dessus, avec $B = 0$ et où, quelle que soit l'entrée u_t, la matrice de transition A est une matrice stochastique (paramétrée par θ). L'hypothèse d'oubli dépend de la norme : par exemple, pour la norme ℓ^1 sur e_t, elle est équivalente à la *condition de Doeblin forte* (à horizon 1), un critère classique de convergence des chaînes de Markov [17].

Les *réseaux de neurones récurrents.* Une des versions les plus simples peut être

$$e_t = \tanh(Ae_{t-1} + B + C\, u_t), \tag{3.7}$$

où la fonction tanh est appliquée coordonnée par coordonnée, où A, B et C sont des matrices et vecteurs de taille appropriée, et où le paramètre à entraîner est $\theta = (A, B, C)$. Comme la fonction tanh est contractante, l'hypothèse d'oubli est satisfaite, comme dans le cas linéaire, dès que la norme d'opérateur de A est inférieure à $1 - \alpha$. De plus elle peut être satisfaite le long d'une trajectoire particulière si cette trajectoire tombe dans les parties « saturées » de la fonction tanh.

17. Nous renvoyons par exemple le lecteur au cours de Eva Löcherbach : Eva Löcherbach. "Ergodicity and speed of convergence to equilibrium for diffusion processes". Cours disponible sur la page web de l'auteur, à l'adresse `https://eloecherbach.u-cergy.fr/cours.pdf`. 2015.

Les *chaînes de Markov cachées finies*[18]. Ici e_t est un vecteur qui contient la probabilités que la chaîne de Markov finie X_t se trouve dans chaque état à l'instant t, étant donné les observations jusqu'à l'instant t. Contrairement au cas des automates probabilistes, la fonction de transition dépend des observations u_t par une mise à jour non linéaire (conditionnement) : le système se compose d'une étape de transition linéaire ne dépendant pas de l'observation, avec matrice de transition $P(j \to i)$, puis d'une étape de mesure et conditionnement. Pour chaque état i, la transition linéaire est

$$e^i_{t|t-1} = \sum_j \Pr\left(X_t = i | \theta, X_{t-1} = j\right) \, e^j_{t-1}, \tag{3.8}$$

et le nouvel état après observation de u_t suit la règle de Bayes

$$e^i_t = \frac{e^i_{t|t-1} \Pr\left(u_t | \theta, X_t = i\right)}{\sum_j e^j_{t|t-1} \Pr(u_t | \theta, X_t = j)}. \tag{3.9}$$

Il est connu que pour la norme de Hilbert sur $\log e$, définie par

$$\|e_1 - e_2\|_{\text{Hilbert}} = \sup_i \log \frac{e^i_1}{e^i_2} - \inf_i \log \frac{e^i_1}{e^i_2}, \tag{3.10}$$

ces opérateurs sont contractants (l'opérateur de Bayes étant en fait isométrique). La stricte contractivité résulte alors de la condition forte de Doeblin, comme pour les automates finis ci-dessus.

Un théorème de convergence de l'algorithme RTRL

Dans le cadre proposé, nous prouvons alors que, pourvu que les états initiaux du système dynamique soient dans une boule bien choisie de l'espace des états, pourvu que le paramètre initial soit suffisamment près d'un paramètre optimal, la suite de paramètres produite par l'algorithme RTRL converge vers ce paramètre optimal.

Les définitions et énoncés mathématiques figurant dans cette section sont tous repris dans le corps de la preuve, dans un ordre un peu différent, et avec quelques modifications d'ordre technique qui n'altèrent pas le sens de notre propos actuel. Nous les présentons ici afin de clarifier l'exposition.

Définition 3.1 (Système dynamique paramétré, pertes)**.** *Nous appellerons* système dynamique à paramètre dans un ensemble $\Theta \simeq \mathbb{R}^{\dim(\Theta)}$*, une famille de fonctions de transition indexée par* $t \in \mathbb{N}$*,*

$$\boldsymbol{T}_t : \mathcal{E}_t \times \Theta \to \mathcal{E}_{t+1},$$

où chaque $\mathcal{E}_t$ *est un espace vectoriel isomorphe à* $\mathbb{R}^{n_t}$*, pour un certain entier* n_t*. La* trajectoire définie par $\theta \in \Theta$ *avec initialisation* $e_0 \in \mathcal{E}_0$ *est définie par récurrence par*

$$e_{t+1} = \boldsymbol{T}_t \left(e_t, \theta\right).$$

Nous noterons $\boldsymbol{e}_t \left(e_0, \theta\right)$ *la fonction qui, à* e_0 *et* θ*, associe la valeur de* e_t *correspondante.*

Nous supposons que sont données, pour $t \geq 0$*, des fonctions de perte sur le couple état-paramètre,*

$$\begin{aligned} p_t : \quad & \mathcal{E}_t \times \Theta \to \mathbb{R} \\ & e, \theta \mapsto p_t \left(e, \theta\right). \end{aligned}$$

18. Cappé, Moulines et Ryden, *Inference in Hidden Markov Models*, op. cit.

Grâce à ces pertes, pour un e_0 fixé, nous pouvons définir les pertes suivantes sur le paramètre :

$$\theta \mapsto p_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right).$$

L'objectif de l'entraînement est de trouver une valeur de θ minimisant la valeur des pertes le long de la trajectoire.

Nous supposerons désormais que le système dynamique est suffisamment régulier pour que les dérivées mentionnées dans les définitions suivantes existent. Les hypothèses de régularité précises sont regroupées ci-dessous dans l'hypothèse 3.9 précédant l'énoncé du théorème.

L'algorithme RTRL ajuste le paramètre en temps réel au cours de la trajectoire, en maintenant une estimation des dérivées des pertes par rapport au paramètre, au travers d'une estimation de la jacobienne de l'état courant par rapport au paramètre.

Définition 3.2 (Algorithme RTRL)**.** *La* trajectoire de l'algorithme RTRL *définie par la donnée de conditions initiales $e_0 \in \mathcal{E}_0$, $\theta_0 \in \Theta$, $J_0 \in \mathcal{L}\left(\Theta,\, \mathcal{E}_0\right)$ et d'un opérateur de mise à jour Φ est la suite d'états (e_t) et de paramètres (θ_t) définie par les relations de récurrence, pour $t \geq 0$,*

$$\begin{cases} e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta_t\right) \\ \theta_{t+1} = \Phi\left(\theta_t,\, -\eta_t\left(\dfrac{\partial\, p_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot J_t + \dfrac{\partial\, p_t}{\partial \theta}\left(e_t,\, \theta_t\right)\right) \cdot J_t\right) \\ J_{t+1} = \dfrac{\partial\, \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot J_t + \dfrac{\partial\, \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right), \end{cases}$$

où (η_t) est une suite de tailles de pas de gradient.

Habituellement J_0 est initialisé à 0, sauf dans certaines situations particulières où l'on souhaite représenter explicitement que l'état initial dépend déjà du paramètre, auquel cas J_0 doit être initialisé à $\partial e_0 / \partial \theta_0$. Nous admettons également, dans la preuve, la présence d'une perturbation additive sur la mise à jour de la différentielle. Sous des conditions peu restrictives, qui seront vérifiées dans la preuve de convergence de l'algorithme « NoBackTrack », celle-ci ne compromet pas la convergence de la descente.

Notons g_t la fonction

$$e_0,\, \theta \mapsto \left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right).$$

Définition 3.3 (Optimum local)**.** *Nous appellerons* optimum local convenable *une valeur $\theta^* \in \Theta$ ainsi qu'un état initial $e_0^* \in \mathcal{E}_0$ possédant les deux propriétés suivantes : le long de la trajectoire associée, la dérivée de la perte moyenne tend vers* 0 *suffisamment vite, et de plus les hessiennes des pertes sont suffisamment positives. Plus exactement : nous supposons qu'il existe une fonction ℓ^1, négligeable devant l'identité en l'infini, telle que*

$$\sum_{s=0}^{\ell^1(t)} \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) = O\left(\ell^1\left(t\right)\right),$$

quand t tend vers l'infini. Nous supposons de plus qu'il existe une fonction ℓ^2, également négligeable devant l'identité en l'infini telle que, pour t suffisamment grand, la plus petite valeur propre des opérateurs

$$\frac{1}{\ell^2\left(t\right)} \sum_{s=t}^{t+\ell^2(t)} \frac{\partial^2\, p_s \circ g_s}{\partial \theta^2}\left(e_0^*,\, \theta^*\right)$$

est supérieure à λ.

Par exemple, pour un problème non-récurrent i.i.d., les dérivées des pertes par rapport à θ en un optimum θ^* sont des variables aléatoires d'espérance nulle. Si elles sont de variance bornée, la loi du logarithme itéré garantit que la moyenne de ces dérivées entre 0 et t décroît comme $O\left(\sqrt{\frac{\ln \ln t}{t}}\right)$. L'hypothèse de décroissance des dérivées est donc satisfaite pour toute fonction $\ell^1(t) = t^{a_1}$, avec $a_1 > 1/2$.

Notre résultat admet des pertes non bornées le long de la trajectoire optimale.

Hypothèse 3.4 (Croissance des pertes)**.** *Nous supposons disposer d'une fonction M_p négligeable devant l'identité en l'infini, telle que*

$$\left\| \frac{\partial\, p_t \circ g_t}{\partial \theta} (e_0^*,\, \theta^*) \right\| = O\left(M_p(t)\right),$$

quand t tend vers l'infini.

Par exemple, nous pouvons avoir $M_p(t) = t^\gamma$, pour un certain $0 \leq \gamma < 1$. Ainsi, le résultat proposé traite en particulier le cas d'une régression linéaire avec perte gaussienne. En effet, le maximum de t variables gaussiennes est dominé par $\log t$ avec grande probabilité.

Une descente de gradient pour une fonction f d'une variable θ s'écrit généralement

$$\theta \leftarrow \theta - \eta \frac{\partial f}{\partial \theta}.$$

Afin de disposer d'erreurs possibles de second ordre (par exemple, sur une variété, ou dans le cas des opérateurs proximaux) nous définissons un opérateur de mise à jour qui peut potentiellement en différer au second ordre.

Définition 3.5 (Opérateur de mise à jour sur Θ)**.** *Nous appelons* opérateur de mise à jour sur Θ *une fonction*

$$\begin{aligned} \Phi : \quad & \Theta \times \Theta \to \Theta \\ & \theta,\, v \mapsto \Phi(\theta,\, v) \end{aligned}$$

telle que

$$\Phi(\theta,\, 0) = \theta,$$

telle que Φ est deux fois continûment différentiable au voisinage du point $(\theta^*,\, 0)$, et telle qu'en tout point $(\theta,\, 0)$ de ce voisinage, la différentielle de Φ est

$$\frac{\partial\, \Phi}{\partial v} (\theta,\, 0) = \mathrm{Id}_\Theta\,.$$

Remarque 3.6. *Le Id_Θ ci-dessus pourrait être remplacé par une forme quadratique sur $T\Theta$, moyennant quelques précautions. En particulier, le cas d'une forme quadratique constante se ramène au cas Id_Θ par changement de coordonnées sur θ.*

Comme nous l'avons dit ci-dessus, dans un système dynamique, la vitesse avec laquelle les gradients tendent vers 0 le long de l'optimum local (qui est quantifiée par les fonctions ℓ^1 et ℓ^2 ci-dessus) détermine une plage de tailles de pas de gradient acceptables, et les taux d'apprentissage ne doivent pas fluctuer rapidement. La vitesse de croissance des pertes contraint également à choisir des pas de descente suffisamment petits pour contrôler celles-ci.

Par exemple, pour $\ell^1(t) = t^{a_1}$, $\ell^2(t) = t^{a_2}$ et $M_p(t) = t^\gamma$, le choix

$$\eta_t = 1/t^{\max(a_1, a_2, \gamma)+\varepsilon},$$

pour ε assez petit, remplira ces critères.

Définition 3.7 (Tailles de pas acceptables)**.** *Soient ℓ^1 et ℓ^2 les fonctions apparaissant dans la définition de l'optimum local (Définition 3.3). Soit M_p la fonction contrôlant la croissance des pertes le long de la trajectoire optimale (l'Hypothèse 3.4). Nous appelons* tailles de pas acceptables *une suite positive $\boldsymbol{\eta} = (\eta_t)$ qui vérifie les propriétés suivantes.*

1. *La série de terme général η_t est divergente.*
2. $$\eta_t = o\left(\frac{1}{\max\left(\ell^1(t),\, \ell^2(t),\, M_p(t)\right)}\right),$$
   quand t tend vers l'infini.
3. *Homogénéité temporelle : pour toute suite d'intervalles $I_t = [g_t,\, d_t[$ tels que g_t tend vers l'infini et $g_t \sim d_t$, quand t tend vers l'infini, nous avons*
   $$\frac{\sup_{I_t} \eta_s}{\inf_{I_t} \eta_s} \to 1,$$
   quand t tend vers l'infini.

Nous avons discuté plus haut l'hypothèse d'oubli des états pour deux trajectoires du système partageant le même paramètre et les mêmes entrées. Cette hypothèse n'est nécessaire qu'au voisinage de la trajectoire optimale considérée.

Hypothèse 3.8 (Oubli uniforme en temps le long de la trajectoire optimale)**.** *Nous supposons que la trajectoire $(e_t^* = \boldsymbol{e}_t(e_0^*,\, \theta^*))$ définie par l'optimum local θ^* considéré satisfait l'hypothèse d'oubli*

$$\sup_{t\geq 0} \left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\,(e_t^*,\, \theta^*) \right\|_{\text{op}} < 1.$$

Hypothèse 3.9 (Régularité des fonctions de transition et de perte)**.** *Nous supposons que les fonctions de transition $\boldsymbol{T}_t$ ainsi que les fonctions de perte p_t sont trois fois continûment différentiables.*

Nous supposons en outre qu'au voisinage de la trajectoire e_t^* définie par l'optimum local θ^* considéré, les dérivées suivantes sont uniformément bornées en temps :

$$\sup_{t\geq 0} \left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta}\,(e_t^*,\, \theta^*) \right\|_{\text{op}} < \infty.$$

Nous supposons enfin qu'il existe une boule $B_\Theta^* \subset \Theta$ centrée en θ^*, de rayon $r_\Theta^* > 0$ ainsi que des boules $B_{\mathcal{E}_t} \subset \mathcal{E}_t$ centrées en les e_t^* et de rayon commun $r_\mathcal{E} > 0$, telles que

$$\sup_{t\geq 0} \sup_{(e,\,\theta)\in B_{\mathcal{E}_t}\times B_\Theta^*} \left\| \frac{\partial^2 \boldsymbol{T}_t}{\partial\,(e,\,\theta)^2}\,(e,\,\theta) \right\|_{\text{bil}} < \infty,$$

$$\sup_{t\geq 0} \sup_{(e,\,\theta)\in B_{\mathcal{E}_t}\times B_\Theta^*} \left\| \frac{\partial^3 \boldsymbol{T}_t}{\partial\,(e,\,\theta)^3}\,(e,\,\theta) \right\|_{\text{tri}} < \infty,$$

$$\sup_{t\geq 0}\ \sup_{(e,\theta)\in B_{\mathcal{E}_t}\times B^*_\Theta} \left\| \frac{\partial^2 p_t}{\partial\,(e,\,\theta)^2}\,(e,\,\theta) \right\|_{\text{bil}} < \infty$$

et

$$\sup_{t\geq 0}\ \sup_{(e,\theta)\in B_{\mathcal{E}_t}\times B^*_\Theta} \left\| \frac{\partial^3 p_t}{\partial\,(e,\,\theta)^3}\,(e,\,\theta) \right\|_{\text{tri}} < \infty.$$

Théorème 3.10 (Convergence de l'algorithme RTRL)**.** *Soit un système dynamique paramétré (Déf. 3.1). Soit $\theta^* \in \Theta$ un optimum local convenable pour un état initial $e_0^* \in \mathcal{E}_0$ (Déf. 3.3). Soit (η_t) une suite de tailles de pas acceptables (Déf. 3.7). Supposons que sont vérifiées l'hypothèse d'oubli (Hyp. 3.8) le long de la trajectoire optimale, ainsi que les hypothèses de régularité (Hyp. 3.9).*

Alors il existe un $\mu_{\max} > 0$ et un voisinage de $(\theta^*,\, e_0^*,\, 0)$ tels que l'algorithme RTRL initialisé avec $(\theta_0,\, e_0,\, J_0)$ dans ce voisinage, et utilisé avec des tailles de pas de gradient $(\mu\,\eta_t)$ pour $\mu \leq \mu_{\max}$, produit un paramètre et une trajectoire convergeant vers l'optimum :

$$d\,(\theta_t,\,\theta^*) \to 0$$

et

$$d\,(e_t,\,e_t^*) \to 0,$$

quand t tend vers l'infini.

En outre, les voisinages obtenus dans la preuve dépendent de manière explicite des constantes et des bornes apparaissant dans les hypothèses de régularité.

Prolongements

Comme toujours lors de la présentation d'un théorème, la question se pose de savoir ce qui ressort du théorème à proprement parler, et ce qui relève des applications de celui-ci. Ainsi, le cadre dans lequel nous nous sommes placés (pertes sur des espaces vectoriels, différentiabilité des fonctions de transition), s'il est pertinent pour l'étude de l'algorithme RTRL, pourrait peut-être être affaibli dans la rédaction du résultat, et ne figurer qu'en application de celui-ci. Nous tentons actuellement d'obtenir une telle version de notre résultat. En particulier, nous pensons être bientôt en mesure de pouvoir prouver la convergence de systèmes qui ne seront pas contractants à chaque itération, mais au bout d'un certain nombre d'itérations. Plus tard, nous considérerons également l'analogue continu des systèmes étudiés, afin de prouver la convergence tant de l'algorithme RTRL que de l'algorithme « NoBackTrack ».

Le choix de formuler les hypothèses d'optimalité pour des fonctions ℓ^1 non spécifiées laisse libre l'utilisateur du théorème de formuler des hypothèses pour des fonctions particulières (notamment, des fonctions puissance). De manière générale, il nous semble que notre résultat décrit un cadre raisonnable relativement large dans lequel parler de l'optimisation d'un système dynamique a du sens, et présente les arguments qui, au sein de ce cadre, permettent d'établir la convergence. Un utilisateur qui voudrait prouver la convergence d'une procédure d'optimisation n'aurait ainsi, pourvu que son problème s'inscrive dans ce cadre, plus qu'à vérifier la validité des hypothèses pour bénéficier des garanties de convergence que nous proposons.

Enfin, nous voudrions également essayer de vérifier expérimentalement la validité des hypothèses d'optimalité que nous avons formulées. Celles-ci peuvent en effet être testées avec le paramètre obtenu lors de l'entraînement.

Présentation de la preuve

Nous décrivons dans ce chapitre, de manière simplifiée, la preuve de convergence obtenue. Tout ce qui y figure ne correspond pas nécessairement de manière exacte à ce qui est fait dans la preuve, mais les explications permettent à notre avis de comprendre le principe des arguments utilisés, plus rapidement que s'ils étaient exposés avec tous les détails comme lors de la preuve, quitte à les voir formuler un peu différemment à la lecture de celle-ci.

Le lecteur peut s'il le souhaite ne pas s'attarder sur ce chapitre. Nous espérons que, dans le cas contraire, ces quelques pages lui faciliteront la lecture de la preuve.

Nous présentons dans un premier temps le déroulement général de la preuve avant de revenir, pour chaque chapitre, sur les résultats principaux qu'il établit, ainsi que sur les arguments invoqués.

Présentation générale

Système dynamique à paramètre

Définition du système Nous étudions un système dynamique à paramètre, défini par son état initial e_0, un paramètre θ et l'équation de récurrence, pour $t \geq 0$,

$$e_{t+1} = \boldsymbol{T}_t\left(e_t, \theta\right).$$

L'état du système à l'instant $t+1$ dépend de son état à l'instant t, e_t, et du paramètre. La dépendance est modélisée par l'opérateur de transition $\boldsymbol{T}_t$. Les e_t prennent leurs valeurs dans des espaces $\mathcal{E}_t$. Le paramètre appartient à un espace Θ. Nous souhaitons pouvoir modifier le paramètre au cours du temps. Ainsi, l'équation de récurrence devient

$$e_{t+1} = \boldsymbol{T}_t\left(e_t, \theta_t\right),$$

où les θ_t sont les paramètres successifs que nous utilisons.

Prise en compte de la non linéarité Comme le système n'est pas linéaire, il est susceptible d'exploser en temps fini. Nous commençons ainsi par définir une trajectoire privilégiée, notée (e_t^*), que nous appelons trajectoire stable, au voisinage de laquelle les trajectoires seront captives. Précisément, nous construisons des boules

$B^*_{\mathcal{E}_t}$ incluses dans les espaces $\mathcal{E}_t$ telles que, pourvu que les paramètres successifs appartiennent à une boule B^*_Θ, et que l'état initial du système soit dans $B^*_{\mathcal{E}_0}$, à chaque instant t, e_t appartienne à $B^*_{\mathcal{E}_t}$. Le fonctionnement du système est illustré par la Figure 3.1.

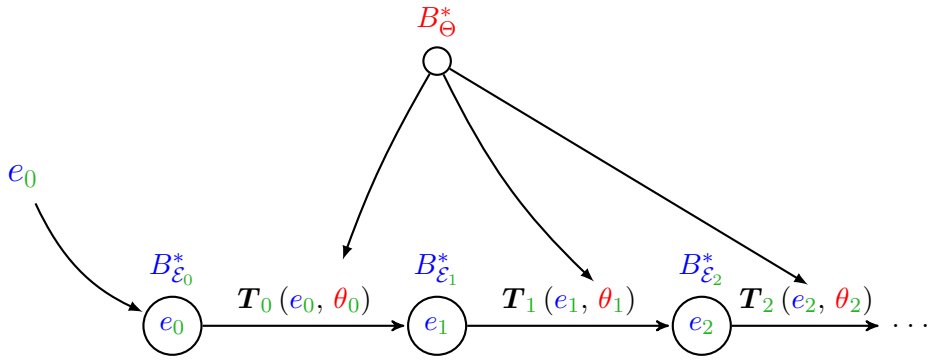

FIGURE 3.1: Schéma du déroulement du système dynamique

Sensibilité des états au paramètre Ceci fait, nous supposons que les opérateurs de transition sont lisses. Ainsi, les états du système sont une fonction lisse du paramètre, dont nous pouvons calculer les différentielles successives. Nous nous assurons également que les différentielles n'explosent pas en temps fini, et restent au voisinage des différentielles J^*_t calculées le long de la trajectoire stable. Ces différentielles vont nous servir pour définir des pertes sur le paramètre.

Pertes sur le paramètre

Construction des pertes sur le paramètre Nous supposons ensuite disposer, sur chaque espace $\mathcal{E}_t \times \Theta$, d'une perte p_t, qui mesure la qualité de l'état du système par rapport à la tâche à laquelle nous le destinons. Grâce à celles-ci, nous pouvons alors définir des pertes sur le paramètre, afin d'estimer la qualité de celui-ci. Notons en effet

$$\boldsymbol{e}_t\,(e_0,\,\theta)$$

l'état du système à l'instant t obtenu à partir d'un état initial e_0 et d'un paramètre θ. Nous définissons alors la perte

$$\theta \in \Theta \mapsto p_t\,(\boldsymbol{e}_t\,(e_0,\,\theta)\,,\,\theta) \in \mathbb{R}.$$

Une des difficultés de la preuve obtenue consiste en ce que la perte ainsi définie sur le paramètre est une composition. Ainsi, disposer par exemple de pertes p_t quadratiques ne garantirait pas que l'optimisation du système dynamique serait possible, car il faut prendre en compte l'influence des $\boldsymbol{e}_t\,(e_0,\,\theta)$.

Vecteurs tangents Notons, pour un état $e_0 \in \mathcal{E}_0$ et un paramètre θ,

$$g_t\,(e_0,\,\theta) = (\boldsymbol{e}_t\,(e_0,\,\theta)\,,\,\theta)\,.$$

Nous supposons les p_t lisses, et pouvons ainsi différentier les pertes sur le paramètre $p_t \circ g_t$ obtenues. Par compostion, leurs différentielles sont égales à

$$\frac{\partial\, p_t \circ g_t}{\partial \theta}\,(e_0,\,\theta) = \frac{\partial\, p_t}{\partial e}\,(\boldsymbol{e}_t\,(e_0,\,\theta)\,,\,\theta) \cdot \frac{\partial\, \boldsymbol{e}_t\,(e_0,\,\theta)}{\partial \theta} + \frac{\partial\, p_t}{\partial \theta}\,(\boldsymbol{e}_t\,(e_0,\,\theta)\,,\,\theta)\,.$$

Nous obtenons ainsi des formes linéaires sur l'espace tangent à Θ. Nous identifions enfin ces formes linéaires à des vecteurs tangents à Θ, en travaillant donc en métrique euclidienne.

Nous devons à présent étudier la dynamique temporelle de ces pertes, le long de la trajectoire utilisant un paramètre privilégié θ^*.

Critère d'optimalité

Afin d'optimiser le système dynamique étudié, nous demandons qu'il vérifie les conditions d'ordre un et deux courantes (annulation de la dérivée, différentielle seconde positive) en moyenne temporelle, le long d'une trajectoire dite optimale selon la notion développée dans l'introduction, et présentée à la Définition 3.3.

Une fois vérifiées les hypothèses sur les pertes de la Définition 3.3, et munis d'une suite de pas de descente convenable, nous construisons une suite d'instants (T_k) qui vérifie les propriétés suivantes. Sur les intervalles de temps $[T_k,\, T_{k+1}[$, la somme des gradients est faible devant la somme des pas, c'est-à-dire

$$\sum_{t=T_k}^{T_{k+1}-1} \eta_t \, \frac{\partial \, p_t \circ g_t}{\partial \theta} \, (e_0^*,\, \theta^*) = o\left(\sum_{t=T_k}^{T_{k+1}-1} \eta_t \right),$$

et la moyenne pondérée des hessiennes est définie positive, soit

$$\frac{\sum_{t=T_k}^{T_{k+1}-1} \eta_t \, \frac{\partial^2 \, p_t \circ g_t}{\partial \theta^2} \, (e_0^*,\, \theta^*)}{\sum_{t=T_k}^{T_{k+1}-1} \eta_t} \succ \lambda_{\min} \, \mathrm{Id}_\Theta,$$

pour un $\lambda_{\min} > 0$, quand k tend vers l'infini. La somme des pas sur les intervalles $[T_k, T_{k+1}[$ se comporte alors comme une sorte de « temps intrinsèque » de l'algorithme. La première équation ci-dessus se lit alors : la somme des gradients sur l'intervalle est petite devant la taille, ou la durée, de celui-ci.

Cette notion étend le cas courant (optimisation d'une perte unique) : en effet, dans ce cas l'échelle de temps (T_k) est l'échelle de temps donnée par les itérations de l'algorithme, et à tout instant t,

$$\eta_t \, \frac{\partial \, p}{\partial \theta} \, (\theta^*) = 0 = o\left(\eta_t\right) \quad \text{et} \quad \frac{\eta_t \, \frac{\partial^2 \, p}{\partial \theta^2} \, (\theta^*)}{\eta_t} = \frac{\partial^2 \, p}{\partial \theta^2} \, (\theta^*) \succ \lambda_{\min},$$

pour un $\lambda_{\min} > 0$.

Une fois définis le système dynamique, les pertes, et vérifiées les hypothèses garantissant l'existence d'un paramètre optimal, nous avons besoin d'une procédure pour mettre à jour le paramètre.

Opérateur de mise à jour du paramètre

Ainsi qu'il a été dit dans l'introduction, le schéma le plus classique de descente de gradient est le suivant. Soit un paramètre θ, un vecteur tangent à Θ que nous notons v, et un pas de descente $\eta \geq 0$. Alors, le paramètre est mis à jour par la relation

$$\theta \leftarrow \theta - \eta \, v.$$

Nous souhaitons travailler dans un cadre un peu plus général. En effet, la convergence de l'algorithme RTRL s'obtient en analysant celui-ci à l'ordre 1, au voisinage de 0,

en le pas de descente η. Plutôt que de supposer l'opérateur de mise à jour linéaire en le pas, nous souhaitons montrer que la convergence est robuste à la présence éventuelle de termes d'ordre 2 dans celui-ci. La mise à jour ci-dessus représente alors un développement à l'ordre 1, au voisinage de $\eta = 0$, de l'opérateur que nous considérons.

Vecteurs tangents utilisés par l'algorithme RTRL

Les vecteurs tangents utilisés par l'algorithme RTRL sont des approximations des vecteurs tangents

$$\frac{\partial\, p_t \circ g_t}{\partial \theta}\,(e_0,\, \theta) = \frac{\partial\, p_t}{\partial e}\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta) \cdot \frac{\partial\, \boldsymbol{e}_t\,(e_0,\, \theta)}{\partial \theta} + \frac{\partial\, p_t}{\partial \theta}\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta)\,.$$

En effet, l'algorithme RTRL maintient une approximation des différentielles

$$\frac{\partial\, \boldsymbol{e}_t\,(e_0,\, \theta)}{\partial \theta}.$$

Notons J_t l'approximation, θ_t le paramètre et e_t l'état de l'algorithme à l'instant t. Alors, l'approximation à l'instant $t+1$ est égale à

$$J_{t+1} = \frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e_t,\, \theta_t) \cdot J_t + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e_t,\, \theta_t)\,.$$

Nous admettons également la présence d'une perturbation additive extérieure bornée ξ_t sur cette mise à jour, de sorte que celle-ci devient

$$J_{t+1} = \frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e_t,\, \theta_t) \cdot J_t + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e_t,\, \theta_t) + \xi_t.$$

Nous formulons enfin une hypothèse garantissant que l'influence du bruit sera négligeable sur les intervalles $[T_k,\, T_{k+1}[$. Cette hypothèse sera vérifiée lors de l'étude de l'algorithme « NoBackTrack ».

Algorithme RTRL, et propriété de structure vérifiée par celui-ci

Munis du système dynamique, d'un optimum local, d'un opérateur de mise à jour du paramètre et d'un procédé pour calculer des vecteurs tangents, nous pouvons considérer l'algorithme RTRL, présenté dans l'introduction à la Définition 3.2. Celui-ci maintient le paramètre, l'état courant et une approximation de la différentielle J_t de l'état courant par rapport au paramètre.

À chaque instant t, il calcule un vecteur tangent à Θ. L'utilisateur lui fournit un pas de descente η_t, et l'algorithme utilise alors l'opérateur Φ pour calculer le nouveau paramètre, θ_{t+1}. Il met ensuite à jour l'état du système (en utilisant l'ancien paramètre) [19].

Propriété de structure de l'algorithme RTRL L'algorithme RTRL vérifie alors la propriété structure suivante. Notons, pour tout paramètre θ, $d\,(\theta,\, \theta^*)$ la

19. Ce dernier point n'est pas nécessaire, mais c'est la convention que nous avons adoptée dans la preuve.

distance de ce dernier au paramètre optimal θ^*. Alors, pour tout $T \geq 0$, le paramètre θ_T produit par l'algorithme RTRL après T itérations vérifie une inégalité de la forme :

$$d\left(\theta_T,\, \theta^*\right) \leq \left(1 - \lambda_T \sum_{t=0}^{T} \eta_t\right) d\left(\theta_0,\, \theta^*\right) + b_T.$$

Le second membre de celle-ci se compose de deux termes : un terme de contractivité, et un terme d'erreur additif.

Terme de contractivité La quantité λ_T du terme de contractivité est un minorant de la plus petite valeur propre de la moyenne, pondérée par les pas de descente, des hessiennes exactes des pertes en θ^*. Pour que l'algorithme puisse apprendre, nous avons besoin que λ_T soit strictement positive.

Terme d'erreur additif Le terme additif b_T est la somme des erreurs effectuées dans les développements de Taylor, et de la moyenne pondérée des gradients exacts calculés avec le paramètre θ^*. Il faut donc que ces termes soient négligeables devant la quantité

$$\lambda_T \sum_{t=0}^{T} \eta_t$$

du terme de contractivité.

Schéma de descente de gradient stochastique vérifié en moyenne Ainsi, l'inégalité vérifiée par l'algorithme RTRL est l'analogue, en moyenne sur l'intervalle de temps $[0,\, T]$, de celle vérifiée dans le cadre d'une descente de gradient stochastique classique. La descente de gradient se comporte donc comme une descente de gradient stochastique, mais sur une échelle de temps (T_k) différente de l'échelle de temps $t, t+1, \ldots$ donnée par les itérations du système dynamique.

Convergence de l'algorithme RTRL

Le comportement moyen de l'algorithme sur les intervalles construits permet alors, en recollant ceux-ci, de prouver la convergence de l'algorithme RTRL. Toutefois, ce contrôle en moyenne n'est vérifié qu'asymptotiquement. Nous devons ainsi nous assurer qu'il est possible de ralentir la descente suffisamment, c'est-à-dire de trouver un pas initial suffisamment petit, pour que l'algorithme n'explose pas avant que les conditions garantissant la convergence ne soient établies. Ceci fait, nous obtenons alors le résultat annoncé.

Récapitulation

Ainsi, la preuve proposée procède en sept temps, correspondant chacun à un chapitre.

1. L'étude du système dynamique.
2. L'étude des pertes.
3. La définition du critère d'optimalité.
4. L'étude de l'opérateur de déplacement sur l'espace des paramètres.
5. La construction des vecteurs tangents utilisés lors de l'optimisation.

6. La définition de l'algorithme RTRL, et l'établissement de la propriété de structure qu'il vérifie.
7. L'établissement de la convergence de l'algorithme RTRL.

Les trois premiers chapitres concernent le système dynamique étudié. Les chapitres quatre à six concernent l'algorithme d'optimisation utilisé.

Les chapitres un, deux, quatre et cinq sont des chapitres techniques. Le chapitre six regroupe les différentes propriétés qui y sont obtenues pour obtenir l'inégalité de structure vérifiée par l'algorithme RTRL.

Le dernier chapitre joint alors les résultats de structure de l'algorithme aux résultats concernant l'optimalité obtenus au chapitre trois, afin d'établir la convergence de l'algorithme.

Résultats et arguments principaux par chapitre

Nous indiquons ici rapidement, pour chaque chapitre, les résultats principaux qu'il établit ainsi que les arguments invoqués.

Système dynamique à paramètre

Supposons un instant le système dynamique linéaire, vérifiant l'équation de récurrence

$$e_{t+1} = A\, e_t + B\, \theta.$$

Notons (e'_t) une trajectoire issue d'un état initial e'_0, et avec un paramètre θ'. Alors, l'écart entre cette trajectoire et la trajectoire (e_t) n'explose pas en temps fini. De plus, si $e_0 = e'_0$, les écarts sont homogène en la différence des paramètres soit, pour $t \geq 1$,

$$e_t - e'_t = \left(\sum_{s=0}^{t-1} A^{t-s}\, B \right) \, (\theta - \theta')$$

et, si $\theta = \theta'$, l'oubli de la différence initiale entre les états est exponentiel soit, pour $t \geq 1$,

$$e_t - e_t = A^{t-1} \, (e_0 - e'_0)\,.$$

En supposant bornées les différentielles successives des opérateurs de transition, nous montrons que le système des écarts est sous-linéaire. Le Lemme 4.41 donne ainsi l'inégalité

$$d\,(e_{t+1},\, e'_{t+1}) \leq (1-\alpha)\; d\,(e_t,\, e'_t) + M_1\, d\,(\theta,\, \theta')\,.$$

Ainsi le système dynamique présente, sur les boules où ces différentielles sont bornées, le même comportement qu'un système linéaire. Nous pouvons alors construire les boules $B^*_{\mathcal{E}_t}$ qui contiendront les trajectoires étudiées, à la section 4.5, puis contrôler les écarts à la section 4.6. Les Figures 3.2 et 3.3 illustrent ceci.

Les différentielles des états se comportent de la même manière, ce que nous établissons également.

Pertes et critère d'optimalité

Dans le chapitre sur l'optimalité, le Fait 6.29 et le Lemme 6.30 établissent les relations asymptotiques vérifiées en l'extremum local θ^*.

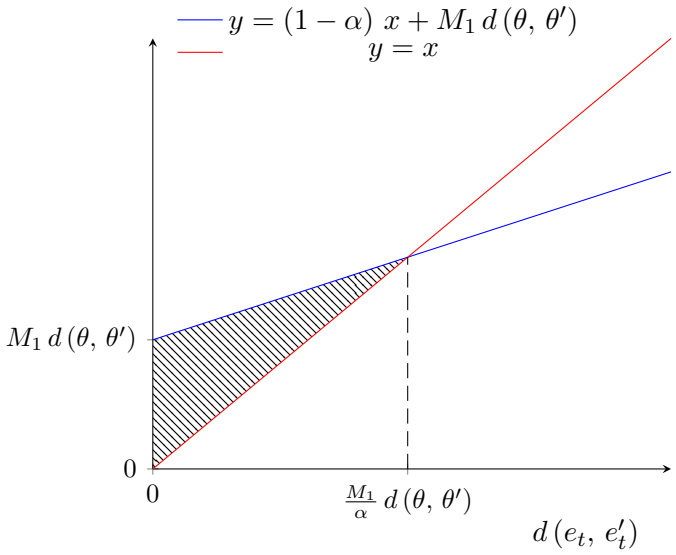

FIGURE 3.2: Contrôle du système sous-linéaire, pour des paramètres différents. La zone hachurée est stable par la fonction tracée en bleu.

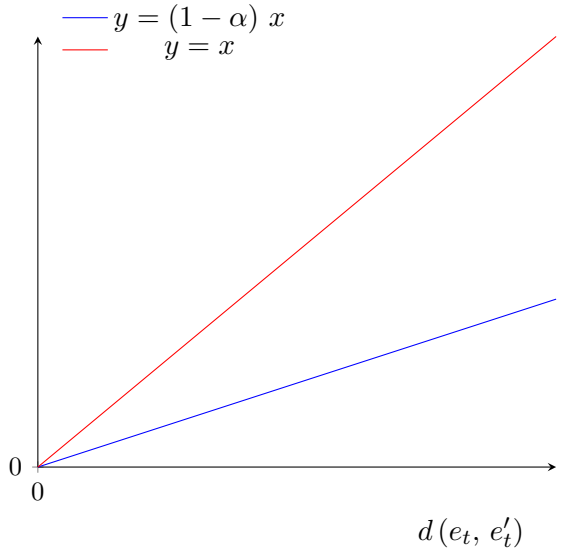

FIGURE 3.3: Oubli exponentiel des états initiaux

Opérateur de mise à jour

Le chapitre sur l'opérateur de mise à jour contrôle les développements à l'ordre 2 de l'opérateur Φ. Le résultat principal est le Lemme 7.10 qui montre que la mise à jour du paramètre est sous-linéaire en le pas utilisé. De même que pour le système dynamique, le comportement de l'opérateur de mise à jour est ainsi le même que celui de la mise à jour

$$\theta \leftarrow \theta - \eta\, v.$$

Vecteurs tangents utilisés lors de l'optimisation

Le chapitre sur les vecteurs tangents tranfère les propriétés d'homogénéité en la différence des états et d'oubli exponentiel aux vecteurs tangents qui seront utilisés lors de l'optimisation, à la section 8.1.2. Il établit également, au Corollaire 8.15, que à l'ordre un en l'écart des paramètres, les vecteurs tangents bruités sont égaux aux vecteurs tangents n'utilisant pas le bruit auxquels est ajouté un terme dont la somme des contributions desquels sur les intervalles $[T_k, T_{k+1}[$ est supposée négligeable devant la somme des pas.

Propriété de structure de l'algorithme RTRL

Trajectoires intermédiaires Une des difficultés de l'étude de l'algorithme RTRL réside dans le fait que toutes les quantités maintenues par celui-ci évoluent simultanément. Afin de remédier à cela, nous introduisons des trajectoires intermédiaires, calculées par l'algorithme fonctionnant en « boucle ouverte » : les mises à jour des paramètres sont calculées en utilisant un paramètre fixe.

Pour des durées courtes, c'est-à-dire des T tels que la somme des pas de 0 à T est petite, la différence entre ces trajectoires et la trajectoire RTRL est négligeable devant la somme des pas, et ces erreurs seront ainsi absorbées par le terme de contractivité. Nous le démontrons à la section 9.4.3. Ceci est la conséquence des contrôles sur les écarts établis précédemment.

Contractivité et propriété de structure Le Fait 9.28 est le résultat clef, qui établit l'inégalité qui permettra d'obtenir la contractivité. La Proposition 9.36 réunit alors tous ces arguments pour établir l'inégalité de structure vérifiée par l'algorithme RTRL. Cette propriété est appelée « Propriété centrale de l'algorithme RTRL ».

Convergence de l'algorithme RTRL

Le chapitre procède en trois temps. La continuité des trajectoires de l'algorithme RTRL par rapport au pas de descente est l'objet de la section 10.1. La section 10.2 montre ensuite qu'il est possible de diminuer suffisamment le pas pour que la descente n'explose pas avant que les hypothèses garantissant la convergence ne soient satisfaites. La convergence d' l'algorithme RTRL est alors établie à la section 10.3.

Le résultat final est énoncé à la section 10.4, ainsi que le corollaire sur la convergence des états et des différentielles de la trajectoire RTRL vers leurs homologues de la trajectoire optimale.

Chapitre

4 Système dynamique à paramètre et trajectoire stable

4.1 Espaces des valeurs, opérateurs

4.1.1 Espaces des valeurs, structure différentiable et topologie

Hypothèse 4.1 (Espaces $\mathcal{E}_t$ des états)**.** *Nous supposons disposer d'une famille, indexée par t entier positif, d'espaces $\mathcal{E}_t$, dans lesquels les états successifs du système dynamique prennent leurs valeurs. Nous supposons que ces espaces sont des espaces $\mathbb{R}^n$, pour des n éventuellement différents.*

Hypothèse 4.2 (Espace Θ du paramètre)**.** *Nous supposons également disposer d'un espace Θ, auquel appartient le paramètre du sytème dynamique. Nous supposons également que c'est un $\mathbb{R}^n$, pour un certain n.*

Pour contrôler le système dynamique étudié, nous souhaitons disposer d'une structure différentiable sur l'espace Θ et les espaces $\mathcal{E}_t$. Or, comme ce sont des espaces vectoriels, nous pouvons leur identifier leurs espaces tangents. Nous utiliserons toutefois, parfois, la notation $T\Theta$ pour distinguer Θ et son espace tangent, afin de préciser le rôle des quantités impliquées. De même, nous utiliserons la notation $T\mathcal{E}_t$.

Afin de prouver la convergence de l'algorithme RTRL appliqué au système dynamique, nous souhaitons disposer d'une topologie sur l'espace Θ et les espaces $\mathcal{E}_t$. Nous procédons alors de la manière suivante.

Définition 4.3 (Normes sur les espaces des états et des paramètres)**.** *Nous munissons chaque espace $\mathcal{E}_t$ de la norme euclidienne relative à la base de travail, que nous notons dans chaque cas $\|\cdot\|$.*

Nous munissons l'espace Θ de la norme euclidienne relative à la base de travail, que nous notons également $\|\cdot\|$.

Enfin, pour des produits cartésiens comme $\mathcal{E}_t \times \Theta$ nous notons, pour $e \in \mathcal{E}_t$ et $\theta \in \Theta$,

$$\|(e, \theta)\| = \left(\|e\|^2 + \|\theta\|^2\right)^{1/2},$$

où les normes de droite sont celles définies ci-dessus.

Dans la suite, nous aurons besoin d'identifier des formes linéaires définies sur Θ à des vecteurs. Pour cela, nous considérons que cet espace est muni de la métrique euclidienne relative à la base de travail.

Remarque 4.4. *Le choix de la norme euclidienne sur les espaces $\mathcal{E}_t$ n'est pas nécessaire, comme nous n'avons pas besoin d'une norme particulière, et que toutes les normes sont équivalentes en dimension finie sur $\mathbb{R}$, mais simplifie la définition de la*

norme d'opérateur sur les dérivées secondes des opérateurs définis sur ces espaces, effectuée ci-dessous.

Le choix de la métrique euclidienne sur l'espace $T\Theta$ n'est pas non plus nécessaire, mais simplifie l'exposition.

Dans la suite, nous exprimerons certains résultats obtenus en termes de distances. En effet, cela permet de mieux mettre en évidence les mécanismes utilisés dans certaines preuves, dont plusieurs ne sont pas tributaires de la structure d'espace vectoriel. Pour cela, nous notons, pour $t \geq 0$ et e et e' appartenant à $\mathcal{E}_t$,

$$d\left(e,\, e'\right) = \left\|e - e'\right\|.$$

De même, pour θ et θ' dans Θ, nous notons

$$d\left(\theta,\, \theta'\right) = \left\|\theta - \theta'\right\|.$$

4.1.2 Choix des normes pour les opérateurs sur les espaces tangents

Définition des normes

Nous choisissons de travailler avec les normes d'opérateurs. Ce choix n'est pas nécessaire, mais il permet des calculs relativement simples.

Définition 4.5 (Norme d'une application linéaire)**.** *Soient $\mathcal{E}_1$ et $\mathcal{E}_2$ deux espaces vectoriels réels. Soit une application linéaire*

$$f \,:\, \mathcal{E}_1 \to \mathcal{E}_2.$$

Munissons $\mathcal{E}_1$ et $\mathcal{E}_2$ des normes euclidiennes relatives aux bases de travail dans ces espaces, notées $\|\cdot\|$ dans chaque cas. Nous définissons alors la norme

$$\|f\|_{\text{op}} = \sup_{v_1 \in \mathcal{E}_1,\, v_1 \neq 0} \frac{\|f(v_1)\|}{\|v_1\|}.$$

Pour une forme linéaire, la norme d'opérateur est égale à la norme du vecteur associé à celle-ci par le produit scalaire sous-jacent à la norme euclidienne (son gradient par rapport à ce produit scalaire). Dans la suite, nous identifierons ainsi ces deux normes.

Définition 4.6 (Norme d'une application bilinéaire)**.** *Soient $\mathcal{E}_1$, $\mathcal{E}_2$ et $\mathcal{E}_3$ trois espaces vectoriels réels. Soit une application bilinéaire*

$$f \,:\, \mathcal{E}_1 \times \mathcal{E}_2 \to \mathcal{E}_3.$$

Munissons $\mathcal{E}_1$, $\mathcal{E}_2$ et $\mathcal{E}_3$ des normes euclidiennes relatives aux choix de bases de travail dans ces espaces, notées $\|\cdot\|$ dans chaque cas. Nous définissons alors la norme

$$\|f\|_{\text{bil}} = \sup_{v_1 \in \mathcal{E}_1,\, v_1 \neq 0,\, v_2 \in \mathcal{E}_2,\, v_2 \neq 0} \frac{\|f\left(v_1,\, v_2\right)\|}{\|v_1\|\ \|v_2\|}.$$

Remarque 4.7. *$\|f\|_{\text{bil}}$ est également la norme d'opérateur de la matrice représentative de f dans les bases considérées.*

Remarque 4.8. *De la sorte, pour une application deux fois différentiable à valeurs réelle, la norme ainsi définie de sa différentielle seconde est la norme d'opérateur de sa matrice (la hessienne) dans les bases de travail.*

Remarque 4.9. *Avec les notations précédentes, pour tous $v_1 \in \mathcal{E}_1$ et $v_2 \in \mathcal{E}_2$,*

$$\|f(v_1,\, v_2)\| \leq \|f\|_{\text{bil}}\, \|v_1\|\, \|v_2\|\,.$$

Définition 4.10 (Norme d'une application trilinéaire)**.** *Soient $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ et $\mathcal{E}_4$ quatre espaces vectoriels réels. Soit une application trilinéaire*

$$f \,:\, \mathcal{E}_1 \times \mathcal{E}_2 \times \mathcal{E}_3 \to \mathcal{E}_4.$$

Munissons $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ et $\mathcal{E}_4$ des normes euclidiennes relatives aux choix de bases de travail dans ces espaces, notées $\|\cdot\|$ dans chaque cas. Nous définissons alors la norme

$$\|f\|_{\text{tri}} = \sup_{v_i \in \mathcal{E}_i,\, v_i \neq 0,\, 1 \leq i \leq 3} \frac{\|f\,(v_1,\, v_2,\, v_3)\|}{\|v_1\|\, \|v_2\|\, \|v_3\|}.$$

Propriétés des normes

Propriétés des normes des applications bilinéaires

Corollaire 4.11 (Norme bilinéaire lors d'une composition à droite par deux applications linéaires)**.** *Nous conservons les notations précédentes. Soit une application bilinéaire f. Considérons de plus une application linéaire g d'un espace $\mathcal{E}_4$ vers $\mathcal{E}_1$, et g' d'un espace $\mathcal{E}_5$ vers $\mathcal{E}_2$. Alors,*

$$\|f\,(g,\, g')\|_{\text{bil}} \leq \|f\|_{\text{bil}}\, \|g\|_{\text{op}}\, \|g'\|_{\text{op}}\,.$$

Démonstration. Soit $v_4 \in \mathcal{E}_4$ et $v_5 \in \mathcal{E}_5$. Alors, d'après la Remarque 4.9,

$$\begin{aligned}\|f\,(g(v_4),\, g'(v_5))\| &\leq \|f\|_{\text{bil}}\, \|g(v_4)\|\, \|g'(v_5)\| \\ &\leq \|f\|_{\text{bil}}\, \|g\|_{\text{op}}\, \|v_4\|\, \|g'\|_{\text{op}}\, \|v_5\|\,.\end{aligned}$$

Donc, pour tout v_4 non nul appartenant à $\mathcal{E}_4$, et tout v_5 non nul appartenant à $\mathcal{E}_5$,

$$\frac{\|f\,(g(v_4),\, g'(v_5))\|}{\|v_4\|\, \|v_5\|} \leq \|f\|_{\text{bil}}\, \|g\|_{\text{op}}\, \|g'\|_{\text{op}}$$

ce qui, par définition de la norme $\|\cdot\|_{\text{bil}}$, donnée à la Définition 4.6, donne le résultat annoncé. □

Corollaire 4.12 (Norme d'opérateur lors d'une composition à droite par une application linéaire)**.** *Nous conservons les notations précédentes. Soit $v_1 \in \mathcal{E}_1$. Alors,*

$$\|f\,(v_1,\, g)\|_{\text{op}} \leq \|f\|_{\text{bil}}\, \|v_1\|\, \|g\|_{\text{op}}\,.$$

Démonstration. Soit $v_4 \in \mathcal{E}_4$. Alors, le Corollaire 4.11 appliqué à $g = \text{Id}$ et $g' = g$ donne :

$$\|f\,(v_1,\, g(v_4))\| \leq \|f\|_{\text{bil}}\, \|g\|_{\text{op}}\, \|v_1\|\, \|v_4\|\,.$$

Ainsi, pour tout v_4 non nul appartenant à $\mathcal{E}_4$,

$$\frac{\|f\,(v_1,\, g(v_4))\|}{\|v_4\|} \leq \|f\|_{\text{bil}}\, \|g\|_{\text{op}}\, \|v_1\|\,,$$

ce qui conclut la preuve. □

Corollaire 4.13 (Norme bilinéaire lors d'une composition à gauche par une application linéaire)**.** *Soit une application linéaire g de $\mathcal{E}_3$ dans un espace vectoriel qu'il n'est pas nécessaire de spécifier. Alors,*

$$\|g \circ f\|_{\text{bil}} \leq \|g\|_{\text{op}} \, \|f\|_{\text{bil}} \, .$$

Démonstration. Soit $v_1 \in \mathcal{E}_1$, et $v_2 \in \mathcal{E}_2$. Alors,

$$\|g\left(f\left(v_1,\, v_2\right)\right)\| \leq \|g\|_{\text{op}} \, \|f\left(v_1,\, v_2\right)\| \leq \|g\|_{\text{op}} \, \|f\|_{\text{bil}} \, \|v_1\| \, \|v_2\| \, ,$$

d'après la Remarque 4.9. Ainsi, pour tous vecteurs non nuls $v_1 \in \mathcal{E}_1$ et $v_2 \in \mathcal{E}_2$,

$$\frac{\|g\left(f\left(v_1,\, v_2\right)\right)\|}{\|v_1\| \, \|v_2\|} \leq \|g\|_{\text{op}} \, \|f\|_{\text{bil}} \, ,$$

ce qui conclut la preuve, par définition de la norme $\|\cdot\|_{\text{bil}}$. □

Propriétés des normes des applications trilinéaires

Corollaire 4.14 (Norme trilinéaire lors d'une composition à droite par trois applications linéaires)**.** *Nous conservons les notations précédentes. Soit une application trilinéaire f. Considérons de plus une application linéaire g d'un espace $\mathcal{E}_5$ vers $\mathcal{E}_1$, g' d'un espace $\mathcal{E}_6$ vers $\mathcal{E}_2$ et g'' d'un espace $\mathcal{E}_7$ vers $\mathcal{E}_3$. Alors,*

$$\|f\left(g,\, g',\, g''\right)\|_{\text{tri}} \leq \|f\|_{\text{tri}} \, \|g\|_{\text{op}} \, \|g'\|_{\text{op}} \, \|g''\| \, .$$

Démonstration. La preuve est similaire à celle du Corollaire 4.11. □

Des résultats analogues à ceux obtenus sur les applications bilinéaires seront utilisés pour des applications trilinéaires. Leurs démonstrations sont identiques aux précédentes, et nous ne les incluons pas, à l'exception de celle ci-dessus.

Distance sur les espaces des opérateurs

Pour f et f' deux applications linéaires de $\mathcal{E}_1 \to \mathcal{E}_2$, nous notons

$$d\left(f,\, f'\right) = \|f - f'\|_{\text{op}} \, .$$

4.2 Système dynamique à paramètre

4.2.1 Définition du système dynamique

Hypothèse 4.15 (Fonctions de transition et régularité)**.** *Nous supposons disposer, pour $t \geq 0$, de fonctions de transition*

$$\boldsymbol{T}_t : \mathcal{E}_t \times \Theta \to \mathcal{E}_{t+1}$$

telles que, si le système étudié est dans l'état e_t à l'instant t, et que le paramètre est θ_t, son état à l'instant $t+1$ est e_{t+1} donné par la relation :

$$e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta_t\right) .$$

Nous supposons de plus que les $\boldsymbol{T}_t$ sont trois fois continûment différentiables.

Remarque 4.16. *Nous avons besoin de la différentiabilité des $\boldsymbol{T}_t$ pour définir des pertes sur le paramètre, de la double différentiabilité pour pouvoir contrôler les différentielles de ces pertes, et de la triple différentiabilité pour contrôler les hessiennes de celles-ci.*

Définition 4.17 (Trajectoire du système dynamique, paramètre fixe)**.** *Soit la suite d'états $(e_t)_{t\geq 0}$ définie par la donnée d'un état initial $e_0 \in \mathcal{E}_0$, d'un paramètre $\theta \in \Theta$ et la relation de récurrence, pour $t \geq 0$,*

$$e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta\right). \tag{4.1}$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \boldsymbol{e}_t :\ & \mathcal{E}_0 \times \Theta \to \mathcal{E}_t \\ & e_0,\, \theta \mapsto \boldsymbol{e}_t\left(e_0,\, \theta\right) = e_t, \end{aligned}$$

qui associent à l'état initial et au paramètre l'état du système à l'instant t.

Notons Θ^∞ l'espace des suites à valeurs dans Θ.

Définition 4.18 (Trajectoire du système dynamique, suite de paramètres)**.** *Soit la suite $(e_t)_{t\geq 0}$ définie par la donnée d'un état initial e_0, d'une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ à valeurs dans Θ et la relation de récurrence, pour $t \geq 0$,*

$$e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta_t\right). \tag{4.2}$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \boldsymbol{e}_t :\ & \mathcal{E}_0 \times \Theta^\infty \to \mathcal{E}_t \\ & e_0,\, \boldsymbol{\theta} \mapsto \boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right) = e_t, \end{aligned}$$

qui associent à l'état initial et à la suite de paramètres l'état du système à l'instant t.

4.2.2 Sensibilité des états au paramètre

Grâce à la régularité des fonctions de transition, les états sont des fonctions régulières des paramètres. Nous définissons à présent leurs différentielles par rapport à ceux-ci. Pour $t \geq 0$, nous notons

$$\mathcal{L}\left(\Theta,\, \mathcal{E}_t\right)$$

l'espace des applications linéaires de l'espace tangent à Θ vers l'espace tangent à $\mathcal{E}_t$ identifiés, conformément à ce qui a été dit précédemment, aux espaces auxquels ils sont tangents.

Conformément à ce qui a été dit à la fin de la section 4.1.2, pour J et J' dans un espace $\mathcal{L}\left(\Theta,\, \mathcal{E}_t\right)$, nous notons

$$d\left(J,\, J'\right) = \left\|J - J'\right\|_{\mathrm{op}}.$$

Définition 4.19 (Équation de récurrence satisfaite par les différentielles des états par rapport au paramètre)**.** *Soit la suite $(J_t)_{t\geq 0}$ définie par la donnée d'un état*

initial e_0, d'une application linéaire initiale $J_0 \in \mathcal{L}(\Theta, \mathcal{E}_0)$, d'un paramètre θ et la relation de récurrence, pour $t \geq 0$,

$$J_{t+1} = \frac{\partial \boldsymbol{T}_t}{\partial e}(\boldsymbol{e}_t(e_0, \theta), \theta) \cdot J_t + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(\boldsymbol{e}_t(e_0, \theta), \theta).$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \boldsymbol{J}_t : \quad & \mathcal{E}_0 \times \mathcal{L}(\Theta, \mathcal{E}_0) \times \Theta \to \mathcal{L}(\Theta, \mathcal{E}_t) \\ & e_0, J_0, \theta \mapsto \boldsymbol{J}_t(e_0, J_0, \theta) = J_t. \end{aligned}$$

Remarque 4.20. *Nous notons J pour « jacobienne ».*

Lemme 4.21 (Justification de l'appellation différentielle)**.**

$$\boldsymbol{J}_t(e_0, J_0, \theta) = \frac{\partial \boldsymbol{e}_t}{\partial e_0}(e_0, \theta) \cdot J_0 + \frac{\partial \boldsymbol{e}_t}{\partial \theta}(e_0, \theta).$$

En particulier, pour $J_0 = 0$,

$$\boldsymbol{J}_t(e_0, 0, \theta) = \frac{\partial \boldsymbol{e}_t(e_0, \theta)}{\partial \theta}.$$

Démonstration. Les $\boldsymbol{e}_t$ sont régulières, par récurrence, car les $\boldsymbol{T}_t$ le sont, d'après l'Hypothèse 4.15. La formule est alors une conséquence, également par récurrence, de la dérivation de l'Équation (4.1) par rapport à θ. □

Définition 4.22 (Équation de récurrence satisfaite par les différentielles des états par rapport à une suite de paramètres)**.** *Soit la suite $(J_t)_{t \geq 0}$ définie par la donnée d'un état initial e_0, d'une application linéaire initiale $J_0 \in \mathcal{L}(\Theta, \mathcal{E}_0)$, d'une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ à valeurs dans Θ et la relation de récurrence, pour $t \geq 0$,*

$$J_{t+1} = \frac{\partial \boldsymbol{T}_t}{\partial e}(\boldsymbol{e}_t(e_0, \boldsymbol{\theta}), \theta_t) \cdot J_t + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(\boldsymbol{e}_t(e_0, \boldsymbol{\theta}), \theta_t).$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \boldsymbol{J}_t : \quad & \mathcal{E}_0 \times \mathcal{L}(\Theta, \mathcal{E}_0) \times \Theta^\infty \to \mathcal{L}(\Theta, \mathcal{E}_t) \\ & e_0, J_0, \boldsymbol{\theta} \mapsto \boldsymbol{J}_t(e_0, J_0, \boldsymbol{\theta}) = J_t. \end{aligned}$$

Définition 4.23 (Injection de l'espace tangent à Θ dans l'espace des suites à valeurs dans cet espace)**.** *Nous définissons la fonction*

$$\begin{aligned} \mathrm{Id}_\Theta^\infty : \quad & T\Theta \to T\Theta^\infty \\ & v \mapsto (v, v, \ldots). \end{aligned}$$

Lemme 4.24 (Justification de l'appellation différentielle)**.**

$$\boldsymbol{J}_t(e_0, J_0, \boldsymbol{\theta}) = \frac{\partial \boldsymbol{e}_t}{\partial e_0}(e_0, \boldsymbol{\theta}) \cdot J_0 + \frac{\partial \boldsymbol{e}_t}{\partial \boldsymbol{\theta}}(e_0, \boldsymbol{\theta}) \cdot \mathrm{Id}_\Theta^\infty.$$

Démonstration. De même que précédemment, les $\boldsymbol{e}_t$ sont régulières, par récurrence, car les $\boldsymbol{T}_t$ le sont, d'après l'Hypothèse 4.15. La formule est alors une conséquence, également par récurrence, de la dérivation de l'Équation (4.2) par rapport à $\boldsymbol{\theta}$. □

Lemme 4.25 (Différentielles deuxièmes des fonctions de transition par rapport au paramètre)**.** *Soit $e_0 \in \mathcal{E}_0$ fixé. Les fonctions*

$$\theta \mapsto \boldsymbol{e}_t\left(e_0,\, \theta\right),$$

pour $t \geq 0$, sont deux fois continûment différentiables. Leurs différentielles secondes vérifient, pour tout $t \geq 0$,

$$\begin{aligned}\frac{\partial^2 \boldsymbol{e}_{t+1}\left(e_0,\, \theta\right)}{\partial \theta^2} &= \mathrm{d}^2 \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left[\left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right),\, \left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2}.\end{aligned}$$

Démonstration. De même que précédemment, les $\boldsymbol{e}_t$ sont deux fois continûment différentiables, par récurrence, car les $\boldsymbol{T}_t$ le sont. Soit $t \geq 0$. Nous savons que

$$\begin{aligned}\frac{\partial \boldsymbol{e}_{t+1}\left(e_0,\, \theta\right)}{\partial \theta} &= \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta} + \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \\ &= \mathrm{d}\, \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right).\end{aligned}$$

Ainsi,

$$\begin{aligned}\frac{\partial^2 \boldsymbol{e}_{t+1}\left(e_0,\, \theta\right)}{\partial \theta^2} &= \mathrm{d}^2 \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left[\left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right),\, \left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \mathrm{d}\, \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left(\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, 0\right) \\ &= \mathrm{d}^2 \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left[\left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right),\, \left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\end{aligned}$$

ce qui est le résultat annoncé. □

Lemme 4.26 (Différentielles troisièmes des fonctions de transition par rapport au paramètre)**.** *Soit $e_0 \in \mathcal{E}_0$ fixé. Les fonctions*

$$\theta \mapsto \boldsymbol{e}_t\left(e_0,\, \theta\right),$$

pour $t \geq 0$, sont trois fois continûment différentiables. Leurs différentielles troisièmes vérifient, pour tout $t \geq 0$,

$$\begin{aligned}\frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0,\, \theta\right)}{\partial \theta^3} &= \mathrm{d}^3 \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right)^{\otimes 3} \\ &\quad + 2\, \mathrm{d}^2 \boldsymbol{T}_t\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left[\left(\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, 0\right),\, \left(\frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta},\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \left[\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, \frac{\partial \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta}\right] \\ &\quad + \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right) \cdot \frac{\partial^3 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^3}.\end{aligned}$$

Démonstration. De même que précédemment, les $\boldsymbol{e}_t$ sont trois fois continûment différentiables, par récurrence, car les $\boldsymbol{T}_t$ le sont. Soit $t \geq 0$. Nous savons que

$$\begin{aligned}\frac{\partial^2 \boldsymbol{e}_{t+1}(e_0,\,\theta)}{\partial\theta^2} &= \mathrm{d}^2 \boldsymbol{T}_t\left(\boldsymbol{e}_t(e_0,\,\theta),\,\theta\right)\cdot\left[\left(\frac{\partial \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta},\,\mathrm{Id}_\Theta\right),\,\left(\frac{\partial \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta},\,\mathrm{Id}_\Theta\right)\right]\\ &\quad+\frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t(e_0,\,\theta),\,\theta\right)\cdot\frac{\partial^2 \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta^2}.\end{aligned}$$

Ainsi,

$$\begin{aligned}\frac{\partial^3 \boldsymbol{e}_{t+1}(e_0,\,\theta)}{\partial\theta^3} &= \mathrm{d}^3 \boldsymbol{T}_t\left(\boldsymbol{e}_t(e_0,\,\theta),\,\theta\right)\cdot\left(\frac{\partial \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta},\,\mathrm{Id}_\Theta\right)^{\otimes 3}\\ &\quad+2\,\mathrm{d}^2 \boldsymbol{T}_t\left(\boldsymbol{e}_t(e_0,\,\theta),\,\theta\right)\cdot\left[\left(\frac{\partial^2 \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta^2},\,0\right),\,\left(\frac{\partial \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta},\,\mathrm{Id}_\Theta\right)\right]\\ &\quad+\frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(\boldsymbol{e}_t(e_0,\,\theta),\,\theta\right)\cdot\left[\frac{\partial^2 \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta^2},\,\frac{\partial \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta}\right]\\ &\quad+\frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t(e_0,\,\theta),\,\theta\right)\cdot\frac{\partial^3 \boldsymbol{e}_t(e_0,\,\theta)}{\partial\theta^3},\end{aligned}$$

ce qui est le résultat annoncé. □

4.3 Contrôle des opérateurs de transition le long de la trajectoire stable

4.3.1 Trajectoire stable

Considérons, pour un état initial $e_0^* \in \mathcal{E}_0$, une application linéaire initiale $J_0^* = 0 \in \mathcal{L}(\Theta,\,\mathcal{E}_0)$, et un paramètre $\theta^* \in \Theta$, les trajectoires (e_t^*) et (J_t^*) qui satisfont les relations de récurrence

$$\begin{cases} e_{t+1}^* = \boldsymbol{T}_t\left(e_t^*,\,\theta^*\right)\\ J_{t+1}^* = \dfrac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t^*,\,\theta^*\right)\cdot J_t^* + \dfrac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t^*,\,\theta^*\right). \end{cases} \tag{4.3}$$

Alors, avec les notations définies précédemment,

$$e_t^* = \boldsymbol{e}_t\left(e_0^*,\,\theta^*\right) \quad \text{et} \quad J_t^* = \boldsymbol{J}_t\left(e_0^*,\,0,\,\theta^*\right).$$

Remarque 4.27. *Il n'est pas nécessaire que J_0^* soit nulle. Cependant, dans l'algorithme RTRL défini dans la suite, il est d'usage d'initialiser la quantité correspondante J_0 à 0. Or, les hypothèses des résultats de convergence présentés par la suite exigent que J_0 et J_0^* soient proches. Ainsi, il est possible de ne pas supposer que J_0^* soit nulle, pourvu que 0 appartienne au voisinage de J_0^* défini à la Définition 4.51, ci-dessous. Il faut de plus rajouter dans les dérivées qui figurent dans l'Hypothèse 6.17 une dérivation par rapport à e_0, et l'évaluation en $(J_0,\,\mathrm{Id}_\Theta)$.*

Nous supposons alors que les hypothèses suivantes sont satisfaites.

Hypothèse 4.28 (Contractivité uniforme en temps le long de la trajectoire stable)**.** *Nous supposons que*

$$\sup_{t\geq 0}\left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t^*,\,\theta^*\right)\right\|_{\mathrm{op}} < 1.$$

Hypothèse 4.29 (Borne uniforme en temps sur les différentielles par rapport au paramètre le long de la trajectoire stable)**.** *Nous supposons que*

$$\sup_{t\geq 0}\left\|\frac{\partial \boldsymbol{T}_t}{\partial\theta}\left(e_t^*,\,\theta^*\right)\right\|_{\mathrm{op}}<\infty.$$

Hypothèse 4.30 (Boules de contrôle uniformes en temps pour le paramètre et les états)**.** *Nous supposons qu'il existe*

1. *une boule $B_\Theta^*\subset\Theta$ centrée en θ^*, de rayon $r_\Theta^*>0$;*
2. *et, pour $t\geq 0$, des boules $B_{\mathcal{E}_t}\subset\mathcal{E}_t$ centrées en*

$$e_t^*=\boldsymbol{e}_t\left(e_0^*,\,\theta^*\right),$$

de rayon commun $r_{\mathcal{E}}$ telles que

$$S=\sup_{t\geq 0}\sup_{(e,\,\theta)\in B_{\mathcal{E}_t}\times B_\Theta^*}\left\|\frac{\partial^2\boldsymbol{T}_t}{\partial\left(e,\,\theta\right)^2}\left(e,\,\theta\right)\right\|_{\mathrm{bil}}<\infty$$

et

$$\sup_{t\geq 0}\sup_{(e,\,\theta)\in B_{\mathcal{E}_t}\times B_\Theta^*}\left\|\frac{\partial^3\boldsymbol{T}_t}{\partial\left(e,\,\theta\right)^3}\left(e,\,\theta\right)\right\|_{\mathrm{tri}}<\infty.$$

Définition 4.31 (Paramètre stable et trajectoire stable)**.** *Un paramètre θ^* tel que la trajectoire associée par l'Équation* (4.3) *satisfasse les Hypothèses 4.28, 4.29 et 4.30 est dit paramètre stable, et la trajectoire associée est dite trajectoire stable.*

Remarque 4.32. *Dans la suite, il nous arrivera d'utiliser l'expression « quitte à diminuer r_Θ^* ». Nous entendrons par celle-ci que la modification peut-être faite dès à présent, même si l'expression intervient plus tardivement dans le texte. À chaque fois, cela sera dû à ce que la diminution demandée est fonction de quantités qui ne dépendent pas de celles fixées selon r_Θ^* (et il n'y a ainsi pas de contrainte « circulaire » formulée sur r_Θ^*).*

Définition 4.33 (Boules de contrôle uniformes en temps pour les différentielles des états)**.** *Fixons un $r_{\mathcal{L}(\Theta,\,\mathcal{E})}>0$. Nous définissons, pour $t\geq 0$, les boules $B_{\mathcal{L}(\Theta,\,\mathcal{E}_t)}\subset\mathcal{L}\left(\Theta,\,\mathcal{E}_t\right)$ centrées en les J_t^*, de rayon commun $r_{\mathcal{L}(\Theta,\,\mathcal{E})}$ pour la norme d'opérateur.*

Remarque 4.34. *Les normes sur des espaces vectoriels de dimension finie sur $\mathbb{R}$ sont équivalentes, mais la constante peut dépendre de la dimension. Ainsi, le choix des normes sur les espaces $\mathcal{L}\left(\Theta,\,\mathcal{E}_t\right)$ serait sans importance par rapport à cet aspect du problème, si nous savions que les dimensions des $\mathcal{E}_t$ étaient bornées en temps.*

Remarque 4.35. *Le choix des normes sur les $\mathcal{L}\left(\Theta,\,\mathcal{E}_t\right)$ importe quoi qu'il en soit pour les propriétés de contractivité : nous voulons que les normes des différentielles par rapport aux états des fonctions de transition soient strictement inférieures à* 1 *le long de la trajectoire stable, et cette propriété dépend des normes choisie.*

4.3.2 Contrôle des différentielles des états le long de la trajectoire stable

Corollaire 4.36 (Borne sur les différentielles des états le long de la trajectoire stable)**.** *Les différentielles*

$$J_t^*=\frac{\partial\,\boldsymbol{e}_t\left(e_0^*,\,\theta^*\right)}{\partial\theta}$$

sont bornées uniformément en temps.

Démonstration. D'après la définition des J_t^* effectuée au Lemme 6.37, pour tout $t \geq 0$,

$$\left\|J_{t+1}^*\right\|_{\mathrm{op}} \leq \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t^*,\, \theta^*\right)\right\|_{\mathrm{op}} \left\|J_t^*\right\|_{\mathrm{op}} + \left\|\frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t^*,\, \theta^*\right)\right\|_{\mathrm{op}}.$$

Donc, pour tout $t \geq 0$,

$$\left\|J_{t+1}^*\right\|_{\mathrm{op}} \leq \left(\sup_{s \geq 0} \left\|\frac{\partial \boldsymbol{T}_s}{\partial e}\left(e_s^*,\, \theta^*\right)\right\|_{\mathrm{op}}\right) \left\|J_t^*\right\|_{\mathrm{op}} + \sup_{s \geq 0} \left\|\frac{\partial \boldsymbol{T}_s}{\partial \theta}\left(e_s^*,\, \theta^*\right)\right\|_{\mathrm{op}}.$$

Or, d'après l'Hypothèse 4.28, le premier supremum est strictement inférieur à 1, et d'après l'Hypothèse 4.29, le second est fini, ce qui donne le résultat annoncé. □

4.4 Contrôle des opérateurs de transition sur les boules de contrôle

4.4.1 Étude des opérateurs de transition sur les boules de contrôle

Lemme 4.37 (Contrôle des différentielles des $\boldsymbol{T}_t$)**.** *Pour tout $t \geq 0$, pour tous e_1, e_2 dans $B_{\mathcal{E}_t}$, et θ_1, θ_2 dans B_Θ^*,*

$$\left\|\mathrm{d}\,\boldsymbol{T}_t\left(e_1,\, \theta_1\right) - \mathrm{d}\,\boldsymbol{T}_t\left(e_2,\, \theta_2\right)\right\|_{\mathrm{op}} \leq S\, \left\|\left(e_1 - e_2,\, \theta_1 - \theta_2\right)\right\|.$$

Démonstration. Soit $t \geq 0$. Soient e_1, e_2 dans $B_{\mathcal{E}_t}$, et θ_1, θ_2 dans B_Θ^*. Soit le chemin inclus dans $B_{\mathcal{E}_t} \times B_\Theta^*$ défini par, pour tout $0 \leq u \leq 1$,

$$\left(e(u),\, \theta(u)\right) = \left(e_2,\, \theta_2\right) + u\, \left(e_1 - e_2,\, \theta_1 - \theta_2\right).$$

Alors,

$$\mathrm{d}\,\boldsymbol{T}_t\left(e_1,\, \theta_1\right) - \mathrm{d}\,\boldsymbol{T}_t\left(e_2,\, \theta_2\right) = \int_0^1 \mathrm{d}^2\,\boldsymbol{T}_t\left(\theta(u),\, e(u)\right) \cdot \left(\dot{e}\left(u\right), \dot{\theta}\left(u\right)\right) du,$$

où, pour $0 \leq u \leq 1$,

$$\mathrm{d}^2\,\boldsymbol{T}_t\left(\theta(u),\, e(u)\right) \cdot \left(\dot{e}\left(u\right), \dot{\theta}\left(u\right)\right)$$

désigne l'application linéaire

$$\mathrm{d}^2\,\boldsymbol{T}_t\left(\theta(u),\, e(u)\right) \cdot \left[\left(\dot{e}\left(u\right), \dot{\theta}\left(u\right)\right),\, \mathrm{Id}_{T\mathcal{E}_t \times T\Theta}\right].$$

Ainsi,

$$\left\|\mathrm{d}\,\boldsymbol{T}_t\left(e_1,\, \theta_1\right) - \mathrm{d}\,\boldsymbol{T}_t\left(e_2,\, \theta_2\right)\right\|_{\mathrm{op}} \leq \int_0^1 \left\|\mathrm{d}^2\,\boldsymbol{T}_t\left(\theta(u),\, e(u)\right) \cdot \left(\dot{e}\left(u\right), \dot{\theta}\left(u\right)\right)\right\|_{\mathrm{op}} du.$$

Posons alors $\mathcal{E}_1 = \mathcal{E}_2 = T\mathcal{E}_t \times T\Theta$ et, pour $0 \leq u \leq 1$,

$$f = f_u = \mathrm{d}^2\,\boldsymbol{T}_t\left(\theta(u),\, e(u)\right)$$

et

$$v_1 = v_1(u) = \left(\dot{e}\left(u\right), \dot{\theta}\left(u\right)\right).$$

Posons également

$$g = \mathrm{Id}_{T\mathcal{E}_t \times T\Theta}.$$

De la sorte,

$$f(v_1, g) = \mathrm{d}^2 \boldsymbol{T}_t(\theta(u),\, e(u))\cdot\Big[\Big(\dot{e}(u), \dot{\theta}(u)\Big),\, \mathrm{Id}_{T\mathcal{E}_t \times T\Theta}\Big] = \mathrm{d}^2 \boldsymbol{T}_t(\theta(u),\, e(u))\cdot\Big(\dot{e}(u), \dot{\theta}(u)\Big).$$

Par conséquent, d'après le Corollaire 4.12,

$$\Big\|\mathrm{d}^2 \boldsymbol{T}_t(\theta(u),\, e(u))\cdot\Big(\dot{e}(u), \dot{\theta}(u)\Big)\Big\|_{\mathrm{op}} \leq \Big\|\mathrm{d}^2 \boldsymbol{T}_t(\theta(u),\, e(u))\Big\|_{\mathrm{bil}} \Big\|\Big(\dot{e}(u), \dot{\theta}(u)\Big)\Big\|.$$

Ainsi,

$$\begin{aligned}\|\mathrm{d}\boldsymbol{T}_t(e_1,\, \theta_1) - \mathrm{d}\boldsymbol{T}_t(e_2,\, \theta_2)\|_{\mathrm{op}} &\leq \int_0^1 \Big\|\mathrm{d}^2 \boldsymbol{T}_t(\theta(u),\, e(u))\Big\|_{\mathrm{bil}} \Big\|\Big(\dot{e}(u), \dot{\theta}(u)\Big)\Big\|\, du \\ &\leq S\, \|(e_1 - e_2,\, \theta_1 - \theta_2)\|,\end{aligned}$$

ce qui conclut la preuve. □

Corollaire 4.38 (Majoration des différentielles des $\boldsymbol{T}_t$)**.** *Pour tout $t \geq 0$, pour tout $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$,*

$$\|\mathrm{d}\boldsymbol{T}_t(e,\, \theta) - \mathrm{d}\boldsymbol{T}_t(e^*,\, \theta^*)\|_{\mathrm{op}} \leq S\sqrt{r_{\mathcal{E}}^2 + {r^*_\Theta}^2}.$$

Démonstration. Soit $t \geq 0$. Soit $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$. D'après le Lemme 4.37 précédent,

$$\|\mathrm{d}\boldsymbol{T}_t(e,\, \theta) - \mathrm{d}\boldsymbol{T}_t(e^*,\, \theta^*)\|_{\mathrm{op}} \leq S\, \|(e - e^*,\, \theta - \theta^*)\|.$$

Or, $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$, donc

$$\|e - e^*\| \leq r_{\mathcal{E}} \quad \text{et} \quad \|\theta - \theta^*\| \leq r^*_\Theta,$$

ce qui conclut la preuve. □

Lemme 4.39 (Contractivité des $\boldsymbol{T}_t$)**.** *Quitte à diminuer $r_{\mathcal{E}}$ et r^*_Θ, nous pouvons trouver $0 < \alpha < 1$ tel que, pour tout $t \geq 0$, pour tout $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$,*

$$\left\|\frac{\partial \boldsymbol{T}_t}{\partial e}(e,\, \theta)\right\|_{\mathrm{op}} \leq 1 - \alpha.$$

Démonstration. Soient $t \geq 0$, et $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$. Nous savons que

$$\left\|\frac{\partial \boldsymbol{T}_t}{\partial e}(e,\, \theta) - \frac{\partial \boldsymbol{T}_t}{\partial e}(e^*_t,\, \theta^*)\right\|_{\mathrm{op}} \leq \|\mathrm{d}\boldsymbol{T}_t(e,\, \theta) - \mathrm{d}\boldsymbol{T}_t(e^*,\, \theta^*)\|_{\mathrm{op}}.$$

Donc, d'après le Corollaire 4.38,

$$\begin{aligned}\left\|\frac{\partial \boldsymbol{T}_t}{\partial e}(e,\, \theta)\right\|_{\mathrm{op}} &\leq \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}(e^*_t,\, \theta^*)\right\|_{\mathrm{op}} + \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}(e,\, \theta) - \frac{\partial \boldsymbol{T}_t}{\partial e}(e^*_t,\, \theta^*)\right\|_{\mathrm{op}} \\ &\leq \sup_{t \geq 0} \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}(e^*_t,\, \theta^*)\right\|_{\mathrm{op}} + S\sqrt{r_{\mathcal{E}}^2 + {r^*_\Theta}^2}.\end{aligned}$$

D'après l'Hypothèse 4.28, le premier terme de la quantité majorante est strictement inférieur à 1, ce qui permet d'obtenir le résultat annoncé. □

Lemme 4.40 (Borne sur les différentielles par rapport au paramètre des $\boldsymbol{T}_t$)**.** *Nous pouvons trouver $M_1 > 0$ tel que, pour tout $t \geq 0$, pour tout $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$,*

$$\left\|\frac{\partial \boldsymbol{T}_t}{\partial \theta}(e,\, \theta)\right\|_{\mathrm{op}} \leq M_1.$$

Démonstration. Soient $t \geq 0$, et $(e,\, \theta) \in B_{\mathcal{E}_t} \times B^*_\Theta$. De même que dans la preuve du Lemme 4.39 précédent,

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta} (e,\, \theta) \right\|_{\text{op}} \leq \sup_{t \geq 0} \left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta} (e_t^*,\, \theta^*) \right\|_{\text{op}} + S \sqrt{r_{\mathcal{E}}^2 + {r_\Theta^*}^2}.$$

Or, d'après l'Hypothèse 4.29, le premier terme de la quantité majorante est fini, ce qui permet d'obtenir le résultat annoncé. □

4.4.2 Contrôle des opérateurs de transition sur les boules de contrôle

Lemme 4.41 (Sous-linéarité des opérateurs de transition sur les boules de contrôle $B_{\mathcal{E}_t} \times B^*_\Theta$)**.** *Soit $t \geq 0$. Soient $e_1,\, e_2 \in B_{\mathcal{E}_t}$ et $\theta_1,\, \theta_2 \in B^*_\Theta$. Alors,*

$$\|\boldsymbol{T}_t (e_1,\, \theta_1) - \boldsymbol{T}_t (e_2,\, \theta_2)\| \leq (1 - \alpha)\, \|e_1 - e_2\| + M_1\, d(\theta_1,\, \theta_2)\,.$$

Démonstration. Soit le chemin inclus dans $B_{\mathcal{E}_t} \times B^*_\Theta$ défini par, pour tout $0 \leq u \leq 1$,

$$(e(u),\, \theta(u)) = (e_2,\, \theta_2) + u\, (e_1 - e_2,\, \theta_1 - \theta_2)\,.$$

Alors,

$$\begin{aligned} &\boldsymbol{T}_t (e_1,\, \theta_1) - \boldsymbol{T}_t (e_2,\, \theta_2) = \\ &\int_0^1 \left(\frac{\partial \boldsymbol{T}_t}{\partial e} (e(u),\, \theta(u)) \cdot \dot{e}\,(u) + \frac{\partial \boldsymbol{T}_t}{\partial \theta} (e(u),\, \theta(u)) \cdot \dot{\theta}\,(u) \right) du. \end{aligned}$$

Ainsi,

$$\begin{aligned} &\|\boldsymbol{T}_t (e_1,\, \theta_1) - \boldsymbol{T}_t (e_2,\, \theta_2)\| \\ &\leq \int_0^1 \left(\left\| \frac{\partial \boldsymbol{T}_t}{\partial e} (e(u),\, \theta(u)) \right\|_{\text{op}} \|\dot{e}\,(u)\| + \left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta} (e(u),\, \theta(u)) \right\|_{\text{op}} \left\| \dot{\theta}\,(u) \right\| \right) du \\ &\leq (1 - \alpha) \int_0^1 \|\dot{e}\,(u)\|\, du + M_1 \int_0^1 \left\| \dot{\theta}\,(u) \right\| du \end{aligned}$$

d'après les Lemmes 4.39 et 4.40. En conséquence,

$$\|\boldsymbol{T}_t (e_1,\, \theta_1) - \boldsymbol{T}_t (e_2,\, \theta_2)\| \leq (1 - \alpha)\, \|e_1 - e_2\| + M_1\, d(\theta_1,\, \theta_2)\,,$$

ce qui est le résultat annoncé. □

Lemme 4.42 (Homogénéité en θ lors d'une transition pour des états initiaux proches)**.** *Soit $t \geq 0$. Soient $e_1,\, e_2 \in B^*_{\mathcal{E}_t}$ et $\theta_1,\, \theta_2 \in B^*_\Theta$. Supposons disposer d'un $r \geq 0$ tel que*

$$d(\theta_1,\, \theta_2) \leq r$$

et

$$\|e_1 - e_2\| \leq \frac{M_1}{\alpha}\, r.$$

Alors,

$$\|\boldsymbol{T}_t (e_1,\, \theta_1) - \boldsymbol{T}_t (e_2,\, \theta_2)\| \leq \frac{M_1}{\alpha}\, r.$$

Démonstration. D'après le Lemme 4.41,

$$\|\boldsymbol{T}_t (e_1,\, \theta_1) - \boldsymbol{T}_t (e_2,\, \theta_2)\| \leq (1 - \alpha)\, \|e_1 - e_2\| + M_1\, d(\theta_1,\, \theta_2)\,,$$

ce qui donne le résultat annoncé. □

Lemme 4.43 (Oubli exponentiel des états initiaux)**.** *Soit $t \geq 0$. Soient $e_1,\, e_2 \in B^*_{\mathcal{E}_t}$ et $\theta \in B^*_\Theta$. Alors,*

$$\|\boldsymbol{T}_t\,(e_1,\,\theta) - \boldsymbol{T}_t\,(e_2,\,\theta)\| \leq (1-\alpha)\; d\,(e_1,\,e_2)\,.$$

Démonstration. C'est une conséquence du Lemme 4.41. □

4.4.3 Contrôle des opérateurs de transition sur les différentielles des états sur les boules de contrôle

Notons

$$M_2 = S\,\sqrt{r^2_{\mathcal{L}(\Theta,\,\mathcal{E})} + 1}.$$

Lemme 4.44 (Sous-linéarité de l'opérateur de transition sur les différentielles sur les boules $B_{\mathcal{L}(\Theta,\,\mathcal{E}_t)}$)**.** *Soient $e_1,\, e_2 \in B_{\mathcal{E}_t}$ et $\theta_1,\, \theta_2 \in B^*_\Theta$. Soient deux applications linéaires du tangent à Θ vers le tangent à $\mathcal{E}_t$ telles que*

$$J_1,\, J_2 \in B_{\mathcal{L}(\Theta,\,\mathcal{E}_t)}.$$

Alors

$$\left\| \frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e_1,\,\theta_1)\cdot J_1 + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e_1,\,\theta_1) - \left(\frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e_2,\,\theta_2)\cdot J_2 + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e_2,\,\theta_2)\right)\right\|_{\mathrm{op}}$$

est majoré par

$$(1-\alpha)\; d\,(J_1,\,J_2) + M_2\,\sqrt{d^2\,(e_1,\,e_2) + d^2\,(\theta_1,\,\theta_2)}.$$

Démonstration. Soit le chemin inclus dans $B_{\mathcal{E}_t} \times B_{\mathcal{L}(\Theta,\,\mathcal{E}_t)} \times B^*_\Theta$ défini par, pour tout $0 \leq u \leq 1$,

$$(e(u),\, J(u),\, \theta(u)) = (e_2,\, \theta_2,\, J_2) + u\;(e_1 - e_2,\, J_1 - J_2,\, \theta_1 - \theta_2)\,.$$

Alors, le chemin

$$c\,:\, u \mapsto \frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e(u),\,\theta(u))\cdot J(u) + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e(u),\,\theta(u))$$

est un chemin reliant

$$\frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e_2,\,\theta_2)\cdot J_2 + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e_2,\,\theta_2)\,.$$

à

$$\frac{\partial\, \boldsymbol{T}_t}{\partial e}\,(e_1,\,\theta_1)\cdot J_1 + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}\,(e_1,\,\theta_1)$$

Donc,

$$\|c(1) - c(2)\|_{\mathrm{op}} \leq \int_0^1 \|\dot{c}(u)\|_{\mathrm{op}}\; du.$$

Or

$$c(u) = \mathrm{d}\,\boldsymbol{T}_t\,(e(u),\,\theta(u))\cdot(J(u),\,\mathrm{Id}_\Theta)\,,$$

donc

$$\begin{aligned}\dot{c}(u) &= \mathrm{d}^2 \boldsymbol{T}_t\left(e(u),\, \theta(u)\right) \cdot \left[\left(\dot{e}\left(u\right),\, \dot{\theta}\left(u\right)\right),\, \left(J(u),\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \mathrm{d}\, \boldsymbol{T}_t\left(e(u),\, \theta(u)\right) \cdot \left(\dot{J}(u),\, 0\right) \\ &= \mathrm{d}^2 \boldsymbol{T}_t\left(e(u),\, \theta(u)\right) \cdot \left[\left(\dot{e}\left(u\right),\, \dot{\theta}\left(u\right)\right),\, \left(J(u),\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e(u),\, \theta(u)\right) \cdot \dot{J}(u).\end{aligned}$$

Pour tout $0 \leq u \leq 1$, $\dot{c}(u)$ est identifé à un élément de $\mathcal{L}\left(\Theta,\, \mathcal{E}_{t+1}\right)$ (comme nous identifions, dans ce cas également, le tangent à l'espace vectoriel $\mathcal{L}\left(\Theta,\, \mathcal{E}_{t+1}\right)$ avec cet espace). Par conséquent,

$$\begin{aligned}\left\|\dot{c}(u)\right\|_{\mathrm{op}} &\leq \left\|\mathrm{d}^2 \boldsymbol{T}_t\left(e(u),\, \theta(u)\right) \cdot \left[\left(\dot{e}\left(u\right),\, \dot{\theta}\left(u\right)\right),\, \left(J(u),\, \mathrm{Id}_\Theta\right)\right]\right\|_{\mathrm{op}} \\ &\quad + \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e(u),\, \theta(u)\right) \cdot \dot{J}(u)\right\|_{\mathrm{op}} \\ &\leq \left\|\mathrm{d}^2 \boldsymbol{T}_t\left(e(u),\, \theta(u)\right)\right\|_{\mathrm{bil}} \left\|\left(\dot{e}\left(u\right), \dot{\theta}\left(u\right)\right)\right\| \left\|\left(J(u),\, \mathrm{Id}_\Theta\right)\right\|_{\mathrm{op}} \\ &\quad + \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e(u),\, \theta(u)\right)\right\|_{\mathrm{op}} \left\|\dot{J}\left(u\right)\right\|_{\mathrm{op}},\end{aligned}$$

d'après le Corollaire 4.12. Or, pour $0 \leq u \leq 1$, $J(u) \in \mathcal{L}\left(\Theta,\, \mathcal{E}_t\right)$ donc, d'après la Définition 4.33,

$$\begin{aligned}\left\|\left(J(u),\, \mathrm{Id}_\Theta\right)\right\|_{\mathrm{op}} &\leq \sqrt{\left\|J(u)\right\|_{\mathrm{op}}^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2} \\ &\leq \sqrt{r_{\mathcal{L}(\Theta,\, \mathcal{E})}^2 + 1}.\end{aligned}$$

Ainsi,

$$\left\|\dot{c}(u)\right\|_{\mathrm{op}} \leq S\, \sqrt{r_{\mathcal{L}(\Theta,\, \mathcal{E})}^2 + 1}\, \left\|(e_1 - e_2,\, \theta_1 - \theta_2)\right\| + (1 - \alpha)\, \left\|J_1 - J_2\right\|_{\mathrm{op}},$$

d'après l'Hypothèse 4.30 et le Lemme 4.39. Donc,

$$\left\|c(1) - c(2)\right\|_{\mathrm{op}} \leq S\, \sqrt{r_{\mathcal{L}(\Theta,\, \mathcal{E})}^2 + 1}\, \left\|(e_1 - e_2,\, \theta_1 - \theta_2)\right\| + (1 - \alpha)\, \left\|J_1 - J_2\right\|_{\mathrm{op}},$$

ce qui conclut la preuve. □

Lemme 4.45 (Homogénéité en θ lors d'une transition pour des états et des différentielles initiaux proches)**.** *Soient $e_1,\, e_2 \in B_{\mathcal{E}_t}$ et $\theta_1,\, \theta_2 \in B_\Theta^*$. Soient deux applications linéaires du tangent à Θ vers le tangent à $\mathcal{E}_t$ telles que*

$$J_1,\, J_2 \in B_{\mathcal{L}(\Theta,\, \mathcal{E}_t)}.$$

Supposons disposer d'un $r \geq 0$ tel que

$$d\left(\theta_1,\, \theta_2\right) \leq r,$$

$$\left\|e_1 - e_2\right\| \leq \frac{M_1}{\alpha}\, r$$

et

$$\left\|J_1 - J_2\right\|_{\mathrm{op}} \leq \frac{M_2}{\alpha}\, \sqrt{\left(\frac{M_1}{\alpha}\right)^2 + 1}\, r.$$

Alors,

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial e}(e_1,\, \theta_1) \cdot J_1 + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e_1,\, \theta_1) - \left(\frac{\partial \boldsymbol{T}_t}{\partial e}(e_2,\, \theta_2) \cdot J_2 + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e_2,\, \theta_2) \right) \right\|_{\mathrm{op}}$$

est majoré par

$$\frac{M_2}{\alpha} \sqrt{\left(\frac{M_1}{\alpha} \right)^2 + 1}\, r.$$

Remarque 4.46. *Dans les majorations ci-dessus, le « r » ne figure pas sous les radicaux, ce qui n'est peut-être pas très visible.*

Démonstration. C'est une conséquence du Lemme 4.44. □

Lemme 4.47 (Transition sur les différentielles pour le même paramètre)**.** *Soient $e_1,\, e_2 \in B_{\mathcal{E}_t}$ et $\theta \in B^*_\Theta$. Soient deux applications linéaires du tangent à Θ vers le tangent à $\mathcal{E}_t$ telles que*

$$J_1,\, J_2 \in B_{\mathcal{L}(\Theta, \mathcal{E}_t)}.$$

Alors

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial e}(e_1,\, \theta) \cdot J_1 + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e_1,\, \theta) - \left(\frac{\partial \boldsymbol{T}_t}{\partial e}(e_2,\, \theta) \cdot J_2 + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e_2,\, \theta) \right) \right\|_{\mathrm{op}}$$

est majoré par

$$(1 - \alpha)\; d\,(J_1,\, J_2) + M_2\, d\,(e_1,\, e_2)\,.$$

Démonstration. C'est une conséquence du Lemme 4.44. □

4.5 Stabilité des trajectoires au voisinage de la trajectoire stable

4.5.1 Définition des boules stables au voisinage de la trajectoire stable

Hypothèse 4.48 (Choix de r^*_Θ)**.** *Quitte à diminuer r^*_Θ, nous pouvons supposer que*

$$\frac{M_1}{\alpha}\, r^*_\Theta \leq r_{\mathcal{E}} \quad \textit{et} \quad \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2}\, r^*_\Theta \leq r_{\mathcal{L}(\Theta, \mathcal{E})}.$$

Remarque 4.49. *Cela sert à garantir les inclusions des boules définies ci-dessous dans les boules de contrôle définies précédemment.*

Définition 4.50 (Boules stables uniformes en temps pour les états)**.** *Nous définissons, pour $t \geq 0$, les boules $B^*_{\mathcal{E}_t} \subset B_{\mathcal{E}_t}$ centrées en les e^*_t, de rayon commun*

$$r^*_{\mathcal{E}} = \frac{M_1}{\alpha}\, r^*_\Theta.$$

Définition 4.51 (Boules stables uniformes en temps pour les différentielles des états)**.** *Nous définissons, pour $t \geq 0$, les boules $B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)} \subset B_{\mathcal{L}(\Theta, \mathcal{E}_t)}$ centrées en les J^*_t, de rayon commun*

$$r^*_{\mathcal{L}(\Theta, \mathcal{E})} = \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2}\, r^*_\Theta.$$

Nous appelons également boule stable pour le paramètre la boule B^*_Θ.

4.5.2 Preuve de stabilité des boules

Corollaire 4.52 (Stabilité des boules $B^*_{\mathcal{E}_t} \times B^*_\Theta$)**.** *Soit $t \geq 0$. Soient $e \in B^*_{\mathcal{E}_t}$ et $\theta \in B^*_\Theta$. Alors,*

$$e' = \boldsymbol{T}_t(e, \theta) \in B^*_{\mathcal{E}_{t+1}}.$$

Démonstration. $\theta \in B^*_\Theta$ donc, d'après l'Hypothèse 4.30,

$$d(\theta, \theta^*) \leq r^*_\Theta.$$

$e \in B^*_{\mathcal{E}_t}$ donc, d'après la Définition 4.50,

$$d(e, e^*_t) \leq r^*_{\mathcal{E}} = \frac{M_1}{\alpha} r^*_\Theta.$$

Ainsi, d'après le Lemme 4.42,

$$\|\boldsymbol{T}_t(e, \theta) - \boldsymbol{T}_t(e^*_t, \theta^*)\| \leq \frac{M_1}{\alpha} r^*_\Theta.$$

Par conséquent,

$$\|e' - e^*_{t+1}\| \leq r^*_{\mathcal{E}},$$

ce qui conclut la preuve. □

Corollaire 4.53 (Trajectoires des états incluses dans les boules stables)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$ et $\boldsymbol{\theta} = (\theta_t)$ incluse dans B^*_Θ. Alors, pour tout $t \geq 0$,*

$$\boldsymbol{e}_t(e_0, \boldsymbol{\theta}) \in B^*_{\mathcal{E}_t}.$$

Démonstration. C'est une conséquence par récurrence du Corollaire 4.52. □

Corollaire 4.54 (Stabilité des boules $B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)}$)**.** *Soient $e \in B^*_{\mathcal{E}_t}$ et $\theta \in B^*_\Theta$. Soit une application linéaire du tangent à Θ vers le tangent à $\mathcal{E}_t$ telle que*

$$J \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)}.$$

Alors

$$\frac{\partial \boldsymbol{T}_t}{\partial e}(e, \theta) \cdot J + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e, \theta) \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_{t+1})}.$$

Démonstration. Nous procédons de la même manière que pour le Corollaire 4.52, en utilisant le Lemme 4.45 et la définition de $r^*_{\mathcal{L}(\Theta, \mathcal{E})}$ à la Définition 4.51. □

Corollaire 4.55 (Trajectoires des différentielles des états incluses dans les boules stables)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$, $J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$ et $\boldsymbol{\theta} = (\theta_t)$ incluse dans B^*_Θ. Alors, pour tout $t \geq 0$,*

$$\boldsymbol{J}_t(e_0, J_0, \boldsymbol{\theta}) \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)}.$$

Démonstration. C'est une conséquence par récurrence du Corollaire 4.54, attendu que d'après le Corollaire 4.53, pour tout $t \geq 0$,

$$\boldsymbol{e}_t(e_0, \boldsymbol{\theta}) \in B^*_{\mathcal{E}_t}.$$

□

Lemme 4.56 (Stabilité des boules en tant que boules ouvertes)**.** *Les boules $B^*_{\mathcal{E}_t}$ et $B_{\mathcal{L}(\Theta, \mathcal{E}_t)}$ sont également stables en tant que boules ouvertes.*

Démonstration. La reprise des calculs aboutissant aux résultats de stabilité en tant que boules fermées, où les inégalités larges sont remplacées par des inégalités strictes, donne le résultat. □

4.6 Évolution de deux systèmes avec les mêmes quantités initiales ou le même paramètre

4.6.1 Homogénéité en θ des distances entre trajectoires issues des mêmes quantités initiales

Lemme 4.57 (Homogénéité en θ des écarts entre états)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$, et deux suites de paramètres $\boldsymbol{\theta} = (\theta_t)$ et $\boldsymbol{\theta'} = (\theta'_t)$ incluses dans B^*_Θ. Alors, pour tout $t \geq 1$,*

$$d\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta'}\right)\right) \leq \frac{M_1}{\alpha} \sup_{s \leq t-1} d\left(\theta_s,\, \theta'_s\right).$$

Démonstration. Prouvons-le par récurrence. Posons

$$r = d\left(\theta_0,\, \theta'_0\right).$$

Or, les deux trajectoires considérées sont issues du même état initial e_0 donc, d'après le Lemme 4.42,

$$\left\|\boldsymbol{T}_0\left(e_0,\, \theta_0\right) - \boldsymbol{T}_0\left(e_0,\, \theta'_0\right)\right\| \leq \frac{M_1}{\alpha}\, r = \frac{M_1}{\alpha}\, d\left(\theta_0,\, \theta'_0\right).$$

Ainsi,

$$\left\|\boldsymbol{e}_1\left(e_0,\, \boldsymbol{\theta}\right) - \boldsymbol{e}_1\left(e_0,\, \boldsymbol{\theta'}\right)\right\| \leq \frac{M_1}{\alpha}\, d\left(\theta_0,\, \theta'_0\right),$$

ce qui établit l'initialisation. Soit alors $t \geq 1$, et supposons la propriété vraie pour t. Alors,

$$\begin{aligned} d\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta'}\right)\right) &\leq \frac{M_1}{\alpha} \sup_{s \leq t-1} d\left(\theta_s,\, \theta'_s\right) \\ &\leq \frac{M_1}{\alpha} \sup_{s \leq t} d\left(\theta_s,\, \theta'_s\right). \end{aligned}$$

Comme, d'après le Corollaire 4.53, les deux états courants appartiennent à la boule $B^*_{\mathcal{E}_t}$, nous pouvons appliquer le Lemme 4.42, en majorant $d\left(\theta_t,\, \theta'_t\right)$ par

$$r = \sup_{s \leq t} d\left(\theta_s,\, \theta'_s\right).$$

Ceci conclut la récurrence, puis la preuve. □

Lemme 4.58 (Homogénéité en θ des écarts entre les différentielles des états)**.** *Soient*

$$e_0 \in B^*_{\mathcal{E}_0}, \quad J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)},$$

et deux suites de paramètres $\boldsymbol{\theta} = (\theta_t)$ et $\boldsymbol{\theta'} = (\theta'_t)$ incluses dans B^*_Θ. Alors, pour tout $t \geq 1$,

$$d\left(\boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta'}\right)\right) \leq \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha}\right)^2} \sup_{s \leq t-1} d\left(\theta_s,\, \theta'_s\right).$$

Démonstration. Prouvons-le par récurrence. Posons

$$r = d\left(\theta_0,\, \theta'_0\right).$$

Or, les deux trajectoires considérées sont issues du même état initial e_0 et de la même différentielle initiale J_0 donc, d'après le Lemme 4.45,

$$\left\| \frac{\partial \boldsymbol{T}_0}{\partial e}(e_0,\, \theta_0) \cdot J_0 + \frac{\partial \boldsymbol{T}_0}{\partial \theta}(e_0,\, \theta_0) - \left(\frac{\partial \boldsymbol{T}_0}{\partial e}(e_0,\, \theta_0') \cdot J_0 + \frac{\partial \boldsymbol{T}_0}{\partial \theta}(e_0,\, \theta_0') \right) \right\|_{\mathrm{op}}$$

est majoré par

$$\frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2}\, r = \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2}\, d(\theta_0,\, \theta_0')\,.$$

Ainsi,

$$\left\| \boldsymbol{J}_1(e_0,\, J_0,\, \boldsymbol{\theta}) - \boldsymbol{J}_1(e_0,\, J_0,\, \boldsymbol{\theta'}) \right\|_{\mathrm{op}} \leq \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2}\, d(\theta_0,\, \theta_0')\,,$$

ce qui établit l'initialisation. Soit alors $t \geq 1$ tel que la propriété est vraie. Alors,

$$\begin{aligned} d\left(\boldsymbol{J}_t(e_0,\, J_0,\, \boldsymbol{\theta}),\, \boldsymbol{J}_t(e_0,\, J_0,\, \boldsymbol{\theta'}) \right) &\leq \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2} \sup_{s \leq t-1} d(\theta_s,\, \theta_s') \\ &\leq \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2} \sup_{s \leq t} d(\theta_s,\, \theta_s')\,. \end{aligned}$$

D'après le Corollaire 4.53, les deux états courants appartiennent à la boule $B^*_{\mathcal{E}_t}$. De plus, d'après le Lemme 4.57, ils vérifient :

$$\begin{aligned} d\left(\boldsymbol{e}_t(e_0,\, \boldsymbol{\theta}),\, \boldsymbol{e}_t(e_0,\, \boldsymbol{\theta'}) \right) &\leq \frac{M_1}{\alpha} \sup_{s \leq t-1} d(\theta,\, \theta') \\ &\leq \frac{M_1}{\alpha} \sup_{s \leq t} d(\theta,\, \theta')\,. \end{aligned}$$

Nous pouvons alors appliquer le Lemme 4.45, en majorant $d(\theta_t,\, \theta_t')$ par

$$r = \sup_{s \leq t} d(\theta_s,\, \theta_s')\,.$$

Ceci conclut la récurrence, puis la preuve. □

4.6.2 Oubli exponentiel des états et des différentielles initiaux

Lemme 4.59 (Oubli exponentiel des états initiaux)**.** *Soient $e_0, e_0' \in B^*_{\mathcal{E}_0}$, et $\theta \in B^*_\Theta$. Alors, pour tout $t \geq 0$,*

$$d\left(\boldsymbol{e}_t(e_0,\, \theta),\, \boldsymbol{e}_t(e_0',\, \theta) \right) \leq (1-\alpha)^t\, d(e_0,\, e_0')\,.$$

Démonstration. C'est une conséquence de l'inclusion des deux trajectoires dans les boules $B^*_{\mathcal{E}_t}$, établie au Corollaire 4.53, et d'une récurrence conjuguée au Lemme 4.43. □

Lemme 4.60 (Oubli exponentiel des différentielles initiales)**.** *Soient e_0, $e_0' \in B^*_{\mathcal{E}_0}$, J_0, $J_0' \in B^*_{\mathcal{L}(\Theta,\, \mathcal{E}_0)}$, et $\theta \in B^*_\Theta$. Alors, pour tout $t \geq 0$,*

$$d\left(\boldsymbol{J}_t(e_0,\, J_0,\, \theta),\, \boldsymbol{J}_t(e_0',\, J_0',\, \theta) \right) \leq (1-\alpha)^t\, d(J_0,\, J_0') + M_2\, t\, (1-\alpha)^{t-1}\, d(e_0,\, e_0')\,.$$

Démonstration. Notons, pour $t \geq 0$,

$$e_t = \boldsymbol{e}_t\left(e_0,\, \theta\right), \quad e'_t = \boldsymbol{e}_t\left(e'_0,\, \theta\right),$$
$$J_t = \boldsymbol{J}_t\left(e_0,\, J_0,\, \theta\right) \quad \text{et} \quad J'_t = \boldsymbol{J}_t\left(e'_0,\, J'_0,\, \theta\right).$$

Les deux trajectoires des états sont incluses dans les boules $B^*_{\mathcal{E}_t}$, grâce au Corollaire 4.53, et les deux trajectoires des différentielles sont incluses dans les boules $B^*_{\mathcal{L}(\Theta,\, \mathcal{E}_t)}$, grâce au Corollaire 4.55. Ainsi, d'après le Lemme 4.47, pour tout $t \geq 0$,

$$d\left(J_{t+1},\, J'_{t+1}\right) \leq (1-\alpha)\; d\left(J_t,\, J'_t\right) + M_2\, d\left(e_t,\, e'_t\right).$$

Par conséquent, pour tout $t \geq 1$,

$$d\left(J_t,\, J'_t\right) \leq (1-\alpha)^t\; d\left(J_0,\, J'_0\right) + M_2 \sum_{s \leq t-1} (1-\alpha)^{t-1-s}\; d\left(e_s,\, e'_s\right).$$

Or, les deux trajectoires des états étant incluses dans les boules $B^*_{\mathcal{E}_t}$, d'après le Lemme 4.59, pour tout $s \geq 0$,

$$d\left(e_s,\, e'_s\right) \leq (1-\alpha)^s\; d\left(e_0,\, e'_0\right).$$

Ainsi, pour tout $t \geq 1$,

$$\begin{aligned} d\left(J_t,\, J'_t\right) &\leq (1-\alpha)^t\; d\left(J_0,\, J'_0\right) + M_2 \sum_{s \leq t-1} (1-\alpha)^{t-1-s}\; (1-\alpha)^s\; d\left(e_0,\, e'_0\right) \\ &= (1-\alpha)^t\; d\left(J_0,\, J'_0\right) + M_2\, t\; (1-\alpha)^{t-1}\; d\left(e_0,\, e'_0\right). \end{aligned}$$

La majoration est toujours vraie pour $t = 0$, ce qui conclut la preuve. □

4.7 Bornes sur les différentielles secondes et troisièmes des états au voisinage de la trajectoire stable

Lemme 4.61 (Borne sur les différentielles secondes des états au voisinage de la trajectoire stable)**.**

$$\sup_{t \geq 0} \sup_{\theta \in B^*_\Theta} \left\| \frac{\partial^2\, \boldsymbol{e}_t\left(e^*_0,\, \theta\right)}{\partial \theta^2} \right\|_{\text{bil}} < \infty.$$

Démonstration. Notons, pour tout $s \geq 0$ et $\theta \in B^*_\Theta$,

$$e_s = \boldsymbol{e}_s\left(e^*_0,\, \theta\right) \quad \text{et} \quad J_s = \boldsymbol{J}_s\left(e^*_0,\, 0,\, \theta\right).$$

Soit $t \geq 0$. D'après le Lemme 4.25, pour $\theta \in B^*_\Theta$,

$$\begin{aligned} \frac{\partial^2\, \boldsymbol{e}_{t+1}\left(e^*_0,\, \theta\right)}{\partial \theta^2} &= \mathrm{d}^2\, \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left[\left(J_t,\, \mathrm{Id}_\Theta\right),\, \left(J_t,\, \mathrm{Id}_\Theta\right)\right] \\ &\quad + \frac{\partial\, \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \cdot \frac{\partial^2\, \boldsymbol{e}_t\left(e^*_0,\, \theta\right)}{\partial \theta^2}. \end{aligned}$$

Donc, pour $\theta \in B^*_\Theta$,

$$\begin{aligned} \left\| \frac{\partial^2\, \boldsymbol{e}_{t+1}\left(e^*_0,\, \theta\right)}{\partial \theta^2} \right\|_{\text{bil}} &\leq \left\| \mathrm{d}^2\, \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left[\left(J_t,\, \mathrm{Id}_\Theta\right),\, \left(J_t,\, \mathrm{Id}_\Theta\right)\right] \right\|_{\text{bil}} \\ &\quad + \left\| \frac{\partial\, \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \cdot \frac{\partial^2\, \boldsymbol{e}_t\left(e^*_0,\, \theta\right)}{\partial \theta^2} \right\|_{\text{bil}}. \end{aligned}$$

Or, d'après le Corollaire 4.11,

$$\left\|\mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left[\left(J_t,\, \mathrm{Id}_\Theta\right),\, \left(J_t,\, \mathrm{Id}_\Theta\right)\right]\right\|_{\mathrm{bil}} \leq \left\|\mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right)\right\|_{\mathrm{bil}} \left\|\left(J_t,\, \mathrm{Id}_\Theta\right)\right\|_{\mathrm{op}}^2$$

et, d'après le Corollaire 4.13,

$$\left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \cdot \frac{\partial^2 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}} \leq \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right)\right\|_{\mathrm{op}} \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}}.$$

Ainsi, pour tout $\theta \in B_\Theta^*$,

$$\begin{aligned}\left\|\frac{\partial^2 \boldsymbol{e}_{t+1}\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}} &\leq \left\|\mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right)\right\|_{\mathrm{bil}} \left\|\left(J_t,\, \mathrm{Id}_\Theta\right)\right\|_{\mathrm{op}}^2 \\ &\quad + \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right)\right\|_{\mathrm{op}} \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}}.\end{aligned}$$

Notons alors

$$S_1 = \sup_{s \geq 0} \sup_{\theta \in B_\Theta^*} \left\|\mathrm{d}^2 \boldsymbol{T}_s\left(e_s,\, \theta\right)\right\|_{\mathrm{bil}}, \quad S_2 = \sup_{s \geq 0} \sup_{\theta \in B_\Theta^*} \left\|J_s\right\|_{\mathrm{op}}$$

et

$$S_3 = \sup_{s \geq 0} \sup_{\theta \in B_\Theta^*} \left\|\frac{\partial \boldsymbol{T}_s}{\partial e}\left(e_s,\, \theta\right)\right\|_{\mathrm{op}}.$$

Dans l'inégalité précédente, nous majorons alors toutes les quantités par leur supremum en $\theta \in B_\Theta^*$, puis nous majorons toutes les quantités, sauf

$$\left\|\frac{\partial^2 \boldsymbol{e}_{t+1}\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}} \quad \text{et} \quad \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}},$$

par leur supremum en temps. Nous obtenons

$$\sup_{\theta \in B_\Theta^*} \left\|\frac{\partial^2 \boldsymbol{e}_{t+1}\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}} \leq S_1 \left(\left(S_2\right)^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2\right) + S_3 \sup_{\theta \in B_\Theta^*} \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}}.$$

Or, d'après le Corollaire 4.53, pour tout $s \geq 0$, $e_s \in B_{\mathcal{E}_s}^*$. D'après le Corollaire 4.55, pour tout $s \geq 0$, $J_s \in B_{\mathcal{L}(\Theta, \mathcal{E}_s)}^*$.

Ainsi, d'après l'Hypothèse 4.30, S_1 est fini. S_2 l'est également par construction des boules $B_{\mathcal{L}(\Theta, \mathcal{E}_s)}^*$. Enfin, d'après le Lemme 4.39, S_3 est inférieur à $1 - \alpha$. De la sorte,

$$\sup_{\theta \in B_\Theta^*} \left\|\frac{\partial^2 \boldsymbol{e}_{t+1}\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}} \leq (1 - \alpha) \sup_{\theta \in B_\Theta^*} \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^2}\right\|_{\mathrm{bil}} + S_1 \left(\left(S_2\right)^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2\right)^{1/2},$$

ce qui donne le résultat annoncé. □

Lemme 4.62 (Borne sur les différentielles troisièmes des états au voisinage de la trajectoire stable)**.**

$$\sup_{t \geq 0} \sup_{\theta \in B_\Theta^*} \left\|\frac{\partial^3 \boldsymbol{e}_t\left(e_0^*,\, \theta\right)}{\partial \theta^3}\right\|_{\mathrm{tri}} < \infty.$$

Démonstration. Notons, pour tout $s \geq 0$ et $\theta \in B_\Theta^*$,

$$e_s = \boldsymbol{e}_s\left(e_0^*,\, \theta\right) \quad \text{et} \quad J_s = \boldsymbol{J}_s\left(e_0^*,\, 0,\, \theta\right).$$

Soit $t \geq 0$. D'après le Lemme 4.26, pour $\theta \in B_\Theta^*$,

$$\begin{aligned}
\frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0,\, \theta\right)}{\partial \theta^3} &= \mathrm{d}^3 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left(J_t,\, \mathrm{Id}_\Theta\right)^{\otimes 3} \\
&\quad + 2\, \mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left[\left(\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, 0\right),\, \left(J_t,\, \mathrm{Id}_\Theta\right)\right] \\
&\quad + \frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(e_t,\, \theta\right) \cdot \left[\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, J_t\right] \\
&\quad + \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \cdot \frac{\partial^3 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^3}.
\end{aligned}$$

Donc, pour $\theta \in B_\Theta^*$,

$$\begin{aligned}
\left\| \frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0^*,\, \theta\right)}{\partial \theta^3} \right\|_{\mathrm{tri}} &\leq \left\| \mathrm{d}^3 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left(J_t,\, \mathrm{Id}_\Theta\right)^{\otimes 3} \right\|_{\mathrm{tri}} \\
&\quad + 2 \left\| \mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left[\left(\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, 0\right),\, \left(J_t,\, \mathrm{Id}_\Theta\right)\right] \right\|_{\mathrm{tri}} \\
&\quad + \left\| \frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(e_t,\, \theta\right) \cdot \left[\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, J_t\right] \right\|_{\mathrm{tri}} \\
&\quad + \left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \cdot \frac{\partial^3 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^3} \right\|_{\mathrm{tri}}.
\end{aligned}$$

Majoration des quatre normes trilinéaires Or, d'après le Corollaire 4.14,

$$\left\| \mathrm{d}^3 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left(J_t,\, \mathrm{Id}_\Theta\right)^{\otimes 3} \right\|_{\mathrm{tri}} \leq \left\| \mathrm{d}^3 \boldsymbol{T}_t\left(e_t,\, \theta\right) \right\|_{\mathrm{tri}} \left\| \left(J_t,\, \mathrm{Id}_\Theta\right) \right\|_{\mathrm{op}}^3.$$

Grâce à des résultats analogues à ceux utilisés dans la preuve du Lemme 4.61, nous obtenons alors les majorations suivantes. Dans un premier temps,

$$\left\| \mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right) \cdot \left[\left(\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, 0\right),\, \left(J_t,\, \mathrm{Id}_\Theta\right)\right] \right\|_{\mathrm{tri}}$$

est inférieur à

$$\left\| \mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right) \right\|_{\mathrm{bil}} \left\| \left(\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, 0\right) \right\|_{\mathrm{bil}} \left\| \left(J_t,\, \mathrm{Id}_\Theta\right) \right\|_{\mathrm{op}},$$

qui est égal à

$$\left\| \mathrm{d}^2 \boldsymbol{T}_t\left(e_t,\, \theta\right) \right\|_{\mathrm{bil}} \left\| \frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2} \right\|_{\mathrm{bil}} \left\| \left(J_t,\, \mathrm{Id}_\Theta\right) \right\|_{\mathrm{op}}.$$

Ensuite,

$$\left\| \frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(e_t,\, \theta\right) \cdot \left[\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2},\, J_t\right] \right\|_{\mathrm{tri}} \leq \left\| \frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(e_t,\, \theta\right) \right\|_{\mathrm{bil}} \left\| \frac{\partial^2 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^2} \right\|_{\mathrm{bil}} \left\| J_t \right\|_{\mathrm{op}}.$$

Enfin,

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \cdot \frac{\partial^3 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^3} \right\|_{\mathrm{tri}} \leq \left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta\right) \right\|_{\mathrm{op}} \left\| \frac{\partial^3 \boldsymbol{e}_t\left(e_0,\, \theta\right)}{\partial \theta^3} \right\|_{\mathrm{tri}}.$$

Conclusion Ainsi, pour tout $\theta \in B^*_\Theta$,

$$\begin{aligned}\left\|\frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}} &\leq \left\|\mathrm{d}^3\,\boldsymbol{T}_t\left(e_t,\,\theta\right)\right\|_{\mathrm{tri}} \left\|\left(J_t,\,\mathrm{Id}_\Theta\right)\right\|_{\mathrm{op}}^3 \\ &+ \left\|\mathrm{d}^2\,\boldsymbol{T}_t\left(e_t,\,\theta\right)\right\|_{\mathrm{bil}} \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\,\theta\right)}{\partial\theta^2}\right\|_{\mathrm{bil}} \left\|\left(J_t,\,\mathrm{Id}_\Theta\right)\right\|_{\mathrm{op}} \\ &+ \left\|\frac{\partial^2 \boldsymbol{T}_t}{\partial e^2}\left(e_t,\,\theta\right)\right\|_{\mathrm{bil}} \left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\,\theta\right)}{\partial\theta^2}\right\|_{\mathrm{bil}} \left\|J_t\right\|_{\mathrm{op}} \\ &+ \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\,\theta\right)\right\|_{\mathrm{op}} \left\|\frac{\partial^3 \boldsymbol{e}_t\left(e_0,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}}.\end{aligned}$$

Notons alors

$$S_1 = \sup_{s\geq 0}\sup_{\theta\in B^*_\Theta}\left\|\mathrm{d}^3\,\boldsymbol{T}_s\left(e_s,\,\theta\right)\right\|_{\mathrm{tri}}, \qquad S_2 = \sup_{s\geq 0}\sup_{\theta\in B^*_\Theta}\left\|J_s\right\|_{\mathrm{op}}$$

$$S_3 = \sup_{s\geq 0}\sup_{\theta\in B^*_\Theta}\left\|\mathrm{d}^2\,\boldsymbol{T}_s\left(e_s,\,\theta\right)\right\|_{\mathrm{bil}}, \quad S_4 = \sup_{s\geq 0}\sup_{\theta\in B^*_\Theta}\left\|\frac{\partial^2 \boldsymbol{e}_t\left(e_0,\,\theta\right)}{\partial\theta^2}\right\|_{\mathrm{bil}}$$

et

$$S_5 = \sup_{s\geq 0}\sup_{\theta\in B^*_\Theta}\left\|\frac{\partial \boldsymbol{T}_s}{\partial e}\left(e_s,\,\theta\right)\right\|_{\mathrm{op}}.$$

Dans l'inégalité précédente, nous majorons alors toutes les quantités par leur supremum en $\theta \in B^*_\Theta$, puis nous majorons toutes les quantités, sauf

$$\left\|\frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}} \quad \text{et} \quad \left\|\frac{\partial^3 \boldsymbol{e}_t\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}},$$

par leur supremum en temps. Nous obtenons

$$\begin{aligned}\sup_{\theta\in B^*_\Theta}\left\|\frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}} &\leq S_1\left(\left(S_2\right)^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2\right)^{3/2} \\ &+ S_3\,S_4\left(\left(S_2\right)^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2\right) + S_3\,S_4\,S_2 \\ &+ S_5 \sup_{\theta\in B^*_\Theta}\left\|\frac{\partial^3 \boldsymbol{e}_t\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}}.\end{aligned}$$

Or, d'après le Corollaire 4.53, pour tout $s \geq 0$, $e_s \in B^*_{\mathcal{E}_s}$. D'après le Corollaire 4.55, pour tout $s \geq 0$, $J_s \in B^*_{\mathcal{L}(\Theta,\,\mathcal{E}_s)}$.

Ainsi, d'après l'Hypothèse 4.30, S_1 et S_3 sont finis. S_2 l'est également par construction des boules $B^*_{\mathcal{L}(\Theta,\,\mathcal{E}_s)}$. S_4 l'est d'après le Lemme 4.61. Enfin, d'après le Lemme 4.39, S_5 est inférieur à $1-\alpha$. Posons alors

$$\kappa = S_1\left(\left(S_2\right)^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2\right)^{3/2} + S_3\,S_4\left(\left(S_2\right)^2 + \left\|\mathrm{Id}_\Theta\right\|_{\mathrm{op}}^2\right) + S_3\,S_4\,S_2.$$

Ainsi,

$$\sup_{\theta\in B^*_\Theta}\left\|\frac{\partial^3 \boldsymbol{e}_{t+1}\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}} \leq (1-\alpha)\sup_{\theta\in B^*_\Theta}\left\|\frac{\partial^3 \boldsymbol{e}_t\left(e_0^*,\,\theta\right)}{\partial\theta^3}\right\|_{\mathrm{tri}} + \kappa.$$

Or, $\kappa < \infty$ d'après ce qui précède, ce qui donne le résultat annoncé. □

Chapitre

5 Pertes sur le paramètre

5.1 Pertes sur le couple état-paramètre

5.1.1 Pertes sur le couple état-paramètre, et hypothèses au voisinage de la trajectoire stable

Hypothèse 5.1 (Pertes sur le couple état-paramètre)**.** *Nous supposons disposer, pour $t \geq 0$, de fonctions de perte*

$$\begin{aligned} p_t : \quad & \mathcal{E}_t \times \Theta \to \mathbb{R} \\ & e,\, \theta \mapsto p_t\left(e,\, \theta\right). \end{aligned}$$

Nous supposons de plus que celles-ci sont trois fois continûment différentiables.

Remarque 5.2. *Ainsi qu'il a été dit lors de la définition des normes, nous identifions les formes linéaires à des vecteurs, en utilisant la métrique euclidienne relative aux bases de travail. Ainsi, nous noterons indistinctement $\|\cdot\|$ et $\|\cdot\|_{\mathrm{op}}$ la norme d'une forme linéaire.*

Hypothèse 5.3 (Différentielles des pertes finies le long de la trajectoire stable)**.** *Nous supposons que, pour tout $t \geq 0$,*

$$\left\| \frac{\partial\, p_t}{\partial\left(e,\, \theta\right)}\left(e_t^*,\, \theta^*\right) \right\| < \infty.$$

Hypothèse 5.4 (Différentielles deuxièmes et troisièmes des pertes uniformément bornées sur les boules de contrôle)**.** *Nous supposons, quitte à diminuer les boules $B_{\mathcal{E}_t}$ et la boule B_Θ^*, que*

$$S_{perte} = \sup_{t \geq 0} \sup_{(e,\theta) \in B_{\mathcal{E}_t} \times B_\Theta^*} \left\| \frac{\partial^2\, p_t}{\partial\left(e,\, \theta\right)^2}\left(e,\, \theta\right) \right\|_{\mathrm{bil}} < \infty$$

et

$$\sup_{t \geq 0} \sup_{(e,\theta) \in B_{\mathcal{E}_t} \times B_\Theta^*} \left\| \frac{\partial^3\, p_t}{\partial\left(e,\, \theta\right)^3}\left(e,\, \theta\right) \right\|_{\mathrm{tri}} < \infty.$$

Remarque 5.5. *Comme nous l'avons dit précédemment, dans le cas des différentielles secondes des pertes, qui sont des formes bilinéaires, la norme d'opérateur de l'endomorphisme associé est égale à la norme $\|\cdot\|_{\mathrm{bil}}$ de la forme bilinéaire.*

5.1.2 Contrôle des différentielles des pertes sur les boules de contrôle

Lemme 5.6 (Contrôle des différentielles des pertes sur les boules de contrôle)**.** *Pour tout $t \geq 0$, pour tous e_1, e_2 dans $B_{\mathcal{E}_t}$, et θ_1, θ_2 dans B^*_Θ,*

$$\|\mathrm{d}\, p_t\,(e_1,\, \theta_1) - \mathrm{d}\, p_t\,(e_2,\, \theta_2)\|_{\mathrm{op}} \leq S_{perte}\, \|(e_1 - e_2,\, \theta_1 - \theta_2)\|\,.$$

Démonstration. La preuve est identique à celle du Lemme 4.37. □

Corollaire 5.7 (Majoration des différentielles des pertes sur les boules de contrôle)**.** *Pour tout $t \geq 0$, pour tous $e \in B_{\mathcal{E}_t}$ et $\theta \in B^*_\Theta$,*

$$\|\mathrm{d}\, p_t\,(e,\, \theta)\|_{\mathrm{op}} \leq \|\mathrm{d}\, p_t\,(e_t^*,\, \theta^*)\|_{\mathrm{op}} + S_{perte}\, \left(r_{\mathcal{E}}^2 + {r_\Theta^*}^2\right)^{1/2}.$$

Démonstration. Soit $t \geq 0$. Soient $e \in B_{\mathcal{E}_t}$ et $\theta \in B^*_\Theta$. D'après le Lemme 5.6,

$$\begin{aligned}\|\mathrm{d}\, p_t\,(e,\, \theta) - \mathrm{d}\, p_t\,(e_t^*,\, \theta^*)\|_{\mathrm{op}} &\leq S_{\mathrm{perte}}\, \|(e - e_t^*,\, \theta - \theta^*)\| \\ &\leq S_{\mathrm{perte}}\, \left(r_{\mathcal{E}}^2 + {r_\Theta^*}^2\right)^{1/2}.\end{aligned}$$

Ainsi,

$$\|\mathrm{d}\, p_t\,(e,\, \theta)\|_{\mathrm{op}} \leq \|\mathrm{d}\, p_t\,(e_t^*,\, \theta^*)\|_{\mathrm{op}} + S_{\mathrm{perte}}\, \left(r_{\mathcal{E}}^2 + {r_\Theta^*}^2\right)^{1/2},$$

ce qui conclut la preuve. □

Corollaire 5.8 (Différentielles des pertes finies sur les boules de contrôle)**.** *Pour tout $t \geq 0$,*

$$\sup_{(e,\,\theta) \in B_{\mathcal{E}_t} \times B^*_\Theta} \|\mathrm{d}\, p_t\,(e,\, \theta)\|_{\mathrm{op}} < \infty.$$

Démonstration. D'après le Corollaire 5.7, pour tout $t \geq 0$,

$$\sup_{(e,\,\theta) \in B_{\mathcal{E}_t} \times B^*_\Theta} \|\mathrm{d}\, p_t\,(e,\, \theta)\|_{\mathrm{op}} \leq \|\mathrm{d}\, p_t\,(e_t^*,\, \theta^*)\|_{\mathrm{op}} + S_{\mathrm{perte}}\, \left(r_{\mathcal{E}}^2 + {r_\Theta^*}^2\right)^{1/2}.$$

D'après les Hypothèses 5.3 et 5.4, pour tout $t \geq 0$, la quantité majorante est finie, ce qui conclut la preuve. □

Définition 5.9 (Majorant des différentielles des pertes sur les boules de contrôle)**.** *Nous définissons la suite $\boldsymbol{m} = (m_t)$ par, pour tout $t \geq 0$,*

$$m_t = \sup_{(e,\,\theta) \in B_{\mathcal{E}_t} \times B^*_\Theta} \|\mathrm{d}\, p_t\,(e,\, \theta)\|_{\mathrm{op}} + 1.$$

Remarque 5.10. *D'après le Corollaire 5.8, pour tout $t \geq 0$, m_t est fini.*

Corollaire 5.11 (Majoration des différentielles partielles des pertes sur les boules de contrôle)**.** *Pour tout $t \geq 0$, pour tout $e \in B_{\mathcal{E}_t}$ et pour tout $\theta \in B^*_\Theta$,*

$$\left\|\frac{\partial\, p_t}{\partial e}\,(e,\, \theta)\right\| \leq m_t \quad \textit{et} \quad \left\|\frac{\partial\, p_t}{\partial \theta}\,(e,\, \theta)\right\| \leq m_t.$$

Démonstration. Pour tout $t \geq 0$,

$$\left\|\frac{\partial\, p_t}{\partial e}\,(e,\, \theta)\right\|_{\mathrm{op}} \quad \text{et} \quad \left\|\frac{\partial\, p_t}{\partial \theta}\,(e,\, \theta)\right\|_{\mathrm{op}}$$

sont inférieures à

$$\|\mathrm{d}\, p_t\,(e,\, \theta)\|_{\mathrm{op}}\,.$$

La Définition 5.9 permet alors d'obtenir le résultat annoncé. □

5.2 Pertes sur le paramètre

Définition 5.12 (Fonction couple état-paramètre)**.** *Pour $t \geq 0$, nous définissons la fonction*

$$\begin{aligned} g_t : \quad & \mathcal{E}_0 \times \Theta \to \mathcal{E}_t \times \Theta \\ & e_0,\, \theta \mapsto (\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta)\,. \end{aligned}$$

Définition 5.13 (Pertes sur le paramètre)**.** *Soit $e_0 \in \mathcal{E}_0$. Pour $t \geq 0$, les fonctions*

$$\theta \in \Theta \mapsto p_t \circ g_t\,(e_0,\, \theta) = p_t\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta) \;\; \in \mathbb{R}$$

définissent des pertes sur Θ.

Corollaire 5.14 (Différentielles des pertes sur le paramètre)**.** *Pour $t \geq 0$,*

$$\begin{aligned} \frac{\partial\, p_t \circ g_t}{\partial \theta}\,(e_0,\, \theta) &= \frac{\partial\, p_t}{\partial e}\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta) \cdot \frac{\partial\, \boldsymbol{e}_t\,(e_0,\, \theta)}{\partial \theta} + \frac{\partial\, p_t}{\partial \theta}\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta) \\ &= \frac{\partial\, p_t}{\partial e}\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta) \cdot \boldsymbol{J}_t\,(e_0,\, 0,\, \theta) + \frac{\partial\, p_t}{\partial \theta}\,(\boldsymbol{e}_t\,(e_0,\, \theta)\,,\, \theta)\,. \end{aligned}$$

Démonstration. Les fonctions coordonnées associées aux états sont régulières, donc les g_t le sont aussi. Les pertes sont également régulières, et la première égalité s'obtient ainsi par dérivation de la composition. La seconde égalité est due au Lemme 4.21. □

Pour $t \geq 0$, les différentielles

$$\frac{\partial\, p_t \circ g_t}{\partial \theta}\,(e_0,\, \theta)$$

sont des formes linéaires. La métrique sur Θ permet de leur associer un vecteur tangent à Θ. Celle-ci est la métrique euclidienne, dont la matrice dans les bases de travail est égale à l'identité, et n'aura ainsi pas d'incidence sur les calculs ultérieurs. Ainsi, pour simplifier les notations, nous omettrons par la suite de faire la différence entre ces formes linéaires et les vecteurs tangents associés. Dans un système de coordonnées, cela revient à omettre le symbole qui dénote la transposition devant les matrices lignes associées à ces différentielles.

Nous aurons également besoin de considérer les formes linéaires sur Θ données, pour des suites $\boldsymbol{\theta}$ à valeurs dans Θ, par

$$\frac{\partial\, p_t}{\partial e}\,(\boldsymbol{e}_t\,(e_0,\, \boldsymbol{\theta})\,,\, \theta) \cdot \boldsymbol{J}_t\,(e_0,\, 0,\, \boldsymbol{\theta}) + \frac{\partial\, p_t}{\partial \theta}\,(\boldsymbol{e}_t\,(e_0,\, \boldsymbol{\theta})\,,\, \theta)\,.$$

De même que ci-dessus, nous omettrons par la suite de marquer la différence entre celles-ci et les vecteurs tangents associés que nous utiliserons en pratique. Ainsi, une égalité de la forme

$$\theta' = \theta - \eta \left(\frac{\partial\, p_t}{\partial e}\,(\boldsymbol{e}_t\,(e_0,\, \boldsymbol{\theta})\,,\, \theta) \cdot \boldsymbol{J}_t\,(e_0,\, 0,\, \boldsymbol{\theta}) + \frac{\partial\, p_t}{\partial \theta}\,(\boldsymbol{e}_t\,(e_0,\, \boldsymbol{\theta})\,,\, \theta) \right)$$

est formellement inexacte, car θ est un vecteur de Θ, et le terme qui lui est soustrait est une forme linéaire. Mais nous considérons implicitement que celle-ci est le vecteur tangent qui lui est associé par la métrique euclidienne, de sorte que l'égalité est bien pourvue de sens.

5.3 Borne sur les différentielles secondes et troisièmes des pertes sur les paramètres sur la boule stable pour le paramètre

Lemme 5.15 (Borne sur les différentielles secondes des pertes sur les paramètres sur la boule stable pour le paramètre)**.** *Les différentielles secondes des pertes sur les paramètres, renormalisées par la suite* $\boldsymbol{m}$*,*

$$\frac{1}{m_t} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta),$$

sont bornées uniformément en temps et en $\theta \in B_\Theta^*$*, pour la norme* $\|\cdot\|_{\text{bil}}$*.*

Remarque 5.16. *Ainsi qu'il a été dit lors de la définition de la norme* $\|\cdot\|_{\text{bil}}$*, comme l'application considérée est une forme bilinéaire, la norme* $\|\cdot\|_{\text{bil}}$ *coïncide en particulier avec la norme d'opérateur de sa matrice représentative dans les bases considérées.*

Démonstration. Notons, pour tout $t \geq 0$ et $\theta \in B_\Theta^*$,

$$e_t = \boldsymbol{e}_t (e_0^*, \theta) \quad \text{et} \quad J_t = \frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta}.$$

D'après le Corollaire 5.14, pour tout $t \geq 0$,

$$\begin{aligned} \frac{\partial p_t \circ g_t}{\partial \theta} (e_0^*, \theta) &= \frac{\partial p_t}{\partial e} (\boldsymbol{e}_t (e_0^*, \theta), \theta) \cdot \frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta} + \frac{\partial p_t}{\partial \theta} (\boldsymbol{e}_t (e_0^*, \theta), \theta) \\ &= \mathrm{d} p_t (\boldsymbol{e}_t (e_0^*, \theta), \theta) \cdot \left(\frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta}, \mathrm{Id}_{T\Theta} \right), \end{aligned}$$

et ainsi

$$\frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta) = \mathrm{d}^2 p_t (e_t, \theta) \cdot (J_t, \mathrm{Id}_{T\Theta})^{\otimes 2} + \frac{\partial p_t}{\partial e} (e_t, \theta) \cdot \frac{\partial^2 \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta^2}.$$

Par conséquent,

$$\begin{aligned} \frac{1}{m_t} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta) = &\frac{1}{m_t} \mathrm{d}^2 p_t (e_t, \theta) \cdot (J_t, \mathrm{Id}_{T\Theta})^{\otimes 2} \\ &+ \frac{1}{m_t} \frac{\partial p_t}{\partial e} (e_t, \theta) \cdot \frac{\partial^2 \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta^2}. \end{aligned}$$

Or, d'après le Corollaire 4.53, pour tout $t \geq 0$ et $\theta \in B_\Theta^*$, nous savons que $e_t \in B_{\mathcal{E}_t}^*$. Ainsi, d'après l'Hypothèse 5.4, les différentielles secondes des pertes, évaluées en (e_t, θ), sont bornées uniformément en temps et en $\theta \in B_\Theta^*$. D'après la Définition 5.9, le terme en $1/m_t$ en facteur de la différentielle seconde est borné.

D'après le Corollaire 5.11, les différentielles des pertes renormalisée par m_t, évaluées en (e_t, θ), sont bornées uniformément en temps et en $\theta \in B_\Theta^*$.

D'après le Corollaire 4.55, pour tout $t \geq 0$ et $\theta \in B_\Theta^*$, $J_t \in B_{\mathcal{L}(\Theta, \mathcal{E}_t)}^*$ donc, par construction de ces boules, définies à la Définition 4.51, les J_t sont bornés uniformément en temps et en $\theta \in B_\Theta^*$.

Enfin, les différentielles secondes des états sont bornées uniformément en temps et en $\theta \in B_\Theta^*$ d'après le Lemme 4.61, ce qui conclut la preuve. □

Corollaire 5.17 (Valeurs propres des différentielles secondes des pertes sur le paramètre bornées)**.** *Nous pouvons trouver $\Lambda^* \geq 0$ tel que, pour tout $t \geq 0$ et $\theta \in B^*_\Theta$, les valeurs propres de l'opérateur*

$$\frac{1}{m_t} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta)$$

sont bornées par Λ^*.

Démonstration. C'est une conséquence du Lemme 5.15 (car la norme $\|\cdot\|_{\text{bil}}$ est la norme d'opérateur de l'application linéaire associée par la métrique euclidienne, donc est la valeur absolue de la plus grande valeur propre). □

Lemme 5.18 (Borne sur les différentielles troisièmes des pertes sur les paramètres sur la boule stable pour le paramètre)**.** *Les différentielles troisièmes des pertes sur les paramètres, renormalisées par la suite $\boldsymbol{m}$,*

$$\frac{1}{m_t} \frac{\partial^3 p_t \circ g_t}{\partial \theta^3} (e_0^*, \theta),$$

sont bornées uniformément en temps et en $\theta \in B^*_\Theta$, pour la norme $\|\cdot\|_{\text{tri}}$.

Démonstration. Notons, pour tout $t \geq 0$ et $\theta \in B^*_\Theta$,

$$e_t = \boldsymbol{e}_t (e_0^*, \theta) \quad \text{et} \quad J_t = \frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta}.$$

Pour tout $t \geq 0$, ainsi qu'il a été vu au Lemme 5.15 précédent,

$$\begin{aligned} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta) = {} & \mathrm{d}^2 p_t (\boldsymbol{e}_t (e_0^*, \theta), \theta) \cdot \left(\frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta}, \mathrm{Id}_{T\Theta} \right)^{\otimes 2} \\ & + \frac{\partial p_t}{\partial e} (\boldsymbol{e}_t (e_0^*, \theta), \theta) \cdot \frac{\partial^2 \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta^2} \end{aligned}$$

et ainsi,

$$\frac{1}{m_t} \frac{\partial^3 p_t \circ g_t}{\partial \theta^3} (e_0^*, \theta)$$

est égal à

$$\begin{aligned} & \frac{1}{m_t} \mathrm{d}^3 p_t (e_t, \theta) \cdot \left(\frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta}, \mathrm{Id}_{T\Theta} \right)^{\otimes 3} \\ & + \frac{3}{m_t} \mathrm{d}^2 p_t (e_t, \theta) \cdot \left[\left(\frac{\partial^2 \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta^2}, 0 \right), \left(\frac{\partial \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta}, \mathrm{Id}_{T\Theta} \right) \right] \\ & + \frac{1}{m_t} \frac{\partial p_t}{\partial e} (e_t, \theta) \cdot \frac{\partial^3 \boldsymbol{e}_t (e_0^*, \theta)}{\partial \theta^3}. \end{aligned}$$

Or, d'après le Corollaire 4.53, pour tout $t \geq 0$ et $\theta \in B^*_\Theta$, nous savons que $e_t \in B^*_{\mathcal{E}_t}$. Ainsi, d'après l'Hypothèse 5.4, les différentielles deuxièmes et troisièmes des pertes, évaluées en (e_t, θ), sont bornées uniformément en temps et en $\theta \in B^*_\Theta$. D'après la Définition 5.9, le terme en $1/m_t$ en facteur de ces différentielles est borné.

D'après le Corollaire 5.11, les différentielles des pertes renormalisée par m_t, évaluées en (e_t, θ), sont bornées uniformément en temps et en $\theta \in B^*_\Theta$.

D'après le Corollaire 4.55, pour tout $t \geq 0$ et $\theta \in B^*_\Theta$, $J_t \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)}$ donc, par construction de ces boules, définies à la Définition 4.51, les J_t sont bornés uniformément en temps et en $\theta \in B^*_\Theta$.

Enfin, les différentielles secondes des états sont bornées uniformément en temps et en $\theta \in B^*_\Theta$ d'après le Lemme 4.61, et les différentielles troisièmes aussi, d'après le Lemme 4.62, ce qui conclut la preuve. □

Lemme 5.19 (Continuité uniforme des différentielles secondes des pertes sur le paramètre au voisinage de θ^*)**.** *La famille de fonctions*

$$\theta \mapsto \frac{1}{m_t} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta) ,$$

pour $t \geq 0$, est uniformément continue sur B^*_Θ.

Démonstration. C'est une conséquence du Lemme 5.18. □

Remarque 5.20. *En fait, nous aurons, lorsque nous utiliserons ce lemme, dans la preuve du Corollaire 10.6, uniquement besoin de l'équicontinuité en θ^*.*

Chapitre

6 Critère d'optimalité et changement d'échelle de temps

Dans ce chapitre uniquement, nous noterons $\sum_{s=g}^{d}$ une somme sur les indices $g \leq s \leq d-1$, afin d'alléger les notations (en retirant les « -1 » sur les bornes supérieures).

6.1 Fonctions d'échelle

Définition 6.1 (Fonction d'échelle)**.** *Nous appelons fonction d'échelle une fonction positive f, définie sur la demi-droite réelle positive, vérifiant les propriétés suivantes.*

1. *$f(t)$ tend vers l'infini, quand t tend vers l'infini.*
2. *f préserve les équivalents en l'infini.*
3. *f est croissante et $f(1) \geq 1$.*

Remarque 6.2. *Le dernier point sert à assurer que les suites considérées à la Définition 6.3 ci-dessous tendent vers l'infini avec k.*

Définition 6.3 (Suite « engendrée » par une fonction d'échelle)**.** *Soit $t \geq 0$. Soit une fonction d'échelle f. Nous appelons suite engendrée par f et initialisée à t, la suite (T_k) de valeur initiale $T_0 = t$ et qui vérifie la relation de récurrence, pour tout $k \geq 0$,*

$$T_{k+1} = T_k + f(T_k).$$

Remarque 6.4. *Afin d'alléger les notations, nous omettons d'expliciter la dépendance en t des suites (T_k).*

Corollaire 6.5 (Les suites « engendrées » par une fonction d'échelle tendent vers l'infini)**.** *Soit une fonction d'échelle f. Pour tout $t \geq 1$, la suite (T_k) engendrée par f, de valeur initiale t, est strictement croissante, et tend vers l'infini avec k.*

Démonstration. Soit $t \geq 1$. D'après la Définition 6.1, f est positive. Ainsi, (T_k) est croissante. Or, $T_0 = t \geq 1$ donc, pour tout $k \geq 0$, $T_k \geq 1$. De plus, toujours d'après la Définition 6.1, f est croissante et $f(1) \geq 1$, de sorte que, pour tout $t \geq 1$, $f(t) \geq 1$. Ainsi, pour tout $k \geq 0$,

$$T_{k+1} = T_k + f(T_k) \geq T_k + 1.$$

Par conséquent, (T_k) est bien strictement croissante, et T_k tend vers l'infini avec k. □

Lemme 6.6 (Équivalent d'une suite)**.** *Soient deux fonctions d'échelle L et ℓ telles que $\ell(t)$ est négligeable devant $L(t)$, et $L(t)$ est négligeable devant t, quand t tend vers l'infini.*

Pour tout $t \geq 0$, nous notons (T_k) la suite engendrée par ℓ, initialisée à t, et nous notons k_t l'unique $k \geq 1$ tel que

$$T_k \leq t + L(t) < T_{k+1}.$$

Alors, $T_{k_t} - t \sim L(t)$, quand t tend vers l'infini.

Démonstration. D'après le Corollaire 6.5, pour tout $t \geq 1$, la suite (T_k) est strictement croissante, et T_k tend vers l'infini avec k, de sorte que k_t est bien défini.

Par hypothèse, $\ell(t)$ est négligeable devant $L(t)$, quand t tend vers l'infini. Nous pouvons donc choisir $T \geq 1$ tel que, pour tout $t \geq T$, $\ell(t) \leq L(t)$. Par conséquent, pour tout $t \geq T$,

$$t + \ell(t) \leq t + L(t),$$

soit

$$T_1 \leq t + L(t).$$

Par conséquent, pour tout $t \geq T$, $k_t > 0$. Pour tout $t \geq T$,

$$T_{k_t} - t \leq L(t). \tag{6.1}$$

De plus, pour tout $t \geq T$,

$$L(t) \leq T_{k_t+1} - t = T_{k_t} - t + \ell(T_{k_t}),$$

d'après la Définition 6.3, de sorte que

$$L(t) - \ell(T_{k_t}) \leq T_{k_t} - t. \tag{6.2}$$

Or, d'après la Définition 6.1, ℓ est croissante. Ainsi,

$$\ell(T_{k_t}) \leq \ell(t + L(t)).$$

D'après la même définition, ℓ préserve les équivalents en l'infini, et est négligeable devant L. De plus, par hypothèse, $L(t)$ est négligeable devant t, quand t tend vers l'infini. Ainsi,

$$\ell(t + L(t)) \sim \ell(t) = o(L(t)),$$

quand t tend vers l'infini. Par conséquent,

$$\ell(T_{k_t}) = o(L(t)),$$

quand t tend vers l'infini. Donc,

$$L(t) - \ell(T_{k_t}) \sim L(t),$$

quand t tend vers l'infini et ainsi, d'après les Équations (6.1) et (6.2),

$$T_{k_t} - t \sim L(t),$$

quand t tend vers l'infini, ce qui conclut la preuve. □

6.2 Croissance des hypothèses de contrôle des sommes des gradients et des hessiennes

Corollaire 6.7 (Croissance de l'hypothèse de contrôle de la somme des termes d'une suite (de gradients))**.** *Soit une suite (u_t) à valeurs dans un espace vectoriel normé muni d'une norme notée $\|\cdot\|$.*

Supposons disposer d'une fonction d'échelle ℓ telle que

$$\sum_{s=0}^{t} u_s = O\left(\ell\left(t\right)\right), \tag{6.3}$$

quand t tend vers l'infini. Soit une fonction d'échelle L devant laquelle ℓ est négligeable. Alors,

$$\sum_{s=0}^{t} u_s = o\left(L\left(t\right)\right),$$

quand t tend vers l'infini.

Remarque 6.8. *Dans la suite, nous formulerons l'hypothèse pour*

$$u_t = \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right).$$

Démonstration. $\ell\left(t\right) = o\left(L\left(t\right)\right)$, quand t tend vers l'infini, donc le terme en $O\left(\ell\left(t\right)\right)$ est négligeable devant $L\left(t\right)$, quand t tend vers l'infini. Nous obtenons ainsi le résultat annoncé. □

Lemme 6.9 (Moyennes sur des intervalles)**.** *Soit un intervalle $I = [g,\, d[$, de longueur L. Soit une suite d'intervalles consécutifs $J_k = [T_k,\, T_{k+1}[$, de longueurs ℓ_k, telle que $T_0 = g$ et $L_0 \leq L$.*

Soit une suite de réels (h_t). Supposons que, pour un certain $\tilde{\lambda}$, pour tout $k \geq 0$,

$$\frac{1}{\ell_k} \sum_{J_k} h_t \geq \tilde{\lambda}.$$

Notons k_I l'unique $k \geq 0$ tel que

$$T_k \leq d = g + L < T_{k+1}.$$

Alors,

$$\frac{1}{L} \sum_{I} h_t \geq \tilde{\lambda}\, \frac{T_{k_I} - g}{L} - \frac{\ell_{k_I}}{L} \sup_{J_{k_I}} \left|h_t\right|.$$

Démonstration. $L_0 \leq L$, donc $T_0 + L_0 \leq T_0 + L$, c'est-à-dire $T_1 \leq d$, car $T_0 = g$, et ainsi $k_I > 0$. Par conséquent,

$$\sum_{I} h_t = \sum_{k=0}^{k_I - 1} \sum_{J_k} h_t + \sum_{t=T_{k_I}}^{d} h_t$$

Or, par hypothèse, pour tout $0 \leq k \leq k_I - 1$,

$$\sum_{J_k} h_t \geq \tilde{\lambda}\, \ell_k.$$

De plus,

$$\left| \sum_{t=T_{k_I}}^{d} h_t \right| \leq (T_{k_I+1} - T_{k_I}) \sup_{J_k} |h_t| = \ell_{k_I} \sup_{J_{k_I}} |h_t| .$$

Par conséquent,

$$\begin{aligned} \sum_I h_t &\geq \tilde{\lambda} \sum_{k=0}^{k_I-1} \ell_k - \ell_{k_I} \sup_{J_{k_I}} |h_t| \\ &= \tilde{\lambda} \, (T_{k_I} - g) - \ell_{k_I} \sup_{J_{k_I}} |h_t| , \end{aligned}$$

ce qui, une fois l'inégalité divisée par L, est le résultat annoncé. □

Lemme 6.10 (Croissance de l'hypothèse de borne inférieure des valeurs propres des sommes des termes d'une suite (de hessiennes))**.** *Soient ℓ et L deux fonctions d'échelle, telles que $\ell(t)$ est négligeable devant $L(t)$, et $L(t)$ est négligeable devant t, quand t tend vers l'infini.*

Supposons disposer d'une suite (h_t) d'opérateurs auto-adjoints, et d'un réel $\tilde{\lambda}$ tels que, pour t suffisamment grand, la plus petite valeur propre des opérateurs

$$\frac{1}{\ell(t)} \sum_{s=t}^{t+\ell(t)} h_s$$

est supérieure à $\tilde{\lambda}$. Supposons également que

$$\frac{\ell(t)}{L(t)} \sup_{s \in [t,\, t+o(t)[} \|h_s\|_{\text{bil}} = o(1) ,$$

quand t tend vers l'infini. Alors, pour t suffisamment grand, la plus petite valeur propre des opérateurs

$$\frac{1}{L(t)} \sum_{s=t}^{t+L(t)} h_s$$

est supérieure à $\tilde{\lambda}/2$.

Remarque 6.11. *Dans la suite, nous formulerons l'hypothèse pour*

$$h_t = \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta^*) .$$

Démonstration. $\ell(t)$ est négligeable devant $L(t)$, quand t tend vers l'infini, donc nous pouvons choisir $T \geq 0$ tel que, pour tout $t \geq T$,

$$\ell(t) \leq L(t) .$$

D'après l'hypothèse vérifiée par les h_t, quitte à augmenter T, nous pouvons également supposer que, pour tout $t \geq T$, la plus petite valeur propre des opérateurs

$$\frac{1}{\ell(t)} \sum_{s=t}^{t+\ell(t)} h_s$$

est supérieure à $\tilde{\lambda}$.

Minoration sur un intervalle Soit v appartenant à la sphère unité de l'espace sur lequel agissent les h_t. Soit alors $t \geq T$. D'après la définition de T ci-dessus, nous savons que

$$\frac{1}{\ell(t)} \sum_{s=t}^{t+\ell(t)} {}^t v\, h_s\, v \geq \tilde{\lambda}.$$

Considérons l'intervalle $I = [t,\, t + L(t)\,[$. Considérons la suite (T_k), engendrée par ℓ et initialisée à t. Considérons alors, pour $k \geq 0$, les intervalles $J_k = [T_k,\, T_{k+1}[$. D'après le Lemme 6.9 appliqué aux réels ${}^t v\, h_s\, v$,

$$\frac{1}{L(t)} \sum_{s=t}^{t+L(t)} {}^t v\, h_s\, v \geq \tilde{\lambda}\, \frac{T_{k_I} - t}{L(t)} - \frac{\ell(T_{k_I})}{L(t)} \sup_{[T_{k_I},\, T_{k_I} + \ell(T_{k_I})[} \left| {}^t v\, h_s\, v \right|.$$

Or, pour tout $s \in [T_{k_I},\, T_{k_I} + \ell(T_{k_I})\,[$,

$$\left| {}^t v\, h_s\, v \right| \leq \|h_s\|_{\text{bil}}\, \|v\|^2 = \|h_s\|_{\text{bil}}.$$

Ainsi,

$$\frac{1}{L(t)}\, {}^t v \left(\sum_{s=t}^{t+L(t)} h_s \right) v \geq \tilde{\lambda}\, \frac{T_{k_I} - t}{L(t)} - \frac{\ell(T_{k_I})}{L(t)} \sup_{[T_{k_I},\, T_{k_I} + \ell(T_{k_I})[} \|h_s\|_{\text{bil}}.$$

Comportement en l'infini Or, pour tout $t \geq 1$, $k_I = k_t$, avec la définition de k_t donnée au Lemme 6.6. Par conséquent, pour tout t suffisamment grand,

$$\frac{1}{L(t)}\, {}^t v \left(\sum_{s=t}^{t+L(t)} h_s \right) v \geq \tilde{\lambda}\, \frac{T_{k_t} - t}{L(t)} - \frac{\ell(T_{k_t})}{L(t)} \sup_{[T_{k_t},\, T_{k_t} + \ell(T_{k_t})[} \|h_s\|_{\text{bil}}.$$

D'après le même lemme,

$$T_{k_t} - t \sim L(t),$$

quand t tend vers l'infini, et le premier terme de la quantité minorante tend vers $\tilde{\lambda}$, quand t tend vers l'infini. De plus, d'après la preuve du lemme,

$$\ell(T_{k_t}) = O(\ell(t)),$$

quand t tend vers l'infini. Pour $t \geq 1$, $T_{k_t} \geq t$. D'après la Définition 6.1, ℓ préserve les équivalents en l'infini. Par hypothèse, ℓ est négligeable devant L, qui est elle-même négligeable devant l'identité et ainsi, d'après ce qui précède,

$$T_{k_t} + \ell(T_{k_t}) = t + o(t),$$

quand t tend vers l'infini. Donc, avec un léger abus de notation,

$$[T_{k_t},\, T_{k_t} + \ell(T_{k_t})\,[\subset [t,\, t + o(t)\,[,$$

quand t tend vers l'infini. Par conséquent, le deuxième terme de la quantité minorante est dominé par

$$\frac{\ell(t)}{L(t)} \sup_{s \in [t,\, t + o(t)[} \|h_s\|_{\text{bil}},$$

quand t tend vers l'infini, et tend donc vers 0, quand t tend vers l'infini. Ainsi, pour tout t assez grand,

$$\frac{1}{L(t)}\, {}^t v \left(\sum_{s=t}^{t+L(t)} h_s \right) v \geq \frac{\tilde{\lambda}}{2}.$$

Conclusion Or, cette minoration est uniforme en les vecteurs v de la sphère unité de l'espace sur lequel agissent les h_t, ce qui conclut la preuve. □

6.3 Transfert des contrôles des sommes aux sommes pondérées, sous l'hypothèse d'homogénéité des pas

Lemme 6.12 (Somme négligeable sur les intervalles)**.** *Soit une suite (u_t) à valeurs dans un espace vectoriel normé. Supposons disposer d'une fonction d'échelle L telle que*

$$\sum_{s=0}^{L(t)} u_t = o\left(L\left(t\right)\right),$$

quand t tend vers l'infini. Supposons disposer d'une suite d'intervalles $I_k = [d_k,\, d_{k+1}[$ telle que d_k tend vers l'infini avec k, et telle que $d_k \sim d_{k+1}$, quand k tend vers l'infini. Alors,

$$\frac{1}{L\left(d_k\right)} \sum_{I_k} u_t = o\left(1\right),$$

quand k tend vers l'infini.

Remarque 6.13. *Les intervalles I_k sont consécutifs. Cette propriété ne sera pas demandée dans les lemmes suivants, et les intervalles s'écriront alors $[g_k,\, d_k[$.*

Démonstration. Pour tout $k \geq 1$,

$$\sum_{I_k} u_t = \sum_{t=0}^{d_{k+1}} u_t - \sum_{t=0}^{d_k} u_t.$$

Or, d_k tend vers l'infini avec k donc, d'après l'hypothèse sur les sommes des u_t, le premier terme de la quantité majorante est négligeable devant $L\left(d_{k+1}\right)$, et le deuxième terme est négligeable devant $L\left(d_k\right)$, quand k tend vers l'infini. Or, par hypothèse, $d_k \sim d_{k+1}$, quand k tend vers l'infini. De plus, L préserve les équivalents en l'infini, en tant que fonction d'échelle. Ainsi,

$$\frac{1}{L\left(d_k\right)} \sum_{I_k} u_t = o\left(1\right),$$

quand k tend vers l'infini, ce qui est le résultat annoncé. □

Lemme 6.14 (Transfert du contrôle des sommes de gradients)**.** *Soit une suite (u_t) à valeurs dans un espace vectoriel normé muni d'une norme notée $\|\cdot\|$.*

Supposons disposer d'une suite d'intervalles (I_k), de longueurs L_k, telle que

$$\frac{1}{L_k} \sum_{I_k} u_t = o\left(1\right),$$

quand k tend vers l'infini.

Supposons également disposer d'une suite $(m_k\left(\boldsymbol{u}\right))$ telle que, pour tout k suffisamment grand,

$$\sup_{I_k} \|u_t\| \leq m_k\left(\boldsymbol{u}\right).$$

Supposons enfin disposer d'une suite $\boldsymbol{\eta} = (\eta_t)$ telle que

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t} = 1 + o\left(\frac{1}{m_k(\boldsymbol{u})}\right),$$

quand k tend vers l'infini. Alors,

$$\sum_{t \in I_k} \eta_t \, u_t = o\left(\sum_{t \in I_k} \eta_t\right),$$

quand k tend vers l'infini.

Démonstration. Soit $k \geq 0$. Notons

$$i_k = \inf_{I_k} \eta_t,$$

et

$$s_k = \sup_{I_k} \eta_t.$$

Alors,

$$\sum_{t \in I_k} \eta_t \, u_t = i_k \sum_{t \in I_k} u_t + i_k \sum_{t \in I_k} \left(\frac{\eta_t}{i_k} - 1\right) u_t,$$

de sorte que

$$\left\| \sum_{I_k} \eta_t \, u_t \right\| \leq i_k \left\| \sum_{I_k} u_t \right\| + i_k \left\| \sum_{I_k} \left(\frac{\eta_t}{i_k} - 1\right) u_t \right\|.$$

Contrôle de la deuxième somme de la quantité majorante

$$\begin{aligned} \left\| \sum_{I_k} \left(\frac{\eta_t}{i_k} - 1\right) u_t \right\| &\leq \sum_{I_k} \left\| \left(\frac{\eta_t}{i_k} - 1\right) u_t \right\| \\ &= \sum_{I_k} \left(\frac{\eta_t}{i_k} - 1\right) \|u_t\| \\ &\leq \left(\frac{s_k}{i_k} - 1\right) \sup_{I_k} \|u_t\| \; L_k. \end{aligned}$$

Conclusion Ainsi,

$$\left\| \sum_{I_k} \eta_t \, u_t \right\| \leq i_k \, o(L_k) + i_k \left(\frac{s_k}{i_k} - 1\right) \sup_{I_k} \|u_t\| \; L_k.$$

Or,

$$\sum_{I_k} \eta_t \geq i_k \, L_k.$$

Par conséquent,

$$\begin{aligned} \frac{\left\| \sum_{I_k} \eta_t \, u_t \right\|}{\sum_{I_k} \eta_t} &\leq o(1) + \left(\frac{s_k}{i_k} - 1\right) \sup_{I_k} \|u_t\| \\ &\leq o(1) + \left(\frac{s_k}{i_k} - 1\right) m_k(\boldsymbol{u}). \end{aligned}$$

Par hypothèse, le deuxième terme de la quantité majorante tend vers 0, quand k tend vers l'infini. Ainsi,

$$\frac{\left\|\sum_{I_k} \eta_t \, u_t\right\|}{\sum_{I_k} \eta_t}$$

tend vers 0 quand k tend vers l'infini, ce qui conclut la preuve. □

Lemme 6.15 (Transfert du contrôle des sommes de hessiennes)**.** *Soit une suite d'opérateurs auto-adjoints (h_t) sur un espace vectoriel.*

Supposons disposer d'une suite d'intervalles (I_k), de longueurs L_k, et d'un réel $\tilde{\lambda}$ tels que, pour k suffisamment grand, la plus petite valeur propre des opérateur

$$\frac{1}{L_k} \sum_{t \in I_k} h_t$$

est supérieure à $\tilde{\lambda}$.

Supposons également disposer d'une suite $(m_k(\boldsymbol{h}))$ telle que, pour tout k suffisamment grand,

$$\sup_{I_k} \|h_t\|_{\mathrm{op}} \leq m_k(\boldsymbol{h}).$$

Supposons enfin disposer d'une suite $\boldsymbol{\eta} = (\eta_t)$ telle que

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t} = 1 + o\left(\frac{1}{m_k(\boldsymbol{h})}\right),$$

quand k tend vers l'infini. Alors, pour k suffisamment grand, la plus petite valeur propre des opérateurs

$$\frac{\sum_{I_k} \eta_t \, h_t}{\sum_{I_k} \eta_t}$$

est supérieure à $\tilde{\lambda}/2$.

Démonstration. Soit $k \geq 0$. Notons, de même qu'à la preuve du Lemme 6.14,

$$i_k = \inf_{I_k} \eta_t,$$

et

$$s_k = \sup_{I_k} \eta_t.$$

Alors,

$$\sum_{I_k} \eta_t \, h_t = s_k \sum_{I_k} h_t + s_k \sum_{I_k} \left(\frac{\eta_t}{s_k} - 1\right) h_t.$$

Soit à présent v appartenant à la sphère unité de l'espace sur lequel agissent les h_t. Alors,

$${}^t v \left(\sum_{I_k} \eta_t \, h_t\right) v = s_k \, {}^t v \left(\sum_{I_k} h_t\right) v + s_k \, {}^t v \left(\sum_{I_k} \left(\frac{\eta_t}{s_k} - 1\right) h_t\right) v.$$

Minoration du premier terme du membre de droite Par hypothèse, pour tout k assez grand, indépendamment de v,

$$\begin{aligned} {}^t v \left(\sum_{I_k} h_t\right) v &\geq \tilde{\lambda} \, L_k \, \|v\|^2 \\ &= \tilde{\lambda} \, L_k. \end{aligned}$$

Majoration du deuxième terme du membre de droite Pour tout $k \geq 0$,

$$\begin{aligned}\left| {}^t v \left(\sum_{I_k} \left(\frac{\eta_t}{s_k} - 1 \right) h_t \right) v \right| &\leq \left\| \sum_{I_k} \left(\frac{\eta_t}{s_k} - 1 \right) h_t \right\|_{\mathrm{op}} \|v\|^2 \\ &\leq \sum_{t \in I_k} \left| \frac{\eta_t}{s_k} - 1 \right| \|h_t\|_{\mathrm{op}} \\ &\leq \left(1 - \frac{i_k}{s_k} \right) \sup_{I_k} \|h_t\|_{\mathrm{op}} \, L_k.\end{aligned}$$

Conclusion Ainsi, pour tout k assez grand,

$${}^t v \left(\sum_{t \in I_k} \eta_t \, h_t \right) v \geq s_k \, \tilde{\lambda} \, L_k - s_k \left(1 - \frac{i_k}{s_k} \right) \sup_{I_k} \|h_t\|_{\mathrm{op}} \, L_k.$$

Or, pour tout $k \geq 0$,

$$i_k \, L_k \leq \sum_{t \in I_k} \eta_t \leq s_k \, L_k.$$

Par conséquent, pour tout k assez grand,

$$\begin{aligned}\frac{{}^t v \left(\sum_{I_k} \eta_t \, h_t \right) v}{\sum_{t \in I_k} \eta_t} &\geq \tilde{\lambda} - \left(1 - \frac{i_k}{s_k} \right) \sup_{I_k} \|h_t\|_{\mathrm{op}} \, \frac{s_k \, L_k}{\sum_{t \in I_k} \eta_t} \\ &\geq \tilde{\lambda} - \left(1 - \frac{i_k}{s_k} \right) \frac{s_k}{i_k} \sup_{I_k} \|h_t\|_{\mathrm{op}} \\ &\geq \tilde{\lambda} - \left(1 - \frac{i_k}{s_k} \right) \frac{s_k}{i_k} \, m_k \left(\boldsymbol{h} \right).\end{aligned}$$

D'après l'hypothèse sur les pas, le deuxième terme de la minoration tend vers 0, quand k tend vers l'infini. Ainsi, pour tout k suffisamment grand,

$$\frac{{}^t v \left(\sum_{I_k} \eta_t \, h_t \right) v}{\sum_{I_k} \eta_t} \geq \frac{\tilde{\lambda}}{2}.$$

Or, cette minoration est uniforme en les vecteurs v de la sphère unité, ce qui conclut la preuve. □

6.4 Contrôle des sommes des pas sur les intervalles

Lemme 6.16 (Contrôle des sommes des pas sur les intervalles)**.** *Soit une suite d'intervalles $I_k = [g_k, d_k[$ de longueurs L_k, telle que g_k tend vers l'infini avec k. Soient de plus une fonction d'échelle M_p, et une suite $\boldsymbol{\eta} = (\eta_t)$, telles que*

$$\sum_{I_k} \eta_t = o\left(\frac{1}{M_p \left(d_k \right)^2} \right),$$

quand k tend vers l'infini. Les propriétés suivantes sont alors vérifiées.

1. $$\left(\sum_{I_k} M_p \left(t \right) \eta_t \right)^2 \quad et \quad \sum_{I_k} M_p \left(t \right)^2 \eta_t^2$$

sont négligeables devant la somme précédente, quand k tend vers l'infini.

2. *Nous supposons de plus que*

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t}$$

est borné, quand k tend vers l'infini, et que

$$M_p(d_k) = o(L_k),$$

quand k tend vers l'infini. Alors,

$$s_k = \sup_{I_k} M_p(t)\ \eta_t$$

est négligeable devant

$$\sum_{I_k} \eta_t,$$

quand k tend vers l'infini.

Démonstration. M_p est une fonction d'échelle donc, d'après la Définition 6.1, elle est croissante. Ainsi, pour tout $k \geq 0$, pour tout $g_k \leq t < d_k$, $M_p(t)\ \eta_t \leq M_p(d_k)\ \eta_t$. Par conséquent,

$$\sum_{I_k} M_p(t)\ \eta_t \leq M_p(d_k) \sum_{I_k} \eta_t = o\left(\frac{1}{M_p(d_k)}\right), \tag{6.4}$$

quand k tend vers l'infini. Or, g_k tend vers l'infini avec k, et par conséquent d_k également. Ainsi,

$$\sum_{I_k} M_p(t)\ \eta_t$$

tend vers 0, quand k tend vers l'infini, et est en particulier borné.

Obtention du premier point Pour tout $k \geq 0$,

$$\left(\sum_{I_k} M_p(t)\ \eta_t\right)^2 \leq M_p(d_k)^2 \left(\sum_{I_k} \eta_t\right)^2$$

donc, d'après l'hypothèse sur la somme des pas sur les intervalles I_k,

$$\left(\sum_{I_k} M_p(t)\ \eta_t\right)^2 \leq o(1) \sum_{I_k} \eta_t,$$

quand k tend vers l'infini. De plus, la somme des carrés des termes $M_p(t)\ \eta_t$ est inférieure au carré de leur somme, car ils sont positifs, et le premier point est ainsi vérifié.

Obtention du deuxième point Notons, pour tout $k \geq 0$,

$$i_k = \inf_{I_k} \eta_t.$$

Alors, pour tout $k \geq 0$,

$$\sum_{I_k} \eta_t \geq i_k\, L_k.$$

De plus, M_p est croissante et ainsi, pour tout $k \geq 0$,

$$\sup_{I_k} M_p(t)\ \eta_t \leq M_p(d_k) \sup_{I_k} \eta_t = M_p(d_k)\ s_k.$$

Ainsi,

$$\sup_{I_k} M_p(t)\ \eta_t \leq \frac{M_p(d_k)}{L_k} \frac{s_k}{i_k} \sum_{I_k} \eta_t.$$

Or, par hypothèse, s_k/i_k est borné et $M_p(d_k)/L_k$ tend vers 0, quand k tend vers l'infini. Par conséquent,

$$\sup_{I_k} M_p(t)\ \eta_t = o\left(\sum_{I_k} \eta_t\right),$$

quand k tend vers l'infini, ce qui conclut la preuve. □

6.5 Critère d'optimalité

6.5.1 Hypothèses pour obtenir l'optimalité

Hypothèse 6.17 (Hypothèses sur les gradients, les hessiennes et les pertes)**.** *Supposons disposer de fonctions d'échelle ℓ^1, ℓ^2 et M_p, négligeables devant l'identité en l'infini, telles que les propriétés suivantes sont vérifiées.*

1. Critère d'optimalité 1/2.

$$\sum_{t=0}^{\ell^1(t)} \frac{\partial\, p_t \circ g_t}{\partial \theta}(e_0^*,\, \theta^*) = O\left(\ell^1(t)\right),$$

   quand t tend vers l'infini.

2. Critère d'optimalité 2/2. *Nous disposons d'un réel $\tilde{\lambda} > 0$ tel que, pour t suffisamment grand, la plus petite valeur propre des opérateurs*

$$\frac{1}{\ell^2(t)} \sum_{s=t}^{t+\ell^2(t)} \frac{\partial^2\, p_s \circ g_s}{\partial \theta^2}(e_0^*,\, \theta^*)$$

   est supérieure à $\tilde{\lambda}$.

3. $$\max\left(\ell^1(t),\, \ell^2(t),\, M_p(t)\right)\ M_p(t)^2$$

   est négligeable devant t, quand t tend vers l'infini.

4. Contrôle des pertes le long de la trajectoire optimale. *Nous supposons que*

$$\left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}(e_0^*,\, \theta^*) \right\| = O\left(M_p(t)\right),$$

   quand t tend vers l'infini.

Hypothèse 6.18 (Suite de pas de descente)**.** *Nous supposons disposer d'une suite de pas $\boldsymbol{\eta} = (\eta_t)$ qui vérifie les propriétés suivantes.*

1. *La série de terme général η_t diverge.*

2.

$$\eta_t = o\left(\frac{1}{\max\left(\ell^1(t)\,,\,\ell^2(t)\,,\,M_p(t)\right)\,M_p(t)^2}\right),$$

quand t tend vers l'infini.

3. *(Homogénéité temporelle) Pour toute suite d'intervalles $I_t = [g_t,\, d_t[$ tels que g_t tend vers l'infini et $g_t \sim d_t$, quand t tend vers l'infini, nous avons*

$$\frac{\sup_{I_t}\eta_s}{\inf_{I_t}\eta_s} = 1 + o\left(\frac{1}{M_p(d_t)}\right),$$

quand t tend vers l'infini.

4. *La suite de terme général $\eta_t^{-1}\,M_p(t)^{-2}$ est une fonction d'échelle.*

Remarque 6.19. *La première hypothèse ne sera utilisée qu'à la preuve du Lemme 10.18.*

Remarque 6.20. *La dernière hypothèse sert à ce que la fonction L du Lemme 6.21 ci-dessous soit une fonction d'échelle. Sa vérification n'est pas très contraignante, comme cela sera vu au Lemme 6.43.*

6.5.2 Changement d'échelle de temps : construction des intervalles pour la convergence

Lemme 6.21 (Fonction longueur d'intervalles adaptée)**.** *Nous pouvons choisir une fonction d'échelle L, devant laquelle ℓ^1, ℓ^2 et M_p sont négligeables, et négligeable devant $\frac{1}{\eta_t\,M_p(t)^2}$, quand t tend vers l'infini.*

Démonstration. Pour tout $t \geq 0$, nous posons

$$L(t) = \left(\min\left(t,\,\frac{1}{\eta_t\,M_p(t)^2}\right)\max\left(\ell^1(t)\,,\,\ell^2(t)\,,\,M_p(t)\right)\right)^{1/2} + 1.$$

D'après l'Hypothèse 6.18, $\max\left(\ell^1(t)\,,\,\ell^2(t)\,,\,M_p(t)\right)$ est négligeable devant $\frac{1}{\eta_t\,M_p(t)^2}$, quand t tend vers l'infini. Il est également négligeable devant la fonction identité, comme ℓ^1, ℓ^2 et M_p. Par conséquent, il l'est devant le minimum de ces deux quantités. Ainsi, la racine du produit est négligeable devant le minimum, et le maximum est négligeable devant elle. Enfin, les minima, maxima, produits et puissances de fonctions d'échelle sont des fonctions d'échelle. L vérifie donc bien les propriétés demandées. □

Définition 6.22 (Construction d'une échelle de temps)**.** *Nous définissons la suite $(T_k)_{\,k\geq 0}$ par sa valeur initiale $T_0 = 1$ et la relation de récurrence, pour $k \geq 0$,*

$$T_{k+1} = T_k + L(T_k)\,.$$

Lemme 6.23 (Propriétés de l'échelle de temps (T_k))**.** 1. *(T_k) est strictement croissante, et T_k tend vers l'infini, quand k tend vers l'infini.*

2. *$T_{k+1} \sim T_k$, quand k tend vers l'infini.*

Démonstration. Prouvons d'abord la première propriété. Comme L est positive, la suite (T_k) est croissante. Or, T_0 vaut 1. Donc, pour tout $k \geq 0$, $T_k \geq 1$. Ainsi, d'après Lemme 6.21, pour tout $k \geq 0$,

$$T_{k+1} \geq T_k + L(T_k) \geq T_k + 1.$$

Par conséquent, (T_k) est strictement croissante, et T_k tend vers l'infini avec k.

Pour la deuxième propriété, nous savons que $L(t) = o(t)$, quand t tend vers l'infini. Par conséquent,

$$T_k + L(T_k) \sim T_k,$$

quand k tend vers l'infini, et ainsi $T_{k+1} \sim T_k$, quand k tend vers l'infini, ce qui conclut la preuve. □

Remarque 6.24. *La première propriété justifie que nous parlions de (T_k) comme d'une échelle de temps.*

Remarque 6.25. *Les intervalles $[T_k, T_{k+1}[$ sont d'autant plus grands que L est grand. Le choix de L est fixé par les contraintes du Lemme 6.21.*

Pour $k \geq 0$, nous posons alors

$$I_k = [T_k, T_{k+1}[.$$

6.5.3 Conséquences

Corollaire 6.26 (Contrôle sur les pertes)**.** *m_t, introduit à la Définition 5.9, vérifie*

$$m_t = O(M_p(t)),$$

quand t tend vers l'infini.

Démonstration. D'après le début de la section 4.3.1 et la Définition 5.12, pour tout $t \geq 0$,

$$(\boldsymbol{e}_t(e_0^*, \theta^*), \theta^*) = (e_t^*, \theta^*).$$

Par conséquent, d'après le Corollaire 5.7 et la Définition 5.9, pour tout $t \geq 0$,

$$m_t \leq \left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}(e_0^*, \theta^*) \right\| + S_{\text{perte}} \left(r_{\mathcal{E}}^2 + {r_{\Theta}^*}^2 \right)^{1/2}.$$

L'Hypothèse 6.18 permet alors d'obtenir le résultat annoncé. □

Corollaire 6.27 (Majoration des sommes des gradients et des hessiennes sur les intervalles)**.**

$$\sup_{I_k} \left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}(e_0^*, \theta^*) \right\| = O(M_p(T_{k+1})),$$

et

$$\sup_{I_k} \left\| \frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}(e_0^*, \theta^*) \right\|_{\text{bil}} = O(M_p(T_{k+1})),$$

quand k tend vers l'infini.

Démonstration. D'après le Lemme 8.3 et le Corollaire 6.26, la suite de terme général

$$\frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right)$$

est dominée par $M_p\left(t\right)$, quand t tend vers l'infini. Ainsi, nous pouvons trouver une constante $\kappa > 0$ telle que, pour tout t suffisamment grand,

$$\left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) \right\| \leq \kappa\, M_p\left(t\right).$$

Or, d'après l'Hypothèse 6.17, M_p est une fonction d'échelle, et est par conséquent croissante. Ainsi, pour tout k suffisamment grand, pour tout $t \in I_k$,

$$\left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) \right\| \leq \kappa\, M_p\left(T_{k+1}\right).$$

De la sorte, pour tout k suffisamment grand,

$$\sup_{I_k} \left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) \right\| \leq \kappa\, M_p\left(T_{k+1}\right).$$

Pour les différentielles secondes, d'après le Lemme 5.15 et le Corollaire 6.26, la suite de terme général

$$\frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e_0^*,\, \theta^*\right)$$

est dominée par $M_p\left(t\right)$, quand t tend vers l'infini. Le résultat s'obtient alors de la même manière que pour les gradients. □

Corollaire 6.28 (Homogénéité temporelle satisfaite)**.**

$$\frac{\sup_{I_k}\, \eta_s}{\inf_{I_k}\, \eta_s} = 1 + o\left(\frac{1}{M_p\left(T_{k+1}\right)}\right),$$

quand k tend vers l'infini.

Démonstration. La suite de pas vérifie l'Hypothèse 6.18, donc en particulier le troisième point de celle-ci. D'après le Lemme 6.23, T_k tend vers l'infini avec k, et $T_k \sim T_{k+1}$, quand k tend vers l'infini. Ainsi, en appliquant ce troisième point à la suite d'intervalles $[T_k,\, T_{k+1}[$, nous obtenons

$$\frac{\sup_{s \in [T_k,\, T_{k+1}[}\, \eta_s}{\inf_{s \in [T_k,\, T_{k+1}[}\, \eta_s} = 1 + o\left(\frac{1}{M_p\left(T_{k+1}\right)}\right),$$

quand k tend vers l'infini. Or, par définition, $I_k = [T_k,\, T_{k+1}[$, ce qui conclut la preuve. □

6.5.4 Existence d'un minimum local

Fait 6.29 (Négligeabilité de la somme des gradients en θ^* sur un intervalle devant la durée de celui-ci)**.**

$$\sum_{I_k} \eta_t \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) = o\left(\sum_{I_k} \eta_t\right),$$

quand k tend vers l'infini.

Démonstration. La suite de terme général

$$\frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right)$$

vérifie le premier point de l'Hypothèse 6.17. Or, d'après le Lemme 6.21, $\ell^1\left(t\right)$ est négligeable devant $L\left(t\right)$, quand t tend vers l'infini. Par conséquent, d'après le Corollaire 6.7,

$$\sum_{s=0}^{t} \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) = o\left(L\left(t\right)\right),$$

quand t tend vers l'infini. De plus, d'après le Lemme 6.23, T_k tend vers l'infini, et $T_{k+1} \sim T_k$, quand k tend vers l'infini. Par conséquent, d'après le Lemme 6.12,

$$\frac{1}{L\left(T_k\right)} \sum_{I_k} u_t = o\left(1\right),$$

quand k tend vers l'infini. Or, pour tout $k \geq 0$,

$$I_k = \left[T_k,\, T_{k+1}\right[$$

et, d'après la Définition 6.22, pour tout $k \geq 0$,

$$T_{k+1} = T_k + L\left(T_k\right),$$

de sorte que $L\left(T_k\right)$ est la longueur de I_k. Enfin, d'après le Corollaire 6.27, nous pouvons trouver $\kappa \geq 0$ tel que, pour tout k suffisamment grand,

$$\sup_{I_k} \left\| \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right) \right\| \leq \kappa\, M_p\left(T_{k+1}\right).$$

D'après le Corollaire 6.28, nous pouvons ainsi utiliser le Lemme 6.14 avec $m_k\left(\boldsymbol{u}\right) = \kappa\, M_p\left(T_{k+1}\right)$ pour obtenir le résultat annoncé. □

Lemme 6.30 (Caractère défini positif en θ^* de la somme des pertes sur un intervalle de temps)**.** *Nous pouvons trouver $\lambda > 0$ tel que, pour k suffisamment grand, la plus petite valeur propre des opérateurs*

$$\frac{\sum_{I_k} \eta_t \frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e_0^*,\, \theta^*\right)}{\sum_{I_k} \eta_t}$$

est supérieure à λ.

Démonstration. La suite de terme général

$$\frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e_0^*,\, \theta^*\right)$$

vérifie le deuxième point de l'Hypothèse 6.17. Or, d'après le Lemme 6.21, $\ell^2\left(t\right)$ est négligeable devant $L\left(t\right)$, quand t tend vers l'infini. Par conséquent, d'après le Lemme 6.10, la plus petite valeur propre des opérateurs

$$\frac{1}{L\left(t\right)} \sum_{s=t}^{t+L(t)} \frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e_0^*,\, \theta^*\right)$$

est supérieure à $\tilde{\lambda}$, quand t tend vers l'infini. Enfin, d'après le Corollaire 6.27, nous pouvons trouver $\kappa \geq 0$ tel que, pour tout k suffisamment grand,

$$\sup_{I_k} \left\| \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta^*) \right\| \leq \kappa M_p (T_{k+1}).$$

D'après le Corollaire 6.28, nous pouvons ainsi utiliser le Lemme 6.15 avec $m_k(\boldsymbol{h}) = \kappa M_p (T_{k+1})$ pour obtenir le résultat annoncé, en posant $\lambda = \tilde{\lambda}/2 > 0$. □

6.5.5 Échelle de temps et pas de descente

Corollaire 6.31 (Contrôle de la somme des pas sur les intervalles)**.**

$$\sum_{I_k} \eta_t = o\left(\frac{1}{M_p (T_{k+1})^2} \right),$$

quand k tend vers l'infini.

Démonstration. Soit $\varepsilon > 0$. D'après le Lemme 6.21, nous pouvons trouver $t \geq T$ tel que, pour tout $t \geq T$,

$$L(t) \leq \frac{\varepsilon}{\eta_t M_p (t)^2}.$$

Or, d'après l'Hypothèse 6.17 et la Définition 6.1, ${M_p}^2$ est croissante. Ainsi, pour tout k tel que $T_k \geq T$, pour tout $t \in I_k$,

$$L(t) \leq \frac{1}{\inf_{I_k} \eta_t} \frac{\varepsilon}{M_p (T_{k+1})^2}.$$

Par conséquent,

$$L(T_k) \leq \frac{1}{\inf_{I_k} \eta_t} \frac{\varepsilon}{M_p (T_{k+1})^2}.$$

D'après la Définition 6.22, $T_{k+1} = T_k + L(T_k)$ et ainsi,

$$\begin{aligned} \sum_{I_k} \eta_t &\leq L(T_k) \sup_{I_k} \eta_t \\ &\leq \frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t} \frac{\varepsilon}{M_p (T_{k+1})^2}. \end{aligned}$$

Or, d'après le Corollaire 6.28, $\sup_{I_k} \eta_t / \inf_{I_k} \eta_t$ tend vers 1, quand k tend vers l'infini. Ainsi, pour tout k suffisamment grand,

$$L(T_k) \leq \frac{2\varepsilon}{M_p (T_{k+1})^2}.$$

Nous avons ainsi établi la négligeabilité de la somme des pas sur les intervalles I_k devant les quantités $1/M_p (T_{k+1})^2$, ce qui est le résultat annoncé. □

Rappelons la définition de la suite $\boldsymbol{m}$, introduite à la Définition 5.9.

Définition 6.32 (Suite de pas renormalisée)**.** *Pour toute suite de pas $\boldsymbol{\eta} = (\eta_t)$, nous notons $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_t)$ la suite définie par, pour tout $t \geq 0$,*

$$\tilde{\eta}_t = m_t \eta_t.$$

Remarque 6.33. *Pour tout $t \geq 0$, m_t est fini d'après le Corollaire 5.8, donc $\tilde{\eta}_t$ est fini. De plus, pour tout $t \geq 0$, m_t est supérieure à 1 par construction et ainsi, pour tout $t \geq 0$,*

$$0 \leq \eta_t \leq \tilde{\eta}_t.$$

Corollaire 6.34 (Échelle de temps et pas de descente renormalisé)**.** *La suite de pas $\boldsymbol{\eta} = (\eta_t)$ vérifie les propriétés suivantes. Les quantités*

$$\sup_{I_k} \tilde{\eta}_t, \quad \left(\sum_{I_k} \tilde{\eta}_t \right)^2 \quad et \quad \sum_{I_k} \tilde{\eta}_t^2$$

sont négligeables devant

$$\sum_{I_k} \eta_t,$$

lorque k tend vers l'infini.

Démonstration. D'après le Lemme 6.23, T_k tend vers l'infini avec k. D'après le Corollaire 6.31,

$$\sum_{I_k} \eta_t = o\left(\frac{1}{M_p\left(T_{k+1}\right)^2} \right),$$

quand k tend vers l'infini. D'après le Corollaire 6.28,

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t}$$

tend vers 1, quand k tend vers l'infini, et est en particulier borné. D'après le Lemme 6.23, $T_{k+1} \sim T_k$, quand k tend vers l'infini. Or, M_p est une fonction d'échelle, et préserve donc les équivalents en l'infini. Ainsi,

$$M_p\left(T_{k+1}\right) \sim M_p\left(T_k\right),$$

quand k tend vers l'infini. Or, d'après le Lemme 6.21, $M_p(t)$ est négligeable devant $L(t)$, lorsque t tend vers l'infini. Ainsi,

$$M_p\left(T_k\right) = o\left(L\left(T_k\right)\right),$$

quand k tend vers l'infini. Enfin, d'après la Définition 6.22, $T_{k+1} = T_k + L\left(T_k\right)$. Ainsi, d'après le Lemme 6.16,

$$\sup_{I_k} M_p(t)\, \eta_t, \quad \left(\sum_{I_k} M_p(t)\, \eta_t \right)^2 \quad \text{et} \quad \sum_{I_k} M_p(t)^2\, \eta_t^2 \tag{6.5}$$

sont négligeables devant

$$\sum_{I_k} \eta_t,$$

lorque k tend vers l'infini. Or, d'après le Corollaire 6.26, $m_t = O\left(M_p(t)\right)$, quand t tend vers l'infini. Ainsi, les quantités de l'Équation (6.5) dominent leurs analogues où $M_p(t)\, \eta_t$ est remplacé par $m_t\, \eta_t = \tilde{\eta}_t$. Nous obtenons alors le résultat annoncé. □

Corollaire 6.35 (Convergence vers 0 des sommes des pas renormalisés sur les intervalles)**.**

$$\sum_{I_k} \tilde{\eta}_t$$

tend vers 0*, quand* k *tend vers l'infini.*

Démonstration. D'après le Corollaire 6.26, m_t est dominé par $M_p(t)$, lorsque t tend vers l'infini. Ainsi,

$$\sum_{I_k} \tilde{\eta}_t = O\left(\sum_{I_k} M_p(t)\ \eta_t\right), \tag{6.6}$$

quand k tend vers l'infini. Or, M_p est croissante en tant que fonction d'échelle. Par conséquent, pour tout $k \geq 0$,

$$\sum_{I_k} M_p(t)\ \eta_t \leq M_p(T_{k+1}) \sum_{I_k} \eta_t.$$

Donc, d'après le Corollaire 6.31,

$$\sum_{I_k} M_p(t)\ \eta_t = o\left(\frac{1}{M_p(T_{k+1})}\right),$$

quand k tend vers l'infini. Or, d'après la Définition 6.22, pour tout $k \geq 0$, $T_k \geq 1$. Ainsi, M_p étant croissante, pour tout $k \geq 0$, $M_p(T_{k+1}) \geq M_p(1)$. Par conséquent,

$$\sum_{I_k} M_p(t)\ \eta_t = o(1),$$

quand k tend vers l'infini. Donc, d'après l'Équation (6.6),

$$\sum_{I_k} \tilde{\eta}_t = o(1),$$

quand k tend vers l'infini. □

Remarque 6.36. *La somme des pas sur les intervalles* I_k *est inférieure à la somme des pas renormalisés, et converge donc elle aussi vers* 0*, lorsque* k *tend vers l'infini.*

6.5.6 Trajectoire optimale

Lemme 6.37 (Trajectoire optimale)**.** *Les trajectoires* (e_t^*) *et* (J_t^*) *satisfont*

$$\frac{\partial p_t}{\partial e}(e_t^*, \theta^*) \cdot J_t^* + \frac{\partial p_t}{\partial \theta}(e_t^*, \theta^*) = \frac{\partial p_t \circ g_t}{\partial \theta}(e_0^*, \theta^*).$$

Démonstration. Ainsi qu'il est dit au début de la section 4.3.1, $J_0^* = 0$, donc le résultat est une conséquence du Corollaire 5.14. □

Remarque 6.38. *Les propriétés établies au Fait 6.29 et au Lemme 6.30 font de* θ^* *un minimum local de la famille de pertes sur le paramètre.*

Définition 6.39 (Paramètre optimal et trajectoire optimale)**.** *Le paramètre* θ^* *est dorénavant dit paramètre optimal, et la trajectoire associée est dite trajectoire optimale.*

6.6 Discussion des hypothèses d'optimalité

6.6.1 Condition d'optimalité sur la somme des gradients

Lemme 6.40 (Cas de variables aléatoires indépendantes et identiquement distribuées)**.** *Dans le cas où les termes*

$$\frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right)$$

sont des variables aléatoires réelles, indépendantes et identiquement distribuées, centrées et de variance finie, la loi du logarithme itéré montre que, avec probabilité 1*, l'équation*

$$\sum_{s=0}^{t} \frac{\partial\, p_s \circ g_s}{\partial \theta}\left(e_0^*,\, \theta^*\right) = O\left(t^{a_1}\right)$$

est vérifiée pour tout $a_1 > 1/2$*.*

Démonstration. Soient des variables aléatoires réelles X_t, indépendantes et identiquement distribuées, centrées et de variance finie. Alors, d'après la loi du logarithme itéré, avec probabilité 1,

$$\frac{1}{t}\sum_{s=0}^{t} X_s = O\left(\frac{\left(\log\log t\right)^{1/2}}{t^{1/2}}\right),$$

quand t tend vers l'infini. Ainsi, avec probabilité 1,

$$\frac{1}{t^{a_1}}\sum_{s=0}^{t} X_s = O\left(\frac{\left(\log\log t\right)^{1/2}}{t^{a_1-1/2}}\right),$$

quand t tend vers l'infini. Or, $a_1 - 1/2 > 0$, ce qui conclut la preuve. □

6.6.2 Cas de la régression linéaire avec bruit gaussien

Définition 6.41 (Définition du modèle de régression linéaire avec bruit gaussien)**.** *Soit un système dynamique paramétré dont l'état au bout de* t *itérations avec le paramètre* θ *est noté*

$$e_t\left(\theta\right).$$

Soit (ξ_t) *une famille de variables aléatoires gaussiennes centrées réduites, indépendantes et identiquement distribuées. Soient les exemples d'entraînement* $(x_t,\, y_t)$ *tels que, pour tout* $t \geq 0$*,*

$$y_t = e_t\left(\theta^*\right) + \xi_t.$$

Soient les pertes sur les états données par, pour $t \geq 0$*,*

$$p_t\left(e\right) = \frac{1}{2}\left\|y_t - e\right\|^2.$$

Les pertes sur le paramètre obtenues avec le système dynamique sont alors les

$$p_t \circ e_t\left(\theta\right) = \frac{1}{2}\left\|y_t - e_t\left(\theta\right)\right\|^2.$$

Lemme 6.42 (Contrôle du modèle)**.** *Les propriétés suivantes sont vérifiées.*

1. *Presque sûrement, pour tout* $t \geq 0$

$$\frac{\partial\, p_t}{\partial e}\left(e_t\left(\theta^*\right)\right) = O\left(\sqrt{\log t}\right).$$

2. *Pour tout* $T \geq 0$*, avec grande probabilité,*

$$\sum_{t \leq T} \frac{\partial\, p_t \circ e_t}{\partial \theta}\left(\theta^*\right) = O\left(\sqrt{T}\right).$$

3. *La majoration précédente est en particulier vérifiée si les* x_t *sont des variables aléatoires indépendantes et identiquement distribuées, de variance finie, indépendantes des* ξ_t *et, pour tout* $t \geq 0$*,*

$$e_t\left(\theta\right) = \theta \cdot x_t.$$

Ainsi, les hypothèses formulées précédemment sont vérifiées pour ce modèle.

Démonstration. Pour tout $t \geq 0$,

$$\frac{\partial\, p_t}{\partial e}\left(e_t\left(\theta^*\right)\right) = \xi_t,$$

et le premier point est alors une conséquence de la loi du supremum de variables gaussiennes. Pour établir les points deux et trois, nous devons montrer que la variance des sommes est dominée par T. Or, pour tout $t \geq 0$,

$$\frac{\partial\, p_t \circ e_t}{\partial \theta}\left(\theta^*\right) = -\xi_t \cdot \frac{\partial\, e_t}{\partial \theta}\left(\theta^*\right).$$

Ainsi, la variance par rapport aux ξ_t de la somme des dérivées précédentes est égale à

$$\begin{aligned}
\mathrm{Var}_\xi\left(\sum_{t \leq T} \frac{\partial\, p_t \circ e_t}{\partial \theta}\left(\theta^*\right)\right) &= \sum_{t \leq T} \mathrm{Var}_\xi\left(\xi_t \cdot \frac{\partial\, e_t}{\partial \theta}\left(\theta^*\right)\right) \\
&= \sum_{t \leq T} \left\| \frac{\partial\, e_t}{\partial \theta}\left(\theta^*\right) \right\|^2.
\end{aligned}$$

Dans le cas du troisième point, la variance totale de la somme est égale à

$$\begin{aligned}
\mathrm{Var}_{\xi,\, x}\left(\sum_{t \leq T} \frac{\partial\, p_t \circ e_t}{\partial \theta}\left(\theta^*\right)\right) &= \sum_{t \leq T} \mathrm{Var}_x\left(\frac{\partial\, e_t}{\partial \theta}\left(\theta^*\right)\right) \\
&= \sum_{t \leq T} \mathrm{Var}_x\left(x_t\right) \\
&= T\, \mathrm{Var}\left(x\right).
\end{aligned}$$

Ceci conclut la preuve. □

6.6.3 Exemples numériques satisfaisant les hypothèses

Lemme 6.43 (Jeu de paramètres satisfaisant les hypothèses pour l'optimalité)**.** *Soient des réels positifs* a_1*,* a_2*,* γ*,* b *et* a'*. Alors, les fonctions* ℓ^1*,* ℓ^2*,* M_p*, la suite* (η_t) *et la fonction* L *définies par*

$$\begin{gathered}
\ell^1\left(t\right) = t^{a_1}, \quad \ell^2\left(t\right) = t^{a_2}, \quad M_p\left(t\right) = t^{\gamma}, \\
\eta_t = t^{-b} \quad et \quad L\left(t\right) = t^{a'}
\end{gathered}$$

satisfont les contraintes spécifiques aux fonctions (les critères d'optimalité 1/2 et 2/2 restent bien sûr en hypothèse) de l'Hypothèse 6.17, de l'Hypothèse 6.18 et du Lemme 6.21 si les conditions suivantes sont vérifiées :

1. $0 \leq a_1, a_2 < 1$;
2. $3\gamma < 1$;
3. $\max(a_1, a_2, \gamma) + 2\gamma < b \leq 1$
4. *et* $\max(a_1, a_2, \gamma) < a' < b - 2\gamma$.

Démonstration. Les contraintes sur les exposants sont compatibles entre-elles.

Nous vérifions d'abord le troisième point de l'Hypothèse 6.17 (ainsi qu'il a été dit ci-dessus, les deux premiers points ne sont pas l'objet de ce lemme).Pour $t \geq 0$,

$$\max\left(\ell^1(t), \ell^1(t), M_p(t)\right) M_p(t)^2 = t^{\max(a_1, a_2, \gamma)+2\gamma} = o(t),$$

quand t tend vers l'infini, d'après les points deux et trois ci-dessus.

Vérification de l'Hypothèse 6.18 La série de terme général η_t diverge, car $b \leq 1$, d'après le troisième point.

Pour $t \geq 1$,

$$\max\left(\ell^1(t), \ell^1(t), M_p(t)\right) M_p(t)^2 \eta_t = t^{\max(a_1, a_2, \gamma)+2\gamma-b} = o(1),$$

quand t tend vers l'infini, d'après le même point.

Soit une suite d'intervalles $I_t = [g_t, d_t[$ tels que g_t tend vers l'infini, et $g_t \sim d_t$, quand t tend vers l'infini. Alors,

$$\frac{\sup_{I_t} \eta_t}{\inf_{I_t} \eta_t} = \left(\frac{g_t}{d_t - 1}\right)^b \sim \left(\frac{g_t}{d_t}\right)^b \sim 1,$$

quand t tend vers l'infini. Enfin, pour tout $t \geq 1$,

$$\frac{1}{\eta_t M_p(t)^2} = t^{b-2\gamma},$$

et d'après le troisième point $b - 2\gamma > 0$, donc $\eta_t^{-1} M_p(t)^{-2}$ est bien une fonction d'échelle.

Vérification des contraintes du Lemme 6.21 D'après les troisième et quatrième points ci-dessus, et le fait que $\gamma \geq 0$, nous avons $a' < 1$. Ainsi, $L(t)$ est négligeable devant t, quand t tend vers l'infini. ℓ^1, ℓ^2 et M_p sont négligeables devant L d'après le quatrième point. Enfin, pour $t \geq 1$,

$$\frac{1}{\eta_t M_p(t)^2} = t^{b-2\gamma}$$

et ainsi, d'après le quatrième point,

$$L(t) = o\left(\frac{1}{\eta_t M_p(t)^2}\right),$$

quand t tend vers l'infini. □

Remarque 6.44. *Avec les choix de fonctions ci-dessus, il se peut par exemple que* $t + \ell^1(t) = t + t^{a_1}$ *ne soit pas entier. Nous conservons la notation, en supposant qu'elle désigne la partie entière de* $t + t^{a_1}$*. Dès que* $t \geq 1$*, cela ne pose pas de problème, car la partie entière de ce nombre est strictement supérieure à* t*. La même convention est adoptée à chaque occurence du même problème.*

Remarque 6.45. *Si les pertes sont bornées, nous pouvons choisir* $\gamma = 0$*.*

Chapitre

7 Mise à jour du paramètre

7.1 Opérateur de mise à jour du paramètre

Afin d'optimiser le paramètre, nous souhaitons lui appliquer des mises à jour de la forme

$$\text{paramètre} \leftarrow \text{paramètre} - \text{pas} \times \text{vecteur tangent}.$$

Pour cela, nous supposons d'abord disposer d'un opérateur Φ qui, à un paramètre θ et à un vecteur tangent v, associe un nouveau paramètre $\Phi(\theta, v)$ puis nous considérons l'opérateur qui, aux mêmes quantités et à un pas $\eta \geq 0$, associe le paramètre

$$\phi(\theta, v, \eta) = \Phi(\theta, -\eta v).$$

L'exemple typique est bien sûr

$$\Phi(\theta, v) = \theta + v$$

et

$$\phi(\theta, v, \eta) = \theta - \eta v.$$

Il est aussi possible de prendre pour Φ l'exponentielle sur une variété.

7.1.1 Opérateur de déplacement sur Θ

Hypothèse 7.1 (Opérateur de déplacement sur Θ)**.** *Nous supposons que l'espace Θ est muni d'un opérateur*

$$\begin{aligned}\Phi : \quad & \Theta \times T\Theta \to \Theta \\ & \theta, v \mapsto \Phi(\theta, v)\end{aligned}$$

tel que

1. $$\Phi(\theta, 0) = \theta;$$
2. *Φ est deux fois continûment différentiable au voisinage du point $(\theta^*, 0)$;*
3. *en tout point $(\theta, 0)$ de ce voisinage, la différentielle de Φ est*
$$\frac{\partial \Phi}{\partial v}(\theta, 0) = \mathrm{Id}_{T\Theta}.$$

Lemme 7.2 (Différentiabilité de Φ sur une boule pour la topologie produit)**.** *Quitte à diminuer B^*_Θ, nous pouvons trouver une boule contenant 0 dans $T\Theta$, que nous notons $B^0_{T\Theta}$, telle que Φ est deux fois continûment différentiable sur $B^*_\Theta \times B^0_{T\Theta}$.*

Démonstration. $\Theta \times T\Theta$ est un espace vectoriel de dimension finie sur $\mathbb{R}$, donc l'énoncé est une conséquence de l'équivalence des normes. □

7.1.2 Opérateur de mise à jour du paramètre

Définition 7.3 (Opérateur de mise à jour du paramètre)**.** *Nous définissons alors l'opérateur de mise à jour du paramètre*

$$\begin{aligned}\phi : \quad & \Theta \times T\Theta \times \mathbb{R}_+ \to \Theta \\ & \theta,\, v,\, \eta \mapsto \phi\,(\theta,\, v,\, \eta)\end{aligned}$$

tel que

$$\phi\,(\theta,\, v,\, \eta) = \Phi\,(\theta,\, -\eta\, v)\,.$$

Définition 7.4 (Compact inclus dans l'ouvert des vecteurs tangents et des pas dont le produit est dans $B^0_{T\Theta}$)**.** *Notons*

$$\Omega_{T\Theta\times\mathbb{R}_+} = \left\{(v,\, \eta) \in T\Theta \times \mathbb{R}_+ \,\middle|\, \eta\, v \in B^0_{T\Theta}\right\}$$

qui est un ouvert de $T\Theta \times \mathbb{R}_+$. Nous fixons un compact quelconque

$$\mathcal{K}_{T\Theta\times\mathbb{R}_+} \subset \Omega_{T\Theta\times\mathbb{R}_+}$$

tel que, pour tout $(v,\, \eta) \in \mathcal{K}_{T\Theta\times\mathbb{R}_+}$, pour tout $u \in [0, 1]$, $(v,\, u\,\eta) \in \mathcal{K}_{T\Theta\times\mathbb{R}_+}$.

Remarque 7.5. *De tels compacts existent. $B^0_{T\Theta} \times [0, 1]$ est l'un d'eux. Mais nous utiliserons les résultats ci-dessous avec un autre compact, défini plus loin.*

7.2 Contrôle d'une application de ϕ

7.2.1 Héritage des propriétés de Φ

Lemme 7.6 (Propriétés de l'opérateur de mise à jour du paramètre)**.** *L'opérateur ϕ de mise à jour du paramètre vérifie les propriétés suivantes.*

1. *ϕ est deux fois continûment différentiable sur $B^*_\Theta \times \Omega_{T\Theta\times\mathbb{R}_+}$.*
2. *Pour tous $\theta \in \Theta$, $v \in T\Theta$ et $\eta \geq 0$,*

$$\phi\,(\theta,\, v,\, 0) = \phi\,(\theta,\, 0,\, \eta) = \phi\,(\theta,\, 0,\, 0) = \theta.$$

3. *Pour tous $\theta \in B^*_\Theta$ et $v \in T\Theta$,*

$$\frac{\partial\,\phi}{\partial\eta}\,(\theta,\, v,\, 0) = -v.$$

Démonstration. L'application

$$\theta,\, v,\, \eta \in B^*_\Theta \times \Omega_{T\Theta\times\mathbb{R}_+} \mapsto (\theta,\, -\eta\, v)$$

est régulière, et à valeurs dans $B^*_\Theta \times B^0_{T\Theta}$. Or, l'opérateur Φ est deux fois continûment différentiable sur cet ensemble, d'après l'Hypothèse 7.1, et la première propriété est ainsi vérifiée.

La deuxième propriété vient de la valeur de l'application ci-dessus aux points considérés et de la relation $\Phi\,(\theta,\, 0) = \theta$.

Enfin, soient $\theta \in B^*_\Theta$ et $v \in T\Theta$. Nous savons que Φ est différentiable en $(\theta,\, 0)$, de différentielle

$$\frac{\partial\,\Phi}{\partial v}\,(\theta,\, 0) = \mathrm{Id}_{T\Theta}\,.$$

En conséquence, par composition, l'application

$$\eta \mapsto \phi(\theta, v, \eta) = \Phi(\theta, -\eta v)$$

et dérivable en 0, et admet pour dérivée

$$\frac{\partial \phi}{\partial \eta}(\theta, v, 0) = \frac{\partial \Phi}{\partial v}(\theta, 0) \cdot (-v) = -v.$$

□

7.2.2 Contrôle à l'ordre 2 d'une mise à jour

Définition 7.7 (Termes en $O(\eta^2)$ et $O\left(\sum_{s\leq t} \eta_s^2\right)$)**.** *Dans la suite, nous utiliserons les symboles $O(\eta^2)$ et $O\left(\sum_{s\leq t} \eta_s^2\right)$. Ces termes désigneront des fonctions définies sur*

$$B_\Theta^* \times \mathcal{K}_{T\Theta\times\mathbb{R}_+},$$

et dominées par les quantités η^2 et $\sum_{s\leq t} \eta_s^2$. De plus, les constantes qui interviendront dans les dominations ne dépendront que de B_Θ^* et $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$.

Lemme 7.8 (Développement à l'ordre 2 de ϕ au voisinage des points $(\theta, v, 0)$)**.** *Pour tout $\theta \in B_\Theta^*$, pour tout vecteur tangent v et tout pas η tels que $(v, \eta) \in \mathcal{K}_{T\Theta\times\mathbb{R}_+}$,*

$$\phi(\theta, v, \eta) = \theta - \eta v + O\left(\eta^2\right).$$

Démonstration. Pour tout $0 \leq u \leq 1$,

$$v, u\eta \in \Omega_{T\Theta\times\mathbb{R}_+}$$

donc, par composition,

$$u \in [0,1] \mapsto \phi(\theta, v, u\eta)$$

est deux fois continûment différentiable et ainsi, d'après la formule de Taylor avec reste intégral,

$$\phi(\theta, v, \eta) = \theta + \eta \frac{\partial \phi}{\partial \eta}(\theta, v, 0) + \eta^2 \int_0^1 u \frac{\partial^2 \phi}{\partial \eta^2}(\theta, v, u\eta)\, du.$$

Or, d'après la Définition 7.4, tout $0 \leq u \leq 1$,

$$v, u\eta \in \mathcal{K}_{T\Theta\times\mathbb{R}_+}.$$

D'après le Lemme 7.6, la différentielle seconde de ϕ est continue sur le compact

$$B_\Theta^* \times \mathcal{K}_{T\Theta\times\mathbb{R}_+} \subset B_\Theta^* \times \Omega_{T\Theta\times\mathbb{R}_+}.$$

Ainsi, l'application

$$\theta, v, \eta \mapsto \int_0^1 u \frac{\partial^2 \phi}{\partial \eta^2}(\theta, v, u\eta)\, du$$

est bornée sur ce compact, ce qui conclut la preuve. □

Lemme 7.9 (Contrôle d'une mise à jour du paramètre pour des paramètres initiaux et des vecteurs différents)**.** *Soient deux paramètres θ et θ' dans B_Θ^*, deux vecteurs tangents v et v' et un pas η tels que (v, η) et (v', η) sont dans $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$. Alors,*

$$d(\phi(\theta, v, \eta), \phi(\theta', v', \eta)) \leq d(\theta, \theta') + \eta\, d(v, v') + O\left(\eta^2\right).$$

Démonstration. D'après le Lemme 7.8,

$$\phi(\theta, v, \eta) - \phi(\theta', v', \eta) = \theta - \theta' - \eta\,(v - v') + O\left(\eta^2\right),$$

donc

$$\begin{aligned} d(\phi(\theta, v, \eta), \phi(\theta', v', \eta)) &= \left\|\theta - \theta' + \eta\,(v - v') + O\left(\eta^2\right)\right\| \\ &\leq \|\theta - \theta'\| + \eta\,\|v - v'\| + O\left(\eta^2\right), \end{aligned}$$

ce qui conclut la preuve. □

Lemme 7.10 (Sous-linéarité des mises à jour sur $B^*_\Theta \times \mathcal{K}_{T\Theta\times\mathbb{R}_+}$)**.** *Nous pouvons trouver $M_4 > 0$ ne dépendant que de $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$ tel que, pour tous θ et θ' appartenant à B^*_Θ, pour tous vecteurs tangents v et v' et tout η tels que (v, η) et (v', η) sont dans $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$,*

$$d(\phi(\theta, v, \eta), \phi(\theta', v', \eta)) \leq d(\theta, \theta') + \eta\, M_4.$$

Démonstration. D'après le Lemme 7.9,

$$d(\phi(\theta, v, \eta), \phi(\theta', v', \eta)) \leq d(\theta, \theta') + \eta\, d(v, v') + O\left(\eta^2\right).$$

Or, (v, η) et (v', η) sont dans $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$ donc, en notant κ son diamètre,

$$d(v, v') \leq \kappa.$$

D'après la Définition 7.7, quitte à augmenter κ, nous avons alors

$$\begin{aligned} d(\phi(\theta, v, \eta), \phi(\theta', v', \eta)) &\leq d(\theta, \theta') + \kappa\,\eta + \kappa\,\eta^2 \\ &\leq d(\theta, \theta') + \kappa\,\eta\,(1 + \eta) \\ &\leq d(\theta, \theta') + \kappa\,(1 + \kappa)\,\eta, \end{aligned}$$

car $\eta \leq \kappa$. Nous posons alors $M_4 = \kappa\,(1 + \kappa)$, ce qui conclut la preuve. □

7.3 Contrôle à l'ordre 2 des itérations de ϕ

Définition 7.11 (Itérations des mises à jour sur le paramètre)**.** *Soit la suite $(\theta_t)_{t\geq 0}$ définie par la donnée de θ, d'une suite de vecteurs tangents $\boldsymbol{v} = (v_s)$, d'une suite de pas $\boldsymbol{\eta} = (\eta_s)$ et la relation de récurrence, pour $t \geq 0$,*

$$\theta_{t+1} = \phi(\theta_t, v_t, \eta_t).$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \phi^t : \quad \Theta \times T\Theta^\infty \times \mathbb{R}_+^\infty &\to \Theta \\ \theta, \boldsymbol{v}, \boldsymbol{\eta} &\mapsto \phi^t(\theta, \boldsymbol{v}, \boldsymbol{\eta}) = \theta_t, \end{aligned}$$

qui associent au paramètre initial, à la suite de vecteurs tangents et à la suite de pas le t-ème paramètre de contrôle.

Lemme 7.12 (Maintien dans un compact)**.** *Soit $\theta \in B^*_\Theta$, tel que de plus*

$$d(\theta, \theta^*) \leq \frac{r^*_\Theta}{3}.$$

Soient une suite de vecteurs tangents $\boldsymbol{v} = (v_s)$, et une suite de pas de descente $\boldsymbol{\eta} = (\eta_s)$, telles que pour tout $s \geq 0$, (v_s, η_s) appartienne à $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$. Notons $\boldsymbol{\theta} = (\theta_s)$ la suite définie par : pour tout $t \geq 0$,

$$\theta_t = \phi^t(\theta, \boldsymbol{v}, \boldsymbol{\eta}).$$

Supposons

$$M_4 \sum_{t\geq 0} \eta_t \leq \frac{r^*_\Theta}{3}.$$

Alors, pour tout $t \geq 0$,

$$d(\theta_t, \theta^*) \leq \frac{2\, r^*_\Theta}{3}.$$

En particulier, pour tout $t \geq 0$,

$$\theta_t \in B^*_\Theta.$$

Démonstration. Considérons tout d'abord la suite (x_s) définie par $x_0 = \frac{r^*_\Theta}{3}$ et la relation de récurrence, pour $t \geq 0$,

$$x_{t+1} = x_t + M_4\, \eta_t.$$

Alors, pour $t \geq 1$,

$$x_t = \frac{r^*_\Theta}{3} + M_4 \sum_{s\leq t-1} \eta_s.$$

En particulier, pour $t \geq 0$,

$$x_t \leq \frac{2\, r^*_\Theta}{3}.$$

Notons à présent, pour $t \geq 0$,

$$d_t = d(\theta_t, \theta^*).$$

Prouvons alors par récurrence sur $t \geq 0$ que

$$d_t \leq x_t.$$

La propriété est vraie pour $t = 0$. Supposons-la alors vérifiée pour un certain $t \geq 0$. Nous savons que,

$$\theta_{t+1} = \phi(\theta_t, v_t, \eta_t).$$

Or, par hypothèse de récurrence,

$$d_t \leq x_t \leq \frac{2\, r^*_\Theta}{3}$$

donc $\theta_t \in B^*_\Theta$. De plus, $(v_t, \eta_t) \in \mathcal{K}_{T\Theta\times\mathbb{R}_+}$ donc, d'après le Lemme 7.10,

$$\begin{aligned} d(\theta_{t+1}, \theta^*) &= d(\phi(\theta_t, v_t, \eta_t), \theta^*) \\ &= d(\phi(\theta_t, v_t, \eta_t), \phi(\theta^*, 0, 0)) \\ &\leq d(\theta_t, \theta^*) + M_4\, \eta_t \\ &\leq d_t + M_4\, \eta_t. \end{aligned}$$

Donc,

$$d_{t+1} \leq d_t + M_4\, \eta_t.$$

Or $d_t \leq x_t$, donc

$$d_t + M_4\,\eta_t \leq x_t + M_4\,\eta_t = x_{t+1}$$

et la propriété est vraie au rang $t+1$. Ceci conclut la récurrence.

Enfin, pour tout $t \geq 0$,

$$d\left(\theta_t,\,\theta^*\right) \leq d_t \leq \frac{2\,r^*_\Theta}{3},$$

et $\boldsymbol{\theta}$ est ainsi bien incluse dans B^*_Θ. □

Lemme 7.13 (Contrôle des itérées de ϕ)**.** *Soit $\theta \in B^*_\Theta$, tel que de plus*

$$d\left(\theta,\,\theta^*\right) \leq \frac{r^*_\Theta}{3}.$$

Soient une suite de vecteurs tangents $\boldsymbol{v} = (v_s)$, et une suite de pas de descente $\boldsymbol{\eta} = (\eta_s)$, telles que pour tout $s \geq 0$, $(v_s,\,\eta_s)$ appartienne à $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$. Notons $\boldsymbol{\theta} = (\theta_s)$ la suite définie par : pour tout $t \geq 0$,

$$\theta_t = \phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right).$$

Supposons

$$M_4 \sum_{t\geq 0} \eta_t \leq \frac{r^*_\Theta}{3}.$$

Alors, pour tout $t \geq 1$,

$$\phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right) = \theta - \sum_{s\leq t-1} \eta_s\,v_s + O\left(\sum_{s\leq t-1} \eta_s^2\right).$$

Ainsi qu'il a été dit à la Définition 7.7, la constante dans le terme en grand O ne dépend que de B^*_Θ et $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$.

Démonstration. Nous procédons par récurrence. L'initialisation est acquise grâce au Lemme 7.8. Supposons alors la propriété vraie pour un certain $t \geq 1$. Nous savons que

$$\theta_{t+1} = \phi\left(\phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right),\,v_t,\,\eta_t\right).$$

Or, d'après le Lemme 7.8, pour tout triplet $\theta,\,v,\,\eta \in B^*_\Theta \times \mathcal{K}_{T\Theta\times\mathbb{R}_+}$,

$$\phi\left(\theta,\,v,\,\eta\right) = \theta - \eta\,v + O\left(\eta^2\right),$$

où la constante du grand O ne dépend que de B^*_Θ et $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$. En particulier, pour le triplet $\phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right),\,v_t,\,\eta_t$,

$$\phi\left(\phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right),\,v_t,\,\eta_t\right) = \phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right) - \eta_t\,v_t + O\left(\eta_t^2\right).$$

Grâce à l'hypothèse de récurrence, nous avons alors :

$$\begin{aligned}\phi\left(\phi^t\left(\theta,\,\boldsymbol{v},\,\boldsymbol{\eta}\right),\,v_t,\,\eta_t\right) &= \theta - \sum_{s\leq t-1} \eta_s\,v_s + O\left(\sum_{s\leq t-1} \eta_s^2\right) - \eta_t\,v_t + O\left(\eta_t^2\right)\\ &= \theta - \sum_{s\leq t} \eta_s\,v_s + O\left(\sum_{s\leq t} \eta_s^2\right).\end{aligned}$$

Toujours grâce à l'hypothèse de récurrence, conjuguée à la remarque précédente sur le terme $O\left(\eta^2\right)$, la constante du terme en grand O ne dépend que des boules B^*_Θ et $\mathcal{K}_{T\Theta\times\mathbb{R}_+}$. Ceci conclut la preuve. □

Chapitre

8 Vecteurs tangents utilisés par l'algorithme RTRL

8.1 Vecteurs tangents utilisant les transitions non bruitées sur les différentielles

8.1.1 Boule contenant les vecteurs tangents

Définition 8.1 (Boule contenant les vecteurs tangents)**.** *Notons $B^*_{T\Theta} \subset T\Theta$, la boule de l'espace tangent à Θ de rayon*

$$\left(r^*_{\mathcal{L}(\Theta,\mathcal{E})}{}^2 + 1\right)^{1/2}.$$

Lemme 8.2 (Vecteurs tangents bornés sur les boules stables)**.** *Notons, pour tout $t \geq 0$, pour tout $e \in B^*_{\mathcal{E}_t}$, pour tout $\theta \in B^*_\Theta$ et pour tout $J \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_t)}$,*

$$v_t = \frac{\partial\, p_t}{\partial e}\,(e,\,\theta)\cdot J + \frac{\partial\, p_t}{\partial \theta}\,(e,\,\theta)\,.$$

Alors,

$$\frac{1}{m_t}\, v_t \in B^*_{T\Theta}.$$

Démonstration. Soient $t \geq 0$, $e \in B^*_{\mathcal{E}_t}$ et $J \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_t)}$. Nous avons

$$v_t = \mathrm{d}\, p_t\,(e,\,\theta)\cdot(J,\,\mathrm{Id}_{T\Theta})\,.$$

Ainsi,

$$\begin{aligned}\|v_t\| &= \|\mathrm{d}\, p_t\,(e,\,\theta)\cdot(J,\,\mathrm{Id}_{T\Theta})\| \\ &\leq \|\mathrm{d}\, p_t\,(e,\,\theta)\|_{\mathrm{op}}\,\|(J,\,\mathrm{Id}_{T\Theta})\| \\ &\leq m_t\,\left(r^*_{\mathcal{L}(\Theta,\mathcal{E})}{}^2 + 1\right)^{1/2},\end{aligned}$$

d'après les Définitions 4.51 et 5.9. Or, d'après la Définition 5.9, pour tout $t \geq 0$, $m_t \geq 1 > 0$. Ainsi, pour tout $t \geq 0$, pour tout $e \in B^*_{\mathcal{E}_t}$, pour tout $\theta \in B^*_\Theta$ et pour tout $J \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_t)}$,

$$\frac{1}{m_t}\,\|v_t\| \leq \left(r^*_{\mathcal{L}(\Theta,\mathcal{E})}{}^2 + 1\right)^{1/2}.$$

Par conséquent, d'après la Définition 8.1,

$$\frac{1}{m_t}\, v_t \in B^*_{T\Theta},$$

ce qui est le résultat annoncé. □

Lemme 8.3 (Vecteurs tangents bornés le long des trajectoires)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$. Soit une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ incluse dans B^*_Θ. Notons, pour $t \geq 0$,*

$$v_t = \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right) + \frac{\partial p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right).$$

Alors, pour tout $t \geq 0$, les vecteurs v_t/m_t appartiennent à $B^*_{T\Theta}$.

Démonstration. D'après le Corollaire 4.53, pour tout $t \geq 0$,

$$\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right) \in B^*_{\mathcal{E}_t}.$$

D'après le Corollaire 4.55, pour tout $t \geq 0$,

$$\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right) \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)}.$$

Le résultat est alors une conséquence du Lemme 8.2 ci-dessus. □

8.1.2 Distances entre vecteurs tangents

Lemme 8.4 (Homogénéité en θ des écarts entre vecteurs tangents le long de trajectoires avec mêmes états et différentielles initiaux)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$. Soient deux suites de paramètres $\boldsymbol{\theta} = (\theta_t)$ et $\boldsymbol{\theta'} = (\theta'_t)$ incluses dans B^*_Θ. Notons, pour $t \geq 0$,*

$$v_t = \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right) + \frac{\partial p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right)$$

et

$$v'_t = \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta'}\right), \theta'_t\right) \cdot \boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right) + \frac{\partial p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta'}\right), \theta'_t\right).$$

Alors, nous pouvons trouver $M_5 > 0$ tel que, pour tout $t \geq 0$,

$$\frac{1}{m_t} d\left(v_t, v'_t\right) \leq M_5 \sup_{s \leq t} d\left(\theta_s, \theta'_s\right).$$

Démonstration. Pour $t \geq 0$,

$$v_t = \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right), \mathrm{Id}_{T\Theta}\right)$$

et

$$v'_t = \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta'}\right), \theta'_t\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right), \mathrm{Id}_{T\Theta}\right).$$

Majoration de la distance entre les vecteurs tangents Ainsi,

$$\begin{aligned}
v_t - v'_t &= \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right), \mathrm{Id}_{T\Theta}\right) \\
&\quad - \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta'}\right), \theta'_t\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right), \mathrm{Id}_{T\Theta}\right) \\
&= \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right) - \boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right), 0\right) \\
&\quad + \left(\mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) - \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta'}\right), \theta'_t\right)\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right), \mathrm{Id}_{T\Theta}\right) \\
&= \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right) - \boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right)\right) \\
&\quad + \left(\mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) - \mathrm{d}\, p_t\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta'}\right), \theta'_t\right)\right) \cdot \left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta'}\right), \mathrm{Id}_{T\Theta}\right).
\end{aligned}$$

Donc

$$\begin{aligned}\|v_t - v'_t\| \leq & \left\| \frac{\partial\, p_t}{\partial e} \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) \right\| \left\| \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta} \right) - \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right) \right\|_{\text{op}} \\ & + \left\| \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) - \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right),\, \theta'_t \right) \right\|_{\text{op}} \left\| \left(\boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right),\, \mathrm{Id}_{T\Theta} \right) \right\|_{\text{op}}.\end{aligned}$$

D'après la Définition 5.9, pour tout $t \geq 0$, $m_t \geq 1$. Par conséquent, pour tout $t \geq 0$,

$$\begin{aligned}\frac{1}{m_t} \left\| v_t - v'_t \right\| \leq & \frac{1}{m_t} \left\| \frac{\partial\, p_t}{\partial e} \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) \right\| \left\| \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta} \right) - \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right) \right\|_{\text{op}} \\ & + \left\| \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) - \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right),\, \theta'_t \right) \right\|_{\text{op}} \left\| \left(\boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right),\, \mathrm{Id}_{T\Theta} \right) \right\|_{\text{op}}.\end{aligned}$$

Fixons à présent un $t \geq 1$.

Majoration des quantités en facteur des différences D'après le Corollaire 4.53,

$$\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right) \quad \text{et} \quad \boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right)$$

appartiennent à la boule $B^*_{\mathcal{E}_t} \subset B_{\mathcal{E}_t}$. Ainsi, d'après le Corollaire 5.11,

$$\frac{1}{m_t} \left\| \frac{\partial\, p_t}{\partial e} \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) \right\|_{\text{op}} \leq 1.$$

D'après le Corollaire 4.55,

$$\left\| \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right) \right\|_{\text{op}} \leq r^*_{\mathcal{L}(\Theta,\, \mathcal{E})},$$

et ainsi

$$\left\| \left(\boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right),\, \mathrm{Id}_{T\Theta} \right) \right\|_{\text{op}} \leq \left({r^*_{\mathcal{L}(\Theta,\, \mathcal{E})}}^2 + 1 \right)^{1/2}.$$

Homogénéité des différences D'après le Lemme 5.6,

$$\left\| \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) - \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right),\, \theta'_t \right) \right\|$$

est majoré par

$$S_{\text{perte}} \left\| \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right) - \boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right),\, \theta_t - \theta'_t \right) \right\|.$$

Or, d'après le Lemme 4.57,

$$d \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right) \right) \leq \frac{M_1}{\alpha} \sup_{s \leq t-1} d \left(\theta_s,\, \theta'_s \right).$$

Notons alors

$$\kappa = \max \left(1,\, \frac{M_1}{\alpha} \right).$$

Nous avons par conséquent

$$\begin{aligned}\left\| \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right) - \boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right),\, \theta_t - \theta'_t \right) \right\| &= \left(\left\| \boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right) - \boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right) \right\|^2 + \left\| \theta_t - \theta'_t \right\| \right)^{1/2} \\ &\leq \left(\left(\frac{M_1}{\alpha} \right)^2 \left(\sup_{s \leq t-1} d \left(\theta_s,\, \theta'_s \right) \right)^2 + d \left(\theta_t,\, \theta'_t \right)^2 \right)^{1/2} \\ &\leq \sqrt{2}\, \kappa \sup_{s \leq t} d \left(\theta_s,\, \theta'_s \right).\end{aligned}$$

Ainsi,

$$\left\| \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta} \right),\, \theta_t \right) - \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \boldsymbol{\theta'} \right),\, \theta'_t \right) \right\| \leq S_{\text{perte}}\, \sqrt{2}\, \kappa \sup_{s \leq t} d\left(\theta_s,\, \theta'_s \right).$$

Enfin, d'après le Lemme 4.58,

$$\left\| \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta} \right) - \boldsymbol{J}_t \left(e_0,\, J_0,\, \boldsymbol{\theta'} \right) \right\|_{\text{op}} \leq \frac{M_2}{\alpha} \sqrt{1 + \left(\frac{M_1}{\alpha} \right)^2} \sup_{s \leq t-1} d\left(\theta_s,\, \theta'_s \right).$$

Conclusion Nous avons ainsi bien établi le résultat souhaité, pour $t \geq 1$. Celui-ci est également vrai en $t = 0$, car

$$d\left(v_0,\, v'_0 \right) = \left\| \frac{\partial\, p_t}{\partial \theta} \left(e_0,\, \theta_0 \right) - \frac{\partial\, p_t}{\partial \theta} \left(e'_0,\, \theta'_0 \right) \right\| \leq S_{\text{perte}}\, \left\| \theta_0 - \theta'_0 \right\|,$$

de sorte que

$$\frac{1}{m_0}\, d\left(v_0,\, v'_0 \right) \leq \frac{S_{\text{perte}}}{m_0}\, \left\| \theta_0 - \theta'_0 \right\|.$$

Ceci conclut la preuve. □

Lemme 8.5 (Écarts exponentiellement faibles entre vecteurs tangents le long de trajectoires avec le même paramètre)**.** *Soient* $e_0,\, e'_0 \in B^*_{\mathcal{E}_0}$, *et* $J_0,\, J'_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$. *Soit un paramètre* $\theta \in B^*_\Theta$. *Notons, pour* $t \geq 0$,

$$v_t = \frac{\partial\, p_t}{\partial e} \left(\boldsymbol{e}_t \left(e_0,\, \theta \right),\, \theta \right) \cdot \boldsymbol{J}_t \left(e_0,\, J_0,\, \theta \right) + \frac{\partial\, p_t}{\partial \theta} \left(\boldsymbol{e}_t \left(e_0,\, \theta \right),\, \theta \right)$$

et

$$v'_t = \frac{\partial\, p_t}{\partial e} \left(\boldsymbol{e}_t \left(e'_0,\, \theta \right),\, \theta \right) \cdot \boldsymbol{J}_t \left(e'_0,\, J'_0,\, \theta \right) + \frac{\partial\, p_t}{\partial \theta} \left(\boldsymbol{e}_t \left(e'_0,\, \theta \right),\, \theta \right).$$

Alors, nous pouvons trouver $M_6 > 0$ *tel que, pour tout* $t \geq 0$,

$$\frac{1}{m_t}\, d\left(v_t,\, v'_t \right) \leq M_6 \left(1 - \frac{\alpha}{2} \right)^t \left(d\left(e_0,\, e'_0 \right) + d\left(J_0,\, J'_0 \right) \right).$$

Démonstration. De même que dans la preuve du Lemme 8.4 précédent, pour tout $t \geq 0$,

$$\begin{aligned} \frac{1}{m_t}\, \left\| v_t - v'_t \right\| \leq{}& \left\| \boldsymbol{J}_t \left(e_0,\, J_0,\, \theta \right) - \boldsymbol{J}_t \left(e'_0,\, J'_0,\, \theta' \right) \right\|_{\text{op}} \\ &+ \left\| \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \theta \right),\, \theta \right) - \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e'_0,\, \theta \right),\, \theta \right) \right\|_{\text{op}} \left({r^*_{\mathcal{L}(\Theta, \mathcal{E})}}^2 + 1 \right)^{1/2}. \end{aligned}$$

Toujours de même que précédemment, d'après le Lemme 5.6, pour tout $t \geq 0$,

$$\left\| \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e_0,\, \theta \right),\, \theta \right) - \mathrm{d}\, p_t \left(\boldsymbol{e}_t \left(e'_0,\, \theta \right),\, \theta \right) \right\|_{\text{op}} \leq S_{\text{perte}}\, \left\| \boldsymbol{e}_t \left(e_0,\, \theta \right) - \boldsymbol{e}_t \left(e'_0,\, \theta \right) \right\|.$$

Or, d'après le Lemme 4.59, pour tout $t \geq 0$,

$$\left\| \boldsymbol{e}_t \left(e_0,\, \theta \right) - \boldsymbol{e}_t \left(e_0,\, \theta \right) \right\| \leq (1 - \alpha)^t\, d\left(e_0,\, e'_0 \right).$$

D'après le Lemme 4.60, pour tout $t \geq 0$,

$$d\left(\boldsymbol{J}_t \left(e_0,\, J_0,\, \theta \right),\, \boldsymbol{J}_t \left(e'_0,\, J'_0,\, \theta \right) \right) \leq (1 - \alpha)^t\, d\left(J_0,\, J'_0 \right) + M_2\, t\, (1 - \alpha)^{t-1}\, d\left(e_0,\, e'_0 \right).$$

Or,

$$\frac{t\,(1-\alpha)^{t-1}}{\left(1-\frac{\alpha}{2}\right)^{t-1}}$$

est borné donc, nous pouvons trouver $\kappa \geq 0$ tel que, pour $t \geq 0$,

$$d\left(\boldsymbol{J}_t\left(e_0,\, J_0,\, \theta\right),\, \boldsymbol{J}_t\left(e_0',\, J_0',\, \theta\right)\right) \leq \kappa \left(1-\frac{\alpha}{2}\right)^t \left(d\left(J_0,\, J_0'\right) + d\left(e_0,\, e_0'\right)\right).$$

Ainsi, nous pouvons bien trouver $M_6 > 0$ tel que, pour $t \geq 0$,

$$\frac{1}{m_t}\, d\left(v_t,\, v_t'\right) \leq M_6 \left(1-\frac{\alpha}{2}\right)^t \left(d\left(e_0,\, e_0'\right) + d\left(J_0,\, J_0'\right)\right).$$

□

8.2 Vecteurs tangents utilisant les transitions bruitées sur les différentielles

8.2.1 Étude des transitions bruitées sur les différentielles

Hypothèse 8.6 (Terme d'erreur sur les différentielles)**.** *Nous supposons disposer, pour $t \geq 0$, d'une famille (ξ_t) à valeurs dans les espaces $\mathcal{L}\left(\Theta,\, \mathcal{E}_t\right)$. Nous supposons de plus que cette famille est bornée en norme d'opérateur, par un réel M_ξ.*

Lemme 8.7 (Bruit total borné sur les boules de contrôle $B_{\mathcal{E}_t} \times B_\Theta^*$)**.** *Soit une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ à valeurs dans B_Θ^*, et une suite d'états $\boldsymbol{e} = (e_t)$ à valeurs dans les boules $B_{\mathcal{E}_t}$. Soit la suite de différentielles (J_t) définie par la donnée d'un $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$ et la relation de récurrence, pour $t \geq 0$,*

$$J_{t+1} = \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot J_t + \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right) + \xi_t.$$

Alors, nous pouvons trouver un réel $r_\xi \geq 0$, ne dépendant que de $r^*_{\mathcal{L}(\Theta,\mathcal{E})}$, α, M_1 et M_ξ tel que, pour tout $t \geq 0$,

$$\left\|J_t\right\|_{\mathrm{op}} \leq r_\xi.$$

Démonstration. Pour tout $t \geq 0$,

$$\left\|J_{t+1}\right\|_{\mathrm{op}} \leq \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right)\right\|_{\mathrm{op}} \left\|J_t\right\|_{\mathrm{op}} + \left\|\frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right)\right\|_{\mathrm{op}} + \left\|\xi_t\right\|_{\mathrm{op}}.$$

Or, d'après les Lemmes 4.39 et 4.40, pour tout $(e,\, \theta) \in B_{\mathcal{E}_t} \times B_\Theta^*$,

$$\left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right)\right\|_{\mathrm{op}} \leq 1-\alpha \quad \text{et} \quad \left\|\frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right)\right\|_{\mathrm{op}} \leq M_1.$$

De plus, d'après l'Hypothèse 8.6, pour tout $t \geq 0$, $\left\|\xi_t\right\|_{\mathrm{op}} \leq M_\xi$. Ainsi, pour tout $t \geq 0$,

$$\left\|J_{t+1}\right\|_{\mathrm{op}} \leq (1-\alpha)\, \left\|J_t\right\|_{\mathrm{op}} + M_1 + M_\xi,$$

ce qui permet d'obtenir le résultat annoncé. □

Lemme 8.8 (Vecteurs tangents utilisant des différentielles bruitées bornés sur les boules stables)**.** *Notons, pour tout $t \geq 0$, pour tout $e \in B^*_{\mathcal{E}_t}$, pour tout $\theta \in B^*_\Theta$ et pour tout $J \in \mathcal{L}(\Theta, \mathcal{E}_t)$ tel que $\|J\|_{\text{op}} \leq r_\xi$,*

$$v_t = \frac{\partial p_t}{\partial e}(e, \theta) \cdot J + \frac{\partial p_t}{\partial \theta}(e, \theta).$$

Alors,

$$\frac{1}{m_t} \|v_t\| \leq \left(r_\xi^2 + 1\right)^{1/2}.$$

Démonstration. La preuve est identique à celle du Lemme 8.2. □

Corollaire 8.9 (Agrandissement du rayon de la boule $B^*_{T\Theta}$)**.** *Nous pouvons remplacer, sans modifier les résultats précédents, le rayon de la boule $B^*_{T\Theta}$ par*

$$\max\left(\left({r^*_{\mathcal{L}(\Theta,\mathcal{E})}}^2 + 1\right)^{1/2}, \left(r_\xi^2 + 1\right)^{1/2}\right).$$

Démonstration. D'après le Lemme 8.7, le rayon r_ξ ne dépend que de $r^*_{\mathcal{L}(\Theta,\mathcal{E})}$, α, M_1 et M_ξ. Or, la deuxième et la troisième quantité sont définies avec la définition de la boule $B^*_{T\Theta}$, et M_ξ est définie indépendamment de celle-ci. Nous avons ainsi bien établi le résultat annoncé. □

Corollaire 8.10 (Vecteurs tangents bruités renormalisés dans $B^*_{T\Theta}$)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$, et $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$. Soit une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ incluse dans B^*_Θ. Soit une suite de différentielles (J_t) vérifiant $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$ et la relation de récurrence, pour $t \geq 0$,*

$$J_{t+1} = \frac{\partial \boldsymbol{T}_t}{\partial e}(\boldsymbol{e}_t(e_0, \boldsymbol{\theta}), \theta_t) \cdot J_t + \frac{\partial \boldsymbol{T}_t}{\partial \theta}(\boldsymbol{e}_t(e_0, \boldsymbol{\theta}), \theta_t) + \xi_t.$$

Posons alors, pour tout $t \geq 0$,

$$v_t = \frac{\partial p_t}{\partial e}(\boldsymbol{e}_t(e_0, \boldsymbol{\theta}), \theta_t) \cdot J_t + \frac{\partial p_t}{\partial \theta}(\boldsymbol{e}_t(e_0, \boldsymbol{\theta}), \theta_t).$$

Alors, pour tout $t \geq 0$, nous avons

$$\frac{1}{m_t} v_t \in B^*_{T\Theta}.$$

Démonstration. D'après le Corollaire 4.53, pour tout $t \geq 0$,

$$\boldsymbol{e}_t(e_0, \boldsymbol{\theta}) \in B^*_{\mathcal{E}_t}.$$

D'après le Lemme 8.7, pour tout $t \geq 0$,

$$\|J_t\|_{\text{op}} \leq r_\xi.$$

Donc, d'après le Lemme 8.8, pour tout $t \geq 0$,

$$\frac{1}{m_t} v_t \leq \left(r_\xi^2 + 1\right).$$

Or, d'après le Corollaire 8.9, cette quantité est inférieure au rayon de la boule $B^*_{T\Theta}$. Le résultat annoncé est ainsi bien vérifié. □

Corollaire 8.11 (Expression des différentielles bruitées en fonction des différentielles non bruitées)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$, et $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$. Soit une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ incluse dans B^*_Θ. Soit une suite de différentielles (J_t) vérifiant $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$ et la relation de récurrence, pour $t \geq 0$,*

$$J_{t+1} = \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot J_t + \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) + \xi_t.$$

Alors, pour tout $t \geq 0$,

$$J_t = \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) + \sum_{s \leq t-1} \left(\prod_{p=s-1}^{t-1} \frac{\partial \boldsymbol{T}_p}{\partial e}\left(\boldsymbol{e}_p\left(e_0,\, \theta_0\right),\, \theta_0\right) \right) \xi_s + O\left(\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right) \right),$$

où la constante du terme en grand O ne dépend que des constantes apparaissant dans les hypothèses de régularité et de la borne sur le bruit.

Démonstration. Pour tout $t \geq 0$,

$$J_{t+1} - \boldsymbol{J}_{t+1}\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) = \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right)\, \left(J_t - \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right)\right) + \xi_t.$$

Or, par hypothèse, $\boldsymbol{\theta}$ est incluse dans B^*_Θ, et $\theta_0 \in B^*_\Theta$. Ainsi, d'après le Lemme 4.57, pour tout $t \geq 0$,

$$d\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{e}_t\left(e_0,\, \theta_0\right)\right) \leq \frac{M_1}{\alpha} \sup_{s \leq t-1} d\left(\theta_s,\, \theta_0\right).$$

De plus, d'après l'Hypothèse 4.30, les différentielles secondes des opérateurs de transition sont bornées sur les boules stables. Par conséquent, pour tout $t \geq 0$,

$$\frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) = \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right) + O\left(\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right) \right).$$

Ainsi, pour tout $t \geq 0$,

$$\begin{aligned} J_{t+1} - \boldsymbol{J}_{t+1}\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) = {} & \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right)\, \left(J_t - \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right)\right) \\ & + O\left(\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right)\, \left(J_t - \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right)\right) \right) + \xi_t. \end{aligned}$$

Or, d'après le Lemme 8.7 et le Corollaire 4.55, pour tout $t \geq 0$,

$$\left\| J_t \right\|_{\mathrm{op}} \leq r_\xi, \quad \left\| \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) \right\|_{\mathrm{op}} \leq r^*_{\mathcal{L}(\Theta,\mathcal{E})}$$

et, d'après le Corollaire 4.53 et le Lemme 4.39, pour tout $t \geq 0$,

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right) \right\|_{\mathrm{op}} \leq 1 - \alpha.$$

Par conséquent, pour tout $t \geq 0$,

$$\begin{aligned} J_t = {} & \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) + \sum_{s \leq t-1} \left(\prod_{p=s-1}^{t-1} \frac{\partial \boldsymbol{T}_p}{\partial e}\left(\boldsymbol{e}_p\left(e_0,\, \theta_0\right),\, \theta_0\right) \right) \xi_s \\ & + O\left(\sum_{s \leq t-1} \left(1-\alpha\right)^{t-1-s} \sup_{q \leq s} d\left(\theta_q,\, \theta_0\right) \right). \end{aligned}$$

Or, le dernier terme est dominé par

$$\sup_{s \leq t-1} d\left(\theta_s,\, \theta_0\right),$$

ce qui conclut la preuve. □

8.2.2 Vecteurs tangents utilisant les transitions bruitées sur les différentielles

Définition 8.12 (Écart entre les vecteurs tangents bruités et ceux non bruités calculés avec les mêmes paramètres)**.** *Notons $\boldsymbol{\theta}$ la suite de paramètres, et (J_t) la suite de différentielles, produites par l'algorithme RTRL bruité. Nous notons alors, pour tout $t \geq 0$,*

$$\begin{aligned} R_t &= \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot J_t + \frac{\partial p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \\ &\quad - \left(\frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) + \frac{\partial p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right)\right) \\ &= \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot \left(J_t - \boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right)\right). \end{aligned}$$

Remarque 8.13. *Pour tout $t \geq 0$, R_t est l'écart entre les vecteurs tangents utilisés par l'algorithme RTRL (bruité), et les vecteurs tangents calculés avec la suite de paramètres produite par l'algorithme RTRL bruité, mais pour la fonction de transition sur les différentielles non bruitée (le terme J_t est remplacé par $\boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right)$).*

Définition 8.14 (Produit des différentielles des opérateurs de transition par rapport aux états)**.** *Soit un temps $t_0 \geq 0$. Soient $\theta_{t_0} \in B^*_{\Theta}$, et $e_{t_0} \in B^*_{\mathcal{E}_{t_0}}$ .Notons, pour tout $t \geq t_0$, $e_t^{t_0}$ l'état obtenu après $t - t_0$ itérations de l'équation*

$$e_{t+1}^{t_0} = \boldsymbol{T}_t\left(e_t^{t_0},\, \theta_{t_0}\right),$$

pour une valeur initiale $e_{t_0}^{t_0} = e_{t_0}$. Nous notons alors, pour $t \geq t_0$,

$$\Pi_{s,\,t}^{t_0} = \frac{1}{m_t}\frac{\partial p_t}{\partial e}\left(e_t^{t_0},\, \theta_{t_0}\right) \cdot \left(\prod_{p=s}^{t-1} \frac{\partial \boldsymbol{T}_p}{\partial e}\left(e_p^{t_0},\, \theta_{t_0}\right)\right).$$

Corollaire 8.15 (Expression des vecteurs tangents bruités en fonction des quantités initiales)**.** *Soient $e_0 \in B^*_{\mathcal{E}_0}$, et $J_0 \in B^*_{\mathcal{L}(\Theta,\,\mathcal{E}_0)}$. Soit une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ incluse dans B^*_{Θ}. Soit une suite de différentielles (J_t) vérifiant $J_0 \in B^*_{\mathcal{L}(\Theta,\,\mathcal{E}_0)}$ et la relation de récurrence, pour $t \geq 0$,*

$$J_{t+1} = \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot J_t + \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) + \xi_t.$$

Alors, pour tout $t \geq 0$,

$$R_t = \sum_{s \leq t-1} m_t\, \Pi_{s,\,t}^0\, \xi_s + O\left(\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right)\right).$$

Démonstration. De même qu'au début de la preuve du Corollaire 8.11, mais en remplaçant l'Hypothèse 4.30 par l'Hypothèse 5.4, qui porte sur les différentielles secondes des pertes, pour tout $t \geq 0$, nous avons

$$\frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) = \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right) + O\left(\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right)\right).$$

Le Corollaire 8.11 nous permet alors d'obtenir le résultat annoncé. □

Hypothèse 8.16 (Négligeabilité du bruit sur les intervalles)**.** *Pour tout $k \geq 0$, pour tout $t \in I_k$, et tout $t \leq s \leq T_{k+1} - 1$, nous considérons le terme*

$$\Pi_{t,\,s}^{T_k},$$

calculé avec le paramètre θ_{T_k} et l'état e_{T_k} qui sont ceux de l'algorithme RTRL à l'instant T_k.

Nous supposons alors que le terme de bruit qui intervient dans la définition de $b_{T_k,\,T_{k+1}}(\boldsymbol{\eta})$, introduit à la Définition 9.32, est négligeable devant la somme des pas sur l'intervalle, c'est-à-dire

$$\sum_{I_k} \left(\sum_{s=t}^{T_{k+1}-1} \tilde{\eta}_s \, \Pi_{t,\,s}^{T_k} \right) \xi_t = o\left(\sum_{I_k} \eta_t \right),$$

quand k tend vers l'infini.

Remarque 8.17. *L'hypothèse peut sembler forte, mais il faudra en fait montrer qu'elle est bien vérifiée pour chaque algorithme particulier de la forme « RTRL bruité ». En particulier, nous montrons que c'est le bien le cas lors de la preuve de convergence de l'algorithme « NoBackTrack ».*

Chapitre

9 Propriété centrale de l'algorithme RTRL

9.1 Définition de l'algorithme RTRL (bruité)

Définition 9.1 (Algorithme RTRL (bruité))**.** *Soit la suite de paramètres* $\boldsymbol{\theta} = (\theta_s)$ *définie par la donnée de* $e_0 \in \mathcal{E}_0$, $\theta_0 \in \Theta$, $J_0 \in \mathcal{L}(\Theta, \mathcal{E}_0)$ *et les relations de récurrence, pour* $t \geq 0$,

$$\begin{cases} \theta_{t+1} = \phi\left(\theta_t, \dfrac{\partial p_t}{\partial e}(e_t, \theta_t) \cdot J_t + \dfrac{\partial p_t}{\partial \theta}(e_t, \theta_t), \eta_t\right) \\ e_{t+1} = \boldsymbol{T}_t(e_t, \theta_t) \\ J_{t+1} = \dfrac{\partial \boldsymbol{T}_t}{\partial e}(e_t, \theta_t) \cdot J_t + \dfrac{\partial \boldsymbol{T}_t}{\partial \theta}(e_t, \theta_t) + \xi_t, \end{cases}$$

où (η_t) *est une suite de pas de descente.*

Lemme 9.2 (Description de RTRL avec des fonctions)**.** *Notons* $\boldsymbol{\theta} = (\theta_t)$ *la suite de paramètres obtenue avec l'algorithme RTRL. Alors, pour* $t \geq 0$,

$$e_t = \boldsymbol{e}_t(e_0, \boldsymbol{\theta}).$$

9.2 Contrôles sur les pas de descente, pas renormalisés, termes en grand O

Définition 9.3 (Borne sur les pas)**.** *Nous définissons*

$$\eta_{\max} = \sup\left\{\eta \geq 0 \,\middle|\, \textit{pour tout } v \in B^*_{T\Theta},\ \eta\, v \in B^0_{T\Theta}\right\}.$$

Remarque 9.4. *Pour tout* $0 \leq \eta \leq \eta_{\max}$, *pour tout* $v \in B^*_{T\Theta}$, *nous avons*

$$\eta\, v \in B^0_{T\Theta}.$$

Corollaire 9.5 (Borne sur les pas strictement positive)**.** $\eta_{\max} > 0$.

Démonstration. C'est une conséquence du fait que la boule $B^*_{T\Theta}$, définie à la Définition 8.1 et modifiée au Corollaire 8.9, est bornée. □

Rappelons que la suite de pas renormalisée a été introduite à la Définition 6.32. Dans la suite de ce chapitre, nous fixons $\theta_0 \in B^*_\Theta$, tel que de plus

$$d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{3},$$

$e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$. Nous supposons de plus que la suite $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$. Ainsi, pour tout $t \geq 0$,

$$0 \leq m_t\, \eta_t = \tilde{\eta}_t \leq \eta_{\max}.$$

Nous appliquerons les résultats sur le contrôle de l'opérateur de mise à jour du paramètre ϕ, au compact

$$\mathcal{K}_{T\Theta\times\mathbb{R}_+} = B^*_{T\Theta} \times [0, \eta_{\max}].$$

En particulier, les bornes des termes en grand O dûs au contrôle de l'opérateur de mise à jour ne dépendront que de B^*_Θ, $B^*_{T\Theta}$ et $\eta_{\max}$. Or, ceux-ci sont bornés dès que les Hypothèses 4.28, 4.29 et 4.30 sur le système dynamique, les Hypothèses 5.3 et 5.4 sur les pertes, l'Hypothèse 7.1 sur l'opérateur de déplacement sur Θ et l'Hypothèse 8.6 sur le bruit sont satisfaites.

De plus, B^*_Θ, $B^*_{T\Theta}$ et $\eta_{\max}$ sont quantitatifs en les constantes apparaissant dans les hypothèses de régularité et dans la majoration du bruit.

Enfin, d'après la fin de l'énoncé du Corollaire 8.11, les termes en grand O dûs au contrôle des différentielles bruitées ne dépendent que des constantes des hypothèses de régularité et de la borne sur le bruit.

9.3 Horizon de maintien dans B^*_Θ

9.3.1 Définition de l'horizon de maintien dans B^*_Θ

Définition 9.6 (Horizon de maintien dans B^*_Θ)**.** *Soit une suite $\boldsymbol{\eta} = (\eta_s)$ de pas de descente. Conformément à la Définition 6.32, la suite renormalisée est notée $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_t)$. Soit un temps $t_0 \geq 0$. Nous appelons*

$$T_{t_0}^{r^*_\Theta} = \inf\left\{ t \geq t_0 + 1 \;\middle|\; M_4 \sum_{s=t_0}^{t-1} \tilde{\eta}_s > \frac{r^*_\Theta}{3} \right\}.$$

Si l'ensemble est vide, nous posons ainsi qu'il est d'usage $T_{t_0}^{r^*_\Theta} = \infty$.

9.3.2 Maintien dans B^*_Θ

Lemme 9.7 (Suite majorant $\sup_{s\leq t} d(\theta_s, \theta^*)$)**.** *Considérons la suite (x_t) définie par $x_0 = \frac{r^*_\Theta}{3}$ et la relation de récurrence, pour $0 \leq t < T_0^{r^*_\Theta}$,*

$$x_{t+1} = x_t + M_4\, \tilde{\eta}_t.$$

Alors, pour tout $0 \leq t < T_0^{r^*_\Theta}$,

$$x_t \leq \frac{2\, r^*}{3}.$$

Démonstration. Pour tout $t \geq 0$,

$$x_t = x_0 + M_4 \sum_{s=0}^{t-1} \tilde{\eta}_s.$$

Or, pour tout $1 \leq t < T_0^{r^*_\Theta}$,

$$M_4 \sum_{s=0}^{t-1} \tilde{\eta}_s \leq \frac{r^*_\Theta}{3}.$$

Par conséquent, pour tout $1 \leq t < T_0^{r_\Theta^*}$,

$$x_t \leq \frac{r_\Theta^*}{3} + \frac{r_\Theta^*}{3} = \frac{2\, r_\Theta^*}{3}.$$

Ceci conclut la preuve. □

Fait 9.8 (Stabilité de l'algorithme RTRL en temps fini)**.** *Notons $\boldsymbol{\theta} = (\theta_t)$ la suite produite par l'algorithme RTRL. Alors, pour tout $0 \leq t < T_0^{r_\Theta^*}$, $\theta_t \in B_\Theta^*$.*

Preuve du Fait 9.8. Notons (J_t) la suite de différentielles produite par l'algorithme RTRL bruité. Notons, pour $t \geq 0$,

$$S_t = \sup_{s \leq t} d\left(\theta_s,\, \theta^*\right).$$

Prouvons alors par récurrence sur $0 \leq t < T_0^{r_\Theta^*}$ que

$$S_t \leq x_t,$$

où la suite (x_t) est celle du Lemme 9.7 précédent. Ceci implique en particulier que, pour tout $t < T_0^{r_\Theta^*}$,

$$\theta_0,\, \theta_1,\, \ldots,\, \theta_t \in B_\Theta^*.$$

La propriété est vraie pour $t = 0$. Supposons la alors vérifiée pour un certain $0 \leq t < T_0^{r_\Theta^*}$. Notons

$$v_t = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot J_t + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right).$$

Alors, d'après la Définition 7.3,

$$\theta_{t+1} = \phi\left(\theta_t,\, v_t,\, \eta_t\right) = \phi\left(\theta_t,\, \frac{v_t}{m_t},\, m_t\, \eta_t\right) = \phi\left(\theta_t,\, \frac{v_t}{m_t},\, \tilde{\eta}_t\right).$$

Or, par hypothèse de récurrence,

$$S_t \leq x_t \leq \frac{2\, r^*}{3}$$

donc, pour $s \leq t$, $\theta_s \in B_\Theta^*$ et « la partie qui compte » (celle d'indice inférieur à t) de la suite $\boldsymbol{\theta}$ est incluse dans B_Θ^*. En particulier, d'après le Corollaire 8.10, v_t/m_t appartient à la boule $B_{T\Theta}^*$. Ainsi,

$$\left(\frac{v_t}{m_t},\, \tilde{\eta}_t\right) \in B_{T\Theta}^* \times [0, \eta_{\max}].$$

De plus, $(0,\, 0)$ appartient également à cet ensemble. Ainsi, d'après le Lemme 7.10 appliqué à $(\theta_t,\, v_t/m_t,\, \tilde{\eta}_t)$ et $(\theta^*,\, 0,\, 0)$,

$$\begin{aligned} d\left(\theta_{t+1},\, \theta^*\right) &= d\left(\phi\left(\theta_t,\, \frac{v_t}{m_t},\, \tilde{\eta}_t\right),\, \theta^*\right) \\ &\leq d\left(\theta_t,\, \theta^*\right) + M_4\, \tilde{\eta}_t \\ &\leq S_t + M_4\, \tilde{\eta}_t. \end{aligned}$$

Or S_{t+1} est le maximum de S_t et de $d(\theta_{t+1}, \theta^*)$. Donc il est inférieur au maximum de S_t et du terme de droite ci-dessus. Or ce dernier est supérieur à S_t. Donc

$$S_{t+1} \leq S_t + M_4 \, \tilde{\eta}_t.$$

Or $S_t \leq x_t$, donc

$$S_t + M_4 \, \tilde{\eta}_t \leq x_t + M_4 \, \tilde{\eta}_t$$

et la propriété est vraie au rang $t+1$. Ceci conclut la récurrence.

Enfin, pour tout $t \geq 0$,

$$d(\theta_t, \theta^*) \leq S_t \leq \frac{2\, r^*}{3},$$

et $\boldsymbol{\theta}$ est ainsi bien incluse dans B^*_Θ. □

Corollaire 9.9 (Maintien des états et des différentielles dans les boules stables)**.** *Notons de nouveau $\boldsymbol{\theta}$ la suite de paramètres produite par l'algorithme RTRL. Alors, pour tout $0 \leq t < T_0^{r^*_\Theta}$,*

$$e_t = \boldsymbol{e}_t(e_0, \boldsymbol{\theta}) \in B^*_{\mathcal{E}_t},$$

et

$$\boldsymbol{J}_t(e_0, J_0, \boldsymbol{\theta}) \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)}.$$

Remarque 9.10. *Les différentielles $\boldsymbol{J}_t(e_0, J_0, \boldsymbol{\theta})$ sont celles calculées avec la suite de paramètres produite par l'algoritme mais la transition* non bruitée*. Ce ne sont pas les différentielles produites par l'algorithme.*

Démonstration. Le résultat est implicitement prouvé dans la preuve du Fait 9.8. Il est dû à ce que, pour tout $0 \leq t < T_0^{r^*_\Theta}$, θ_t appartient à B^*_Θ et aux Corollaires 4.53 et 4.55. □

9.4 Trajectoires intermédiaires, en boucle ouverte

9.4.1 Définition des trajectoires intermédiaires

Nous appelons l'algorithme ci-dessous algorithme RTRL en boucle ouverte : la mise à jour du paramètre est calculée mais les états successifs ainsi que les différentielles successives, et par conséquent les vecteurs tangents, sont calculés en utilisant le paramètre initial.

Définition 9.11 (Algorithme RTRL en boucle ouverte)**.** *Soit la suite de paramètres (θ_t) définie par la donnée de $e_0 \in \mathcal{E}_0$, $\theta_0 = \theta \in \Theta$, $J_0 \in \mathcal{L}(\Theta, \mathcal{E}_0)$ et les relations de récurrence, pour $t \geq 0$,*

$$\begin{cases} \theta_{t+1} = \phi\left(\theta_t, \dfrac{\partial\, p_t}{\partial e}(e_t, \theta) \cdot J_t + \dfrac{\partial\, p_t}{\partial \theta}(e_t, \theta), \eta_t\right) \\ e_{t+1} = \boldsymbol{T}_t(e_t, \theta) \\ J_{t+1} = \dfrac{\partial\, \boldsymbol{T}_t}{\partial e}(e_t, \theta) \cdot J_t + \dfrac{\partial\, \boldsymbol{T}_t}{\partial \theta}(e_t, \theta), \end{cases}$$

où (η_t) est une suite de pas de descente.

Lemme 9.12 (Propriété de l'algorithme RTRL en boucle ouverte)**.** *Les quantités définies à la Définition 9.11 ci-dessus vérifient :*

$$e_t = \boldsymbol{e}_t\left(e_0,\, \theta\right) \quad \textit{et} \quad J_t = \boldsymbol{J}_t\left(e_0,\, J_0,\, \theta\right).$$

Définition 9.13 (Fonction RTRL en boucle ouverte)**.** *Nous définissons les fonctions*

$$\begin{aligned}\boldsymbol{\theta}_t^{\mathrm{o}} :\ \ & \Theta \times \mathcal{E}_0 \times \mathcal{L}\left(\Theta,\, \mathcal{E}_0\right) \times \mathbb{R}_+^\infty \to \Theta \\ & \theta, e_0,\, J_0,\, \boldsymbol{\eta} \mapsto \boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta,\, e_0,\, J_0,\, \boldsymbol{\eta}\right) = \theta_t,\end{aligned}$$

où les θ_t sont ceux de la Définition 9.11, pour $\theta_0 = \theta$.

Il n'est pas besoin de définir les fonctions coordonnées associées aux états et aux différentielles successifs, car celles-ci existent déjà, ainsi que le montre le Lemme 9.12.

Trajectoires intermédiaires entre la trajectoire RTRL et la trajectoire stable Dans la suite, nous allons considérer trois trajectoires intermédiaires entre la trajectoire produite par l'algorithme RTRL et la trajectoire fixe qui reste en θ^*. Nous donnons ci-dessous les termes généraux de ces trajectoires.

1. Trajectoire partageant l'initialisation de la trajectoire RTRL, mais produite par l'algorithme en boucle ouverte.

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta_0,\, e_0,\, J_0,\, \boldsymbol{\eta}\right).$$

2. Trajectoire partageant l'état initial et la différentielle initiale de la trajectoire stable, mais issue du même paramètre que la trajectoire RTRL.

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta_0,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right).$$

3. Trajectoire en boucle ouverte issue des quantités stables.

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right).$$

9.4.2 Stabilité des trajectoires intermédiaires

Lemme 9.14 (Stabilité des états et des différentielles des trajectoires intermédiaires)**.** *Pour tout $t \geq 0$, les états et différentielles associés aux quantités intermédiaires, c'est-à-dire pour les états*

$$\boldsymbol{e}_t\left(e_0,\, \theta_0\right), \quad \boldsymbol{e}_t\left(e_0^*,\, \theta_0\right) \quad \textit{et} \quad \boldsymbol{e}_t\left(e_0^*,\, \theta^*\right)$$

et, pour les différentielles,

$$\boldsymbol{J}_t\left(e_0,\, J_0,\, \theta_0\right), \quad \boldsymbol{J}_t\left(e_0^*,\, 0,\, \theta_0\right) \quad \textit{et} \quad \boldsymbol{J}_t\left(e_0^*,\, 0,\, \theta^*\right),$$

appartiennent aux boules stables $B^*_{\mathcal{E}_t}$ et $B^*_{\mathcal{L}(\Theta,\, \mathcal{E}_t)}$.

Démonstration. C'est une conséquence des Corollaires 4.53 et 4.55. □

Lemme 9.15 (Maintien des paramètres mis à jour dans B^*_Θ)**.** *Pour tout $0 \leq t < T_0^{r^*_\Theta}$,*

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta_0,\, e_0,\, J_0,\, \boldsymbol{\eta}\right), \quad \boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta_0,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right)$$

et

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right).$$

appartiennent à B^*_Θ.

Démonstration. Nous effectuons la démonstration pour le premier cas uniquement, comme les deux autres se traitent de manière analogue. Posons, pour tout $t \geq 0$,

$$v_t = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right) \cdot \boldsymbol{J}_t\left(e_0,\, J_0,\, \theta_0\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right).$$

Notons $\boldsymbol{v} = (v_t)$, et $\tilde{\boldsymbol{v}} = (v_t / m_t)$. Alors, pour tout $t \geq 0$,

$$\boldsymbol{\theta}^{\mathrm{o}}_t\left(\theta_0,\, e_0,\, J_0,\, \boldsymbol{\eta}\right) = \phi^t\left(\theta_0,\, \boldsymbol{v},\, \boldsymbol{\eta}\right) = \phi^t\left(\theta_0,\, \tilde{\boldsymbol{v}},\, \tilde{\boldsymbol{\eta}}\right).$$

Or, d'après le Lemme 8.3, pour tout $t \geq 0$, $v_t / m_t \in B^*_{T\Theta}$. Ainsi, pour tout $t \geq 0$, $(v_t / m_t,\, \tilde{\eta}_t) \in B^*_{T\Theta} \times [0, \eta_{\max}]$. Nous pouvons alors appliquer le Lemme 7.12 pour conclure la preuve, en remplaçant l'hypothèse

$$M_4 \sum_{t \geq 0} \tilde{\eta}_t \leq \frac{r^*_\Theta}{3}$$

par la condition satisfaite par $T_0^{r^*_\Theta}$. □

9.4.3 Écarts entre les trajectoires intermédiaires

Nous contrôlons d'abord les écarts entre la trajectoire RTRL et la trajectoire RTRL en boucle ouverte, puis entre celle-ci et la trajectoire issue des quantités initiales stables. La dernière trajectoire, issue des quantités initiales stables, et calculée avec le paramètre stable, sera utilisée dans la section 9.5 ci-dessous.

Lemme 9.16 (Contrôle de l'écart entre les vecteurs tangents bruités et ceux non bruités calculés avec les mêmes paramètres)**.** *Notons $\boldsymbol{\theta}$ la suite de paramètres, et (J_t) la suite de différentielles, produites par l'algorithme RTRL bruité. Considérons la suite (R_t) introduite à la Définition 8.12. Alors, pour tout $1 \leq t < T_0^{r^*_\Theta}$,*

$$R_t = \sum_{s \leq t-1} m_t\, \Pi^0_{s,\,t}\, \xi_s + O\left(\sum_{s \leq t-1} \tilde{\eta}_s\right).$$

Démonstration. D'après le Fait 9.8, pour tout $0 \leq t < T_0^{r^*_\Theta}$, $\theta_t \in B^*_\Theta$. Donc, d'après la Définition 9.1, les suites $(\theta_t)_{t < T_0^{r^*_\Theta}}$ et $(J_t)_{t < T_0^{r^*_\Theta}}$ vérifient les hypothèses du Corollaire 8.15. Par conséquent, pour tout $1 \leq t < T_0^{r^*_\Theta}$,

$$R_t = \sum_{s \leq t-1} m_t\, \Pi^0_{s,\,t}\, \xi_s + O\left(\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right)\right).$$

Or, en reprenant la preuve du Fait 9.8, et en remplaçant θ^* par θ_0 qui appartient également à B^*_Θ (et est à une distance initiale $0 \leq x_0 = r^*_\Theta / 3$ « de lui-même »), nous obtenons, pour tout $1 \leq t < T_0^{r^*_\Theta}$,

$$\sup_{s \leq t} d\left(\theta_s,\, \theta_0\right) \leq M_4 \sum_{s \leq t-1} \tilde{\eta}_s,$$

ce qui conclut la preuve. □

Lemme 9.17 (Contrôle de la somme des écarts entre les vecteurs tangents bruités et ceux non bruités calculés avec les mêmes paramètres)**.** *Pour tout* $1 \leq t < T_0^{r_\Theta^*}$,

$$\sum_{s \leq t-1} \eta_s \, R_s = \sum_{p \leq t-1} \left(\sum_{s=p}^{t-1} \tilde{\eta}_s \, \Pi_{p,\,s}^0 \right) \xi_p + O\left(\sum_{s \leq t-1} \eta_s \sum_{s \leq t-1} \tilde{\eta}_s \right).$$

Démonstration. D'après le Lemme 9.16, pour tout $1 \leq t < T_0^{r_\Theta^*}$,

$$\sum_{s \leq t-1} \eta_s \, R_s = \sum_{s \leq t-1} \eta_s \sum_{p \leq s-1} m_s \, \Pi_{p,\,s}^0 \, \xi_p + O\left(\sum_{s \leq t-1} \eta_s \sum_{p \leq s-1} \tilde{\eta}_p \right). \tag{9.1}$$

Or, pour tout $1 \leq t < T_0^{r_\Theta^*}$,

$$\sum_{s \leq t-1} \eta_s \sum_{p \leq s-1} m_s \, \Pi_{p,\,s}^0 \, \xi_p = \sum_{p \leq t-1} \left(\sum_{s=p}^{t-1} \eta_s \, m_s \, \Pi_{p,\,s}^0 \right) \xi_p.$$

Nous majorons de plus, dans le terme en grand O de l'Équation (9.1), les sommes jusqu'à $s-1$ des $\tilde{\eta}_p$ par la somme jusqu'à $t-1$ des $\tilde{\eta}_s$, ce qui conclut la preuve. □

Fait 9.18 (Écart entre la trajectoire RTRL bruitée et la trajectoire RTRL en boucle ouverte issue des mêmes quantités initiales)**.** *Notons* $\boldsymbol{\theta} = (\theta_t)$ *la trajectoire produite par l'algorithme RTRL. Alors, pour* $1 \leq t < T_0^{r_\Theta^*}$,

$$d\left(\theta_t,\, \boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta_0,\, e_0,\, J_0,\, \boldsymbol{\eta}\right)\right) \leq \left\| \sum_{p \leq t-1} \left(\sum_{s=p}^{t-1} \tilde{\eta}_s \, \Pi_{p,\,s}^0 \right) \xi_p \right\| + O\left(\left(\sum_{s \leq t-1} \tilde{\eta}_s \right)^2 \right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2 \right).$$

Remarque 9.19. *Ainsi qu'il a été dit à la Définition 7.7 et à la section 9.2, les constantes des termes en grand O ne dépendent que de* B_Θ^*, $B_{T\Theta}^0$ *et* $\eta_{\max}$, *des constantes des hypothèses de régularité et de la borne sur le bruit.*

Preuve du Fait 9.18. Notons (J_t) la suite de différentielles produite par l'algorithme RTRL bruité. Pour $t \geq 0$ notons, afin d'alléger l'écriture,

$$\boldsymbol{\theta}_t^{\mathrm{o}} = \boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta_0,\, e_0,\, J_0,\, \boldsymbol{\eta}\right).$$

Notons

$$v_t = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot J_t + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right)$$

les vecteurs tangents utilisés par l'algorithme RTRL bruité, et

$$v_t' = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right) \cdot \boldsymbol{J}_t\left(e_0,\, J_0,\, \theta_0\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right)$$

les vecteurs tangents utilisés par l'algorithme en boucle ouverte. Nous savons que, pour $t \geq 0$,

$$\theta_{t+1} = \phi\left(\theta_t,\, v_t,\, \eta_t\right) = \phi\left(\theta_t,\, \frac{v_t}{m_t},\, \tilde{\eta}_t\right)$$

et

$$\boldsymbol{\theta}_{t+1}^{\mathrm{o}} = \phi\left(\boldsymbol{\theta}_t^{\mathrm{o}},\, v_t',\, \eta_t\right) = \phi\left(\boldsymbol{\theta}_t^{\mathrm{o}},\, \frac{v_t'}{m_t},\, \tilde{\eta}_t\right).$$

Majoration à l'ordre 1 **en la somme des pas** Dans un premier temps, prouvons que, pour $t \geq 1$,

$$d\left(\theta_t,\, \boldsymbol{\theta}_t^{\mathrm{o}}\right) = O\left(\sum_{s \leq t-1} \tilde{\eta}_s\right). \tag{9.2}$$

D'après le Fait 9.8 et le Lemme 9.15, pour $0 \leq t < T_0^{r_\Theta^*}$, θ_t et $\boldsymbol{\theta}_t^{\mathrm{o}}$ sont dans B_Θ^*. En particulier, d'après le Lemme 8.3 et le Corollaire 8.10, pour tout $t \geq 0$, v_t/m_t et v_t'/m_t appartiennent à $B_{T\Theta}^*$. Par conséquent, pour tout $t \geq 0$,

$$\left(\frac{v_t}{m_t},\, \tilde{\eta}_t\right) \in B_{T\Theta}^* \times [0, \eta_{\max}] \quad \text{et} \quad \left(\frac{v_t'}{m_t},\, \tilde{\eta}_t\right) \in B_{T\Theta}^* \times [0, \eta_{\max}].$$

Ainsi, d'après le Lemme 7.10,

$$\begin{aligned} d\left(\theta_{t+1},\, \boldsymbol{\theta}_{t+1}^{\mathrm{o}}\right) &= d\left(\phi\left(\theta_t,\, v_t,\, \eta\right),\, \phi\left(\boldsymbol{\theta}_t^{\mathrm{o}},\, v_t',\, \eta_t\right)\right) \\ &= d\left(\phi\left(\theta_t,\, \frac{v_t}{m_t},\, \tilde{\eta}_t\right),\, \phi\left(\theta_t,\, \frac{v_t'}{m_t},\, \tilde{\eta}_t\right)\right) \\ &\leq d\left(\theta_t,\, \boldsymbol{\theta}_t^{\mathrm{o}}\right) + M_4\, \tilde{\eta}_t. \end{aligned}$$

Donc, pour tout $1 \leq t < T_0^{r_\Theta^*}$,

$$\begin{aligned} d\left(\theta_t,\, \boldsymbol{\theta}_t^{\mathrm{o}}\right) &\leq d\left(\theta_0,\, \boldsymbol{\theta}_0^{\mathrm{o}}\right) + M_4 \sum_{s \leq t-1} \tilde{\eta}_s \\ &= M_4 \sum_{s \leq t-1} \tilde{\eta}_s, \end{aligned}$$

car $\boldsymbol{\theta}_0^{\mathrm{o}} = \theta_0$. Ceci établit l'Équation (9.2).

Majoration à l'ordre 2 **en la somme des pas** Prouvons à présent que, pour $1 \leq t < T_0^{r_\Theta^*}$,

$$d\left(\theta_t,\, \boldsymbol{\theta}_t^{\mathrm{o}}\right) = O\left(\left(\sum_{s \leq t-1} \tilde{\eta}_s\right)^2\right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right) + \sum_{s \leq t-1} \eta_s\, R_s.$$

De même que précédemment, en invoquant cette fois le Lemme 7.13 nous avons, pour tout $1 \leq t < T_0^{r_\Theta^*}$,

$$\theta_t - \boldsymbol{\theta}_t^{\mathrm{o}} = \sum_{s \leq t-1} \eta_s\, \left(v_s - v_s'\right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right).$$

Or, d'après la Définition 8.12, pour tout $0 \leq s \leq t-1$,

$$v_s = \frac{\partial\, p_s}{\partial e}\left(\boldsymbol{e}_s\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_s\right) \cdot \boldsymbol{J}_s\left(e_0,\, J_0,\, \boldsymbol{\theta}\right) + \frac{\partial\, p_s}{\partial \theta}\left(\boldsymbol{e}_s\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_s\right) + R_s.$$

Par conséquent, pour tout $1 \leq t < T_0^{r_\Theta^*}$,

$$\sum_{s \leq t-1} \eta_s\, \left(v_s - v_s'\right)$$

est égal à

$$\sum_{s\leq t-1}\eta_s\left(\frac{\partial\, p_s}{\partial e}\left(\boldsymbol{e}_s\left(e_0,\,\boldsymbol{\theta}\right),\,\theta_s\right)\cdot\boldsymbol{J}_s\left(e_0,\,J_0,\,\boldsymbol{\theta}\right)+\frac{\partial\, p_s}{\partial\theta}\left(\boldsymbol{e}_s\left(e_0,\,\boldsymbol{\theta}\right),\,\theta_s\right)-v'_s\right)$$
$$+\sum_{s\leq t-1}\eta_s\,R_s.$$

Or, d'après le Lemme 8.4, pour $s\geq 1$,

$$\frac{1}{m_s}\,d\left(\frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_s\left(e_0,\,\boldsymbol{\theta}\right),\,\theta_s\right)\cdot\boldsymbol{J}_s\left(e_0,\,J_0,\,\boldsymbol{\theta}\right)+\frac{\partial\, p_t}{\partial\theta}\left(\boldsymbol{e}_s\left(e_0,\,\boldsymbol{\theta}\right),\,\theta_s\right),\,v'_s\right)$$

est inférieur à

$$M_5\,\sup_{p\leq s}d\left(\theta_p,\,\theta_0\right),$$

et d'après l'Équation (9.2), pour tout $1\leq s\leq t$, pour tout $1\leq p\leq s$,

$$d\left(\theta_p,\,\boldsymbol{\theta}^{\mathrm{o}}_p\right)=O\left(\sum_{p'\leq p-1}\tilde{\eta}_{p'}\right),$$

de sorte que

$$\frac{1}{m_s}\,d\left(\frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_s\left(e_0,\,\boldsymbol{\theta}\right),\,\theta_s\right)\cdot\boldsymbol{J}_s\left(e_0,\,J_0,\,\boldsymbol{\theta}\right)+\frac{\partial\, p_t}{\partial\theta}\left(\boldsymbol{e}_s\left(e_0,\,\boldsymbol{\theta}\right),\,\theta_s\right),\,v'_s\right)$$

est dominé par

$$\sum_{p\leq s-1}\tilde{\eta}_p.$$

Ainsi, pour tout $1\leq t<T_0^{r^*_\Theta}$, en majorant les sommes des pas jusqu'à s par des sommes jusqu'à t,

$$d\left(\theta_t,\,\boldsymbol{\theta}^{\mathrm{o}}_t\right)=O\left(\left(\sum_{s\leq t-1}\tilde{\eta}_s\right)^2\right)+O\left(\sum_{s\leq t-1}\tilde{\eta}_s^2\right)+\left\|\sum_{s\leq t-1}\eta_s\,R_s\right\|.$$

Le Lemme 9.17 permet alors d'obtenir le résultat annoncé (en utilisant le fait que, pour tout $s\geq 0$, $\eta_s\leq\tilde{\eta}_s$, afin d'absorber le terme d'erreur de ce lemme dans le carré de la somme des $\tilde{\eta}_s$). □

Fait 9.20 (Écart en temps fini entre des trajectoires RTRL boucle ouverte avec le même paramètre mais issues de quantités initiales différentes)**.** *Pour* $1\leq t<T_0^{r^*_\Theta}$,

$$d\left(\boldsymbol{\theta}^{\mathrm{o}}_t\left(\theta_0,\,e_0,\,J_0,\,\boldsymbol{\eta}\right),\,\boldsymbol{\theta}^{\mathrm{o}}_t\left(\theta_0,\,e_0^*,\,0,\,\boldsymbol{\eta}\right)\right)=O\left(\sup_{0\leq s\leq t-1}\tilde{\eta}_s\right)+O\left(\sum_{s\leq t-1}\tilde{\eta}_s^2\right).$$

Preuve du Fait 9.20. Pour $t\geq 0$ notons, afin d'alléger l'écriture,

$$\theta_t=\boldsymbol{\theta}^{\mathrm{o}}_t\left(\theta_0,\,e_0,\,J_0,\,\boldsymbol{\eta}\right)$$

et

$$\theta'_t=\boldsymbol{\theta}^{\mathrm{o}}_t\left(\theta_0,\,e_0^*,\,0,\,\boldsymbol{\eta}\right).$$

Notons également

$$v_t = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right) \cdot \boldsymbol{J}_t\left(e_0,\, J_0,\, \theta_0\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \theta_0\right),\, \theta_0\right)$$

et

$$v'_t = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta_0\right),\, \theta_0\right) \cdot \boldsymbol{J}_t\left(e_0^*,\, 0,\, \theta_0\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta_0\right),\, \theta_0\right).$$

De même que dans la preuve du Fait 9.20 précédent, d'après le Lemme 7.13, pour tout $1 \leq t < T_0^{r_\Theta^*}$, nous avons

$$\theta_t - \theta'_t = \sum_{s \leq t-1} \eta_s\, \left(v_s - v'_s\right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right),$$

et ainsi,

$$d\left(\theta_t,\, \theta'_t\right) \leq \sum_{s \leq t-1} \tilde{\eta}_s \frac{1}{m_s}\, d\left(v_s,\, v'_s\right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right).$$

Or, d'après le Lemme 8.5, pour tout $s \leq t$,

$$\frac{1}{m_s}\, d\left(v_s,\, v'_s\right) \leq M_6\, \left(1 - \frac{\alpha}{2}\right)^s\, \left(d\left(e_0,\, e_0^*\right) + d\left(J_0,\, 0\right)\right).$$

Ceci implique, pour $1 \leq t < T_0^{r_\Theta^*}$,

$$d\left(\theta_t,\, \theta'_t\right) = O\left(\sum_{s \leq t-1} \tilde{\eta}_s\, \left(1 - \frac{\alpha}{2}\right)^s\right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right).$$

Or,

$$\sum_{s \leq t-1} \tilde{\eta}_s\, \left(1 - \frac{\alpha}{2}\right)^s \leq \sup_{0 \leq s \leq t-1} \tilde{\eta}_s \sum_{s \leq t-1} \left(1 - \frac{\alpha}{2}\right)^s.$$

La somme dans la quantité majorante est bornée indépendamment de t, en tant que somme partielle d'une série convergente, donc

$$d\left(\theta_t,\, \theta'_t\right) = O\left(\sup_{0 \leq s \leq t-1} \tilde{\eta}_s\right) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right),$$

ce qui conclut la preuve. □

9.5 Contractivité sur le paramètre

9.5.1 Trajectoire intermédiaire issue des quantités initiales stables

Pour les deux trajectoires intermédiaires de termes généraux $\boldsymbol{\theta}_t^{\circ}\left(\theta_0,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right)$ et $\boldsymbol{\theta}_t^{\circ}\left(\theta^*,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right)$, comme $J_0 = 0$, le Corollaire 5.14 donne

$$\frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta_0\right),\, \theta_0\right) \cdot \boldsymbol{J}_t\left(e_0^*,\, 0,\, \theta_0\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta_0\right),\, \theta_0\right) = \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta_0\right).$$

D'après le même corollaire, comme la dernière trajectoire partage toutes les quantités de la trajectoire stable, sauf le paramètre qui est mis à jour (mais dont la mise à jour n'est pas utilisée pour les mises à jour sur les états et les différentielles),

$$\frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta^*\right),\, \theta^*\right) \cdot \boldsymbol{J}_t\left(e_0^*,\, 0,\, \theta^*\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta^*\right),\, \theta^*\right)$$

est égal à

$$\frac{\partial\, p_t}{\partial e}\left(e_t^*,\, \theta^*\right) \cdot J_t^* + \frac{\partial\, p_t}{\partial \theta}\left(e_t^*,\, \theta^*\right) = \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right).$$

Pour simplifier les notations, dans la suite, nous utiliserons $\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta,\, \boldsymbol{\eta}\right)$ au lieu de $\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta,\, e_0^*,\, 0,\, \boldsymbol{\eta}\right)$.

9.5.2 Horizon de contrôle de la plus grande valeur propre le long de la trajectoire stable

Définition 9.21 (Valeurs propres extrémales sur B_Θ^*)**.** *Soit un instant $t_0 \geq 0$. Soit $\theta \in \Theta$. Pour tout $t \geq t_0$, nous notons $\Lambda_{t_0,\, t}\left(\theta\right)$ la plus grande valeur propre de l'opérateur*

$$\frac{\sum_{s=t_0}^{t} \eta_s \frac{\partial^2\, p_s \circ g_s}{\partial \theta^2}\left(e_0^*,\, \theta\right)}{\sum_{s=t_0}^{t} \eta_s},$$

et $\lambda_{t_0,\, t}\left(\theta\right)$ sa plus petite. Nous notons alors

$$\Lambda_{t_0,\, t}^* = \sup_{\theta \in B_\Theta^*} \Lambda_{t_0,\, t}\left(\theta\right)$$

et

$$\lambda_{t_0,\, t}^* = \inf_{\theta \in B_\Theta^*} \lambda_{t_0,\, t}\left(\theta\right).$$

Définition 9.22 (Horizon de contrôle de la plus grande valeur propre)**.** *Soit une suite $\boldsymbol{\eta} = \left(\eta_s\right)$ de pas de descente. Soit un temps $t_0 \geq 0$. Nous appelons*

$$T_{t_0}^{\Lambda^*} = \inf\left\{ t \geq t_0 + 1 \,\middle|\, \Lambda_{t_0,\, t-1}^* \sum_{s=t_0}^{t-1} \tilde{\eta}_s > 1 \right\}.$$

Si l'ensemble est vide, nous posons ainsi qu'il est l'usage $T_{t_0}^{\Lambda^*} = \infty$.

Remarque 9.23. *Nous devons, pour contrôler les valeurs propres des opérateurs*

$$\frac{\sum_{s=t_0}^{t} \eta_s \frac{\partial^2\, p_s \circ g_s}{\partial \theta^2}\left(e_0^*,\, \theta\right)}{\sum_{s=t_0}^{t} \eta_s},$$

formuler des hypothèses sur la suite $\tilde{\boldsymbol{\eta}}$ et non directement la suite $\boldsymbol{\eta}$. Cela est dû à ce que, d'après le Corollaire 5.17, nous contrôlons les $\frac{1}{m_t} \frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e_0^*,\, \theta\right)$, et non directement les différentielles secondes.

Lemme 9.24 (Contrôle des valeurs propres entre t_0 et $T_{t_0}^{\Lambda^*}$)**.** *Soit une suite $\boldsymbol{\eta} = \left(\eta_s\right)$ de pas de descente. Soit un temps $t_0 \geq 0$. Soit $\theta \in B_\Theta^*$. Alors, pour $t \geq t_0$, les valeurs propres de l'opérateur*

$$\mathrm{Id}_{T\Theta} - \sum_{s=t_0}^{t} \eta_s \frac{\partial^2\, p_s \circ g_s}{\partial \theta^2}\left(e_0^*,\, \theta\right)$$

sont comprises entre

$$1 - \Lambda_{t_0,\, t}^* \sum_{s=t_0}^{t} \eta_s \quad \textit{et} \quad 1 - \lambda_{t_0,\, t}^* \sum_{s=t_0}^{t} \eta_s.$$

Démonstration. En effet, $\theta \in B^*_\Theta$ donc, d'après la Définition 9.21, pour $t \geq t_0$, les valeurs propre de l'opérateur

$$\sum_{s=t_0}^{t} \eta_s \frac{\partial^2 p_s \circ g_s}{\partial \theta^2} (e_0^*, \theta)$$

sont comprises entre

$$\lambda^*_{t_0,t} \sum_{s=t_0}^{t} \eta_s \quad \text{et} \quad \Lambda^*_{t_0,t} \sum_{s=t_0}^{t} \eta_s,$$

ce qui donne le résultat annoncé. □

Corollaire 9.25 (Valeurs propres positives entre t_0 et $T_{t_0}^{\Lambda^*}$)**.** *Soit $\theta \in B^*_\Theta$. Alors, pour $t_0 \leq t < T_{t_0}^{\Lambda^*}$, les valeurs propres de l'opérateur*

$$\mathrm{Id}_{T\Theta} - \sum_{s=t_0}^{t} \eta_s \frac{\partial^2 p_s \circ g_s}{\partial \theta^2} (e_0^*, \theta)$$

sont positives.

Démonstration. D'après la Définition 9.22, pour $t_0 \leq t < T_{t_0}^{\Lambda^*}$,

$$\Lambda^*_{t_0,t} \sum_{s=t_0}^{t} \eta_s \leq \Lambda^*_{t_0,t} \sum_{s=t_0}^{t} \tilde{\eta}_s \leq 1$$

et ainsi,

$$1 - \Lambda^*_{t_0,t} \sum_{s=t_0}^{t} \eta_s \geq 0,$$

ce qui conclut la preuve, grâce au Lemme 9.24. □

Remarque 9.26. *Il se pourrait que $\lambda^*_{t_0,t}$ soit négative, auquel cas le terme*

$$1 - \lambda^*_{t_0,t} \sum_{s=t_0}^{t} \eta_s$$

est plus grand que 1*, malgré ce que suggèrent les notations. Toutefois, dans la preuve de convergence de l'algorithme RTRL, nous rajouterons des conditions qui impliqueront que $\lambda^*_{t_0,t}$ est positive, ce qui justifie les notations utilisées.*

9.5.3 Contractivité sur le paramètre

Définition 9.27 (Infimum des deux horizons)**.** *Nous définissons $T_{t_0}^{\infty}$ le minimum de $T_{t_0}^{r^*_\Theta}$ et $T_{t_0}^{\Lambda^*}$.*

Fait 9.28 (Contractivité sur le paramètre)**.** *Pour tout $\theta \in B^*_\Theta$, pour tout $1 \leq t < T_0^{\infty}$,*

$$d\left(\boldsymbol{\theta}^{\circ}_t(\theta, \boldsymbol{\eta}), \boldsymbol{\theta}^{\circ}_t(\theta^*, \boldsymbol{\eta})\right) \leq \left(1 - \lambda^*_{0,t-1} \sum_{s \leq t-1} \eta_s\right) d(\theta, \theta^*) + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right).$$

Preuve du Fait 9.28. Soit $\theta \in B^*_\Theta$. Pour $0 \leq t < T_0^\infty$ notons, conformément à la remarque figurant à la fin de la section 9.5.1,

$$\boldsymbol{\theta}^{\mathrm{o}}_t(\theta, \boldsymbol{\eta}) = \boldsymbol{\theta}^{\mathrm{o}}_t(\theta, e^*_0, 0, \boldsymbol{\eta}),$$

et

$$\boldsymbol{\theta}^{\mathrm{o}}_t(\theta^*, \boldsymbol{\eta}) = \boldsymbol{\theta}^{\mathrm{o}}_t(\theta^*, e^*_0, 0, \boldsymbol{\eta}).$$

Notons

$$\begin{aligned} v_t &= \frac{\partial p_t}{\partial e}(\boldsymbol{e}_t(e^*_0, \theta), \theta) \cdot \boldsymbol{J}_t(e^*_0, 0, \theta) + \frac{\partial p_t}{\partial \theta}(\boldsymbol{e}_t(e^*_0, \theta), \theta) \\ &= \frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta) \end{aligned}$$

et

$$\begin{aligned} v^*_t &= \frac{\partial p_t}{\partial e}(\boldsymbol{e}_t(e^*_0, \theta^*), \theta^*) \cdot \boldsymbol{J}_t(e^*_0, 0, \theta^*) + \frac{\partial p_t}{\partial \theta}(\boldsymbol{e}_t(e^*_0, \theta^*), \theta^*) \\ &= \frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta^*). \end{aligned}$$

Notons enfin $\boldsymbol{v} = (v_t)$, $\boldsymbol{v^*} = (v^*_t)$ et $\tilde{\boldsymbol{v}} = \left(\frac{v_t}{m_t}\right)$, $\tilde{\boldsymbol{v^*}} = \left(\frac{v^*_t}{m_t}\right)$. Alors, pour tout $0 \leq t < T_0^\infty \leq T_0^{r^*_\Theta}$,

$$\boldsymbol{\theta}^{\mathrm{o}}_t(\theta, \boldsymbol{\eta}) = \phi^t(\theta, \boldsymbol{v}, \boldsymbol{\eta}) = \phi^t(\theta, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{\eta}})$$

et

$$\boldsymbol{\theta}^{\mathrm{o}}_t(\theta^*, \boldsymbol{\eta}) = \phi^t(\theta^*, \boldsymbol{v^*}, \boldsymbol{\eta}) = \phi^t(\theta^*, \tilde{\boldsymbol{v^*}}, \tilde{\boldsymbol{\eta}}).$$

D'après le Lemme 8.3, pour tout $t \geq 0$, v_t/m_t et v^*_t/m_t sont dans $B^*_{T\Theta}$. Ainsi, pour tout $t \geq 0$,

$$\left(\frac{v_t}{m_t}, \tilde{\eta}_t\right) \in B^*_{T\Theta} \times [0, \eta_{\max}] \quad \text{et} \quad \left(\frac{v^*_t}{m_t}, \tilde{\eta}_t\right) \in B^*_{T\Theta} \times [0, \eta_{\max}].$$

Par conséquent, d'après le Lemme 7.13, pour tout $1 \leq t < T_0^\infty \leq T_0^{r^*_\Theta}$,

$$\begin{aligned} \boldsymbol{\theta}^{\mathrm{o}}_t(\theta, \boldsymbol{\eta}) - \boldsymbol{\theta}^{\mathrm{o}}_t(\theta^*, \boldsymbol{\eta}) &= \phi^t(\theta, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{\eta}}) - \phi^t(\theta^*, \tilde{\boldsymbol{v^*}}, \tilde{\boldsymbol{\eta}}) \\ &= \theta - \theta^* - \sum_{s \leq t-1} \tilde{\eta}_s \frac{1}{m_s}(v_s - v^*_s) + O\left(\sum_{s \leq t-1} \tilde{\eta}^2_s\right) \\ &= \theta - \theta^* - \sum_{s \leq t-1} \eta_s (v_s - v^*_s) + O\left(\sum_{s \leq t-1} \tilde{\eta}^2_s\right). \end{aligned}$$

Soit alors le chemin inclus dans la boule B^*_Θ, et reliant θ et θ^*,

$$\theta(\cdot) : u \in [0,1] \mapsto \theta + u(\theta^* - \theta).$$

Alors, le chemin :

$$u \in [0,1] \mapsto \frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta(u))$$

est un chemin qui relie $\frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta)$ et $\frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta^*)$. Donc, pour $1 \leq t < T_0^{r^*_\Theta}$,

$$\begin{aligned} v_t - v^*_t &= \frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta) - \frac{\partial p_t \circ g_t}{\partial \theta}(e^*_0, \theta^*) \\ &= \int_0^1 \frac{\partial^2 p_t \circ g_t}{\partial \theta^2}(e^*_0, \theta(u)) \cdot \dot{\theta}(u)\, du \\ &= \int_0^1 \frac{\partial^2 p_t \circ g_t}{\partial \theta^2}(e^*_0, \theta(u)) \cdot (\theta^* - \theta)\, du. \end{aligned}$$

Ainsi,

$$\begin{aligned}\sum_{s\leq t-1}\eta_s\left(v_s-v_s^*\right)&=\sum_{s\leq t-1}\eta_s\int_0^1\frac{\partial^2 p_s\circ g_s}{\partial\theta^2}\left(e_0^*,\,\theta(u)\right)\cdot\left(\theta^*-\theta\right)\,du\\&=\int_0^1\sum_{s\leq t-1}\eta_s\frac{\partial^2 p_s\circ g_s}{\partial\theta^2}\left(e_0^*,\,\theta(u)\right)\cdot\left(\theta^*-\theta\right)\,du,\end{aligned}$$

et

$$\begin{aligned}\theta-\theta^*-\sum_{s\leq t-1}\eta_s\left(v_s-v_s^*\right)&=\int_0^1\left(\theta-\theta^*-\sum_{s\leq t-1}\eta_s\frac{\partial^2 p_s\circ g_s}{\partial\theta^2}\left(e_0^*,\,\theta(u)\right)\cdot\left(\theta^*-\theta\right)\right)du\\&=\int_0^1\left(\mathrm{Id}_{T\Theta}-\sum_{s\leq t-1}\eta_s\frac{\partial^2 p_s\circ g_s}{\partial\theta^2}\left(e_0^*,\,\theta(u)\right)\right)\cdot\left(\theta-\theta^*\right)\,du.\end{aligned}$$

Or, pour $1\leq u\leq 1$, $\theta(u)\in B_\Theta^*$ donc, d'après le Corollaire 9.25, pour tout $0\leq t<T_0^\infty\leq T_0^{\Lambda^*}$, les valeurs propres de l'opérateur

$$\mathrm{Id}_{T\Theta}-\sum_{s\leq t-1}\eta_s\frac{\partial^2 p_s\circ g_s}{\partial\theta^2}\left(e_0^*,\,\theta(u)\right)$$

sont positives et, d'après le Lemme 9.24, la plus grande est inférieure à

$$1-\lambda_{0,\,t-1}^*\sum_{s\leq t-1}\eta_s.$$

En conséquence,

$$\begin{aligned}\left\|\theta-\theta^*-\sum_{s\leq t-1}\eta_s\left(v_s-v_s^*\right)\right\|&\leq\int_0^1\left\|\mathrm{Id}_{T\Theta}-\sum_{s\leq t-1}\eta_s\frac{\partial^2 p_s\circ g_s}{\partial\theta^2}\left(e_0^*,\,\theta(u)\right)\right\|_{\mathrm{op}}\left\|\theta-\theta^*\right\|\,du\\&\leq\left(1-\lambda_{0,\,t-1}^*\sum_{s\leq t-1}\eta_s\right)\left\|\theta-\theta^*\right\|.\end{aligned}$$

Ainsi,

$$\left\|\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta,\,\boldsymbol{\eta}\right)-\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\,\boldsymbol{\eta}\right)\right\|\leq\left(1-\lambda_{0,\,t-1}^*\sum_{s\leq t}\eta_s\right)\left\|\theta-\theta^*\right\|+O\left(\sum_{s\leq t-1}\tilde{\eta}_s^2\right),$$

ce qui conclut la preuve. □

9.6 Contrôle de la « trajectoire stable en boucle ouverte »

Fait 9.29 (Contrôle de $\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\,\boldsymbol{\eta}\right)$)**.** *Pour* $1\leq t<T_0^{r_\Theta^*}$,

$$d\left(\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\,\boldsymbol{\eta}\right),\,\theta^*\right)\leq\left\|\sum_{s\leq t-1}\eta_s\frac{\partial\, p_s\circ g_s}{\partial\theta}\left(e_0^*,\,\theta^*\right)\right\|+O\left(\sum_{s\leq t-1}\tilde{\eta}_s^2\right).$$

Remarque 9.30. *Remarquons que, pour* $t\geq 0$, $\theta^*=\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\,\mathbf{0}\right)$.

Preuve du Fait 9.29. Pour $t \geq 0$, notons

$$\begin{aligned} v_t^* &= \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta^*\right),\, \theta^*\right) \cdot \boldsymbol{J}_t\left(e_0^*,\, 0,\, \theta^*\right) + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0^*,\, \theta^*\right),\, \theta^*\right) \\ &= \frac{\partial\, p_t \circ g_t}{\partial \theta}\left(e_0^*,\, \theta^*\right). \end{aligned}$$

Notons également $\boldsymbol{v^*} = (v_t^*)$, et $\tilde{\boldsymbol{v}}^* = (v_t^*/m_t)$. Alors, pour tout $t \geq 0$,

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\, \boldsymbol{\eta}\right) = \phi^t\left(\theta^*,\, \boldsymbol{v^*},\, \boldsymbol{\eta}\right) = \phi^t\left(\theta^*,\, \tilde{\boldsymbol{v}}^*,\, \tilde{\boldsymbol{\eta}}\right).$$

Ainsi, de même qu'à la preuve du Fait 9.28, nous pouvons utiliser le Lemme 7.13 pour obtenir, pour $1 \leq t < T_0^{r_\Theta^*}$,

$$\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\, \boldsymbol{\eta}\right) = \theta^* - \sum_{s \leq t-1} \eta_s\, v_s^* + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right).$$

Par conséquent,

$$\left\|\boldsymbol{\theta}_t^{\mathrm{o}}\left(\theta^*,\, \boldsymbol{\eta}\right) - \theta^*\right\| \leq \left\|\sum_{s \leq t-1} \eta_s \frac{\partial\, p_s \circ g_s}{\partial \theta}\left(e_0^*,\, \theta^*\right)\right\| + O\left(\sum_{s \leq t-1} \tilde{\eta}_s^2\right),$$

ce qui conclut la preuve. □

9.7 Propriété centrale de l'algorithme RTRL dans le cas $t_0 = 0$

Corollaire 9.31 (Constante majorant les termes en grand O)**.** *Nous pouvons trouver une constante $M_7 > 0$, ne dépendant que du système dynamique et des constantes dans les hypothèses, qui majore les constantes des termes en grand O des faits 9.18, 9.20, 9.28 et 9.29. Nous imposons de plus $M_7 \geq 1$.*

Démonstration. C'est une conséquence de la discussion à la fin de la section 9.2. Nous voulons que M_7 soit supérieur à 1 pour « faire rentrer les termes de sommes de gradients et de bruit » dans la majoration avec M_7. □

Rappelons que les $\Pi_{s,\,p}^{t_0}$ sont introduits à la Définition 8.14.

Définition 9.32 (Terme additif fonction du pas)**.** *Soit un temps $t_0 \geq 0$. Soient $\theta_{t_0} \in B_\Theta^*$, et $e_{t_0} \in B_{\mathcal{E}_{t_0}}^*$. Nous notons, pour $t \geq t_0 + 1$,*

$$\begin{aligned} b_{t_0,\,t}\left(\boldsymbol{\eta}\right) = M_7 \Bigg(&\left\|\sum_{s=t_0}^{t-1} \eta_s \frac{\partial\, p_s \circ g_s}{\partial \theta}\left(e_0^*,\, \theta^*\right)\right\| + \left\|\sum_{s=t_0}^{t-1} \left(\sum_{p=s}^{t-1} \tilde{\eta}_p\, \Pi_{s,\,p}^{t_0}\right) \xi_s\right\| \\ &+ \sup_{t_0 \leq s \leq t-1} \tilde{\eta}_s + \left(\sum_{s=t_0}^{t-1} \tilde{\eta}_s\right)^2 + \sum_{s=t_0}^{t-1} \tilde{\eta}_s^2 \Bigg). \end{aligned}$$

La dépendance des termes $\Pi_{s,\,p}^{t_0}$, et donc de $b_{t_0,\,t}\left(\boldsymbol{\eta}\right)$, en θ_{t_0} et e_{t_0} est gardée implicite, ce qui ne sera pas gênant dans la suite.

Remarque 9.33. *La définition de $b_{t_0,\,t}(\boldsymbol{\eta})$ fait intervenir e_0^* mais, comme celui-ci est l'état initial de la trajectoire stable, et est donc indépendant de la trajectoire produite par l'algorithme RTRL, ne pas expliciter cette dépendance ne pose pas de problème.*

Fait 9.34 (Propriété centrale de l'algorithme RTRL, énoncé pour $t_0 = 0$)**.** *Soit $\theta_0 \in B^*_\Theta$, tel que de plus*

$$d(\theta_0,\,\theta^*) \leq \frac{r^*_\Theta}{3}.$$

Soient $e_0 \in B^*_{\mathcal{E}_0}$, et $J_0 \in B^*_{\mathcal{L}(\Theta,\,\mathcal{E}_0)}$. Soit une suite $\boldsymbol{\eta} = (\eta_t)$ telle que la suite $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0,\,\eta_{\max}]$. Notons alors $\boldsymbol{\theta} = (\theta_s)$ la suite de paramètres produite par l'algorithme RTRL. Alors, pour tout $0 \leq t < T_0^\infty$, $\theta_t \in B^*_\Theta$ et, pour tout $1 \leq t < T_0^\infty$,

$$d(\theta_t,\,\theta^*) \leq \left(1 - \lambda^*_{0,\,t-1} \sum_{s=0}^{t-1} \eta_s\right) d(\theta_0,\,\theta^*) + b_{0,\,t}(\boldsymbol{\eta}).$$

Remarque 9.35. *Le paramètre et l'état à l'instant $t_0 = 0$ qui rentrent dans la définition des $b_{0,\,t}(\boldsymbol{\eta})$, et sont gardés implicites, ainsi qu'il a été dit à la Définition 9.32, sont ceux de l'énoncé.*

Preuve du Fait 9.34. $T_0^\infty \leq T_0^{r^*_\Theta}$ donc, d'après le Fait 9.8, pour $1 \leq t < T_0^\infty$, $\theta_t \in B^*_\Theta$. Toujours pour $1 \leq t < T_0^\infty$,

$$\begin{aligned} d(\theta_t,\,\theta^*) &\leq d(\theta_t,\,\boldsymbol{\theta}^{\mathrm{o}}_t(\theta_0,\,e_0,\,J_0,\,\boldsymbol{\eta})) + d(\boldsymbol{\theta}^{\mathrm{o}}_t(\theta_0,\,e_0,\,J_0,\,\boldsymbol{\eta}),\,\boldsymbol{\theta}^{\mathrm{o}}_t(\theta_0,\,e_0^*,\,0,\,\boldsymbol{\eta})) \\ &\quad + d(\boldsymbol{\theta}^{\mathrm{o}}_t(\theta_0,\,e_0^*,\,0,\,\boldsymbol{\eta}),\,\boldsymbol{\theta}^{\mathrm{o}}_t(\theta^*,\,e_0^*,\,0,\,\boldsymbol{\eta})) + d(\boldsymbol{\theta}^{\mathrm{o}}_t(\theta^*,\,e_0^*,\,0,\,\boldsymbol{\eta}),\,\theta^*) \\ &\leq \left(1 - \lambda^*_{0,\,t-1} \sum_{s \leq t-1} \eta_s\right) d(\theta_0,\,\theta^*) + b_{0,\,t}(\boldsymbol{\eta}), \end{aligned}$$

d'après le Corollaire 9.31 et les Faits 9.18, 9.20, 9.28 et 9.29. □

9.8 Propriété centrale de l'algorithme RTRL

Proposition 9.36 (Propriété centrale de l'algorithme RTRL)**.** *Soit un instant t_0 quelconque. Soit $\theta_{t_0} \in B^*_\Theta$, tel que de plus*

$$d(\theta_{t_0},\,\theta^*) \leq \frac{r^*_\Theta}{3}.$$

Soient $e_{t_0} \in B^*_{\mathcal{E}_{t_0}}$ et $J_{t_0} \in B^*_{\mathcal{L}(\Theta,\,\mathcal{E}_{t_0})}$. Soit une suite $\boldsymbol{\eta} = (\eta_t)$ telle que la suite $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0,\,\eta_{\max}]$. Notons alors $\boldsymbol{\theta} = (\theta_s)$ la suite de paramètres produite par l'algorithme RTRL initialisé à l'instant t_0. Alors, pour tout $t_0 \leq t < T_{t_0}^\infty$, $\theta_t \in B^*_\Theta$ et, pour tout $t_0 + 1 \leq t < T_{t_0}^\infty$,

$$d(\theta_t,\,\theta^*) \leq \left(1 - \lambda^*_{t_0,\,t-1} \sum_{s=t_0}^{t-1} \eta_s\right) d(\theta_{t_0},\,\theta^*) + b_{t_0,\,t}(\boldsymbol{\eta}).$$

Remarque 9.37. *Le paramètre et l'état à l'instant t_0 qui rentrent dans la définition des $b_{t_0,\,t}(\boldsymbol{\eta})$, et sont gardés implicites, ainsi qu'il a été dit à la Définition 9.32, sont ceux de l'énoncé.*

Preuve de la Proposition 9.36. L'algorithme RTRL initialisé à t_0 est l'algorithme RTRL appliqué aux pertes $p_{t_0+\cdot}$, avec les pas $\eta_{t_0+\cdot}$, les opérateurs de transition $\boldsymbol{T}_{t_0+\cdot}$ et les termes de bruit $\xi_{t_0+\cdot}$. □

Chapitre

10 Convergence de l'algorithme RTRL

10.1 Continuité des trajectoires obtenues par l'algorithme RTRL par rapport à la suite de pas

10.1.1 Renormalisation de la suite de pas, paramétrage des suites

Lemme 10.1 (Renormalisations de la suite de pas)**.** *Quitte à la multiplier par une constante, nous pouvons supposer que la suite de pas de descente $\boldsymbol{\eta} = (\eta_t)$, dont nous disposons grâce à l'Hypothèse 6.18, vérifie pour tout $t \geq 0$, $\tilde{\eta}_t = m_t \, \eta_t \leq 1$. Ceci ne modifie pas les propriétés requises dans l'hypothèse.*

Démonstration. Les propriétés de l'Hypothèse 6.18 sont homogènes en la suite $\boldsymbol{\eta}$. Elles sont ainsi vérifiées pour toute suite $\mu \, \boldsymbol{\eta}$, où $\mu > 0$. Or, d'après le deuxième point de l'Hypothèse 6.18, $\tilde{\boldsymbol{\eta}}$ est convergente, donc bornée. Nous choisissons alors μ supérieur au maximum des $\tilde{\eta}_t$, et remplaçons la suite de pas $\boldsymbol{\eta}$ par la suite $\mu^{-1} \, \boldsymbol{\eta}$. Or, pour tout $t \geq 0$,

$$\widetilde{(\mu^{-1} \, \boldsymbol{\eta})}_t = m_t \, \mu^{-1} \, \eta_t = \mu^{-1} \, m_t \, \eta_t = \mu^{-1} \, \tilde{\eta}_t,$$

et ainsi, pour tout $t \geq 0$,

$$\widetilde{(\mu^{-1} \, \boldsymbol{\eta})}_t \leq 1,$$

ce qui conclut la preuve. □

Dans la suite, nous considérons alors la demi-droite dirigée par la suite $\boldsymbol{\eta}$, que nous notons $\bar{\boldsymbol{\eta}}$. Nous paramétrons celle-ci par les réels $\mu \geq 0$. Ainsi, nous considérons des suites $\mu \, \boldsymbol{\eta}$ égales à

$$\mu \, \eta_0, \, \mu \, \eta_1, \, \mu \, \eta_2, \ldots, \mu \, \eta_t, \ldots$$

Dans la preuve du Lemme 10.1 précédent, nous avons montré en particulier que, pour tout réel μ,

$$\widetilde{\mu \, \boldsymbol{\eta}} = \mu \, \tilde{\boldsymbol{\eta}}.$$

Rappelons que nous demandions jusqu'à présent que les suites de pas de descente $\boldsymbol{\eta}$ utilisées soient telles que les suites $\tilde{\boldsymbol{\eta}}$ associées soient incluses dans un segment $[0, \eta_{\max}]$, introduit à la Définition 9.3.

Corollaire 10.2 (Contrôle des suites de pas de descente)**.** *Pour $\mu \leq \eta_{\max}$, la suite $\mu \, \tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$.*

Démonstration. D'après le Lemme 10.1 précédent, pour tout $t \geq 0$, $\tilde{\eta}_t \leq 1$, donc

$$\mu\, \tilde{\eta}_t \leq \mu.$$

Ainsi, la suite $\widetilde{\mu\, \boldsymbol{\eta}} = \mu\, \tilde{\boldsymbol{\eta}} = (\mu\, \tilde{\eta}_t)$ est incluse dans le segment $[0, \mu]$, ce qui conclut la preuve. □

10.1.2 Continuité des trajectoires obtenues par l'algorithme RTRL par rapport à la suite de pas

Lemme 10.3 (Continuité d'une application de l'algorithme RTRL)**.** *Pour tout $t \geq 0$, l'application qui, à $\theta \in B^*_\Theta$, $e \in B^*_{\mathcal{E}_t}$, $J \in \mathcal{L}(\Theta, \mathcal{E}_t)$ tel que $\|J\|_{\text{op}} \leq r_\xi$ et $\eta \in [0, \frac{\eta_{\max}}{m_t}]$ associe leurs images par une application de l'algorithme RTRL à l'instant t est continue.*

Démonstration. Soit $t \geq 0$. Soient $\theta \in B^*_\Theta$, $e \in B^*_{\mathcal{E}_t}$ et $J \in \mathcal{L}(\Theta, \mathcal{E}_t)$ tel que $\|J\|_{\text{op}} \leq r_\xi$. Notons

$$v_t = \frac{\partial\, p_t}{\partial e}(e, \theta) \cdot J + \frac{\partial\, p_t}{\partial \theta}(e, \theta).$$

D'après l'Hypothèse 5.1, les différentielles dépendent continûment de e et θ. Par composition, v_t est donc continu en ses arguments. De plus, d'après le Lemme 8.8 et le Corollaire 8.9, $v_t / m_t \in B^*_{T\Theta}$. Or, d'après le Lemme 7.6, l'opérateur de mise à jour du paramètre est continu sur le produit $B^*_\Theta \times B^*_{T\Theta} \times [0, \eta_{\max}]$. Ainsi, l'application

$$\begin{aligned} B^*_\Theta \times B^*_{\mathcal{E}_t} \times B^*_{\mathcal{L}(\Theta, \mathcal{E}_t)} \times [0, \frac{\eta_{\max}}{m_t}] &\to \Theta \\ \theta,\, e,\, J,\, \eta &\mapsto \phi\left(\theta,\, \frac{v_t}{m_t},\, m_t\, \eta\right), \end{aligned}$$

est continue. Or, elle coïncide avec l'application qui, aux mêmes arguments, associe

$$\phi(\theta,\, v_t,\, \eta),$$

qui est le nouveau paramètre. Ce dernier dépend ainsi continûment des variables considérées.

L'opérateur de transition sur les états est continu d'après l'Hypothèse 4.15.

L'opérateur de transition sur les différentielles des états est

$$\theta,\, e,\, J \mapsto \frac{\partial\, \boldsymbol{T}_t}{\partial e}(e, \theta) \cdot J + \frac{\partial\, \boldsymbol{T}_t}{\partial \theta}(e, \theta) + \xi_t.$$

Le terme ξ_t est constant en θ, e et J. Les différentielles de $\boldsymbol{T}_t$ dépendent continûment de leurs arguments, toujours d'après l'Hypothèse 4.15 donc, de même que ci-dessus, par composition, la nouvelle différentielle dépend continûment des variables considérées. Ceci conclut la preuve. □

Lemme 10.4 (Continuité des coordonnées de l'algorithme RTRL par rapport à μ)**.** *Fixons $\theta_0 \in B^*_\Theta$, $e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$. Alors, pour tout $t \geq 0$, nous pouvons trouver un $\eta^t_{\max} \leq \eta_{\max}/m_t$ suffisamment petit pour que, pour $\mu \leq \eta^t_{\max}$, l'application*

$$\mu \mapsto (\theta_t,\, e_t,\, J_t)$$

soit continue.

Démonstration. Prouvons-le par récurrence. L'initialisation est acquise en posant par exemple $\eta_{\max}^0 = 1$, car θ_0, e_0 et J_0 ne dépendent pas du pas de descente. Soit alors $t \geq 0$, et supposons la propriété vraie pour t. L'image de 0 par l'application

$$\mu \mapsto (\theta_t, e_t, J_t)$$

est

$$(\theta_0, \boldsymbol{e}_t(e_0, \theta_0), J_t(0)),$$

où $J_t(0)$ désigne la coordonnée en J de celle-ci, sans qu'il soit nécessaire de spécifier plus avant ses propriétés. Cette image appartient ainsi au produit (ouvert, d'après le Lemme 4.56 ; de même, pour le bruit, nous pouvions choisir r_ξ tel que l'inégalité $\|J\|_{\text{op}} \leq r_\xi$ soit stricte dès que $\|J_0\|_{\text{op}}$ était dans la boule ouverte $B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$)

$$B^*_\Theta \times B^*_{\mathcal{E}_t} \times \left\{ J \in \mathcal{L}(\Theta, \mathcal{E}_t) \,\middle|\, \|J\|_{\text{op}} < r_\xi \right\}$$

d'après les Corollaires 4.53 et 4.55, et le Lemme 8.7. Donc, par hypothèse de récurrence, nous pouvons trouver

$$\eta_{\max}^{t+1} \leq \min\left(\eta_{\max}^t, \frac{m_t}{m_{t+1}} \eta_{\max}^t\right)$$

tel que, pour $\mu \leq \eta_{\max}^{t+1}$, l'image de η_0 appartient également à ce produit. Or, pour $\mu \leq \eta_{\max}^{t+1}$, l'application

$$\mu \mapsto (\theta_{t+1}, e_{t+1}, J_{t+1})$$

est la composée de RTRL à l'instant t, qui est continu sur le produit ci-dessus d'après le Lemme 10.3, car $\eta_{\max}^{t+1} \leq \eta_{\max}^t \leq \eta_{\max}/m_t$, avec l'application

$$\mu \mapsto (\theta_t, e_t, J_t),$$

qui est continue par hypothèse de récurrence, car $\eta_{\max}^{t+1} \leq \eta_{\max}^t$, donc elle est également continue. Enfin, par construction, $\eta_{\max}^{t+1} \leq \eta_{\max}/m_{t+1}$. Ceci clôt la récurrence, et la preuve. □

Corollaire 10.5 (Continuité des trajectoires obtenues par l'algorithme RTRL par rapport à μ)**.** *Fixons $\theta_0 \in B^*_\Theta$, $e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$. Alors, pour tout $t \geq 0$, nous pouvons trouver un $\eta_{\max}^t \leq \eta_{\max}/m_t$ suffisamment petit pour que, pour $\mu \leq \eta_{\max}^t$, l'application*

$$\mu \mapsto (\theta_s, e_s, J_s)_{s \leq t}$$

soit continue.

Démonstration. C'est une conséquence du Lemme 10.4 précédent, car l'application considérée est continue en tant que vecteur de dimension finie d'applications continues. □

10.2 Établissement des conditions pour la convergence

10.2.1 Établissement de la condition homogène en la suite de pas de descente

Corollaire 10.6 (Opérateurs définis positifs au voisinage de θ^*)**.** *Nous pouvons trouver $\lambda_{\min} > 0$ tel que, quitte à restreindre B^*_Θ, nous pouvons trouver $k_0 = k_0(\bar{\boldsymbol{\eta}})$*

tel que, pour tout $k \geq k_0$, pour tout $\theta \in B_\Theta^*$, la plus petite valeur propre des opérateurs

$$\frac{\sum_{I_k} \eta_t \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta)}{\sum_{I_k} \eta_t}$$

est supérieure à $\lambda_{\min}$.

Démonstration. Posons, pour $k \geq 0$,

$$H_k(\theta) = \frac{\sum_{I_k} \eta_t \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta)}{\sum_{I_k} \eta_t}.$$

L'énoncé à prouver est équivalent à : nous pouvons trouver $\lambda_{\min} > 0$ tel que, pour tout $k \geq k_0$, pour tout $\theta \in B_\Theta^*$, pour tout v sur la sphère unité $\mathbb{S}_\Theta$ de l'espace tangent à Θ,

$$^t v \, H_k(\theta) \, v \geq \lambda_{\min}.$$

D'après le Lemme 6.30, nous pouvons trouver k_0 tel que, pour $k \geq k_0$, la plus petite valeur propre de $H_k(\theta^*)$ est supérieure à λ. Ainsi, pour tout $v \in \mathbb{S}_\Theta$,

$$^t v \, H_k(\theta^*) \, v \geq \lambda.$$

D'après le Lemme 5.19, les fonctions

$$\theta \mapsto \frac{1}{m_t} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta)$$

sont équicontinues en θ^*. Or, pour tout $k \geq 0$,

$$H_k(\theta) = \frac{\sum_{I_k} m_t \eta_t \frac{1}{m_t} \frac{\partial^2 p_t \circ g_t}{\partial \theta^2} (e_0^*, \theta)}{\sum_{I_k} \eta_t}.$$

Ainsi, les H_k sont également équicontinues en θ^*. Soit alors $\varepsilon > 0$. Nous pouvons trouver un voisinage de θ^* tel que, pour tout paramètre θ dans ce voisinage, pour tout $k \geq 0$,

$$\|H_k(\theta) - H_k(\theta^*)\|_{\text{op}} \leq \varepsilon.$$

Quitte à restreindre B_Θ^*, nous supposons dorénavant que cette inégalité est vérifiée sur celle-ci. Alors, pour tout $k \geq k_0$, pour tout $\theta \in B_\Theta^*$, pour tout $v \in \mathbb{S}_\Theta$,

$$\begin{aligned} ^t v \, (H_k(\theta) - H_k(\theta^*)) \, v &\leq \|H_k(\theta) - H_k(\theta^*)\|_{\text{op}} \, \|v\|^2 \\ &\leq \varepsilon. \end{aligned}$$

La première inégalité est due au fait que $H_k(\theta) - H_k(\theta^*)$ est auto-adjoint pour le produit scalaire euclidien.

Alors, pour tout $k \geq k_0$, pour tout $\theta \in B_\Theta^*$, pour tout $v \in \mathbb{S}_\Theta$,

$$^t v \, H_k(\theta) \, v \geq \lambda - \varepsilon > 0,$$

pourvu que ε soit choisi assez petit indépendamment de k et θ. Nous posons alors

$$\lambda_{\min} = \lambda - \varepsilon.$$

Enfin, la propriété vérifiée au Lemme 6.30 est homogène en $\boldsymbol{\eta}$, donc k_0 ne dépend que de $\bar{\boldsymbol{\eta}}$, et le résultat annoncé est bien démontré. □

10.2.2 Établissement des conditions non homogènes en la suite de pas de descente

Rappelons que les quantités $b_{t_0,\,t}(\boldsymbol{\eta})$ ont été introduites à la Définition 9.32.

Définition 10.7 (Terme additif le long de la suite (T_k))**.** *Notons, pour $\mu \geq 0$ et $k \geq 0$,*

$$b^k(\mu) = b_{T_k,\,T_{k+1}}(\mu\,\boldsymbol{\eta}).$$

Lemme 10.8 (Contrôle de $b^k(\mu)$)**.** *Pour tout $\mu \geq 0$, $b^k(\mu)$ est négligeable devant*

$$\sum_{I_k} \mu\,\eta_t$$

quand k tend vers l'infini.

Démonstration. C'est une conséquence du Fait 6.29, du Corollaire 6.34 et de l'Hypothèse 8.16. □

Définition 10.9 (Infimum pour le maintien de $\boldsymbol{\theta}$ dans B^*_Θ)**.** *Pour tout $\mu \geq 0$, notons*

$$k_1(\mu) = \inf\left\{ k \geq 0 \;\middle|\; b^k(\mu) \leq \frac{1}{2}\,\lambda_{\min}\,\frac{r^*_\Theta}{3}\,\sum_{I_k} \mu\,\eta_t \right\}.$$

Corollaire 10.10 (Propriétés de l'infimum pour le maintien de $\boldsymbol{\theta}$ dans B^*_Θ)**.** *k_1 vérifie les propriétés suivantes.*

1. *Pour tout $\mu \geq 0$, $k_1(\mu) < \infty$.*
2. *k_1 est une fonction croissante de μ.*

Démonstration. La première propriété est une conséquence du Lemme 10.8.

Soient $0 \leq \mu_1 \leq \mu_2$. Alors, pour tout $k \geq 0$, comme $b^k(\mu)$ ne comporte que des termes homogènes de degré supérieur à 1 en μ,

$$\begin{aligned} b^k(\mu_1) &= b^k\left(\mu_2\,\frac{\mu_1}{\mu_2}\right) \\ &\leq \frac{\mu_1}{\mu_2}\,b^k(\mu_2). \end{aligned}$$

Or, pour tout $k \geq k_1(\mu_2)$,

$$b^k(\mu_2) \leq \frac{1}{2}\,\lambda_{\min}\,\frac{r^*_\Theta}{3}\,\sum_{I_k} \mu_2\,\eta_t,$$

donc

$$\begin{aligned} b^k(\mu_1) &\leq \frac{\mu_1}{\mu_2}\,\frac{1}{2}\,\lambda_{\min}\,\frac{r^*_\Theta}{3}\,\sum_{I_k} \mu_2\,\eta_t \\ &= \frac{1}{2}\,\lambda_{\min}\,\frac{r^*_\Theta}{3}\,\sum_{I_k} \mu_1\,\eta_t. \end{aligned}$$

Ainsi, $k_1(\mu_1) \leq k_1(\mu_2)$, ce qui conclut la preuve. □

10.2.3 Choix d'une suite de pas

Notons $\theta_t(\mu)$ le t-ème paramètre obtenu par l'algorithme RTRL avec la suite de pas de descente $\mu\,\boldsymbol{\eta}$, avec des quantités initiales qui seront précisées par la suite.

Rappelons que la constante M_4 a été définie au Lemme 7.10.

Corollaire 10.11 (Choix d'une suite de pas et attente des conditions de convergence)**.** *Nous pouvons trouver $\mu_{\max} > 0$ et $0 \leq k_2 < \infty$ tels que, pour tout $\mu \leq \mu_{\max}$, les propriétés suivantes sont vérifiées.*

1. $k_2 \geq k_0(\bar{\boldsymbol{\eta}})$.
2. *La suite $\mu\,\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$.*
3. *Pour tout $k \geq k_2$,*
$$M_4 \sum_{I_k} \mu\,\tilde{\eta}_t \leq \frac{r^*_\Theta}{3}.$$
4. *Pour tout $k \geq k_2$,*
$$\Lambda^* \sum_{I_k} \mu\,\tilde{\eta}_t \leq 1.$$
5. $k_2 \geq k_1(\mu)$.
6. *Pour tous $\theta_0 \in B^*_\Theta$ tel que de plus*
$$d(\theta_0,\, \theta^*) \leq \frac{r^*_\Theta}{4},$$
$e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$, pour tout $0 \leq \mu' \leq \mu$, pour tout $0 \leq t \leq T_{k_2}$,
$$d(\theta_t(\mu'),\, \theta^*) \leq \frac{r^*_\Theta}{3}.$$

Démonstration. Montrons d'abord que nous pouvons trouver un $\mu > 0$ vérifiant les propriétés.

Les deux premières propriétés sont vérifiées en posant $k_2 = k_0(\bar{\boldsymbol{\eta}})$, et en choisissant μ assez petit strictement positif (car $\eta_{\max} > 0$), d'après le Corollaire 10.2.

D'après le Corollaire 6.35
$$\sum_{I_k} \tilde{\eta}_t$$
tend vers 0 avec k donc, quitte à augmenter k_2, les troisième et quatrième propriétés sont vérifiées.

Pour le μ choisi jusqu'à présent, si $k_2 < k_1(\mu)$, nous posons $k_2 = k_1(\mu)$, ce qui ne remet pas en cause les propriétés précédentes (et en particulier les deux précédentes, qui étaient valables pour tous les k au-delà d'un certain rang).

Enfin, pour ce k_2, d'après le Corollaire 10.5, nous pouvons trouver $\mu' > 0$ tel que, pour tout $\mu'' \leq \mu'$, pour tout $0 \leq t \leq T_{k_2}$,
$$d(\theta_t(\mu''),\, \theta^*) \leq \frac{r^*_\Theta}{3}.$$

Si $\mu' \geq \mu$, nous conservons alors le μ choisi jusqu'à présent. Sinon, nous posons $\mu = \mu'$, ce qui ne remet pas en cause les propriétés précédentes. En particulier, cela ne remet pas en cause la cinquième propriété, car k_1 est une fonction croissante de μ, d'après le Corollaire 10.10.

Enfin, nous posons $\mu_{\max} = \mu$, et la propriété est alors bien vérifiée par tout $\mu \leq \mu_{\max}$, car les contraintes sur μ sont croissantes : si elles sont vérifiées en un point μ, elles le sont en tout point compris entre 0 et μ. La croissance des contraintes est vraie pour les contraintes 2 à 4, elle l'est par construction pour la contrainte 6, et elle l'est d'après le Corollaire 10.10 pour la contrainte 5. Ceci conclut la preuve. □

Lemme 10.12 (Horizons et échelle de temps)**.** *Fixons μ et k_2 vérifiant (ce qui est possible) les cinq premières propriétés du Corollaire 10.11. Soit $k \geq k_2$. Soient θ_{T_k} tel que*

$$d\left(\theta_{T_k},\, \theta^*\right) \leq \frac{r^*_\Theta}{3},$$

$e_{T_k} \in B^*_{\mathcal{E}_{T_k}}$ et $J_{T_k} \in B^*_{\mathcal{L}(\Theta,\, \mathcal{E}_{T_k})}$. Notons (θ_t) la suite de paramètres produite par l'algorithme RTRL initialisé à l'instant T_k. Alors, pour tout $T_k \leq t \leq T_{k+1}$, θ_t appartient à B^*_Θ, et

$$d\left(\theta_{T_{k+1}},\, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\, \eta_t\right)\, d\left(\theta_{T_k},\, \theta^*\right) + b^k\left(\mu\right).$$

Remarque 10.13. *Comme précédemment, le terme $b^k\left(\mu\right)$ est calculé avec le θ_{T_k} de l'énoncé et un e_{T_k} dans $B^*_{\mathcal{E}_{T_k}}$.*

Démonstration.

$$d\left(\theta_{T_k},\, \theta^*\right) \leq \frac{r^*_\Theta}{3},$$

$e_{T_k} \in B^*_{\mathcal{E}_{T_k}}$, et $J_{T_k} \in B^*_{\mathcal{L}(\Theta,\, \mathcal{E}_{T_k})}$ donc, d'après la Proposition 9.36, pour tout $T_k + 1 \leq t < T^\infty_{T_k}$, nous savons que $\theta_t \in B^*_\Theta$ et

$$d\left(\theta_t,\, \theta^*\right) \leq \left(1 - \lambda^*_{T_k,\, t-1} \sum_{s=T_k}^{t-1} \mu\, \eta_s\right)\, d\left(\theta_{T_k},\, \theta^*\right) + b_{T_k,\, t}\left(\mu\, \boldsymbol{\eta}\right).$$

Nous allons prouver que $T^\infty_{T_k}$ est strictement supérieur à T_{k+1}. D'après la troisième propriété du Corollaire 10.11,

$$M_4 \sum_{t=T_k}^{T_{k+1}-1} \mu\, \tilde{\eta}_t = M_4 \sum_{I_k} \mu\, \tilde{\eta}_t \leq \frac{r^*_\Theta}{3},$$

et ainsi, d'après la Définition 9.6,

$$T^{r^*_\Theta}_{T_k} \geq T_{k+1} + 1.$$

Pour tous $0 \leq t_1 \leq t_2$,

$$\sum_{t=t_1}^{t_2} \eta_t\, \frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e^*_t,\, \theta^*\right) = \sum_{t=t_1}^{t_2} \tilde{\eta}_t\, \frac{1}{m_t}\, \frac{\partial^2\, p_t \circ g_t}{\partial \theta^2}\left(e^*_t,\, \theta^*\right).$$

Ainsi, d'après le Corollaire 5.17, nous savons que

$$\Lambda^*_{t_1,\, t_2} \leq \Lambda^* \sum_{t=t_1}^{t_2} \tilde{\eta}_t.$$

Donc,

$$\Lambda^*_{T_k,\,T_{k+1}-1} \leq \Lambda^* \sum_{t=T_k}^{T_{k+1}-1} \tilde{\eta}_t = \Lambda^* \sum_{I_k} \tilde{\eta}_t.$$

Par conséquent, d'après la quatrième propriété du Corollaire 10.11, et la Définition 9.22,

$$T^{\Lambda^*}_{T_k} \geq T_{k+1} + 1.$$

Par conséquent,

$$T^{\infty}_{T_k} \geq T_{k+1} + 1,$$

soit

$$T_{k+1} < T^{\infty}_{T_k}.$$

Ainsi, pour tout $T_k + 1 \leq t \leq T_{k+1}$, θ_t appartient à B^*_Θ, et

$$d\left(\theta_t,\, \theta^*\right) \leq \left(1 - \lambda^*_{T_k,\,t-1} \sum_{s=T_k}^{t-1} \mu\, \eta_s\right) d\left(\theta_{T_k},\, \theta^*\right) + b_{T_k,\,t}\left(\mu\, \boldsymbol{\eta}\right).$$

Enfin, d'après la première propriété du Corollaire 10.11 et le Corollaire 10.6,

$$\lambda^*_{T_k,\,T_{k+1}-1} \geq \lambda_{\min},$$

et ainsi

$$d\left(\theta_{T_{k+1}},\, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{t=T_k}^{T_{k+1}-1} \mu\, \eta_t\right) d\left(\theta_{T_k},\, \theta^*\right) + b^k\left(\mu\right).$$

□

10.3 Convergence de l'algorithme RTRL

Fixons $\theta_0 \in B^*_\Theta$, tel que de plus

$$d\left(\theta_0,\, \theta^*\right) \leq \frac{r^*_\Theta}{4},$$

$e_0 \in B^*_{\mathcal{E}_0}$ et $J_0 \in B^*_{\mathcal{L}(\Theta,\mathcal{E}_0)}$. Fixons également $\mu_{\max} > 0$ et $k_2 \geq 0$ donnés par le Corollaire 10.11. Considérons alors un $0 \leq \mu \leq \mu_{\max}$.

Fait 10.14 (Contrôle de la sous-suite (θ_{T_k}))**.** *Pour tout $k \geq k_2$,*

$$d\left(\theta_{T_k},\, \theta^*\right) \leq \frac{r^*_\Theta}{3}.$$

Démonstration. Prouvons-le par récurrence. L'initialisation est une conséquence du choix de k_2 et de μ donné par le Corollaire 10.11. Soit alors $k \geq k_2$, et supposons la propriété vraie pour k. D'après le Lemme 10.12,

$$d\left(\theta_{T_{k+1}},\, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\, \eta_t\right) d\left(\theta_{T_k},\, \theta^*\right) + b^k\left(\mu\right).$$

Or, d'après la cinquième propriété du Corollaire 10.11,

$$b^k\left(\mu\right) \leq \frac{1}{2}\, \lambda_{\min}\, \frac{r^*_\Theta}{3} \sum_{I_k} \mu\, \eta_t$$

et ainsi,

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\, \eta_t\right) \frac{r_\Theta^*}{3} + \frac{1}{2}\, \lambda_{\min} \frac{r_\Theta^*}{3} \sum_{I_k} \mu\, \eta_t$$
$$\leq \frac{r_\Theta^*}{3},$$

ce qui conclut la récurrence, et la preuve. □

Corollaire 10.15 (Maintien de $\boldsymbol{\theta}$ dans B_Θ^*)**.** *Pour tout $t \geq 0$, θ_t appartient à B_Θ^*.*

Démonstration. D'après le Corollaire 10.11, pour tout $t \leq T_{k_2}$, $\theta_t \in B_\Theta^*$. D'après le Fait 10.14, pour tout $k \geq k_2$,

$$d\left(\theta_{T_k}, \theta^*\right) \leq \frac{r_\Theta^*}{3}$$

donc, d'après le Lemme 10.12, pour tout $T_k \leq t \leq T_{k+1}$, $\theta_t \in B_\Theta^*$. Or, d'après le Lemme 6.23, T_k tend vers l'infini avec k, ce qui conclut la preuve. □

Corollaire 10.16 (Inégalité vérifiée par la sous-suite (θ_{T_k}))**.** *Pour tout $k \geq k_2$,*

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\, \eta_t\right) d\left(\theta_{T_k}, \theta^*\right) + b^k\left(\mu\right).$$

Démonstration. C'est une conséquence du Lemme 10.12 et du Fait 10.14. □

Lemme 10.17 (Convergence d'une suite vérifiant un schéma de récurrence de type gradient stochastique)**.** *Soit une suite positive (r_k), majorée par 1, et une suite positive (b_k). Nous considérons alors une suite (x_k) définie par la donnée d'un $x_0 \geq 0$ et la relation de récurrence, pour $k \geq 0$,*

$$x_{k+1} = (1 - r_k)\, x_k + b_k.$$

Supposons que $b_k = o\left(r_k\right)$ quand k tend vers l'infini, et que la série de terme général r_k diverge. Alors, la suite (x_k) converge vers 0.

Démonstration. La suite (r_k) est majorée par 1 et la suite (b_k) est positive donc la suite (x_k) est positive.

Soit $\varepsilon > 0$. Soit $K \geq 0$ tel que, pour $k \geq K$, $b_k \leq \varepsilon\, r_k$. Alors, pour $k \geq K$, le segment $[0, \varepsilon]$ est stable par la transformation

$$x \mapsto (1 - r_k)\, x + b_k.$$

En effet, pour tout $k \geq K$, pour tout $x \geq \varepsilon$,

$$(1 - r_k)\, x + b_k \leq (1 - r_k)\, \varepsilon + \varepsilon\, r_k = \varepsilon.$$

Alors, soit il existe $k' \geq K$ tel que $x_{k'} \leq \varepsilon$, soit pour tout $k \geq K$, $x_k \geq \varepsilon$.

Dans le premier cas, d'après la remarque précédente, pour tout $k \geq k'$, x_k est inférieur à ε.

Dans le deuxième cas, pour tout $k \geq K$,

$$x_{k+1} - \varepsilon = (1 - r_k)\, x_k + b_k - \varepsilon$$
$$\leq (1 - r_k)\, x_k + \varepsilon\, r_k - \varepsilon$$
$$= (1 - r_k)\, (x_k - \varepsilon).$$

Ainsi, pour tout $k \geq K$,

$$0 \leq x_k - \varepsilon \leq \left(\prod_{k'=K}^{k-1} (1 - r_{k'}) \right) (x_K - \varepsilon) .$$

Or, la série de terme général r_k diverge, donc le produit ci-dessus tend vers 0 avec k. Ainsi, nous pouvons trouver $K' \geq K$ tel que, pour $k \geq K'$, $x_k \leq 2\,\varepsilon$.

Ainsi, dans tous les cas, nous pouvons trouver $K' \geq K$ tel que, pour $k \geq K'$, $x_k \leq 2\,\varepsilon$. Or, ε était quelconque. Nous avons donc établi la convergence vers 0 de la suite (x_k). □

Lemme 10.18 (Suite majorant $d(\theta_{T_k}, \theta^*)$)**.** *Considérons la suite $(x_k)_{k \geq k_2}$ définie par $x_{k_2} = \frac{r^*_\Theta}{3}$ et la relation de récurrence, pour $k \geq k_2$,*

$$x_{k+1} = \left(1 - \lambda_{\min} \sum_{I_k} \mu\, \eta_t \right) x_k + b^k(\mu) .$$

Alors, la suite (x_k) converge vers 0.

Démonstration. Posons, pour $k \geq k_2$,

$$r_k = \lambda_{\min} \sum_{I_k} \mu\, \eta_t.$$

Alors, d'après la quatrième condition du Corollaire 10.11, pour tout $k \geq k_2$, $r_k \leq 1$. D'après le Lemme 10.8, la suite de terme général $b^k(\mu)$ est négligeable devant r_k, quand k tend vers l'infini.

D'après l'Hypothèse 6.18, la série de terme général η_t diverge. Or, d'après le Lemme 6.23, T_k tend vers l'infini, quand k tend vers l'infini. Donc, la série de terme général

$$\sum_{I_k} \eta_t$$

diverge, et ainsi la série de terme général r_k diverge. Nous pouvons alors utiliser le Lemme 10.17 pour obtenir le résultat annoncé. □

Fait 10.19 (Convergence vers 0 de $d(\theta_{T_k}, \theta^*)$)**.** *La suite de terme général $d(\theta_{T_k}, \theta^*)$ converge vers 0.*

Démonstration.

$$d\left(\theta_{T_{k_2}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}$$

donc, par récurrence, grâce au Corollaire 10.16, pour tout $k \geq k_2$,

$$d(\theta_{T_k}, \theta^*) \leq x_k.$$

Ainsi, d'après le Lemme 10.18,

$$d(\theta_{T_k}, \theta^*) \to 0$$

quand k tend vers l'infini. □

Lemme 10.20 (Écart à la sous-suite tendant vers 0)**.**

$$\sup_{t \in I_k} d\left(\theta_t,\, \theta_{T_k}\right) \to 0,$$

quand k tend vers l'infini.

Démonstration. Notons (J_t) la suite de différentielles produite par l'algorithme RTRL. D'après le Corollaire 10.15, pour tout $t \geq 0$,

$$\theta_t \in B_\Theta^*.$$

Donc, d'après le Corollaire 8.10, pour tout $t \geq 0$, en posant

$$v_t = \frac{\partial\, p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \cdot J_t + \frac{\partial\, p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right),$$

nous avons

$$\frac{1}{m_t}\, v_t \in B_{T\Theta}^*.$$

Or, d'après le Corollaire 10.11, pour tout $t \geq 0$, $\mu\, \tilde{\eta}_t \leq \eta_{\max}$ donc, d'après le Lemme 7.10, de même qu'à la preuve du Fait 9.18, pour tout $k \geq 0$, pour tout $T_k \leq t < T_{k+1}$, (en posant $v'_s = 0$ pour $T_k \leq s \leq t$),

$$d\left(\theta_t,\, \theta_{T_k}\right) \leq M_4 \sum_{s=T_k+1}^{t} \tilde{\eta}_s \leq M_4 \sum_{t=T_k}^{T_{k+1}-1} \tilde{\eta}_s = M_4 \sum_{I_k} \tilde{\eta}_s.$$

Soit alors $\varepsilon > 0$. D'après le Corollaire 6.35,

$$\sum_{I_k} \tilde{\eta}_s$$

tend vers 0 avec k, donc nous pouvons trouver $K \geq 0$ tel que, pour $k \geq K$,

$$\sum_{I_k} \tilde{\eta}_s \leq \varepsilon.$$

Soit alors $k \geq K$. Pour tout

$$T_k \leq t < T_{k+1},$$

$$d\left(\theta_t,\, \theta_{T_k}\right) \leq M_4 \sum_{I_k} \tilde{\eta}_s \leq M_4\, \varepsilon.$$

Ainsi, pour tout $k \geq K$,

$$\sup_{T_k \leq t < T_{k+1}} d\left(\theta_t,\, \theta_{T_k}\right) \leq M_4\, \varepsilon,$$

ce qui conclut la preuve. □

Fait 10.21 (Convergence de l'algorithme RTRL)**.** *θ_t tend vers θ^* quand t tend vers l'infini.*

Démonstration. Soit $\varepsilon > 0$. D'après le Lemme 10.20, nous pouvons trouver $k_3 \geq 0$ tel que, pour $k \geq k_3$,

$$\sup_{s \in I_k} d\left(\theta_s, \theta_{T_k}\right) \leq \varepsilon.$$

D'après le Fait 10.19, quitte à augmenter k_3, nous pouvons également supposer que, pour $k \geq k_3$,

$$d\left(\theta_{T_k}, \theta^*\right) \leq \varepsilon.$$

Soit alors $t \geq T_{k_3}$. Soit $k \geq T_{k_3}$ tel que

$$t \in I_k.$$

Alors,

$$\begin{aligned} d\left(\theta_t, \theta^*\right) &\leq d\left(\theta_t, \theta_{T_k}\right) + d\left(\theta_{T_k}, \theta^*\right) \\ &\leq \sup_{s \in I_k} d\left(\theta_s, \theta_{T_k}\right) + d\left(\theta_{T_k}, \theta^*\right) \\ &\leq \varepsilon + \varepsilon. \end{aligned}$$

Nous avons donc établi la convergence vers 0 de la suite de terme général

$$d\left(\theta_t, \theta^*\right),$$

et l'algorithme RTRL produit bien ainsi une suite de paramètres convergeant vers le paramètre optimal. □

10.4 Énoncé du théorème de convergence

Théorème 10.22 (Convergence de l'algorithme RTRL)**.** *Supposons vérifiées les Hypothèses 4.28, 4.29 et 4.30 sur le système dynamique, les Hypothèses 5.3 et 5.4 sur les pertes, l'Hypothèse 7.1 sur l'opérateur de déplacement sur Θ, l'Hypothèse 6.17 sur les gradients et les hessiennes des pertes sur le paramètre ainsi que sur les pertes, l'Hypothèse 6.18 sur la suite de pas et l'Hypothèse 8.16 sur le bruit. Soient $\theta_0 \in B^*_\Theta$, tel que de plus*

$$d\left(\theta_0, \theta^*\right) \leq \frac{r^*_\Theta}{4},$$

$e_0 \in B^*_{\mathcal{E}_0}$, et $J_0 \in B^*_{\mathcal{L}(\Theta, \mathcal{E}_0)}$. Alors, nous pouvons trouver $\mu_{\max} > 0$ tel que, pour tout $\mu \leq \mu_{\max}$, la suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ produite par l'algorithme RTRL qui utilise la suite de pas $\mu\,\boldsymbol{\eta}$ converge vers le paramètre optimal θ^*.

Corollaire 10.23 (Convergence des états et des différentielles vers les états et différentielles optimaux)**.** *Sous les mêmes hypothèses que le Théorème 10.22, notons $\boldsymbol{\theta}$ la suite de paramètres, et (e_t) la suites d'états, produits par l'algorithme RTRL. Alors,*

$$d\left(e_t, e^*_t\right) \to 0 \quad \textit{et} \quad d\left(\boldsymbol{J}_t\left(e_0, J_0, \boldsymbol{\theta}\right), J^*_t\right) \to 0,$$

quand t tend vers l'infini.

Remarque 10.24. *Il n'y a pas de raison que la distance entre les J_t produites par l'algorithme RTRL bruité et les différentielles optimales tende vers 0, car les J_t sont bruitées.*

Démonstration. Pour tout $t \geq 0$,

$$e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta_t\right) \quad \text{et} \quad e^*_{t+1} = \boldsymbol{T}_t\left(e^*_t,\, \theta^*\right).$$

Or, pour tout $t \geq 0$, e_t et e^*_t sont dans $B^*_{\mathcal{E}_t}$, donc dans $B_{\mathcal{E}_t}$. De plus, θ_t et θ^* sont dans B^*_Θ. Ainsi, d'après le Lemme 4.41, pour tout $t \geq 0$,

$$\left\|\boldsymbol{T}_t\left(e_t,\, \theta_t\right) - \boldsymbol{T}_t\left(e^*_t,\, \theta^*\right)\right\| \leq (1-\alpha)\, \left\|e_t - e^*_t\right\| + M_1\, d\left(\theta_t,\, \theta^*\right),$$

c'est-à-dire

$$d\left(e_{t+1},\, e^*_{t+1}\right) \leq (1-\alpha)\, d\left(e_t,\, e^*_t\right) + M_1\, d\left(\theta_t,\, \theta^*\right).$$

Soit $\varepsilon > 0$. D'après le Théorème 10.22, θ_t tend vers θ^* quand t tend vers l'infini, donc nous pouvons trouver $T \geq 0$ tel que, pour tout $t \geq T$,

$$M_1\, d\left(\theta_t,\, \theta^*\right) \leq \varepsilon.$$

Ainsi, pour tout $t \geq T$,

$$d\left(e_{t+1},\, e^*_{t+1}\right) \leq (1-\alpha)\, d\left(e_t,\, e^*_t\right) + \varepsilon.$$

Par conséquent, pour tout $t \geq T+1$,

$$\begin{aligned} d\left(e_t,\, e^*_t\right) &\leq (1-\alpha)^{t-T}\, d\left(e_T,\, e^*_T\right) + \varepsilon \sum_{s=T}^{t-1} (1-\alpha)^{t-T} \\ &\leq (1-\alpha)^{t-T}\, r^*_{\mathcal{E}} + \frac{\varepsilon}{\alpha}. \end{aligned}$$

La quantité majorante converge vers $\frac{\varepsilon}{\alpha}$ quand t tend vers l'infini, donc est inférieure à $\frac{2\,\varepsilon}{\alpha}$ pour t suffisamment grand. Ainsi, nous pouvons trouver $T_1 \geq T$ tel que, pour tout $T \geq T_1$,

$$d\left(e_t,\, e^*_t\right) \leq \frac{2\,\varepsilon}{\alpha}.$$

Nous avons démontré :

$$d\left(e_t,\, e^*_t\right) \to 0,$$

quand t tend vers l'infini. Nous procédons alors de la même manière pour prouver la convergence vers 0 de

$$d\left(\boldsymbol{J}_t\left(e_0,\, J_0,\, \boldsymbol{\theta}\right),\, J^*_t\right),$$

en utilisant le Lemme 4.44. Nous obtenons ainsi le résultat annoncé. □

Deuxième partie

Convergence de l'algorithme « NoBackTrack »

Sommaire de la Partie II

Introduction

L'algorithme « NoBackTrack » a été introduit dans l'article "Training recurrent networks online without backtracking" [1]. Ainsi que nous le disions en introduction, il préserve la propriété d'être en ligne de l'algorithme RTRL, mais requiert une capacité mémoire bien plus faible. En effet, l'algorithme RTRL maintient une différentielle (ou une approximation d'une différentielle) de taille la dimension de l'espace des états multipliée par la dimension du paramètre, ce qui est prohibitif dès que le système d'apprentissage est moyennement grand (dans l'échelle des systèmes utilisés actuellement).

L'algorithme « NoBackTrack » réduit drastiquement le coût en mémoire en maintenant une estimation de rang 1 de la différentielle mentionnée. L'estimée est calculée de manière probabiliste, de manière à ce que la mise à jour de la différentielle soit une estimée non biaisée de la mise à jour RTRL. Nous prouvons ci-dessous que le recours à cette estimation ne modifie pas le comportement asymptotique de l'algorithme, par rapport à celui de RTRL.

Une version plus simple algorithmiquement de « NoBackTrack » a été produite, appelée « UORO », pour « Unbiased Online Recurrent Optimization » [2]. La preuve que nous présentons, même si elle traite le cas « NoBackTrack », est presque immédiatement transposable à « UORO », de sorte que les garanties qu'elle apporte sont également valables pour ce dernier.

1. Ollivier, Tallec et Charpiat, "Training recurrent networks online without backtracking", art. cit.
2. Tallec et Ollivier, "Unbiased Online Recurrent Optimization", art. cit.

Présentation de la preuve

L'algorithme « NoBackTrack » est appliqué au même système dynamique que l'algorithme RTRL. Il utilise les mêmes pertes, et le même opérateur de mise à jour du paramètre. Il remplace les estimées des différentielles des états par rapport au paramètre utilisées par RTRL, par des estimées aléatoires de rang un. Pour cela, l'algorithme « NoBackTrack » maintient un couple de vecteurs que nous appelons « vecteurs spécifiques à « NoBackTrack » », dont le produit tensoriel fournit l'approximation de rang un. Ces vecteurs sont mis à jour grâce à un opérateur que nous appelons opérateur de réduction. Celui-ci est construit de telle sorte que la mise à jour de l'approximation de la différentielle utilisée par « NoBackTrack » est égale, en moyenne sur le choix d'un ensemble de signes aléatoires, à la mise à jour utilisée par l'algorithme RTRL.

La différence entre la mise à jour de la différentielle utilisée par RTRL et celle de l'estimée calculée par « NoBackTrack » peut ainsi s'écrire sous la forme d'un bruit perturbant la mise à jour des différentielles dans RTRL. « NoBackTrack » s'interprète donc comme un algorithme RTRL bruité. Les calculs algébriques impliquant le bruit de la preuve de convergence de l'algorithme RTRL sont donc toujours valables pour le bruit particulier produit par « NoBackTrack ». En revanche, nous ne savons pas *a priori* que celui-ci est borné.

L'objet de la preuve de convergence de « NoBackTrack » est ainsi d'établir que la perturbation aléatoire qu'il introduit par rapport à RTRL n'empêche pas la convergence. Tant que le bruit est borné, les résultats valables pour l'algorithme RTRL bruité sont directement applicables à « NoBackTrack ». Le fait que le bruit reste effectivement borné est en revanche une propriété spécifique de l'algorithme « NoBackTrack », que nous établissons au cours de la preuve. L'argument est que les écarts successifs aux mises à jour RTRL sont non corrélés, conditionnellement à l'instant courant. Ainsi, par exemple, pour un pas de descente η constant, la somme des écarts au bout d'un temps T est de l'ordre de grandeur de $\eta\sqrt{T}$, alors que le terme de contractivité est en ηT, de sorte que l'erreur (la somme des écarts) sera absorbée par celui-ci.

Afin de contrôler le terme de bruit dû à la mise à jour des différentielles de « NoBackTrack », nous effectuons quelques modifications des hypothèses d'optimalité, introduites au chapitre 6 pour l'algorithme RTRL. En effet, les pertes doivent être un peu plus petites, afin que le terme de bruit introduit par « NoBackTrack » soit

bien négligeable devant la somme des pas. Il faut de plus que les intervalles I_k soient un peu plus longs, afin que la probabilité de convergence soit arbitrairement proche de un [3].

La preuve procède ainsi en six temps, correspondant chacun à un chapitre.

1. La modification des hypothèses d'optimalité.
2. L'étude de l'opérateur de réduction.
3. L'étude des vecteurs spécifiques à « NoBackTrack ».
4. L'application de la propriété centrale à l'algorithme « NoBackTrack ».
5. La convergence sur un ensemble bien choisi de l'algorithme « NoBackTrack ».
6. Le contrôle probabiliste du bruit de « NoBackTrack ».

Les chapitres sur l'opérateur de réduction et l'étude des vecteurs spécifiques sont des chapitres techniques. Les deux chapitres suivants sont une reprise des chapitres correspondant de RTRL. En particulier, seules les modifications par rapport à RTRL sont détaillées. Enfin, le dernier chapitre est le chapitre où les arguments cruciaux qui justifient la convergence de « NoBackTrack » sont exposés.

3. Dans la preuve de convergence de RTRL, nous avions supposé que la contribution du bruit sur les intervalles I_k était négligeable devant la somme des pas. Ici, nous démontrons que cela est bien le cas, au prix d'un léger renforcement des hypothèses.

Chapitre

11 Modifications pour le changement d'échelle de temps

11.1 Jonction avec la preuve sur RTRL

Nous étudions le même système dynamique que dans la partie sur l'algorithme RTRL. Ainsi, nous supposons valables les chapitres 4 à 7. Nous effectuons toutefois quelques modifications pour le chapitre 6 : nous supposons les pertes un peu plus petites que pour la partie RTRL. Nous devons le faire afin de contrôler des termes d'erreur un peu plus gros, qui sont ceux dûs au bruit généré par « NoBackTrack ». Le chapitre courant expose alors les modifications du chapitre 6 que nous considérons. Le reste du chapitre 6 est inchangé.

Toutes les notations utilisées dans cette partie, et qui ne sont pas définies dans celle-ci, sont celles de la partie sur l'algorithme RTRL.

11.2 Modifications du contrôle des sommes des pas sur les intervalles

Définition 11.1 (Exposant pour le contrôle du bruit)**.** *Nous fixons un réel* $0 < \nu < 1$.

Remarque 11.2. *Après le chapitre courant, le réel* ν *est utilisé à partir de la Définition 15.2.*

Lemme 11.3 (Contrôle des sommes des pas sur les intervalles pour l'algorithme « NoBackTrack »)**.** *Soit une suite d'intervalles* $I_k = [g_k, d_k[$*, de longueurs* L_k*. Soit une fonction d'échelle* M_p*. Soit une suite* $\boldsymbol{\eta} = (\eta_t)$ *telle que*

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t}$$

est borné, quand k *tend vers l'infini. Nous supposons que*

$$M_p(d_k) = o\left(L_k^{\frac{1}{1+\nu}-\frac{1}{2}}\right),$$

quand k *tend vers l'infini. Alors,*

$$\frac{\left(\sum_{I_k} M_p(t)^2 \, \eta_t^2\right)^{1/2}}{\left(\sum_{I_k} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}}}$$

tend vers 0*, quand* k *tend vers l'infini.*

Remarque 11.4. *D'après la Définition 11.1,* $0 \leq \nu < 1$*, donc* $0 \leq 1+\nu < 2$ *et* $1/(1+\nu) - \frac{1}{2} > 0$.

Démonstration. Notons, pour tout $k \geq 0$,

$$i_k = \inf_{I_k} \eta_t \quad \text{et} \quad s_k = \sup_{I_k} \eta_t.$$

M_p est croissante donc, pour tout $k \geq 0$, pour tout $t \in I_k$,

$$M_p(t)\ \eta_t \leq M_p(d_k)\ s_k.$$

Ainsi, pour tout $k \geq 0$,

$$\sum_{I_k} M_p(t)^2\ \eta_t^2 \leq \left(\sup_{I_k} M_p(t)\ \eta_t\right)^2 L_k \leq M_p(d_k)^2\ s_k^2\, L_k,$$

de sorte que

$$\left(\sum_{I_k} M_p(t)^2\ \eta_t^2\right)^{1/2} \leq M_p(d_k)\, s_k\, L_k^{1/2}.$$

De plus, pour tout $k \geq 0$,

$$\sum_{I_k} \eta_t^{1+\nu} \geq i_k^{1+\nu}\, L_k,$$

de sorte que

$$\left(\sum_{I_k} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}} \geq i_k\, L_k^{\frac{1}{1+\nu}}.$$

Par conséquent, pour tout $k \geq 0$,

$$\frac{\left(\sum_{I_k} M_p(t)^2\ \eta_t^2\right)^{1/2}}{\left(\sum_{I_k} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}}} \leq \frac{M_p(d_k)}{L_k^{\frac{1}{1+\nu}-\frac{1}{2}}}\ \frac{s_k}{i_k}.$$

Or, par hypothèse, s_k/i_k est borné, et

$$\frac{M_p(d_k)}{L_k^{\frac{1}{1+\nu}-\frac{1}{2}}}$$

tend vers 0, quand k tend vers l'infini. Par conséquent,

$$\frac{\left(\sum_{I_k} M_p(t)^2\ \eta_t^2\right)^{1/2}}{\left(\sum_{I_k} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}}}$$

tend vers 0, quand k tend vers l'infini, ce qui est le résultat annoncé. □

Lemme 11.5 (Convergence d'une série pour le contrôle de la probabilité de convergence)**.** *Soit une suite d'intervalles* $I_k = [g_k,\, d_k[$*, de longueurs* L_k*. Soit une suite de pas* (η_t) *telle que le rapport*

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k}\ \eta_t}$$

est borné, quand k tend vers l'infini. Nous supposons que la série de terme général

$$\frac{1}{L_k^{\frac{\nu}{1+\nu}}}$$

est convergente. Alors, la série de terme général

$$\frac{\left(\sum_{I_k} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}}}{\sum_{I_k} \eta_t}$$

converge également.

Démonstration. Pour tout $k \geq 0$, notons

$$i_k = \inf_{I_k} \eta_t \quad \text{et} \quad s_k = \sup_{I_k} \eta_t.$$

Soit alors $k \geq 0$. Pour tout $t \in I_k$,

$$\eta_t^{1+\nu} \leq s_k^{1+\nu},$$

donc

$$\sum_{I_k} \eta_t^{1+\nu} \leq s_k^{1+\nu} L_k.$$

Or,

$$i_k L_k \leq \sum_{I_k} \eta_t,$$

et ainsi

$$\frac{\left(\sum_{I_k} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}}}{\sum_{I_k} \eta_t} \leq \frac{s_k}{i_k} \frac{1}{L_k^{\frac{\nu}{1+\nu}}}.$$

Or, par hypothèse, s_k/i_k est borné, et la quantité majorante est le terme général d'une série convergente, ce qui conclut la preuve. □

11.3 Critère d'optimalité

11.3.1 Modifications des hypothèses pour obtenir l'optimalité

Nous renforçons légèrement les hypothèses formulées pour RTRL. Les pertes doivent être un peu plus petites, et les taux d'apprentissage aussi. Les nouvelles hypothèses sont les suivantes.

Remarque 11.6. *Par rapport à RTRL, seuls le troisième point de l'Hypothèse 6.17 et le deuxième point de l'Hypothèse 6.18 sont modifiés.*

Hypothèse 11.7 (Hypothèses sur les gradients, les hessiennes et les pertes pour « NoBackTrack »)**.** *Supposons disposer de fonctions d'échelle ℓ^1, ℓ^2 et M_p, négligeables devant l'identité en l'infini, telles que les propriétés suivantes sont vérifiées.*

1. Critère d'optimalité 1/2.

$$\sum_{t=0}^{\ell^1(t)} \frac{\partial\, p_t \circ g_t}{\partial \theta} (e_0^*,\, \theta^*) = O\left(\ell^1(t)\right),$$

quand t tend vers l'infini.

2. Critère d'optimalité 2/2. *Nous disposons d'un réel* $\tilde{\lambda} > 0$ *tel que, pour* t *suffisamment grand, la plus petite valeur propre des opérateurs*

$$\frac{1}{\ell^2(t)} \sum_{s=t}^{t+\ell^2(t)} \frac{\partial^2 p_s \circ g_s}{\partial \theta^2}(e_0^*, \theta^*)$$

est supérieure à $\tilde{\lambda}$.

3. $$\max\left(\ell^1(t), \ell^2(t), M_p(t)^{\left(\frac{1}{1+\nu}-\frac{1}{2}\right)^{-1}}\right) M_p(t)^2$$

est négligeable devant t, *quand* t *tend vers l'infini.*

4. Contrôle des pertes le long de la trajectoire optimale. *Nous supposons que*

$$\left\| \frac{\partial p_t \circ g_t}{\partial \theta}(e_0^*, \theta^*) \right\| = O\left(M_p(t)\right),$$

quand t *tend vers l'infini.*

Hypothèse 11.8 (Suite de pas de descente pour « NoBackTrack »)**.** *Nous supposons disposer d'une suite de pas* $\boldsymbol{\eta} = (\eta_t)$ *qui vérifie les propriétés suivantes.*

1. *La série de terme général* η_t *diverge.*
2. $$\eta_t = o\left(\frac{1}{\max\left(\ell^1(t), \ell^2(t), M_p(t)^{\left(\frac{1}{1+\nu}-\frac{1}{2}\right)^{-1}}\right) M_p(t)^2}\right),$$

quand t *tend vers l'infini.*

3. *(Homogénéité temporelle) Pour toute suite d'intervalles* $I_t = [g_t, d_t[$ *tels que* g_t *tend vers l'infini et* $g_t \sim d_t$, *quand* t *tend vers l'infini, nous avons*

$$\frac{\sup_{I_t} \eta_s}{\inf_{I_t} \eta_s} = 1 + o\left(\frac{1}{M_p(d_t)}\right),$$

quand t *tend vers l'infini.*

4. *La suite de terme général* $\eta_t^{-1} M_p(t)^{-2}$ *est une fonction d'échelle.*

11.3.2 Changement d'échelle de temps : construction des intervalles pour la convergence

Lemme 11.9 (Fonction longueur d'intervalles adaptée pour « NoBackTrack »)**.** *Nous pouvons choisir une fonction d'échelle* L, *devant laquelle* ℓ^1, ℓ^2 *et* $M_p^{\left(\frac{1}{1+\nu}-\frac{1}{2}\right)^{-1}}$ *sont négligeables, et négligeable devant* $\frac{1}{\eta_t M_p(t)^2}$, *quand* t *tend vers l'infini.*

Démonstration. La preuve est identique à celle du Lemme 6.21. □

Hypothèse 11.10 (Convergence de la série des $1/L(T_k)^{\frac{\nu}{1+\nu}}$)**.** *Nous supposons que la série de terme général*

$$\frac{1}{L(T_k)^{\frac{\nu}{1+\nu}}}$$

est convergente.

Remarque 11.11. *L'hypothèse sert à démontrer le Lemme 11.12 ci-dessous. Il faut que les intervalles de temps soient suffisamment long pour que le bruit de « NoBackTrack » soit négligeable devant la somme des pas. Les exemples numériques donnés au Lemme 11.15 et le Lemme 11.17 montrent que son obtention n'est pas très contraignante.*

Dans la suite, nous notons, pour $k \geq 0$,

$$u_k = M_7 \left(\sum_{I_k} \eta_t \right)^{-1} \left(\sum_{I_k} \eta_t^{1+\nu} \right)^{\frac{1}{1+\nu}}.$$

Ces termes apparaissent au Lemme 16.14.

Lemme 11.12 (Convergence du produit infini des $1 - u_k$)**.** *Le produit infini des $1 - u_k$ est convergent.*

Démonstration. D'après le Corollaire 6.28,

$$\frac{\sup_{I_k} \eta_t}{\inf_{I_k} \eta_t}$$

tend vers 1, et est en particulier borné. D'après l'Hypothèse 11.10, la série de terme général

$$\frac{1}{L(T_k)^{\frac{\nu}{1+\nu}}}$$

est convergente. Par conséquent, d'après le Lemme 11.5, la série de terme général u_k est convergente. Ceci implique la convergence du produit infini des $1 - u_k$. □

Corollaire 11.13 (Produit infini arbitrairement proche de 1)**.** *Pour tout $\varepsilon > 0$, nous pouvons trouver $K \geq 0$ tel que, pour tout $k \geq K$,*

$$\prod_{k'=k}^{\infty} (1 - u_{k'}) \geq 1 - \varepsilon.$$

Démonstration. C'est une conséquence du Lemme 11.12 précédent. □

11.3.3 Échelle de temps et pas de descente

Corollaire 11.14 (Échelle de temps et pas de descente renormalisé pour l'algorithme « NoBackTrack »)**.**

$$\frac{\left(\sum_{I_k} \tilde{\eta}_t^2 \right)^{1/2}}{\left(\sum_{I_k} \eta_t^{1+\nu} \right)^{\frac{1}{1+\nu}}}$$

tend vers 0, lorsque k tend vers l'infini.

Démonstration. La démonstration est identique à celle du Corollaire 6.34, sauf pour deux points. D'une part, le Lemme 6.21 est remplacé par le Lemme 11.9. En particulier, la propriété

$$M_p(t) = o(L(t)),$$

quand t tend vers l'infini, est remplacée par la propriété

$$M_p(t)^{\left(\frac{1}{1+\nu} - \frac{1}{2}\right)^{-1}} = o(L(t)),$$

quand t tend vers l'infini. Or, d'après la Définition 11.1, $0 \leq \nu < 1$, donc $1/(1+\nu) - 1/2 > 0$ et ainsi, en élevant à la puissance $1/(1+\nu) - 1/2$ l'inégalité précédente, nous obtenons

$$M_p(t) = o\left(L(t)^{\frac{1}{1+\nu}-\frac{1}{2}}\right),$$

quand t tend vers l'infini. D'autre part, le Lemme 6.16 est remplacé par le Lemme 11.3. □

11.4 Exemples numériques satisfaisant les hypothèses

Lemme 11.15 (Jeu de paramètres satisfaisant les hypothèses pour l'optimalité pour « NoBackTrack »)**.** *Soient des réels positifs a_1, a_2, γ, b et a'. Alors, les fonctions ℓ^1, ℓ^2, M_p, la suite (η_t) et la fonction L définies par*

$$\ell^1(t) = t^{a_1}, \quad \ell^2(t) = t^{a_2}, \quad M_p(t) = t^{\gamma},$$
$$\eta_t = t^{-b} \quad et \quad L(t) = t^{a'}$$

satisfont les contraintes spécifiques aux fonctions (les critères d'optimalité 1/2 et 2/2 restent bien sûr en hypothèse) de l'Hypothèse 11.7, de l'Hypothèse 11.8 et du Lemme 11.9 si les conditions suivantes sont vérifiées :

1. $0 \leq a_1, a_2 < 1$;
2. $\left(2 + \left(\frac{1}{1+\nu} - \frac{1}{2}\right)^{-1}\right) \gamma < 1$;
3. $\max\left(a_1, a_2, \left(\frac{1}{1+\nu} - \frac{1}{2}\right)^{-1} \gamma\right) + 2\gamma < b \leq 1$
4. *et* $\max\left(a_1, a_2, \left(\frac{1}{1+\nu} - \frac{1}{2}\right)^{-1} \gamma\right) < a' < b - 2\gamma$.

Remarque 11.16. *Pour $\nu = 0$, $(1/(1+\nu) - 1/2)^{-1} = 4$, et nous demandons donc que $4\gamma < 1$. Pour $\nu > 0$, cette quantité est strictement supérieure à 4.*

Démonstration. La preuve est identique à celle du Lemme 6.43. Seules les conditions sur γ sont modifiées (le 3 du deuxième point est remplacé par $2+(1/(1+\nu) - 1/2)^{-1}$, et les γ entre parenthèses des points trois et quatre sont remplacés par $(1/(1+\nu) - 1/2)^{-1}$). □

Lemme 11.17 (Convergence de la somme des $1/L(T_k)^{\frac{\nu}{1+\nu}}$)**.** *Pour $a' > \left(1 + \frac{\nu}{1+\nu}\right)^{-1}$, la série de terme général*

$$\frac{1}{L(T_k)^{\frac{\nu}{1+\nu}}} = \frac{1}{{T_k}^{a'\frac{\nu}{1+\nu}}}$$

est convergente.

Démonstration. En effet, pour tout $k \geq 0$, à une constante multiplicative près, T_k est équivalent à

$$k^{\frac{1}{1-a'}},$$

quand k tend vers l'infini. Donc, $L(T_k)^{\frac{\nu}{1+\nu}} = {T_k}^{a'\frac{\nu}{1+\nu}}$ est équivalent, à une constante multiplicative près, à

$$k^{\frac{a'}{1-a'}\frac{\nu}{1+\nu}},$$

quand k tend vers l'infini. Or, pour $a' > \left(1 + \frac{\nu}{1+\nu}\right)^{-1}$, l'exposant est strictement supérieur à 1, ce qui conclut la preuve. □

Pour $0 < \nu < 1$,

$$\frac{2}{3} < \left(1 + \frac{\nu}{1+\nu}\right)^{-1} < 1.$$

Chapitre

12 Opérateur d'égalisation des normes, opérateur de réduction et produit tensoriel

12.1 Opérateur d'égalisation des normes sur des espaces vectoriels normés

12.1.1 Opérateur d'égalisation des normes

Définition 12.1 (Espaces de travail)**.** *Dans ce chapitre, nous considérons trois espaces vectoriels réels, $\mathcal{E}_1$, $\mathcal{E}_2$ et $\mathcal{E}_3$, de dimension finie.*

Pour $1 \leq i \leq \dim \mathcal{E}_3$, nous notons $\mathfrak{e}_i$ le i-ème vecteur de la base de travail sur $\mathcal{E}_3$.

Ces trois espaces sont munis des normes euclidiennes relatives aux bases de travail.

De même que dans la partie sur l'algorithme RTRL, pour des produits cartésiens comme $\mathcal{E}_1 \times \mathcal{E}_2$ nous notons, pour $v_1 \in \mathcal{E}_1$ et $v_2 \in \mathcal{E}_2$,

$$\|(v_1,\, v_2)\| = \left(\|v_1\|^2 + \|v_2\|^2\right)^{1/2}.$$

Nous utiliserons de plus souvent la majoration

$$\|v_1\| = \sqrt{\|v_1\|^2} \leq \sqrt{\|v_1\|^2 + \|v_2\|^2},$$

qui implique l'inégalité

$$\|v_1\| + \|v_2\| \leq 2\, \|(v_1,\, v_2)\|\,.$$

Définition 12.2 (Opérateur d'égalisation des normes)**.** *Nous définissons l'opérateur*

$$\begin{aligned} \odot :\ & \mathcal{E}_1 \times \mathcal{E}_2 \to \mathcal{E}_1 \times \mathcal{E}_2 \\ & v_1,\, v_2 \mapsto (v_1 \odot v_2)\,, \end{aligned}$$

tel que

$$(v_1 \odot v_2) = \begin{cases} \left(\sqrt{\dfrac{\|v_2\|}{\|v_1\|}}\, v_1,\, \sqrt{\dfrac{\|v_1\|}{\|v_2\|}}\, v_2\right) & \textit{si} \quad v_1 \neq 0 \quad \textit{et} \quad v_2 \neq 0, \\ (0,\, 0) & \textit{sinon.} \end{cases}$$

Définition 12.3 (Norme du maximum sur un produit cartésien)**.** *Pour $(v_1,\, v_2) \in \mathcal{E}_1 \times \mathcal{E}_2$, nous posons*

$$\|(v_1,\, v_2)\|_{max} = \max\left(\|v_1\|\,,\, \|v_2\|\right).$$

Remarque 12.4. *Chaque membre des couples de la forme* $(v_1 \odot v_2)$ *a la même norme. Ainsi,* $\|(v_1 \odot v_2)\|_{max}$ *est la norme de chacun des membres.*

Lemme 12.5 (Norme de l'image de vecteurs par l'opérateur d'égalisation des normes)**.** *Pour tous* $v_1 \in \mathcal{E}_1$ *et* $v_2 \in \mathcal{E}_2$*,*

$$\|(v_1 \odot v_2)\|_{max} = \sqrt{\|v_1\| \, \|v_2\|}.$$

Démonstration. Cette égalité est vérifiée pour tout (v_1, v_2) distinct de l'origine. Elle l'est également en $(0, 0)$, ce qui conclut la preuve. □

12.1.2 Continuité de l'opérateur d'égalisation des normes

Définition 12.6 (Fonctions auxiliaires qui décomposent la première coordonnée de l'opérateur d'égalisation des normes)**.** *Nous définissons les fonctions*

$$\begin{aligned} f_1 : \quad & \mathcal{E}_1 \to \mathcal{E}_1 \\ & v_1 \mapsto f_1(v_1), \end{aligned}$$

telle que

$$f_1(v_1) = \begin{cases} \dfrac{v_1}{\sqrt{\|v_1\|}} & si \quad v_1 \neq 0 \\ 0 & sinon, \end{cases}$$

et

$$\begin{aligned} f_2 : \quad & \mathcal{E}_2 \to \mathcal{E}_2 \\ & v_2 \mapsto \sqrt{\|v_2\|}. \end{aligned}$$

Lemme 12.7 (Décomposition de la première coordonnée de l'opérateur d'égalisation des normes)**.** *La première coordonnée de l'opérateur d'égalisation des normes coïncide avec l'application qui, à tout* $(v_1, v_2) \in \mathcal{E}_1 \times \mathcal{E}_2$*, associe*

$$f_1(v_1) \; f_2(v_2).$$

Démonstration. Soit $(v_1, v_2) \in \mathcal{E}_1 \times \mathcal{E}_2$. Le résultat s'obtient en distinguant le cas où aucun vecteur n'est nul, celui où v_2 est nul et celui où v_1 est nul. □

Lemme 12.8 (Continuité de l'opérateur d'égalisation des normes)**.** *L'opérateur d'égalisation des normes est continu.*

Démonstration. f_2 est continue. f_1 est continue sur $\mathcal{E}_1$ privé de 0. Elle l'est également en 0 en tant que fonction homogène de degré $1/2 > 0$, et est donc continue sur $\mathcal{E}_1$. Ainsi, la première coordonnée de l'opérateur d'égalisation des normes est continue. De même, sa deuxième coordonnée l'est, ce qui conclut la preuve. □

12.2 Opérateur de réduction

12.2.1 Opérateur de réduction

Définition 12.9 (Suite de signes)**.** *Dans la suite de ce chapitre, nous considérons une suite finie* $\boldsymbol{\varepsilon}$ *de signes*

$$\varepsilon_i, \quad 1 \leq i \leq \dim \mathcal{E}_3.$$

Définition 12.10 (Opérateurs multiplicatifs et additifs)**.** *Dans la suite de ce chapitre, nous considérons une application linéaire*

$$a : \mathcal{E}_1 \to \mathcal{E}_3.$$

Nous considérons également une application linéaire

$$b : \mathcal{E}_2 \to \mathcal{E}_3,$$

dont les lignes dans sa matrice représentative dans les bases de travail des espaces $\mathcal{E}_2$ *et* $\mathcal{E}_3$ *sont les*

$$b_i, \quad 1 \leq i \leq \dim \mathcal{E}_3.$$

Les b_i *sont des éléments du dual de* $\mathcal{E}_2$*. Nous identifions celui-ci à* $\mathcal{E}_2$*, en considérant sur ce dernier la métrique égale à l'identité. Ainsi, nous pouvons considérer les* b_i *comme des éléments de* $\mathcal{E}_2$*.*

Définition 12.11 (Opérateur de réduction)**.** *Nous définissons l'opérateur de réduction*

$$\begin{aligned} \mathcal{R}(\cdot, \cdot\,; a, b, \boldsymbol{\varepsilon}) : \quad & \mathcal{E}_1 \times \mathcal{E}_2 \to \mathcal{E}_3 \times \mathcal{E}_2 \\ & v_1, v_2 \mapsto \mathcal{R}(v_1, v_2\,; a, b, \boldsymbol{\varepsilon}), \end{aligned}$$

tel que

$$\mathcal{R}(v_1, v_2\,; a, b, \boldsymbol{\varepsilon}) = (a(v_1) \odot v_2) + \sum_{i=1}^{\dim \mathcal{E}_3} \varepsilon_i \, (\mathfrak{e}_i \odot b_i).$$

Lemme 12.12 (Absence de biais de l'opérateur de réduction)**.** *Supposons que les signes* ε_i*, pour* $1 \leq i \leq \dim \mathcal{E}_3$*, sont des variables de Bernoulli indépendantes prenant les valeurs* 1 *et* -1 *avec probabilité* $1/2$*. Supposons que toutes les autres quantités sont mesurables pour la tribu engendrée par ces variables. Notons*

$$(w_1, w_2) = \mathcal{R}(v_1, v_2\,; a, b, \boldsymbol{\varepsilon}).$$

Alors,

$$\mathrm{E}[w_1 \otimes w_2] = a(v_1) \otimes v_2 + b,$$

où l'espérance est prise par rapport à la loi des variables de Bernoulli.

Démonstration. Nous savons que, pour

$$\rho = \sqrt{\frac{\|v_2\|}{\|a(v_1)\|}}$$

et, pour $1 \leq i \leq \dim \mathcal{E}_3$,

$$\rho_i = \sqrt{\frac{\|b_i\|}{\|\mathfrak{e}_i\|}},$$

$$\begin{cases} w_1 = \rho \, a(v_1) + \displaystyle\sum_{i=1}^{\dim \mathcal{E}_3} \varepsilon_i \, \rho_i \, \mathfrak{e}_i \\ w_2 = \rho^{-1} \, v_2 + \displaystyle\sum_{i=1}^{\dim \mathcal{E}_3} \varepsilon_i \, \rho_i^{-1} \, b_i. \end{cases}$$

Donc, d'après le « coup du rang 1 » de l'article présentant l'algorithme « NoBack-Track » [1],

$$\mathrm{E}\left[w_1 \otimes w_2\right] = a\left(v_1\right) \otimes v_2 + \sum_{i=1}^{\dim \mathcal{E}_3} \mathfrak{e}_i \otimes b_i.$$

Or, d'après la Définition 12.10, la somme dans le membre de droite est égale à b, et nous obtenons ainsi le résultat annoncé. □

12.2.2 Majoration des images de l'opérateur de réduction

Corollaire 12.13 (Majoration de l'image de l'opérateur de réduction)**.** *Soient $v_1 \in \mathcal{E}_1$ et $v_2 \in \mathcal{E}_2$. Notons*

$$(w_1,\, w_2) = \mathcal{R}\left(v_1,\, v_2\,;a,\, b,\, \boldsymbol{\varepsilon}\right).$$

Alors,

$$\left\|(w_1,\, w_2)\right\|_{max} \leq \|a\|_{\mathrm{op}}^{1/2} \left(\|v_1\|\, \|v_2\|\right)^{1/2} + \sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2}.$$

Démonstration.

$$\left\|(w_1,\, w_2)\right\|_{\max} \leq \left\|\left(a\left(v_1\right) \odot v_2\right)\right\|_{\max} + \sum_{i=1}^{\dim \mathcal{E}_3} \left\|\left(\mathfrak{e}_i \odot b_i\right)\right\|_{\max}.$$

Or, d'après le Lemme 12.5,

$$\begin{aligned}\left\|\left(a\left(v_1\right) \odot v_2\right)\right\|_{\max} &= \left(\left\|a\left(v_1\right)\right\|\, \|v_2\|\right)^{1/2} \\ &\leq \|a\|_{\mathrm{op}}^{1/2} \left(\|v_1\|\, \|v_2\|\right)^{1/2},\end{aligned}$$

et

$$\left\|\left(\mathfrak{e}_i \odot b_i\right)\right\|_{\max} = \left(\|\mathfrak{e}_i\|\, \|b_i\|\right) = \left(\|b_i\|\right)^{1/2},$$

car la famille des $\mathfrak{e}_i$ est orthonormée par construction, d'après la Définition 12.1. □

Corollaire 12.14 (Majoration du produit des normes des vecteurs produits par l'opérateur de réduction)**.** *Soient $v_1 \in \mathcal{E}_1$ et $v_2 \in \mathcal{E}_2$. Notons*

$$(w_1,\, w_2) = \mathcal{R}\left(v_1,\, v_2\,;a,\, b,\, \boldsymbol{\varepsilon}\right).$$

Alors,

$$\begin{aligned}\|w_1\|\, \|w_2\| \leq \|a\|_{\mathrm{op}}\, \|v_1\|\, \|v_2\| + 2\, \|a\|_{\mathrm{op}}^{1/2} \left(\|v_1\|\, \|v_2\|\right)^{1/2} \sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2} \\ + \left(\sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2}\right)^2.\end{aligned}$$

Démonstration.

$$\|w_1\|\, \|w_2\| \leq \|(w_1,\, w_2)\|_{\max}^2.$$

1. Ollivier, Tallec et Charpiat, "Training recurrent networks online without backtracking", art. cit.

Or, d'après le Corollaire 12.13,

$$\|(w_1,\, w_2)\|_{\max} \leq \|a\|_{\mathrm{op}}^{1/2}\, (\|v_1\|\, \|v_2\|)^{1/2} + \sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2}\,.$$

Ainsi,

$$\begin{aligned}\|w_1\|\, \|w_2\| &\leq \left(\|a\|_{\mathrm{op}}^{1/2}\, (\|v_1\|\, \|v_2\|)^{1/2} + \sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2} \right)^2 \\ &= \|a\|_{\mathrm{op}}\, \|v_1\|\, \|v_2\| + 2\, \|a\|_{\mathrm{op}}^{1/2}\, (\|v_1\|\, \|v_2\|)^{1/2} \sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2} \\ &\quad + \left(\sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2} \right)^2 .\end{aligned}$$

□

Lemme 12.15 (Intervalle stable par l'application majorant le produit des normes des images de l'opérateur de réduction)**.** *Soient des constantes* $0 \leq \gamma_1 < 1$ *et* $\gamma_2 \geq 0$. *Considérons la fonction*

$$\begin{aligned} g_{\gamma_1,\, \gamma_2} :\ \ & \mathbb{R}_+ \to \mathbb{R}_+ \\ & x \mapsto g_{\gamma_1,\, \gamma_2}(x)\,, \end{aligned}$$

telle que

$$g_{\gamma_1,\, \gamma_2}(x) = \gamma_1\, x + 2\, x\, \gamma_1^{1/2}\, \gamma_2 + \gamma_2^2.$$

Notons

$$r(\gamma_1,\, \gamma_2) = \gamma_2\, \frac{\gamma_1 + 2\, \gamma_1{}^{1/2} + 1}{\gamma_1{}^2 - 2\, \gamma_1 + 1}.$$

Alors, le segment

$$[0,\, r(\gamma_1,\, \gamma_2)]$$

est stable par $g_{\gamma_1,\, \gamma_2}$.

Démonstration. $r(\gamma_1,\, \gamma_2)$ est l'unique point fixe de l'application $g_{\gamma_1,\, \gamma_2}$. (Si γ_2 est nul, le segment stable est réduit au singleton $\{0\}$.) □

Corollaire 12.16 (Zone sous hyperbole stable par l'opérateur de réduction)**.** *Soient* $v_1 \in \mathcal{E}_1$ *et* $v_2 \in \mathcal{E}_2$. *Notons*

$$(w_1,\, w_2) = \mathcal{R}(v_1,\, v_2\,;a,\, b,\, \boldsymbol{\varepsilon})\,.$$

Supposons disposer de deux réels $0 \leq M_{mul} < 1$ *et* $M_{add} \geq 0$ *tels que*

$$\|a\|_{\mathrm{op}} \leq M_{mul},$$

et

$$\sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2} \leq M_{add}.$$

Supposons de plus que

$$\|v_1\|\, \|v_2\| \leq r(M_{mul},\, M_{add})\,.$$

Alors,

$$\|w_1\|\, \|w_2\| \leq r(M_{mul},\, M_{add})\,.$$

Démonstration. D'après le Corollaire 12.14 et les hypothèses du lemme,

$$\begin{aligned}\|w_1\| \, \|w_2\| &\leq g_{\|a\|_{\mathrm{op}}, \sum_{i=1}^{\dim \mathcal{E}_3} \|b_i\|^{1/2}} \left(\|v_1\| \, \|v_2\|\right) \\ &\leq g_{M_{\mathrm{mul}}, M_{\mathrm{add}}} \left(\|v_1\| \, \|v_2\|\right).\end{aligned}$$

Le Lemme 12.15 permet alors de conclure. □

12.3 Produit tensoriel

Définition 12.17 (Produit tensoriel des espaces)**.** *Nous considérons l'espace*

$$\mathcal{E}_3 \otimes \mathcal{E}_2.$$

Nous l'identifions à l'espace $\mathcal{L}(\mathcal{E}_2, \mathcal{E}_3)$ *des applications linéaires de* $\mathcal{E}_2$ *dans* $\mathcal{E}_3$*, que nous munissons de la norme d'opérateur.*

Définition 12.18 (Produit tensoriel des vecteurs)**.** *Nous considérons l'application*

$$\begin{aligned}\otimes : \quad \mathcal{E}_3 \times \mathcal{E}_2 &\to \mathcal{E}_3 \otimes \mathcal{E}_2 \\ v_3, \, v_2 &\mapsto v_3 \otimes v_2.\end{aligned}$$

La norme d'opérateur d'un produit $v_3 \otimes v_2$, identifié à un élément de $\mathcal{L}(\mathcal{E}_2, \mathcal{E}_3)$, est égale au produit des normes des vecteurs v_3 et v_2.

Corollaire 12.19 (Continuité du produit tensoriel)**.** *Le produit tensoriel est continu.*

Démonstration. Le produit tensoriel est une application bilinéaire et l'espace de départ est de dimension finie. Il est ainsi continu. □

Chapitre

13 Vecteurs spécifiques à « NoBackTrack »

13.1 Trajectoires des vecteurs spécifiques à « NoBackTrack »

13.1.1 Opérateur de réduction sur les vecteurs spécifiques à « NoBackTrack »

Rappelons que $T\Theta$ désigne l'espace tangent à Θ et que, pour $t \geq 0$, $T\mathcal{E}_t$ désigne le tangent à $\mathcal{E}_t$.

Définition 13.1 (Signes aléatoires)**.** *Nous considérons des variables de Bernoulli indépendantes, et identiquement distribuées*

$$\varepsilon_i(t), \quad t \geq 1, \quad 1 \leq i \leq \dim T\mathcal{E}_t,$$

prenant les valeurs 1 *ou* -1 *avec probabilité* $1/2$*.*

Pour tout $t \geq 1$*, nous notons* $\varepsilon(t)$ *le vecteur des* $\varepsilon_i(t)$*, pour* $1 \leq i \leq \dim T\mathcal{E}_t$*.*

Pour tout $t \geq 1$*, nous notons* $\mathcal{F}_t$ *la tribu engendrée par les* $\varepsilon(s)$*, pour* $1 \leq s \leq t$*.*

Nous notons également $\mathcal{F}_0$ *une tribu quelconque, incluse dans toutes les* $\mathcal{F}_t$*, pour* $t \geq 1$*.*

Remarque 13.2. *Ainsi, calculer l'espérance conditionnelle par rapport à* $\mathcal{F}_t$ *signifie que nous intégrons pour les lois des* $\varepsilon(s)$*, avec* $s > t$*.*

Remarque 13.3. *Dans la suite, nous démontrerons des résultats pour l'algorithme « NoBackTrack » initialisé à l'instant* 0*, puis nous les appliquerons à l'algorithme intialisé à un instant* t_0 *quelconque, mais dont les quantités initiales seront celles produites par l'algorithme « NoBackTrack ». Ces quantités dépendront alors des* $\varepsilon(t)$*, pour* $1 \leq t \leq t_0$*. Afin de pouvoir appliquer les résultats obtenus à l'instant* t_0*, nous aurons besoin de prendre en compte cette dépendance. La tribu* $\mathcal{F}_0$ *répond à ce besoin, et représente donc le passé de l'algorithme.*

Définition 13.4 (Opérateur de réduction à l'instant t)**.** *Soit* $t \geq 0$*. Nous définissons l'opérateur de réduction à l'instant* t*,*

$$\begin{aligned} \mathcal{R}_t : \quad & T\mathcal{E}_t \times T\Theta \times \mathcal{E}_t \times \Theta \to T\mathcal{E}_{t+1} \times T\Theta \\ & v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta \mapsto \mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right), \end{aligned}$$

tel que

$$\mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right) = \mathcal{R}\left(v^{\mathcal{E}},\, v^{\Theta};\, \frac{\partial \boldsymbol{T}_t}{\partial e}(e,\, \theta),\, \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e,\, \theta),\, \varepsilon(t+1)\right).$$

Ainsi,

$$\mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right) = \left(\left(\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e,\, \theta\right)\, v^{\mathcal{E}}\right)\, \odot\, v^{\Theta}\right) + \sum_{i=1}^{\dim T\mathcal{E}_t} \varepsilon_i\left(t+1\right)\, \left(\mathfrak{e}_i\, \odot\, \frac{\partial \boldsymbol{T}_t^i}{\partial \theta}\left(e,\, \theta\right)\right).$$

Remarque 13.5. *L'opérateur de réduction est l'opérateur de transition sur les $v_t^{\mathcal{E}}$ et v_t^{Θ}.*

Remarque 13.6. *Nous appliquons la remarque faite à la fin de la Définition 12.10 aux $\frac{\partial \boldsymbol{T}_t^i}{\partial \theta}\left(e,\, \theta\right)$, qui sont des formes linéaires sur le dual de $T\Theta$. Nous les considérons ainsi comme des vecteurs de $T\Theta$. De même, nous considérons les v^{Θ} comme des vecteurs de $T\Theta$.*

13.1.2 Trajectoires des vecteurs spécifiques à « NoBackTrack »

Définition 13.7 (Trajectoire des vecteurs de « NoBackTrack », paramètre fixe)**.** *Soit la suite de vecteurs $\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right)_{t\geq 0}$ tels que, pour tout $t \geq 0$,*

$$\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right) \in T\mathcal{E}_t \times T\Theta,$$

définie par la donnée de vecteurs initiaux $\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in T\mathcal{E}_0 \times T\Theta$, d'un état initial e_0, d'un paramètre $\theta \in \Theta$ et la relation de récurrence, pour $t \geq 0$,

$$\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right) = \mathcal{R}_t\left(v_t^{\mathcal{E}},\, v_t^{\Theta},\, \boldsymbol{e}_t\left(e_0,\, \theta\right),\, \theta\right).$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \boldsymbol{V}^{\mathcal{E}}:\ & T\mathcal{E}_0 \times T\Theta \times \mathcal{E}_0 \times \Theta \to T\mathcal{E}_t \\ & v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \theta \mapsto \boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \theta\right) = v_t^{\mathcal{E}}, \end{aligned}$$

et

$$\begin{aligned} \boldsymbol{V}^{\Theta}:\ & T\mathcal{E}_0 \times T\Theta \times \mathcal{E}_0 \times \Theta \to T\mathcal{E}_t \\ & v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \theta \mapsto \boldsymbol{V}_t^{\Theta}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \theta\right) = v_t^{\Theta}, \end{aligned}$$

qui associent aux vecteurs initiaux, à l'état initial et au paramètre les vecteurs à l'instant t.

Définition 13.8 (Trajectoire des vecteurs de « NoBackTrack », suite de paramètres)**.** *Soit la suite de vecteurs $\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right)_{t\geq 0}$ tels que, pour tout $t \geq 0$,*

$$\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right) \in T\mathcal{E}_t \times T\Theta,$$

définie par la donnée de vecteurs initiaux $\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in T\mathcal{E}_0 \times T\Theta$, d'un état initial e_0, d'une suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ à valeurs dans Θ et la relation de récurrence, pour $t \geq 0$,

$$\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right) = \mathcal{R}_t\left(v_t^{\mathcal{E}},\, v_t^{\Theta},\, \boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right).$$

Soient pour $t \geq 0$ les fonctions coordonnées

$$\begin{aligned} \boldsymbol{V}^{\mathcal{E}}:\ & T\mathcal{E}_0 \times T\Theta \times \mathcal{E}_0 \times \Theta^{\infty} \to T\mathcal{E}_t \\ & v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta} \mapsto \boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right) = v_t^{\mathcal{E}}, \end{aligned}$$

et

$$\begin{aligned}\boldsymbol{V}^{\Theta}: \quad & T\mathcal{E}_0 \times T\Theta \times \mathcal{E}_0 \times \Theta^{\infty} \to T\mathcal{E}_t \\ & v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta} \mapsto \boldsymbol{V}_t^{\Theta}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right) = v_t^{\Theta},\end{aligned}$$

qui associent aux vecteurs initiaux, à l'état initial et à la suite de paramètres les vecteurs à l'instant t.

13.2 Contrôle des opérateurs de réduction

13.2.1 Stabilité des ensembles $B_{T\mathcal{E}_t \times T\Theta} \cap \mathcal{Z}_t$

Hypothèse 13.9 (Dimensions des espaces des états bornées)**.** *Nous supposons que*

$$\sup_{t\geq 0} \dim \mathcal{E}_t = \sup_{t\geq 0} \dim T\mathcal{E}_t < \infty.$$

Lemme 13.10 (Majorant des sommes des racines des normes des dérivées des opérateurs de transition par rapport au paramètre)**.** *Nous pouvons trouver $M_8 \geq 0$ tel que, pour tout $t \geq 0$, pour tous $e \in B_{\mathcal{E}_t}$ et $\theta \in B_{\Theta}^*$,*

$$\sum_{i=1}^{\dim T\mathcal{E}_{t+1}} \left\| \frac{\partial \boldsymbol{T}_t^i}{\partial \theta}(e,\, \theta) \right\|^{1/2} \leq M_8.$$

Démonstration. Soit $t \geq 0$ et $1 \leq i \leq \dim T\mathcal{E}_{t+1}$. Alors,

$$\left\| \frac{\partial \boldsymbol{T}_t^i}{\partial \theta}(e,\, \theta) \right\| \leq \left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e,\, \theta) \right\|_{\text{op}}.$$

Or, d'après le Lemme 4.40,

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta}(e,\, \theta) \right\|_{\text{op}} \leq M_1.$$

Ainsi,

$$\begin{aligned}\sum_{i=1}^{\dim T\mathcal{E}_{t+1}} \left\| \frac{\partial \boldsymbol{T}_t^i}{\partial \theta}(e,\, \theta) \right\|^{1/2} &\leq \sum_{i=1}^{\dim T\mathcal{E}_{t+1}} M_1^{1/2} \\ &= M_1^{1/2} \dim T\mathcal{E}_{t+1} \\ &\leq M_1^{1/2} \sup_{s\geq 0} \dim T\mathcal{E}_s.\end{aligned}$$

Or, d'après l'Hypothèse 13.9, le supremum dans le membre de droite est fini, ce qui conclut la preuve. □

Rappelons que $r(\gamma_1,\, \gamma_1)$ a été introduit au Lemme 12.15.

Définition 13.11 (Zones sous hyperbole stable par les opérateurs de réduction)**.** *Nous définissons, pour $t \geq 0$, la « zone sous l'hyperbole »*

$$\mathcal{Z}_t = \left\{ \left(v^{\mathcal{E}},\, v^{\Theta}\right) \in T\mathcal{E}_t \times T\Theta \,\middle|\, \left\|v^{\mathcal{E}}\right\| \left\|v^{\Theta}\right\| \leq r(1-\alpha,\, M_8) \right\}.$$

Lemme 13.12 (Zone sous hyperbole stable par les opérateurs de réduction)**.** *Soit $t \geq 0$. Soient $\left(v^{\mathcal{E}},\, v^{\Theta}\right) \in \mathcal{Z}_t$, $e \in B_{\mathcal{E}_t}$ et $\theta \in B_{\Theta}^{*}$. Notons*

$$\left(w^{\mathcal{E}},\, w^{\Theta}\right) = \mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right).$$

Alors,

$$\left(w^{\mathcal{E}},\, w^{\Theta}\right) \in \mathcal{Z}_{t+1}.$$

Démonstration.

$$\mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right) = \mathcal{R}\left(v^{\mathcal{E}},\, v^{\Theta};\, \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e,\, \theta\right),\, \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e,\, \theta\right),\, \varepsilon\left(t\right)\right).$$

Posons

$$M_{\text{mul}} = 1 - \alpha \quad \text{et} \quad M_{\text{add}} = M_8.$$

Alors, d'après le Lemme 4.39, $0 \leq M_{\text{mul}} < 1$ et

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e,\, \theta\right) \right\|_{\text{op}} \leq M_{\text{mul}},$$

et d'après le Lemme 13.10, $M_{\text{add}} \geq 0$ et

$$\sum_{i=1}^{\dim T\mathcal{E}_{t+1}} \left\| \frac{\partial \boldsymbol{T}_t^i}{\partial \theta}\left(e,\, \theta\right) \right\|^{1/2} \leq M_{\text{add}}.$$

Or, par hypothèse,

$$\left\|v^{\mathcal{E}}\right\| \left\|v^{\Theta}\right\| \leq r\left(1 - \alpha,\, M_8\right) = r\left(M_{\text{mul}},\, M_{\text{add}}\right).$$

Par conséquent, d'après le Corollaire 12.16,

$$\left\|w^{\mathcal{E}}\right\| \left\|w^{\Theta}\right\| \leq r\left(M_{\text{mul}},\, M_{\text{add}}\right),$$

et ainsi

$$\left(w^{\mathcal{E}},\, w^{\Theta}\right) \in \mathcal{Z}_{t+1},$$

ce qui conclut la preuve. □

Définition 13.13 (Boules contenant les vecteurs « NoBackTrack »)**.** *Pour $t \geq 0$, nous définissons les boules*

$$B_{T\mathcal{E}_t \times T\Theta} \subset T\mathcal{E}_t \times T\Theta,$$

de rayon commun

$$r_{T\mathcal{E} \times T\Theta} = \sqrt{2}\left(\left(1 - \alpha\right)^{1/2} r\left(1 - \alpha,\, M_8\right)^{1/2} + M_8\right).$$

Corollaire 13.14 (Intersection de $B_{T\mathcal{E}_t \times T\Theta}$ et $\mathcal{Z}_t$ non vide)**.** *Pour tout $t \geq 0$, l'intersection de $B_{T\mathcal{E}_t \times T\Theta}$ et $\mathcal{Z}_t$ est non vide.*

Démonstration. Soit $t \geq 0$. Soit

$$\left(v^{\mathcal{E}},\, v^{\Theta}\right) \in B_{T\mathcal{E}_t \times T\Theta},$$

tel que de plus

$$\left\|\left(v^{\mathcal{E}},\, v^{\Theta}\right)\right\| \leq r\left(1-\alpha,\, M_8\right)^{1/2}.$$

Alors,

$$\left\|v^{\mathcal{E}}\right\| \left\|v^{\Theta}\right\| \leq \left\|\left(v^{\mathcal{E}},\, v^{\Theta}\right)\right\|^2 \leq r\left(1-\alpha,\, M_8\right).$$

Par conséquent, d'après la Définition 13.11,

$$\left(v^{\mathcal{E}},\, v^{\Theta}\right) \in \mathcal{Z}_t,$$

ce qui conclut la preuve. □

Lemme 13.15 (Stabilité des ensembles $B_{T\mathcal{E}_t \times T\Theta} \cap \mathcal{Z}_t$ par l'opérateur de réduction)**.** *Soient*

$$\left(v^{\mathcal{E}},\, v^{\Theta}\right) \in B_{T\mathcal{E}_t \times T\Theta} \cap \mathcal{Z}_t,$$

$e \in B_{\mathcal{E}_t}$, *et* $\theta \in B_{\Theta}^{*}$. *Alors,*

$$\mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right) \in B_{T\mathcal{E}_{t+1} \times T\Theta} \cap \mathcal{Z}_{t+1}.$$

Remarque 13.16. *D'après le Corollaire 13.14, l'intersection de la boule $B_{T\mathcal{E}_t \times T\Theta}$ et de $\mathcal{Z}_t$ n'est pas vide, ce qui assure un contenu au lemme.*

Démonstration. D'après la Définition 13.4, nous savons que

$$\mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right) = \left(\left(\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e,\, \theta\right)\, v^{\mathcal{E}}\right) \odot v^{\Theta}\right) + \sum_{i=1}^{\dim T\mathcal{E}_t} \varepsilon_i\left(t\right)\left(\mathfrak{e}_i \odot \frac{\partial \boldsymbol{T}_t^i}{\partial \theta}\left(e,\, \theta\right)\right).$$

Notons alors

$$\left(w^{\mathcal{E}},\, w^{\Theta}\right) = \mathcal{R}_t\left(v^{\mathcal{E}},\, v^{\Theta},\, e,\, \theta\right).$$

D'après le Corollaire 12.13, nous savons que

$$\begin{aligned}\left\|\left(w^{\mathcal{E}},\, w^{\Theta}\right)\right\|_{\max} \leq{}& \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e,\, \theta\right)\right\|_{\mathrm{op}}^{1/2} \left(\left\|v^{\mathcal{E}}\right\| \left\|v^{\Theta}\right\|\right)^{1/2} \\ &+ \sum_{i=1}^{\dim T\mathcal{E}_{t+1}} \left\|\frac{\partial \boldsymbol{T}_t^i}{\partial \theta}\left(e,\, \theta\right)\right\|^{1/2}.\end{aligned}$$

Or, $\left(v^{\mathcal{E}},\, v^{\Theta}\right) \in \mathcal{Z}_t$ donc, d'après la Définition 13.11,

$$\left\|v^{\mathcal{E}}\right\| \left\|v^{\Theta}\right\| \leq r\left(1-\alpha,\, M_8\right).$$

Ainsi, d'après les Lemmes 4.39 et 13.10,

$$\left\|\left(w^{\mathcal{E}},\, w^{\Theta}\right)\right\|_{\max} \leq \left(1-\alpha\right)^{1/2} r\left(1-\alpha,\, M_8\right)^{1/2} + M_8.$$

Par conséquent,

$$\begin{aligned}\left\|\left(w^{\mathcal{E}},\, w^{\Theta}\right)\right\| &\leq \sqrt{2}\,\left\|\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right)\right\|_{\max} \\ &\leq \sqrt{2}\,\left((1-\alpha)^{1/2}\, r\,(1-\alpha,\, M_8)^{1/2} + M_8\right) \\ &= r_{T\mathcal{E}\times T\Theta}.\end{aligned}$$

Ainsi,

$$\left(w^{\mathcal{E}},\, w^{\Theta}\right) \in B_{T\mathcal{E}_{t+1}\times T\Theta}.$$

Enfin, d'après le Lemme 13.12,

$$\left(w^{\mathcal{E}},\, w^{\Theta}\right) \in \mathcal{Z}_{t+1},$$

ce qui conclut la preuve. □

13.2.2 Stabilité des vecteurs spécifiques à « NoBackTrack »

Lemme 13.17 (Stabilité de la zone sous l'hyperbole)**.** *Soient* $\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in \mathcal{Z}_0$, $e_0 \in B^*_{\mathcal{E}_0}$ *et une suite de paramètres* $\boldsymbol{\theta}$ *incluse dans* B^*_{Θ}*. Alors, pour tout* $t \geq 0$*,*

$$\left(\boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{V}_t^{\Theta}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right)\right) \in \mathcal{Z}_t.$$

Démonstration. C'est une conséquence par récurrence du Lemme 13.12 car, d'après le Corollaire 4.53 et la Définition 4.50, pour tout $t \geq 0$,

$$\boldsymbol{e}_t\,(e_0,\, \boldsymbol{\theta}) \in B^*_{\mathcal{E}_t} \subset B_{\mathcal{E}_t}.$$

□

Lemme 13.18 (Vecteurs « NoBackTrack » contenus dans les ensembles $B_{T\mathcal{E}_t\times T\Theta}\cap\mathcal{Z}_t$)**.** *Soient*

$$\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in B_{T\mathcal{E}_0\times T\Theta} \cap \mathcal{Z}_0,$$

$e_0 \in B^*_{\mathcal{E}_0}$*, et une suite de paramètres* $\boldsymbol{\theta} = (\theta_t)$ *incluse dans* B^*_{Θ}*. Alors, pour tout* $t \geq 0$*,*

$$\left(\boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{V}_t^{\Theta}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right)\right) \in B_{T\mathcal{E}_t\times T\Theta} \cap \mathcal{Z}_t.$$

Démonstration. Notons, pour $t \geq 0$,

$$\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right) = \left(\boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right),\, \boldsymbol{V}_t^{\Theta}\left(v_0^{\mathcal{E}},\, v_0^{\Theta},\, e_0,\, \boldsymbol{\theta}\right)\right).$$

$$\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in \mathcal{Z}_0$$

donc, d'après le Lemme 13.17, pour tout $t \geq 0$,

$$\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right) \in \mathcal{Z}_t.$$

D'après la Définition 13.8, pour tout $t \geq 0$,

$$\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right) = \mathcal{R}_t\left(v_t^{\mathcal{E}},\, v_t^{\Theta},\, \boldsymbol{e}_t\,(e_0,\, \boldsymbol{\theta}),\, \theta_t\right).$$

Prouvons alors par récurrence sur $t \geq 0$ que

$$\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right) \in B_{T\mathcal{E}_t \times T\Theta}.$$

L'initialisation est vérifiée par hypothèse. Soit alors $t \geq 0$ et supposons la propriété vraie pour t. Or, $e_0 \in B^*_{\mathcal{E}_0}$ donc, d'après le Corollaire 4.53 et la Définition 4.50,

$$\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right) \in B^*_{\mathcal{E}_t} \subset B_{\mathcal{E}_t}.$$

Ainsi, d'après le Lemme 13.15,

$$\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right) \in B_{T\mathcal{E}_{t+1} \times T\Theta},$$

ce qui conclut la récurrence, puis la preuve. □

Lemme 13.19 (« Image » des boules $B_{T\mathcal{E}_t \times T\Theta}$ par le produit tensoriel bornée)**.** *Soit* $t \geq 0$. *Soit*

$$\left(v^{\mathcal{E}},\, v^{\Theta}\right) \in B_{T\mathcal{E}_t \times T\Theta}.$$

Alors,

$$\left\| v^{\mathcal{E}} \otimes v^{\Theta} \right\|_{\text{op}} \leq r^2_{T\mathcal{E} \times T\Theta}.$$

Démonstration. En effet,

$$\left\| v^{\mathcal{E}} \otimes v^{\Theta} \right\|_{\text{op}} = \left\| v^{\mathcal{E}} \right\| \left\| v^{\Theta} \right\| \leq \left\| \left(v^{\mathcal{E}},\, v^{\Theta}\right) \right\|^2 \leq r^2_{T\mathcal{E} \times T\Theta}.$$

□

Chapitre

14 Application de la propriété centrale à l'algorithme « NoBackTrack »

14.1 Contrôle « déterministe » du bruit produit par l'algorithme « NoBackTrack »

Définition 14.1 (Écart à la mise à jour RTRL)**.** *Soit une suite $\boldsymbol{\theta}$ de paramètres. Pour $t \geq 0$, notons*

$$e_t = \boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right)$$

et

$$\left(v_t^{\mathcal{E}}, v_t^{\Theta}\right) = \left(\boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}}, v_0^{\Theta}, e_0, \boldsymbol{\theta}\right), \boldsymbol{V}_t^{\Theta}\left(v_0^{\mathcal{E}}, v_0^{\Theta}, e_0, \boldsymbol{\theta}\right)\right).$$

Pour tout $t \geq 0$, nous posons

$$\xi_t\left(\boldsymbol{\theta}\right) = v_{t+1}^{\mathcal{E}} \otimes v_{t+1}^{\Theta} - \left(\frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t, \theta_t\right) \cdot \left(v_t^{\mathcal{E}} \otimes v_t^{\Theta}\right) + \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t, \theta_t\right)\right).$$

Définition 14.2 (Rayon de la boule contenant les estimées « NoBackTrack » et borne sur les écarts à la mise à jour RTRL)**.** *Nous posons*

$$r_\xi = r_{T\mathcal{E}\times T\Theta}^2$$

et

$$M_\xi = r_{T\mathcal{E}\times T\Theta}^2 + \left(1-\alpha\right) r_{T\mathcal{E}\times T\Theta}^2 + M_1.$$

Remarque 14.3. *Dans la preuve de convergence de RTRL, nous définissions r_ξ à partir de M_ξ. Ce n'est pas le cas ici car nous ne savons pas* a priori *que le bruit est borné. Mais cela ne posera pas problème dans la suite de la preuve.*

Corollaire 14.4 (Contrôle du bruit sur B_Θ^*)**.** *Soit $\boldsymbol{\theta} = (\theta_t)$ une suite de paramètres incluse dans la boule B_Θ^*. Alors, pour tout $t \geq 0$,*

$$\left\|\xi_t\left(\boldsymbol{\theta}\right)\right\|_{\mathrm{op}} \leq M_\xi.$$

Démonstration. D'après la Définition 14.1, pour tout $t \geq 0$,

$$\begin{aligned}\left\|\xi_t\right\|_{\mathrm{op}} \leq{}& \left\|v_{t+1}^{\mathcal{E}} \otimes v_{t+1}^{\Theta}\right\|_{\mathrm{op}} + \left\|\frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right)\right\|_{\mathrm{op}} \left\|v_t^{\mathcal{E}} \otimes v_t^{\Theta}\right\|_{\mathrm{op}} \\ &+ \left\|\frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right)\right\|_{\mathrm{op}}.\end{aligned}$$

D'après le Lemme 13.18, pour tout $t \geq 0$,

$$\left(v_t^{\mathcal{E}}, v_t^{\Theta}\right) \in B_{T\mathcal{E}_t\times T\Theta}$$

donc, d'après le Lemme 13.19, pour tout $t \geq 0$,

$$\left\| v_t^{\mathcal{E}} \otimes v_t^{\Theta} \right\|_{\mathrm{op}} \leq r^2_{T\mathcal{E}\times T\Theta}.$$

De plus, d'après le Corollaire 4.53, pour tout $t \geq 0$,

$$\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right) \in B^*_{\mathcal{E}_t}.$$

Par conséquent, d'après les Lemmes 4.39 et 4.40, pour tout $t \geq 0$,

$$\left\| \frac{\partial \boldsymbol{T}_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \right\|_{\mathrm{op}} \leq 1-\alpha \quad \text{et} \quad \left\| \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0,\, \boldsymbol{\theta}\right),\, \theta_t\right) \right\|_{\mathrm{op}} \leq M_1.$$

D'après la Définition 14.2, nous avons établi le résultat annoncé. □

14.2 Définition de l'algorithme « NoBackTrack »

Définition 14.5 (Algorithme « NoBackTrack »)**.** *Soit la suite de paramètres $\boldsymbol{\theta} = (\theta_s)$ définie par la donnée de $e_0 \in \mathcal{E}_0$, $\theta_0 \in \Theta$, $\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in T\mathcal{E}_0 \times T\Theta$ et les relations de récurrence, pour $t \geq 0$,*

$$\begin{cases} \theta_{t+1} = \phi\left(\theta_t,\, \dfrac{\partial p_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot \left(v_t^{\mathcal{E}} \otimes v_t^{\Theta}\right) + \dfrac{\partial p_t}{\partial \theta}\left(e_t,\, \theta_t\right),\, \eta_t\right) \\ e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta_t\right) \\ \left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right) = \mathcal{R}_t\left(v_t^{\mathcal{E}},\, v_t^{\Theta},\, e_t,\, \theta_t\right), \end{cases}$$

où (η_t) est une suite de pas de descente.

Les trajectoires de l'algorithme « NoBackTrack » sont aléatoires, et dépendent du tirage des signes aléatoires. Dans la suite, nous n'expliciterons pas ceci systématiquement. Ainsi, nous parlerons par exemple, par abus de langage, de « la suite de paramètres » produite par l'algorithme « NoBackTrack », au lieu de parler de la suite de paramètres produite par l'algorithme le long d'un certain tirage, ou de la suite aléatoire de paramètres produite par l'algorithme.

14.3 Algorithme « NoBackTrack » comme « RTRL bruité »

Lemme 14.6 (Réécriture de « NoBackTrack » comme « RTRL bruité »)**.** *Pour tout $t \geq 0$, avec les notations de la Définition 14.5, les quantités maintenues par l'algorithme « NoBackTrack » vérifient :*

$$\begin{cases} \theta_{t+1} = \phi\left(\theta_t,\, \dfrac{\partial p_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot \left(v_t^{\mathcal{E}} \otimes v_t^{\Theta}\right) + \dfrac{\partial p_t}{\partial \theta}\left(e_t,\, \theta_t\right),\, \eta_t\right) \\ e_{t+1} = \boldsymbol{T}_t\left(e_t,\, \theta_t\right) \\ v_{t+1}^{\mathcal{E}} \otimes v_{t+1}^{\Theta} = \dfrac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot \left(v_t^{\mathcal{E}} \otimes v_t^{\Theta}\right) + \dfrac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right) + \xi_t\left(\boldsymbol{\theta}\right). \end{cases}$$

Démonstration. C'est une conséquence des Définitions 14.5 et 14.1. □

14.4 Application de la propriété centrale à l'algorithme « NoBackTrack »

Lemme 14.7 (Maintien dans les boules de contrôle)**.** *Soient $\theta_0 \in B^*_\Theta$, tel que de plus*

$$d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{3},$$

$e_0 \in B^*_{\mathcal{E}_0}$ *et*

$$\left(v_0^{\mathcal{E}}, v_0^\Theta\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0.$$

Soit une suite $\boldsymbol{\eta}$ telle que $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$. Notons $\boldsymbol{\theta}$ la suite de paramètres produite par l'algorithme « NoBackTrack » utilisant ces quantités. Alors, pour tout $0 \leq t < T_0^{r^*_\Theta}$, $\theta_t \in B^*_\Theta$.

Remarque 14.8. *La preuve n'est pas une conséquence du Fait 9.8 car alors, nous avions supposé le bruit borné* a priori. *Ce n'est pas le cas ici. Elle est cependant presque identique.*

Démonstration. La preuve est presque identique à celle du Fait 9.8. La seule différence est la suivante. Notons, pour $0 \leq t < T_0^{r^*_\Theta}$,

$$\left(v_t^{\mathcal{E}}, v_t^\Theta\right) = \left(\boldsymbol{V}_t^{\mathcal{E}}\left(v_0^{\mathcal{E}}, v_0^\Theta, e_0, \boldsymbol{\theta}\right), \boldsymbol{V}_t^\Theta\left(v_0^{\mathcal{E}}, v_0^\Theta, e_0, \boldsymbol{\theta}\right)\right)$$

et

$$v_t = \frac{\partial p_t}{\partial e}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right) \cdot \left(v_t^{\mathcal{E}} \otimes v_t^\Theta\right) + \frac{\partial p_t}{\partial \theta}\left(\boldsymbol{e}_t\left(e_0, \boldsymbol{\theta}\right), \theta_t\right).$$

À chaque étape de la récurrence, nous savons que « la partie qui compte » de la suite $\boldsymbol{\theta}$ est incluse dans la boule B^*_Θ. Par conséquent, d'après le Lemme 13.18,

$$\left(v_t^{\mathcal{E}}, v_t^\Theta\right) \in B_{T\mathcal{E}_t \times T\Theta}$$

et ainsi, d'après le Lemme 13.19 et la Définition 14.2,

$$\left\|v_t^{\mathcal{E}} \otimes v_t^\Theta\right\| \leq r_\xi.$$

Donc, d'après le Corollaire 8.10, v_t/m_t appartient à la boule $B^*_{T\Theta}$. Ceci achève l'exposition des différences avec la preuve du Fait 9.8. □

14.4.1 Application de la propriété centrale à la trajectoire « NoBackTrack »

Corollaire 14.9 (Application de la propriété centrale à la trajectoire « NoBackTrack », énoncé pour $t_0 = 0$)**.** *Soit $\theta_0 \in B^*_\Theta$, tel que de plus*

$$d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{3},$$

Soient $e_0 \in B^*_{\mathcal{E}_0}$ et

$$\left(v_0^{\mathcal{E}}, v_0^\Theta\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0.$$

Soit une suite $\boldsymbol{\eta} = (\eta_t)$ telle que la suite $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$. Notons alors $\boldsymbol{\theta} = (\theta_t)$ la suite de paramètres produite par l'algorithme « NoBackTrack ». Alors, pour tout $0 \leq t < T_0^\infty$, nous avons $\theta_t \in B^*_\Theta$ et, pour tout $1 \leq t < T_0^\infty$,

$$d(\theta_t, \theta^*) \leq \left(1 - \lambda^*_{0,t-1} \sum_{s=0}^{t-1} \eta_s\right) d(\theta_0, \theta^*) + b_{0,t}(\boldsymbol{\eta}).$$

Démonstration. D'après le Lemme 14.7 et le Corollaire 14.4, pour tout $0 \leq t < T_0^{r^*_\Theta}$,

$$\|\xi_t(\boldsymbol{\theta})\| \leq M_\xi.$$

Or, $T_0^\infty \leq T_0^{r^*_\Theta}$, donc cette inégalité est vraie pour tout $0 \leq t < T_0^\infty$. La preuve est alors une conséquence du Fait 9.34. □

Corollaire 14.10 (Application de la propriété centrale à la trajectoire « NoBack-Track »)**.** *Soit $\theta_{t_0} \in B^*_\Theta$, tel que de plus*

$$d(\theta_{t_0}, \theta^*) \leq \frac{r^*_\Theta}{3}.$$

Soient $e_{t_0} \in B^*_{\mathcal{E}_{t_0}}$ et

$$\left(v^{\mathcal{E}}_{t_0}, v^{\Theta}_{t_0}\right) \in B_{T\mathcal{E}_{t_0} \times T\Theta} \cap \mathcal{Z}_{t_0}.$$

Soit une suite $\boldsymbol{\eta} = (\eta_t)$ telle que la suite $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$. Notons alors $\boldsymbol{\theta} = (\theta_t)$ la suite de paramètres produite par l'algorithme « NoBack-Track » initialisé à l'instant t_0. Alors, pour tout $t_0 \leq t < T^\infty_{t_0}$, $\theta_t \in B^*_\Theta$ et, pour tout $t_0 + 1 \leq t < T^\infty_{t_0}$,

$$d(\theta_t, \theta^*) \leq \left(1 - \lambda^*_{t_0, t-1} \sum_{s=t_0}^{t-1} \eta_s\right) d(\theta_{t_0}, \theta^*) + b_{t_0, t}(\boldsymbol{\eta}).$$

Chapitre

15 Ensemble de convergence de l'algorithme « NoBackTrack »

Dans la suite, nous considérons la suite de pas $\boldsymbol{\eta} = (\eta_t)$ dont nous disposons grâce à l'Hypothèse 11.8. Comme les modifications ne changent pas le fait que les hypothèses formulées sont homogènes en la suite de pas, nous pouvons, comme cela avait été établi au Lemme 10.1 et au Corollaire 10.2, supposer que, pour tout $t \geq 0$, $\tilde{\eta}_t = m_t \, \eta_t \leq \eta_{\max}$, de sorte que la suite $\mu \, \tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \, \eta_{\max}]$.

Dans la suite, nous considérons alors la demi-droite dirigée par la suite $\boldsymbol{\eta}$. Nous paramétrons celle-ci par les réels $\mu \geq 0$. Ainsi, nous considérons des suites $\mu \, \boldsymbol{\eta}$ égales à

$$\mu \, \eta_0, \, \mu \, \eta_1, \, \mu \, \eta_2, \ldots, \mu \, \eta_t, \ldots$$

15.1 Continuité des trajectoires obtenues par l'algorithme « NoBackTrack » par rapport au pas initial

Corollaire 15.1 (Continuité des trajectoires obtenues par l'algorithme « NoBackTrack » par rapport à μ)**.** *Fixons $\theta_0 \in B^*_\Theta$, $e_0 \in B^*_{\mathcal{E}_0}$ et*

$$\left(v_0^{\mathcal{E}}, \, v_0^{\Theta}\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0.$$

Alors, pour tout $t \geq 0$, nous pouvons trouver un $\eta^t_{\max} \leq \eta_{\max}/m_t$ suffisamment petit pour que, pour $\mu \leq \eta^t_{\max}$, l'application

$$\mu \mapsto \left(\theta_s, \, e_s, \, \left(v_s^{\mathcal{E}}, \, v_s^{\Theta}\right)\right)_{s \leq t}$$

soit continue.

Démonstration. La preuve est analogue à celle correspondante pour l'algorithme RTRL, qui est effectuée à la section 10.1.2. La seule différence réside dans l'établissement du résultat analogue au Lemme 10.3. Nous devons justifier que les vecteurs tangents dépendent bien continûment de θ, de e et de $\left(v^{\mathcal{E}}, \, v^{\Theta}\right)$, et que l'opérateur de transition sur les vecteurs spécifiques à « NoBackTrack » est bien continu. Or, d'après le Corollaire 12.19, le produit tensoriel sur $T\mathcal{E}_t \times T\Theta$ est continu et ainsi, de même que pour RTRL, l'application

$$\theta, \, e, \, \left(v^{\mathcal{E}}, \, v^{\Theta}\right) \mapsto \frac{\partial \, p_t}{\partial e} \, (e, \, \theta) \cdot \left(v^{\mathcal{E}} \otimes v^{\Theta}\right) + \frac{\partial \, p_t}{\partial \theta} \, (e, \, \theta)$$

est continue. Enfin, d'après le Lemme 12.8, l'opérateur d'égalisation des normes est continu, et l'opérateur de transition sur les vecteurs spécifiques à « NoBackTrack », introduit à la Définition 13.4, est continu. Nous avons ainsi bien établi les deux points annoncés. □

15.2 Établissement des conditions pour la convergence

15.2.1 Établissement des conditions non homogènes en la suite de pas de descente pour « NoBackTrack »

Rappelons que le réel ν a été introduit à la Définition 11.1.

Définition 15.2 (Terme majorant l'espérance du terme additif fonction du pas)**.** *Soit M_{12} majorant M_7 et la constante uniquement fonction de α et M_ξ du terme en grand O du Lemme 16.13. Soit un réel $0 < \nu < 1$. Soit un temps $t_0 \geq 0$. Nous notons, pour $t \geq t_0 + 1$,*

$$\tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right) = M_{12}\left(\left\|\sum_{s=t_0}^{t-1}\eta_s \frac{\partial\, p_s \circ g_s}{\partial\theta}\left(e_0^*,\,\theta^*\right)\right\| + \frac{\left(\sum_{s=t_0}^{t-1}\tilde{\eta}_s^2\right)^{1/2}}{\left(\sum_{s=t_0}^{t-1}\eta_s^{1+\nu}\right)^{\frac{1}{1+\nu}}}\sum_{s=t_0}^{t-1}\eta_s\right.$$
$$\left. + \sup_{t_0\leq s\leq t-1}\tilde{\eta}_s + \left(\sum_{s=t_0}^{t-1}\tilde{\eta}_s\right)^2 + \sum_{s=t_0}^{t-1}\tilde{\eta}_s^2\right).$$

Remarque 15.3. *Le terme de droite de la première ligne pourrait en fait être n'importe quelle quantité dont la somme sur les intervalles I_k serait négligeable devant la suite des pas, si nous n'utilisions pas qu'il est homogène en le pas, dans la preuve du Corollaire 15.8. C'est pour cela que nous utilisons un terme de cette forme.*

Remarque 15.4. *Nous avons besoin que $\nu > 0$ afin que l'énoncé du Lemme 11.5 ne soit pas vide, et que $\nu < 1$ à la fin de la preuve du Corollaire 11.14.*

Définition 15.5 (Majorant du terme additif)**.** *Notons, pour $\mu \geq 0$ et $k \geq 0$,*

$$\tilde{b}^k\left(\mu\right) = \tilde{b}_{T_k,\,T_{k+1}}\left(\mu\,\boldsymbol{\eta}\right).$$

Lemme 15.6 (Contrôle de $\tilde{b}^k\left(\mu\right)$)**.** *Pour tout $\mu \geq 0$, $\tilde{b}^k\left(\mu\right)$ est négligeable devant*

$$\sum_{I_k}\mu\,\eta_t$$

quand k tend vers l'infini.

Démonstration. D'après le Fait 6.29 et le Corollaire 6.34, il suffit de montrer que c'est le cas pour les termes

$$\frac{\left(\sum_{I_k}\tilde{\eta}_s^2\right)^{1/2}}{\left(\sum_{I_k}\eta_s^{1+\nu}\right)^{\frac{1}{1+\nu}}}\sum_{I_k}\eta_t.$$

Or, d'après le Corollaire 11.14, le terme en facteur de la somme des pas tend vers 0, quand k tend vers l'infini, et le résultat annoncé est ainsi établi. □

Définition 15.7 (Infimum pour le maintien de $\boldsymbol{\theta}$ dans B_Θ^* pour « NoBackTrack »)**.** *Pour tout $\mu \geq 0$, notons*

$$\tilde{k}_1\left(\mu\right) = \inf\left\{k \geq 0 \,\middle|\, \tilde{b}^k\left(\mu\right) \leq \frac{1}{2}\,\lambda_{\min}\,\frac{r_\Theta^*}{3}\sum_{I_k}\mu\,\eta_t\right\}.$$

Corollaire 15.8 (Propriétés de l'infimum pour le maintien de $\boldsymbol{\theta}$ dans B^*_Θ pour « NoBackTrack »)**.** *$\tilde{k}_1$ vérifie les propriétés suivantes.*

1. *Pour tout $\mu \geq 0$, $\tilde{k}_1(\mu) < \infty$.*
2. *$\tilde{k}_1$ est une fonction croissante de μ.*

Démonstration. La première propriété est une conséquence du Lemme 15.6.

La deuxième s'établit de même qu'au Corollaire 10.10, en utilisant le fait que $\tilde{b}^k(\mu)$ ne comporte que des termes homogènes de degré supérieur à 1 en μ. □

15.2.2 Choix d'une suite de pas

Notons $\theta_t(\mu)$ le t-ème paramètre obtenu par l'algorithme « NoBackTrack » avec la suite de pas de descente $\mu\,\boldsymbol{\eta}$, avec des quantités initiales qui seront précisées par la suite.

Rappelons que la constante M_4 a été définie au Lemme 7.10.

Corollaire 15.9 (Choix d'une suite de pas et attente des conditions de convergence pour « NoBackTrack »)**.** *Pour tout $\varepsilon > 0$, nous pouvons trouver $\tilde{\mu}_{\max} > 0$ et $0 \leq \tilde{k}_2 < \infty$ tels que, pour tout $\mu \leq \tilde{\mu}_{\max}$, les propriétés suivantes sont vérifiées.*

1. *$\tilde{k}_2 \geq k_0(\bar{\boldsymbol{\eta}})$.*
2. *La suite $\mu\,\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$.*
3. *Pour tout $k \geq \tilde{k}_2$,*
$$M_4 \sum_{I_k} \mu\,\tilde{\eta}_t \leq \frac{r^*_\Theta}{3}.$$
4. *Pour tout $k \geq \tilde{k}_2$,*
$$\Lambda^* \sum_{I_k} \mu\,\tilde{\eta}_t \leq 1.$$
5. *$\tilde{k}_2 \geq \tilde{k}_1(\mu)$.*
6. $$\prod_{k=\tilde{k}_2}^{\infty} (1 - u_k) \geq 1 - \varepsilon.$$
7. *Pour tous $\theta_0 \in B^*_\Theta$ tel que de plus*
$$d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{4},$$
$e_0 \in B^*_{\mathcal{E}_0}$ et
$$\left(v_0^{\mathcal{E}}, v_0^{\Theta}\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0,$$
pour tout $0 \leq \mu' \leq \mu$, pour tout $0 \leq t \leq T_{\tilde{k}_2}$,
$$d\left(\theta_t(\mu'), \theta^*\right) \leq \frac{r^*_\Theta}{3}.$$

Remarque 15.10. *$\tilde{k}_2$ dépend de $\tilde{\mu}_{\max}$, mais nous ne l'expliciterons pas dans la suite.*

Remarque 15.11. *Le sixième point est utilisé au Lemme 16.18.*

Démonstration. La démonstration est identique à celle du Corollaire 10.11 pour l'algorithme RTRL. Nous considérons juste $\tilde{k}_1(\mu)$ au lieu de $k_1(\mu)$ (et par conséquent le Corollaire 15.8 à la place du Corollaire 10.10) et, pour le septième point (qui était le sixième pour RTRL), nous utilisons le Corollaire 15.1 à la place du Corollaire 10.5.

La seule différence réside dans l'obtention du (nouveau) sixième point. Le produit des $1 - u_k$, pour $k \geq K$, est arbitrairement proche de 1 pour de grandes valeurs de K d'après le Corollaire 11.13. Par conséquent, pour un K tel que ce produit est supérieur à $1 - \varepsilon$, nous pouvons, si le $\tilde{k}_2$ obtenu à l'issue du cinquième point est inférieur à K, poser $\tilde{k}_2 = K$, ce qui ne remet pas en cause les points précédents. □

15.3 Convergence sur un ensemble de l'algorithme « NoBackTrack »

Corollaire 15.12 (T_{k+1} inférieur à $T^{\infty}_{T_k}$ pour k assez grand)**.** *Pour tout $k \geq \tilde{k}_2$ et $\mu \leq \tilde{\mu}_{\max}$,*

$$T_{k+1} < T^{\infty}_{T_k}.$$

Démonstration. L'argument est identique à celui formulé dans la preuve du Lemme 10.12, en remplaçant le Corollaire 10.11 par le Corollaire 15.9. □

Lemme 15.13 (Horizons et échelle de temps pour « NoBackTrack »)**.** *Soit $0 \leq \mu \leq \tilde{\mu}_{\max}$ et $k \geq \tilde{k}_2$. Soient θ_{T_k} tel que*

$$d(\theta_{T_k}, \theta^*) \leq \frac{r^*_{\Theta}}{3},$$

$e_{T_k} \in B^*_{\mathcal{E}_{T_k}}$ et

$$\left(v^{\mathcal{E}}_{T_k}, v^{\Theta}_{T_k}\right) \in B_{T\mathcal{E}_{T_k} \times T\Theta} \cap \mathcal{Z}_{T_k}.$$

Notons (θ_t) la suite de paramètres produite par l'algorithme « NoBackTrack » initialisé à l'instant T_k, avec la suite de pas $\mu\,\boldsymbol{\eta}$. Alors, pour tout $T_k \leq t \leq T_{k+1}$, θ_t appartient à B^*_{Θ}, et

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\,\eta_t\right) d(\theta_{T_k}, \theta^*) + b^k(\mu).$$

Démonstration. D'après le Corollaire 14.10 appliqué à $t_0 = T_k$, pour tout $T_k + 1 \leq t < T^{\infty}_{T_k}$, θ_t appartient à B^*_{Θ} et

$$d(\theta_t, \theta^*) \leq \left(1 - \lambda^*_{T_k, t-1} \sum_{s=t_0}^{t-1} \mu\,\eta_s\right) d(\theta_{t_0}, \theta^*) + b_{t_0,t}(\mu\,\boldsymbol{\eta}).$$

Or, d'après le Corollaire 15.12, $T_{k+1} < T^{\infty}_{T_k}$. Par conséquent, pour tout $T_k \leq t \leq T_{k+1}$, θ_t appartient à B^*_{Θ}, et

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda^*_{T_k, T_{k+1}-1} \sum_{I_k} \mu\,\eta_t\right) d(\theta_{T_k}, \theta^*) + b^k(\mu).$$

Enfin, $k \geq \tilde{k}_2$ donc, d'après la première propriété du Corollaire 15.9 et le Corollaire 10.6,

$$\lambda^*_{T_k, T_{k+1}-1} \geq \lambda_{\min},$$

et nous obtenons ainsi le résultat annoncé. □

Définition 15.14 (Ensemble sur lequel converge l'algorithme « NoBackTrack »)**.** *Soit, pour* $0 \leq \mu \leq \mu_{\max}$*, l'ensemble aléatoire*

$$\mathcal{S}(\mu) = \bigcap_{k \geq \tilde{k}_2} \left\{ b^k(\mu) \leq \tilde{b}^k(\mu) \right\}.$$

Celui-ci dépend de $\tilde{k}_2$*, mais nous n'explicitons pas la dépendance.*

Lemme 15.15 (Convergence de l'algorithme « NoBackTrack » sur $\mathcal{S}(\mu)$)**.** *Soient* $\theta_0 \in B^*_\Theta$*, tel que de plus*

$$d(\theta_0, \theta^*) < \frac{r^*_\Theta}{4},$$

$e_0 \in B^*_{\mathcal{E}_0}$ *et*

$$\left(v_0^{\mathcal{E}}, v_0^{\Theta}\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0.$$

Alors, pour tout $0 \leq \mu \leq \tilde{\mu}_{\max}$*, sur l'ensemble* $\mathcal{S}(\mu)$*, le paramètre* θ_t*, obtenu par l'algorithme « NoBackTrack » qui utilise la suite de pas* $\mu\,\boldsymbol{\eta}$*, converge vers* θ^**, quand* t *tend vers l'infini.*

Démonstration. La preuve est analogue à la preuve de convergence de l'algorithme RTRL qui figure à la section 10.3. Nous indiquons ici, pour chaque étape, les modifications à effectuer.

Initialisation Soit $\tilde{k}_2$ relatif à $\tilde{\mu}_{\max}$ donné par le Corollaire 15.9.

Analogue du Fait 10.14 Le Corollaire 10.11 est remplacé par le Corollaire 15.9. Le Lemme 10.12 est remplacé par le Lemme 15.13, et nous obtenons, à chaque étape de la récurrence,

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\,\eta_t\right) d(\theta_{T_k}, \theta^*) + b^k(\mu).$$

Or, nous sommes sur l'ensemble $\mathcal{S}(\mu)$, et ainsi

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\,\eta_t\right) d(\theta_{T_k}, \theta^*) + \tilde{b}^k(\mu).$$

Nous pouvons alors utiliser la cinquième propriété du Corollaire 15.9 pour conclure.

Analogue du Corollaire 10.15 Il faut remplacer le Corollaire 10.11 par le Corollaire 15.9, et le Fait 10.14 par son analogue.

Analogue du Corollaire 10.16 Les modifications sont les mêmes qu'au point précédent.

Analogue du Lemme 10.18 Le terme $b^k(\mu)$ est remplacé par $\tilde{b}^k(\mu)$. Le Lemme 10.8 est remplacé par le Lemme 15.6.

Analogue du Fait 10.19 Le Corollaire 10.16 et le Lemme 10.18 sont remplacés par leurs analogues ci-dessus.

Analogue du Lemme 10.20 (J_t) est désormais la suite de différentielles produite par l'algorithme « NoBackTrack ». Le Corollaire 10.15 est remplacé par son analogue ci-dessus. Par conséquent, d'après le Corollaire 14.4, pour tout $t \geq 0$,

$$\|\xi_t(\boldsymbol{\theta})\|_{\mathrm{op}} \leq M_\xi.$$

La fin de la preuve est alors inchangée.

Analogue du Fait 10.21 Le Lemme 10.20 et le Fait 10.19 sont remplacés par leurs analogues ci-dessus. □

Chapitre

16 Contrôle probabiliste des trajectoires de l'algorithme « NoBackTrack »

16.1 Ensemble de convergence de l'algorithme « NoBackTrack »

Définition 16.1 (Ensembles de maintien dans les boules de contrôle, et de maîtrise du bruit)**.** *Notons, pour* $k \geq 0$*,*

$$A_k = \left\{ d(\theta, \theta^*) \leq \frac{r^*_\Theta}{3} \right\} \times B^*_{\mathcal{E}_{T_k}} \times \left(B_{T\mathcal{E}_{T_k} \times T\Theta} \cap \mathcal{Z}_{T_k} \right).$$

Pour $k \geq 0$ *et* $\mu \geq 0$*, nous notons* $\theta_{T_k}(\mu)$ *le paramètre,* $e_{T_k}(\mu)$ *l'état, et* $\left(v^{\mathcal{E}}_{T_k}, v^{\Theta}_{T_k} \right)(\mu)$ *les vecteurs spécifiques, à l'instant* T_k*, produits par l'algorithme « NoBackTrack » avec la suite de pas de descente* $\mu \boldsymbol{\eta}$*, avec des quantités initiales appartenant à l'ensemble* A_0*, et un paramètre initial vérifiant*

$$d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{4}.$$

Ces quantités dépendent du tirage des signes aléatoires.

Pour $k \geq 0$*, nous définissons alors les ensembles aléatoires*

$$\mathcal{A}_k(\mu) = \left\{ \left(\theta_{T_k}(\mu), e_{T_k}(\mu), \left(v^{\mathcal{E}}_{T_k}, v^{\Theta}_{T_k} \right)(\mu) \right) \in A_k \right\}$$

et

$$\mathcal{C}_k(\mu) = \mathcal{A}_k(\mu) \cap \left\{ b^k(\mu) \leq \tilde{b}^k(\mu) \right\}.$$

Définition 16.2 (Ensemble de convergence)**.** *Notons alors, pour tout* $0 \leq K < K'$*,*

$$\mathcal{E}_{K,K'}(\mu) = \bigcap_{k=K}^{K'-1} \mathcal{C}_k(\mu) = \bigcap_{k=K}^{K'-1} \mathcal{A}_k(\mu) \cap \left\{ b^k(\mu) \leq \tilde{b}^k(\mu) \right\},$$

et

$$\mathcal{E}_{K,\infty}(\mu) = \bigcap_{k \geq K} \mathcal{C}_k(\mu)$$

l'intersection décroissante des $\mathcal{E}_{K,K'}(\mu)$*, pour* $K' > K$*.*

Corollaire 16.3 (Convergence sur l'ensemble $\mathcal{E}_{\tilde{k}_2,\infty}(\mu)$)**.** *Pour tout* $0 \leq \mu \leq \tilde{\mu}_{\max}$*, l'algorithme « NoBackTrack » converge sur l'ensemble* $\mathcal{E}_{\tilde{k}_2,\infty}(\mu)$*.*

Démonstration. Pour tout $\mu \geq 0$, l'ensemble $\mathcal{E}_{\tilde{k}_2,\infty}(\mu)$ est inclus dans l'ensemble $\mathcal{S}(\mu)$ par construction. Le résultat est alors une conséquence du Lemme 15.15. □

Corollaire 16.4 (Inclusion de $\mathcal{C}_k(\mu)$ dans $\mathcal{A}_{k+1}(\mu)$)**.** *Pour tous* $0 \leq \mu \leq \tilde{\mu}_{\max}$ *et* $k \geq \tilde{k}_2$*,*

$$\mathcal{C}_k(\mu) = \mathcal{A}_k(\mu) \cap \left\{ b^k(\mu) \leq \tilde{b}^k(\mu) \right\} \subset \mathcal{A}_{k+1}(\mu).$$

Démonstration. Soient $0 \leq \mu \leq \tilde{\mu}_{\max}$ et $k \geq \tilde{k}_2$. Notons (θ_t) la suite de paramètres produite par l'algorithme « NoBackTrack » avec la suite de pas $\mu\,\boldsymbol{\eta}$.

D'après la définition de $\mathcal{A}_k(\mu)$, les hypothèses du Lemme 15.13 sont satisfaites par θ_{T_k}, $e_{T_k}(\mu)$ et $\left(v^{\mathcal{E}}_{T_k}, v^{\Theta}_{T_k}\right)(\mu)$. Ainsi, pour tout $T_k \leq t \leq T_{k+1}$, θ_t appartient à B^*_Θ, et

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \mu\,\eta_t\right) d\left(\theta_{T_k}, \theta^*\right) + b^k(\mu).$$

Par conséquent, sur $\mathcal{C}_k(\mu)$, comme $b^k(\mu) \leq \tilde{b}^k(\mu)$, nous obtenons

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \left(1 - \lambda_{\min} \sum_{I_k} \eta_t\right) d\left(\theta_{T_k}, \theta^*\right) + \tilde{b}^k(\mu).$$

Or, $k \geq \tilde{k}_2$ et, d'après le Corollaire 15.9, $\tilde{k}_2 \geq \tilde{k}_1(\mu)$. Ainsi, le terme majorant est inférieur à $r^*_\Theta/3$, de sorte que

$$d\left(\theta_{T_{k+1}}, \theta^*\right) \leq \frac{r^*_\Theta}{3}.$$

Enfin comme, pour tout $T_k \leq t < T_{k+1}$, θ_t appartient à B^*_Θ, des récurrences conjuguées au Corollaire 4.52 et au Lemme 13.15 nous permettent d'obtenir

$$e_{T_{k+1}}(\mu) \in B^*_{\mathcal{E}_{T_{k+1}}} \quad \text{et} \quad \left(v^{\mathcal{E}}_{T_{k+1}}, v^{\Theta}_{T_{k+1}}\right)(\mu) \in B_{T\mathcal{E}_{T_{k+1}} \times T\Theta} \cap \mathcal{Z}_{T_{k+1}}.$$

Nous avons ainsi établi l'inclusion

$$\mathcal{C}_k(\mu) \subset \mathcal{A}_{k+1}(\mu),$$

ce qui est le résultat annoncé. □

16.2 Mesurabilité des quantités maintenues par l'algorithme « NoBackTrack »

Lemme 16.5 (Mesurabilité le long de la trajectoire « NoBackTrack »)**.** *Soient* $\theta_0 \in \Theta$*,* $e_0 \in \mathcal{E}_0$ *et*

$$\left(v^{\mathcal{E}}_0, v^{\Theta}_0\right) \in T\mathcal{E}_0 \times T\Theta.$$

Nous supposons que θ_0*,* e_0 *et* $\left(v^{\mathcal{E}}_0, v^{\Theta}_0\right)$ *sont* $\mathcal{F}_0$*–mesurable. Soit une suite* $\boldsymbol{\eta}$ *de pas de descente. Soient* (θ_t) *la suite de paramètres,* (e_t) *la suite d'états, et* $\left(\left(v^{\mathcal{E}}_t, v^{\Theta}_t\right)\right)$ *la suite de vecteurs spécifiques produits par l'algorithme « NoBackTrack ». Alors, pour tout* $t \geq 0$*,* θ_t*,* e_t *et* $\left(v^{\mathcal{E}}_t, v^{\Theta}_t\right)$ *sont* $\mathcal{F}_t$*–mesurables.*

Démonstration. Prouvons le résultat par récurrence sur $t \geq 0$. L'initialisation est acquise par hypothèse. Soit alors $t \geq 0$, et supposons le résultat vrai pour t. Alors,

d'après la Définition 14.5, θ_{t+1} et e_{t+1} sont des fonctions Borel mesurables des quantités à l'instant t. Or, ces dernières sont $\mathcal{F}_t$–mesurables, donc θ_{t+1} et e_{t+1} sont $\mathcal{F}_{t+1}$–mesurables. De plus, d'après les Définitions 14.5 et 13.4,

$$\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right) = \mathcal{R}\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\,;\, \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right),\, \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right),\, \varepsilon\left(t+1\right)\right),$$

donc $\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right)$ est une fonction Borel mesurable des quantités à l'instant t et de $\varepsilon\left(t+1\right)$, et est ainsi $\mathcal{F}_{t+1}$–mesurable, ce qui conclut la récurrence, puis la preuve. □

Remarque 16.6. *Le fait que, pour tout $t \geq 1$, θ_t et e_t sont $\mathcal{F}_{t-1}$–mesurables, ce qui se voit dans la preuve ci-dessus, est une conséquence de ce que nous utilisons $\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right)$ dans la mise à jour de θ_t. Effectuer celle-ci avec $\left(v_{t+1}^{\mathcal{E}},\, v_{t+1}^{\Theta}\right)$ ne modifierait pas les résultats de convergence obtenus.*

Remarque 16.7. *Pour tout $t \geq 0$, $\xi_t\left(\boldsymbol{\theta}\right)$ est $\mathcal{F}_{t+1}$–mesurable.*

Rappelons que le terme $\tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right)$ a été introduit à la Définition 15.2.

Corollaire 16.8 (Mesurabilité des ensembles $\left\{b_{t_0,\,t}\left(\boldsymbol{\eta}\right) \leq \tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right)\right\}$)**.** *Soient $0 \leq t_0 < t$. Alors, l'ensemble aléatoire*

$$\left\{b_{t_0,\,t}\left(\boldsymbol{\eta}\right) \leq \tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right)\right\}$$

est $\mathcal{F}_t$–mesurable.

Remarque 16.9. *Notons $\boldsymbol{\theta}$ la suite de paramètres produite par l'algorithme « NoBackTrack ». Comme, pour tout $t \geq 0$, $\xi_t\left(\boldsymbol{\theta}\right)$ est $\mathcal{F}_{t+1}$–mesurable, à l'exception des « quantités à l'instant t_0 », $b_{t_0,\,t}\left(\boldsymbol{\eta}\right)$ ne fait intervenir que les $\varepsilon\left(s\right)$, pour $t_0 + 1 \leq s \leq t$.*

Démonstration. Notons $\boldsymbol{\theta}$ la suite (aléatoire) de paramètres produite par l'algorithme « NoBackTrack ». D'après le Lemme 16.5, les quantités produites par l'algorithme « NoBackTrack » à l'instant t_0 sont $\mathcal{F}_{t_0}$–mesurables. De plus, pour tout $s \geq 0$, $\xi_s\left(\boldsymbol{\theta}\right)$ est $\mathcal{F}_{s+1}$–mesurable. Or, d'après les Définitions 8.14 et 9.32, $b_{t_0,\,t}\left(\boldsymbol{\eta}\right)$ est une fonction Borel mesurable des quantités produites par l'algorithme « NoBackTrack » à l'instant t_0 et des $\xi_s\left(\boldsymbol{\theta}\right)$, pour $t_0 \leq s \leq t-1$. Or, l'ensemble

$$\left]-\infty,\, \tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right)\right]$$

est un borélien, ce qui conclut la preuve. □

16.3 Contrôle du bruit

Pour les deux prochains lemmes (Lemmes 16.10 et 16.11), nous fixons θ_0 dans B_{Θ}^{*}, tel que de plus

$$d\left(\theta_0,\, \theta^*\right) \leq \frac{r_{\Theta}^{*}}{3},$$

$e_0 \in B_{\mathcal{E}_0}^{*}$ et

$$\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0.$$

Nous supposons de plus que ces quantités sont $\mathcal{F}_0$–mesurables. Nous supposons enfin disposer d'une suite $\boldsymbol{\eta}$ telle que la suite $\tilde{\boldsymbol{\eta}}$ est incluse dans le segment $[0, \eta_{\max}]$. Nous notons enfin $\boldsymbol{\theta} = (\theta_t)$ la suite de paramètres, (e_t) la suite d'états, et $\left(\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right)\right)$ la suite de vecteurs spécifiques produits par l'algorithme « NoBackTrack ».

Lemme 16.10 (Absence de biais de la mise à jour « NoBackTrack »)**.** *Pour tout* $t \geq 0$*,*

$$\mathrm{E}\left[v_{t+1}^{\mathcal{E}} \otimes v_{t+1}^{\Theta} \middle| \mathcal{F}_t\right] = \frac{\partial \boldsymbol{T}_t}{\partial e}\left(e_t,\, \theta_t\right) \cdot \left(v_t^{\mathcal{E}} \otimes v_t^{\Theta}\right) + \frac{\partial \boldsymbol{T}_t}{\partial \theta}\left(e_t,\, \theta_t\right),$$

et

$$\mathrm{E}\left[\xi_t\left(\boldsymbol{\theta}\right) \middle| \mathcal{F}_t\right] = 0.$$

Démonstration. D'après le début de la section 16.3, θ_0, e_0 et $\left(v_0^{\mathcal{E}},\, v_0^{\Theta}\right)$ sont $\mathcal{F}_0$–mesurables donc, d'après le Lemme 16.5, pour tout $t \geq 0$, θ_t, e_t et $\left(v_t^{\mathcal{E}},\, v_t^{\Theta}\right)$ sont $\mathcal{F}_t$–mesurables. Le premier point est alors une conséquence du Lemme 12.12, et le deuxième en découle d'après la Définition 14.1. □

Lemme 16.11 (Bruits à des instants différents non corrélés)**.** *Pour tous instants* $s \neq t$*,*

$$\mathrm{E}\left[\left\langle \xi_t\left(\boldsymbol{\theta}\right), \xi_s\left(\boldsymbol{\theta}\right)\right\rangle \middle| \mathcal{F}_0\right] = 0.$$

Démonstration. Sans perte de généralité, nous pouvons supposer $0 \leq s < t$. D'après le Lemme 16.5 et la Définition 14.1, $\xi_s\left(\boldsymbol{\theta}\right)$ est $\mathcal{F}_{s+1}$–mesurable, donc $\mathcal{F}_t$–mesurable car $s + 1 \leq t$. Par conséquent,

$$\begin{aligned}
\mathrm{E}\left[\left\langle \xi_t\left(\boldsymbol{\theta}\right), \xi_s\left(\boldsymbol{\theta}\right)\right\rangle \middle| \mathcal{F}_0\right] &= \mathrm{E}\left[\mathrm{E}\left[\left\langle \xi_t\left(\boldsymbol{\theta}\right), \xi_s\left(\boldsymbol{\theta}\right)\right\rangle \middle| \mathcal{F}_t\right] \middle| \mathcal{F}_0\right] \\
&= \mathrm{E}\left[\left\langle \mathrm{E}\left[\xi_t\left(\boldsymbol{\theta}\right) \middle| \mathcal{F}_t\right], \xi_s\left(\boldsymbol{\theta}\right)\right\rangle \middle| \mathcal{F}_0\right] \\
&= 0,
\end{aligned}$$

d'après le Lemme 16.10, ce qui conclut la preuve. □

Lemme 16.12 (Contrôle d'une somme double des pas)**.** *Pour tout* $t \geq 0$*,*

$$\sum_{0 \leq p,\, q \leq t} \tilde{\eta}_p\, \tilde{\eta}_q\, \left(1 - \alpha\right)^{|p-q|} \leq \left(1 + \frac{2}{\alpha}\right) \sum_{p=0}^{t} \tilde{\eta}_p^2.$$

Démonstration. Soit $t \geq 0$.

$$\begin{aligned}
\sum_{0 \leq p,\, q \leq t} \tilde{\eta}_p\, \tilde{\eta}_q\, \left(1 - \alpha\right)^{|p-q|} &= \sum_{0 \leq p,\, q \leq t} \tilde{\eta}_p\, \tilde{\eta}_q\, \left(1 - \alpha\right)^{\frac{|p-q|}{2}}\, \left(1 - \alpha\right)^{\frac{|p-q|}{2}} \\
&\leq \sum_{0 \leq p,\, q \leq t} \tilde{\eta}_p^2\, \left(1 - \alpha\right)^{|p-q|},
\end{aligned}$$

d'après l'inégalité de Cauchy-Schwarz. Or,

$$\begin{aligned}
\sum_{0 \leq p,\, q \leq t} \tilde{\eta}_p^2\, \left(1 - \alpha\right)^{|p-q|} &= \sum_{p=0}^{t} \tilde{\eta}_p^2 + 2 \sum_{p=0}^{t-1} \tilde{\eta}_p^2 \sum_{q=p+1}^{t} \left(1 - \alpha\right)^{q-p} \\
&= \sum_{p=0}^{t} \tilde{\eta}_p^2 + 2 \sum_{p=0}^{t-1} \tilde{\eta}_p^2 \sum_{q=1}^{t-p} \left(1 - \alpha\right)^{q} \\
&\leq \left(1 + \frac{2}{\alpha}\right) \sum_{p=0}^{t} \tilde{\eta}_p^2,
\end{aligned}$$

car la série de terme général $\left(1 - \alpha\right)^q$ est convergente, ce qui établit bien le résultat annoncé. □

Lemme 16.13 (Espérance du terme d'erreur dû au bruit « NoBackTrack »)**.** *Soit* $t_0 \geq 0$*. Soit* $\theta_{t_0} \in B^*_\Theta$*, tel que de plus*

$$d\left(\theta_{t_0},\, \theta^*\right) \leq \frac{r^*_\Theta}{3}.$$

Soient $e_{t_0} \in B^*_{\mathcal{E}_{t_0}}$ *et*

$$\left(v^{\mathcal{E}}_{t_0},\, v^{\Theta}_{t_0}\right) \in B_{T\mathcal{E}_{t_0}\times T\Theta} \cap \mathcal{Z}_{t_0}.$$

Nous supposons que θ_{t_0}*,* e_{t_0} *et* $\left(v^{\mathcal{E}}_{t_0},\, v^{\Theta}_{t_0}\right)$ *sont* $\mathcal{F}_{t_0}$*–mesurables. Soit une suite* $\boldsymbol{\eta} = (\eta_t)$ *telle que la suite* $\tilde{\boldsymbol{\eta}}$ *est incluse dans le segment* $[0,\, \eta_{\max}]$*. Notons alors* $\boldsymbol{\theta} = (\theta_t)$ *la suite de paramètres produite par l'algorithme « NoBackTrack » initialisé à l'instant* t_0*. Alors, pour tout* $t_0 + 1 \leq t < T^{r^*_\Theta}_{t_0}$*,*

$$\mathrm{E}\left[\left\|\sum_{s=t_0}^{t-1}\left(\sum_{p=s}^{t-1}\tilde{\eta}_p\,\Pi^{t_0}_{s,\,p}\right)\xi_s\left(\boldsymbol{\theta}\right)\right\|_{\mathrm{op}}\middle|\,\mathcal{F}_{t_0}\right] = O\left(\left(\sum_{s=t_0}^{t-1}\tilde{\eta}_s^2\right)^{1/2}\right),$$

avec des constantes qui ne dépendent que de α *et* M_ξ*.*

Démonstration. Nous prouvons le résultat dans le cas $t_0 = 0$, la preuve est identique dans le cas général (ou se déduit du cas particulier en considérant l'algorithme « NoBackTrack » initialisé à t_0).

Soit $1 \leq t < T^{r^*_\Theta}_0$. D'après la Définition 8.14, pour tous $0 \leq s,\, p \leq t-1$, $\Pi^0_{s,\,p}$ est $\mathcal{F}_0$–mesurable. Donc, d'après le Lemme 16.11, les bruits à des instants différents sont non corrélés, conditionnellement à $\mathcal{F}_0$, soit pour la loi des $\varepsilon\left(s\right)$, avec $s \geq 1$. Donc,

$$\begin{aligned}\mathrm{E}\left[\left\|\sum_{s=0}^{t-1}\left(\sum_{p=s}^{t-1}\tilde{\eta}_p\,\Pi^0_{s,\,p}\right)\xi_s\left(\boldsymbol{\theta}\right)\right\|^2_{\mathrm{op}}\middle|\,\mathcal{F}_0\right] &= \sum_{s=0}^{t-1}\mathrm{E}\left[\left\|\left(\sum_{p=s}^{t-1}\tilde{\eta}_p\,\Pi^0_{s,\,p}\right)\xi_s\left(\boldsymbol{\theta}\right)\right\|^2_{\mathrm{op}}\middle|\,\mathcal{F}_0\right] \\ &\leq \sum_{s=0}^{t-1}\left\|\sum_{p=s}^{t-1}\tilde{\eta}_p\,\Pi^0_{s,\,p}\right\|^2_{\mathrm{op}}\mathrm{E}\left[\left\|\xi_s\left(\boldsymbol{\theta}\right)\right\|^2_{\mathrm{op}}\middle|\,\mathcal{F}_0\right].\end{aligned} \tag{16.1}$$

Contrôle de la norme d'opérateur Or, pour tous $0 \leq s \leq p \leq t-1$,

$$\left\|\sum_{p=s}^{t-1}\tilde{\eta}_p\,\Pi^0_{s,\,p}\right\|^2_{\mathrm{op}} \leq \left(\sum_{p=s}^{t-1}\tilde{\eta}_p\,\left\|\Pi^0_{s,\,p}\right\|_{\mathrm{op}}\right)^2.$$

De plus, d'après la Définition 8.14, le Corollaire 4.53, le Corollaire 5.11 et le Lemme 4.39, pour tous $0 \leq s \leq p \leq t-1$,

$$\left\|\Pi^0_{s,\,p}\right\|_{\mathrm{op}} \leq (1-\alpha)^{p-s}$$

donc, pour tous $0 \leq s \leq p \leq t-1$,

$$\left\|\sum_{p=s}^{t-1}\tilde{\eta}_p\,\Pi^0_{s,\,p}\right\|^2_{\mathrm{op}} \leq \left(\sum_{p=s}^{t-1}\tilde{\eta}_p\,(1-\alpha)^{p-s}\right)^2.$$

Contrôle des espérances Soit $0 \leq s \leq t-1$. Pour tout $0 \leq p \leq s$, nous savons que $p < T_0^{r_\Theta^*}$, donc $\theta_p \in B_\Theta^*$ et ainsi, d'après le Corollaire 14.4, nous avons

$$\|\xi_s(\boldsymbol{\theta})\|_{\text{op}} \leq M_\xi.$$

Par conséquent, pour tout $0 \leq s \leq t-1$,

$$\mathrm{E}\left[\|\xi_s(\boldsymbol{\theta})\|^2 \middle| \mathcal{F}_0\right] \leq M_\xi^2.$$

Contrôle du terme majorant de l'Équation (16.1) D'après les deux points précédents, le terme majorant de l'Équation (16.1) est inférieur à

$$M_\xi^2 \sum_{s=0}^{t-1} \sum_{s \leq p,\, q \leq t-1} \tilde{\eta}_p\, \tilde{\eta}_q \left\|\Pi_{s,\,p}^0\right\|_{\text{op}} \left\|\Pi_{s,\,q}^0\right\|_{\text{op}},$$

donc à

$$M_\xi^2 \sum_{0 \leq p,\, q \leq t-1} \tilde{\eta}_p\, \tilde{\eta}_q \sum_{s=0}^{\min(p,\,q)} (1-\alpha)^{p-s} (1-\alpha)^{q-s},$$

soit encore à

$$M_\xi^2 \sum_{0 \leq p,\, q \leq t-1} \tilde{\eta}_p\, \tilde{\eta}_q\, (1-\alpha)^{\max(p,\,q)-\min(p,\,q)} \sum_{s=0}^{\min(p,\,q)} (1-\alpha)^{2\,(\min(p,\,q)-s)},$$

et enfin, en effectuant le changement de variable $s' = \min(p,\, q) - s$ dans la somme de droite, puis en notant $s' = s$,

$$M_\xi^2 \sum_{0 \leq p,\, q \leq t-1} \tilde{\eta}_p\, \tilde{\eta}_q\, (1-\alpha)^{|p-q|} \sum_{s=0}^{\min(p,\,q)} (1-\alpha)^{2\,s}.$$

Or, la somme de droite est bornée indépendamment de p et q, car $0 \leq 1-\alpha < 1$. Enfin, d'après le Lemme 16.12,

$$\sum_{0 \leq p,\, q \leq t-1} \tilde{\eta}_p\, \tilde{\eta}_q\, (1-\alpha)^{|p-q|} \leq \left(1 + \frac{2}{\alpha}\right) \sum_{s=0}^{t-1} \tilde{\eta}_s^2,$$

ce qui conclut la preuve. □

Lemme 16.14 (Probabilité de majoration du terme additif)**.** *Soit $t_0 \geq 0$. Soit $\theta_{t_0} \in B_\Theta^*$, tel que de plus*

$$d(\theta_{t_0},\, \theta^*) \leq \frac{r_\Theta^*}{3}.$$

Soient $e_{t_0} \in B_{\mathcal{E}_{t_0}}^*$ et

$$\left(v_{t_0}^{\mathcal{E}},\, v_{t_0}^{\Theta}\right) \in B_{T\mathcal{E}_{t_0} \times T\Theta} \cap \mathcal{Z}_{t_0}.$$

Nous supposons de plus que ces quantités sont $\mathcal{F}_{t_0}$–mesurables. Alors, pour tout $t_0 + 1 \leq t < T_{t_0}^{r_\Theta^*}$,

$$P\left(b_{t_0,\,t}(\boldsymbol{\eta}) \leq \tilde{b}_{t_0,\,t}(\boldsymbol{\eta})\right) \geq 1 - M_7 \left(\sum_{s=t_0}^{t-1} \eta_s\right)^{-1} \left(\sum_{s=t_0}^{t-1} \eta_s^{1+\nu}\right)^{\frac{1}{1+\nu}}.$$

Démonstration. Notons $\boldsymbol{\theta} = (\theta_s)$ la suite de paramètres produite par l'algorithme « NoBackTrack » initialisé à l'instant t_0. Soit $t_0 + 1 \leq t < T_{t_0}^{r_\Theta^*}$. D'après le Corollaire 16.8 précédent, l'ensemble

$$\left\{ b_{t_0,\,t}\left(\boldsymbol{\eta}\right) \leq \tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right) \right\}$$

est $\mathcal{F}_t$–mesurable, donc sa probabilité est bien définie. Alors, d'après les Définitions 9.32 et 15.2,

$$P\left(b_{t_0,\,t}\left(\boldsymbol{\eta}\right) \geq \tilde{b}_{t_0,\,t}\left(\boldsymbol{\eta}\right) \right)$$

est inférieur à

$$P\left(\left\| \sum_{s=t_0}^{t-1} \left(\sum_{p=s}^{t-1} \tilde{\eta}_p \, \Pi_{s,\,p}^{t_0} \right) \xi_s\left(\boldsymbol{\theta}\right) \right\| \geq \frac{M_{12}}{M_7} \frac{\left(\sum_{s=t_0}^{t-1} \tilde{\eta}_s^2 \right)^{\frac{1}{2}}}{\left(\sum_{s=t_0}^{t-1} \eta_s^{1+\nu} \right)^{\frac{1}{1+\nu}}} \sum_{s=t_0}^{t-1} \eta_s \right).$$

Or, le terme majorant est inférieur à

$$\frac{M_7}{M_{12}} \left(\sum_{s=t_0}^{t-1} \eta_s \right)^{-1} \frac{\left(\sum_{s=t_0}^{t-1} \eta_s^{1+\nu} \right)^{\frac{1}{1+\nu}}}{\left(\sum_{s=t_0}^{t-1} \tilde{\eta}_s^2 \right)^{\frac{1}{2}}} \mathrm{E}\left[\left\| \sum_{s=t_0}^{t-1} \left(\sum_{p=s}^{t-1} \tilde{\eta}_p \, \Pi_{s,\,p}^{t_0} \right) \xi_s\left(\boldsymbol{\theta}\right) \right\| \right].$$

De plus, $t < T_{t_0}^{r_\Theta^*}$ donc, d'après le Lemme 14.7, pour tout $t_0 \leq s \leq t-1$, $\theta_s \in B_\Theta^*$. Par conséquent, d'après le Lemme 16.13 et la définition de M_{12} à la Définition 15.2, le terme ci-dessus est majoré par

$$M_7 \left(\sum_{s=t_0}^{t-1} \eta_s \right)^{-1} \frac{\left(\sum_{s=t_0}^{t-1} \eta_s^{1+\nu} \right)^{\frac{1}{1+\nu}}}{\left(\sum_{s=t_0}^{t-1} \tilde{\eta}_s^2 \right)^{\frac{1}{2}}} \left(\sum_{s=t_0}^{t-1} \tilde{\eta}_s^2 \right)^{1/2},$$

qui est égal à

$$M_7 \left(\sum_{s=t_0}^{t-1} \eta_s \right)^{-1} \left(\sum_{s=t_0}^{t-1} \eta_s^{1+\nu} \right)^{\frac{1}{1+\nu}},$$

ce qui conclut la preuve. □

16.4 Probabilité de l'ensemble de convergence

Corollaire 16.15 (Mesurabilité des ensembles étudiés)**.** *Pour tout $\mu \geq 0$, pour tout $k \geq 0$, les ensembles $\mathcal{A}_k\left(\mu\right)$ et $\mathcal{C}_k\left(\mu\right)$ sont $\mathcal{F}_{T_k}$–mesurables, et l'ensemble*

$$\left\{ b^k\left(\mu\right) \leq \tilde{b}^k\left(\mu\right) \right\}$$

est $\mathcal{F}_{T_{k+1}}$–mesurable. Enfin, pour tous $0 \leq K < K'$, les ensembles $\mathcal{E}_{K,\,K'}\left(\mu\right)$ sont $\mathcal{F}_{T_{K'}-1}$–mesurables.

Démonstration. C'est une conséquence du Lemme 16.5 et du Corollaire 16.8. □

Lemme 16.16 (Probabilités des $\mathcal{E}_{K,\,K'}\left(\mu\right)$)**.** *Pour tous $0 \leq \mu \leq \tilde{\mu}_{\max}$ et $\tilde{k}_2 \leq K < K'$,*

$$P\left(\mathcal{E}_{K,\,K'}\left(\mu\right)\right) \geq \left(\prod_{k=K}^{K'-1} \left(1 - u_k\right) \right) P\left(\mathcal{A}_K\left(\mu\right)\right).$$

Démonstration. Soient $0 \leq \mu \leq \tilde{\mu}_{\max}$ et $\tilde{k}_2 \leq K$. Soit également $K' > K$, (qui n'est pas « celui » de l'énoncé mais un entier strictement supérieur à K « générique »).

Minoration de $P\left(\mathcal{E}_{K,\,K'+1}\left(\mu\right)\right)$

$$\begin{aligned}\mathcal{E}_{K,\,K'+1}\left(\mu\right) &= \mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{C}_{K'}\left(\mu\right) \\ &= \mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{A}_{K'}\left(\mu\right) \cap \left\{b^{K'}\left(\mu\right) \leq \tilde{b}^{K'}\left(\mu\right)\right\}\end{aligned}$$

donc, comme tous les ensembles concernés sont mesurables pour $\mathcal{F}_{T_{K'}}$ (par exemple), d'après le Corollaire 16.15,

$$P\left(\mathcal{E}_{K,\,K'+1}\left(\mu\right)\right) = P\left(\mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{A}_{K'}\left(\mu\right) \cap \left\{b^{K'}\left(\mu\right) \leq \tilde{b}^{K'}\left(\mu\right)\right\}\right).$$

Or, d'après le même corollaire, $\mathcal{E}_{K,\,K'}\left(\mu\right)$ est $\mathcal{F}_{T_{K'}-1}$–mesurable, et $\mathcal{A}_{K'}\left(\mu\right)$ est $\mathcal{F}_{T_{K'}}$–mesurable, donc

$$\begin{aligned}P\left(\mathcal{E}_{K,\,K'+1}\left(\mu\right)\right) &= \mathrm{E}\left[\mathbf{1}\left\{\mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{A}_{K'}\left(\mu\right)\right\}\, \mathrm{E}\left[\mathbf{1}\left\{b^{K'}\left(\mu\right) \leq \tilde{b}^{K'}\left(\mu\right)\right\}\middle|\, \mathcal{F}_{T_{K'}}\right]\right] \\ &= \mathrm{E}\left[\mathbf{1}\left\{\mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{A}_{K'}\left(\mu\right)\right\}\; P\left(b^{K'}\left(\mu\right) \leq \tilde{b}^{K'}\left(\mu\right)\middle|\, \mathcal{F}_{T_{K'}}\right)\right].\end{aligned}$$

Or, d'après le Corollaire 15.12, nous savons que

$$T_{K'+1} < T^{\infty}_{T_{K'}}.$$

De plus, d'après la Définition 15.5,

$$\tilde{b}^{K'}\left(\mu\right) = \tilde{b}_{T_{K'},\,T_{K'+1}}\left(\mu\,\boldsymbol{\eta}\right).$$

Par conséquent, d'après le Lemme 16.14, sur $\mathcal{A}_{K'}\left(\mu\right)$,

$$\begin{aligned}P\left(b^{K'}\left(\mu\right) \leq \tilde{b}^{K'}\left(\mu\right)\middle|\, \mathcal{F}_{T_{K'}}\right) &\geq 1 - M_7\left(\sum_{I_{K'}} \eta_t\right)^{-1}\left(\sum_{I_{K'}} \eta_t^{1+\nu}\right)^{\frac{1}{1+\nu}} \\ &= 1 - u_{K'}.\end{aligned}$$

Ainsi,

$$\begin{aligned}P\left(\mathcal{E}_{K,\,K'+1}\left(\mu\right)\right) &\geq \mathrm{E}\left[\mathbf{1}\left\{\mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{A}_{K'}\left(\mu\right)\right\}\right]\,\left(1 - u_{K'}\right) \\ &= P\left(\mathcal{E}_{K,\,K'}\left(\mu\right) \cap \mathcal{A}_{K'}\left(\mu\right)\right)\,\left(1 - u_{K'}\right).\end{aligned}$$

Formule de récurrence pour $P\left(\mathcal{E}_{K,\,K'}\left(\mu\right)\right)$ De plus, d'après le Corollaire 16.4,

$$\mathcal{C}_{K'-1}\left(\mu\right) \subset \mathcal{A}_{K'}\left(\mu\right).$$

Or, d'après la Définition 16.2,

$$\mathcal{E}_{K,\,K'}\left(\mu\right) \subset \mathcal{C}_{K'-1}\left(\mu\right),$$

de sorte que

$$\mathcal{A}_{K'}\left(\mu\right) \cap \mathcal{E}_{K,\,K'}\left(\mu\right) = \mathcal{E}_{K,\,K'}\left(\mu\right),$$

ce qui implique

$$P\left(\mathcal{A}_{K'}\left(\mu\right) \cap \mathcal{E}_{K,\,K'}\left(\mu\right)\right) = P\left(\mathcal{E}_{K,\,K'}\left(\mu\right)\right).$$

Ainsi,

$$P\left(\mathcal{E}_{K,\,K'+1}\left(\mu\right)\right) \geq P\left(\mathcal{E}_{K,\,K'}\left(\mu\right)\right)\,\left(1 - u_{K'}\right).$$

Formule pour $P\left(\mathcal{E}_{K,\,K'}(\mu)\right)$ Par conséquent, pour tout $K' > K$,

$$P\left(\mathcal{E}_{K,\,K'}(\mu)\right) \geq \left(\prod_{k=K+1}^{K'-1}(1-u_k)\right) P\left(\mathcal{E}_{K,\,K+1}(\mu)\right).$$

Enfin,

$$\mathcal{E}_{K,\,K+1}(\mu) = \mathcal{C}_K(\mu) = \mathcal{A}_K(\mu) \cap \left\{b^K(\mu) \leq \tilde{b}^K(\mu)\right\}$$

donc, de même que précédemment,

$$P\left(\mathcal{E}_{K,\,K+1}(\mu)\right) \geq P\left(\mathcal{A}_K(\mu)\right)(1-u_K).$$

Donc, pour tout $K' > K$,

$$P\left(\mathcal{E}_{K,\,K'}(\mu)\right) \geq \left(\prod_{k=K}^{K'-1}(1-u_k)\right) P\left(\mathcal{A}_K(\mu)\right),$$

ce qui est bien le résultat annoncé. □

Lemme 16.17 (Minoration de la probabilité de $\mathcal{E}_{\tilde{k}_2,\,\infty}(\mu)$)**.** *Pour tout* $0 \leq \mu \leq \tilde{\mu}_{\max}$,

$$P\left(\mathcal{E}_{\tilde{k}_2,\,\infty}(\mu)\right) \geq \left(\prod_{k=\tilde{k}_2}^{\infty}(1-u_k)\right) P\left(\mathcal{A}_{\tilde{k}_2}(\mu)\right).$$

Démonstration. Faisons tendre K' vers l'infini dans l'inégalité du Lemme 16.16. Alors, par convergence monotone décroissante pour une mesure finie (car la mesure est une mesure de probabilité), le terme de gauche tend vers

$$P\left(\mathcal{E}_{\tilde{k}_2,\,\infty}(\mu)\right).$$

Le produit du terme de droite tend vers le produit infini des $1-u_k$, d'après le Lemme 11.12. □

Lemme 16.18 (Probabilité arbitrairement grande de l'ensemble de convergence)**.** *Pour tout* $\varepsilon > 0$*, nous pouvons trouver* $\tilde{\mu}_{\max} > 0$ *et* $\tilde{k}_2 \geq 0$ *tels que, pour tout* $0 \leq \mu \leq \tilde{\mu}_{\max}$*, nous avons*

$$P\left(\mathcal{E}_{\tilde{k}_2,\,\infty}(\mu)\right) \geq 1-\varepsilon.$$

Démonstration. D'après le Lemme 16.17 précédent, pour tout $0 \leq \mu \leq \tilde{\mu}_{\max}$,

$$P\left(\mathcal{E}_{\tilde{k}_2,\,\infty}(\mu)\right) \geq \left(\prod_{k=\tilde{k}_2}^{\infty}(1-u_k)\right) P\left(\mathcal{A}_{\tilde{k}_2}(\mu)\right).$$

Or, d'après le sixième point du Corollaire 15.9,

$$\prod_{k=\tilde{k}_2}^{\infty}(1-u_k) \geq 1-\varepsilon$$

et, d'après le septième point du même lemme et la Définition 16.1, pour tout $0 \leq \mu \leq \tilde{\mu}_{\max}$,

$$P\left(\mathcal{A}_{\tilde{k}_2}(\mu)\right) = 1,$$

ce qui conclut la preuve. □

Corollaire 16.19 (Probabilité arbitrairement grande de l'intersection des ensembles de convergence)**.** *Pour tout $\varepsilon > 0$, nous pouvons trouver $\tilde{\mu}_{\max} > 0$ et $\tilde{k}_2 \geq 0$ tels que*

$$P\left(\bigcap_{0 \leq \mu \leq \mu_{\max}} \mathcal{E}_{\tilde{k}_2, \infty}(\mu)\right) \geq 1 - \varepsilon.$$

Démonstration. Les ensembles $\mathcal{C}_k(\mu)$ sont croissants en μ, donc les ensembles $\mathcal{E}_{\tilde{k}_2, \infty}(\mu)$ le sont également. Ainsi, l'intersection de l'énoncé est égale à l'intersection prise sur les μ rationnels, de sorte qu'elle est mesurable et que sa probabilité est bien définie.

De plus, toujours par croissance des $\mathcal{E}_{\tilde{k}_2, \infty}(\mu)$, pour tout $n \geq 0$,

$$\bigcap_{\frac{1}{n} \leq \mu \leq \mu_{\max}} \mathcal{E}_{\tilde{k}_2, \infty}(\mu) = \mathcal{E}_{\tilde{k}_2, \infty}\left(\frac{1}{n}\right),$$

de sorte que, d'après le Lemme 16.18,

$$P\left(\bigcap_{\frac{1}{n} \leq \mu \leq \mu_{\max}} \mathcal{E}_{\tilde{k}_2, \infty}(\mu)\right) = P\left(\mathcal{E}_{\tilde{k}_2, \infty}\left(\frac{1}{n}\right)\right) \geq 1 - \varepsilon.$$

Enfin, l'intersection pour $\mu \geq 1/n$ ci-dessus tend vers l'intersection de l'énoncé, lorsque n tend vers l'infini. Par conséquent, par convergence monotone décroissante pour une mesure finie (comme la mesure est une mesure de probabilité), le terme de gauche de l'inégalité ci-dessus tend vers le terme de gauche de l'énoncé, lorsque n tend vers l'infini, ce qui conclut la preuve. □

16.5 Énoncé du résultat de convergence de l'algorithme « NoBackTrack »

Proposition 16.20 (Convergence de l'algorithme « NoBackTrack »)**.** *Supposons vérifiées les Hypothèses 4.28, 4.29 et 4.30 sur le système dynamique, les Hypothèses 5.3 et 5.4 sur les pertes et l'Hypothèse 7.1 sur l'opérateur de déplacement sur Θ. Supposons de plus vérifiée l'Hypothèse 13.9 sur les dimensions des espaces des états.*

Supposons également vérifiées l'Hypothèse 11.7 sur les gradients et les hessiennes des pertes sur le paramètre ainsi que sur les pertes, et l'Hypothèse 11.8 sur la suite de pas. Supposons enfin vérifiée l'Hypothèse 11.10 sur la convergence des inverses des racines des durées des intervalles.

Soient $\theta_0 \in B^*_\Theta$, tel que de plus

$$d(\theta_0, \theta^*) \leq \frac{r^*_\Theta}{4},$$

$e_0 \in B^*_{\mathcal{E}_0}$, et

$$\left(v_0^{\mathcal{E}}, v_0^{\Theta}\right) \in B_{T\mathcal{E}_0 \times T\Theta} \cap \mathcal{Z}_0.$$

Alors, pour tout $\varepsilon > 0$, nous pouvons trouver $\mu_{\max} > 0$ tel que, avec probabilité supérieure à $1 - \varepsilon$, pour tout $\mu \leq \mu_{\max}$, la suite de paramètres $\boldsymbol{\theta} = (\theta_t)$ produite par l'algorithme « NoBackTrack » qui utilise la suite de pas $\mu\,\boldsymbol{\eta}$ converge vers le paramètre optimal θ^* et, en notant (e_t) la suite d'états produite par l'algorithme « NoBackTrack », nous avons

$$d(e_t, e_t^*) \to 0 \quad et \quad d\left(\boldsymbol{J}_t\left(e_0, v_0^{\mathcal{E}} \otimes v_0^{\Theta}, \boldsymbol{\theta}\right), J_t^*\right) \to 0,$$

quand t tend vers l'infini.

Démonstration. C'est une conséquence des Corollaires 16.3, 16.19 et 10.23. □

Troisième partie

Adaptation en temps réel du pas d'apprentissage d'une descente de gradient

Chapter

17 Speed learning on the fly

Pierre-Yves Massé, Yann Ollivier

The practical performance of online stochastic gradient descent algorithms is highly dependent on the chosen step size, which must be tediously hand-tuned in many applications. The same is true for more advanced variants of stochastic gradients, such as SAGA, SVRG, or AdaGrad. Here we propose to adapt the step size by performing a gradient descent on the step size itself, viewing the whole performance of the learning trajectory as a function of step size. Importantly, this adaptation can be computed online at little cost, without having to iterate backward passes over the full data.

Introduction

This work aims at improving gradient ascent procedures for use in machine learning contexts, by adapting the step size of the descent as it goes along.

Let $\ell_0, \ell_1, \ldots, \ell_t, \ldots$ be functions to be maximised over some parameter space Θ. At each time t, we wish to compute or approximate the parameter $\theta_t^* \in \Theta$ that maximizes the sum

$$L_t(\theta) := \sum_{s \leq t} \ell_s(\theta). \tag{17.1}$$

In the experiments below, as in many applications, $\ell_t(\theta)$ writes $\ell(x_t, \theta)$ for some data $x_0, x_1, \ldots, x_t, \ldots$

A common strategy, especially with large data size or dimensionality [1], is the online stochastic gradient ascent (SG)

$$\theta_{t+1} = \theta_t + \eta \, \partial_\theta \ell_t(\theta_t) \tag{17.2}$$

with step size η, where $\partial_\theta \ell_t$ stands for the Euclidean gradient of ℓ_t with respect to θ.

Such an approach has become a mainstay of both the optimisation and machine learning communities [2]. Various conditions for convergence exist, starting with the celebrated article of Robbins and Monro [3], or later Kushner and Clark [4]. Other types of results are proved in convex settings,

1. Bottou, "Large-scale machine learning with stochastic gradient descent", op. cit.
2. Ibid.
3. Robbins and Monro, "A Stochastic Approximation Method", op. cit.
4. Kushner and Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, op. cit.

Several variants have since been introduced, in part to improve the convergence of the algorithm, which is much slower in stochastic than than in deterministic settings. For instance, algorithms such as SAGA, Stochastic Variance Reduced Gradient (SVRG) or Stochastic Average Gradient (SAG) [5], perform iterations using a comparison between the latest gradient and an average of past gradients. This reduces the variance of the resulting estimates and allows for nice convergence theorems [6], provided a reasonable step size η is used.

Influence of the step size. The ascent requires a parameter, the step size η, usually called "learning rate" in the machine learning community. Empirical evidence highligting the sensitivity of the ascent to its actual numerical value exists aplenty; see for instance the graphs in Section 17.3.2. Slow and tedious hand-tuning is therefore mandatory in most applications. Moreover, admittable values of η depend on the parameterisation retained—except for descents described in terms of Riemannian metrics [7], which provide some degree of parameterisation-invariance.

Automated procedures for setting reasonable value of η are therefore of much value. For instance, AdaGrad [8] divides the derivative $\partial_\theta \ell_t$ by a root mean square average of the magnitude of its recent values, so that the steps are of size approximately 1; but this still requires a "master step size" η.

Shaul, Zhang and LeCun in "No More Pesky Learning Rates" [9] study a simple separable quadratic loss model and compute the value of η which minimises the expected loss after each parameter update. This value can be expressed in terms of computable quantities depending on the trajectory of the descent. These quantities still make sense for non-quadratic models, making this idea amenable to practical use.

More recently, Maclaurin, Douglas and Duvenaud [10] propose to directly conduct a gradient ascent on the hyperparameters (such as the learning rate η) of any algorithm. The gradients with respect to the hyperparameters are computed exactly by "chaining derivatives backwards through the entire training procedure" [11]. Consequently, this approach is extremely impractical in an online setting, as it optimizes the learning rate by performing several passes, each of which goes backwards from time t to time 0.

5. Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives". In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Ed. by Zoubin Ghahramani et al. 2014; Rie Johnson and Tong Zhang. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Ed. by Christopher J. C. Burges et al. 2013; Mark Schmidt, Nicolas Le Roux, and Francis Bach. *Minimizing finite sums with the Stochastic Average Gradient*. Tech. rep. 00860051. HAL, 2013.

6. Defazio, Bach, and Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives", op. cit.; Schmidt, Le Roux, and Bach, *Minimizing finite sums with the Stochastic Average Gradient*, op. cit.

7. Amari, "Natural gradient works efficiently in learning", op. cit.

8. John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2121–2159.

9. Schaul, Zhang, and LeCun, "No More Pesky Learning Rates", op. cit.

10. Douglas Maclaurin, David Duvenaud, and Ryan Adams. "Gradient-based Hypermarameter Optimization through Reversible Learning". In: *Proceedings of The 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. 2015.

11. Ibid.

Finding the best step size. The ideal value of the step size η would be the one that maximizes the cumulated objective function (17.1). Write $\theta_t(\eta)$ for the parameter value obtained after t iterations of the gradient step (17.2) using a given value η, and consider the sum

$$\sum_{s \leq t} \ell_s(\theta_s(\eta)). \tag{17.3}$$

Our goal is to find an online way to approximate the value of η that provides the best value of this sum. This can be viewed as an ascent on the space of stochastic ascent algorithms.

We suggest to update η through a stochastic gradient ascent on this sum:

$$\eta \leftarrow \eta + \alpha \frac{\partial}{\partial \eta} \ell_t(\theta_t(\eta)) \tag{17.4}$$

and then to use, at each time, the resulting value of η for the next gradient step (17.2).

The ascent (17.4) on η depends, in turn, on a step size α. Hopefully, the dependance on α of the whole procedure is somewhat lower than that of the original stochastic gradient scheme on its step size η.

This approach immediately extends to other stochastic gradient algorithms; in what follows we apply it both to the standard SG ascent and to the SVRG algorithm.

The main point in this approach is to find efficient ways to compute or approximate the derivatives $\frac{\partial}{\partial \eta} \ell_t(\theta_t(\eta))$. Indeed, the value $\theta_t(\eta)$ after t steps depends on the whole trajectory of the algorithm, and so does its derivative with respect to η.

After reviewing the setting for gradient ascents in Section 17.1, in Section 17.2.1 we provide an exact but impractical way of computing the derivatives $\frac{\partial}{\partial \eta} \ell_t(\theta_t(\eta))$. Sections 17.2.2–17.2.3 contain the main contribution: SG/SG and SG/AG, practical algorithms to adjust η based on two approximations with respect to these exact derivatives.

Section 17.2.4 extends this to other base algorithms such as SVRG. In Section 17.4 one of the approximations is justified by showing that it computes a derivative, not with respect to a fixed value of η as in (17.4), but with respect to the sequences of values of η effectively used along the way. This also suggests improved algorithms.

Section 17.3 provides experimental comparisons of gradient ascents using traditional algorithms with various values of η, and the same algorithms where η is self-adjusted according to our scheme. The comparisons are done on three sets of synthetic data: a one-dimensional Gaussian model, a one-dimensional Bernoulli model and a 50-dimensional linear regression model: these simple models already exemplify the strong dependence of the traditional algorithms on the value of η.

Terminology. We say that an algorithm is of type "LLR" for "Learning the Learning Rate" when it updates its step size hyperparameter η as it unfolds. We refer to LLR algorithms by a compound abbreviation: "SVRG/SG", for instance, for an algorithm which updates its parameter θ through SVRG and its hyperparameter η through an SG algorithm on η.

17.1 The Stochastic Gradient algorithm

To fix ideas, we define the Stochastic Gradient (SG) algorithm as follows. In all that follows, $\Theta = \mathbb{R}^n$ for some n.[12] The functions ℓ_t are assumed to be smooth. In all our algorithms, the index t starts at 0.

Algorithm 1 (Stochastic Gradient)**.** *We maintain $\theta_t \in \Theta$ (current parameter), initialised at some arbitrary $\theta_0 \in \Theta$. We fix $\eta \in \mathbb{R}$. At each time t, we fix a rate $f(t) \in \mathbb{R}$. The update equation reads:*

$$\theta_{t+1} = \theta_t + \frac{\eta}{f(t)} \partial_\theta \ell_t(\theta_t). \tag{17.5}$$

The chosen rate $f(t)$ usually satisfies the well-known Robbins–Monro conditions[13]:

$$\sum_{t \geq 0} f(t)^{-1} = \infty, \quad \sum_{t \geq 0} f(t)^{-2} < \infty. \tag{17.6}$$

The divergence of the sum of the rates allows the ascent to go anywhere in parameter space, while the convergence of the sum of the squares ensures that variance remains finite. Though custom had it that small such rates should be chosen, such as $f(t) = 1/t$, recently the trend bucked towards the use of large ones, to allow for quick exploration of the parameter space. Throughout the article and experiments we use one such rate:

$$f(t) = \sqrt{t+2} \log(t+3). \tag{17.7}$$

17.2 Learning the learning rate on a stochastic gradient algorithm

17.2.1 The loss as a function of step size

To formalise what we said in the introduction, let us define, for each $\eta \in \mathbb{R}$, the sequence

$$(\theta_0, \theta_1, \theta_2, \ldots) \tag{17.8}$$

obtained by iterating (17.5) from some initial value θ_0. Since they depend on η, we introduce, for each $t > 0$, the operator

$$T_t : \eta \in \mathbb{R} \mapsto T_t(\eta) \in \Theta, \tag{17.9}$$

which maps any $\eta \in \mathbb{R}$ to the parameter θ_t obtained after t iterations of (17.5). T_0 maps every η to θ_0. For each $t \geq 0$, the map T_t is a regular function of η. As explained in the introduction, we want to optimise η according to the function:

$$\mathcal{L}_t(\eta) := \sum_{s \leq t} \ell_s(T_s(\eta)), \tag{17.10}$$

by conducting an online stochastic gradient ascent on it. We therefore need to compute the derivative in (17.4):

$$\frac{\partial}{\partial \eta} \ell_t(T_t(\eta)). \tag{17.11}$$

12. Θ may also be any Riemannian manifold, a natural setting when dealing with gradients. Most of the text is written in this spirit.

13. Robbins and Monro, "A Stochastic Approximation Method", op. cit.

To act more decisively on the order of magnitude of η, we perform an ascent on its logarithm, so that we actually need to compute [14]:

$$\frac{\partial}{\partial \log \eta} \ell_t(T_t(\eta)). \tag{17.12}$$

Now, the derivative of the loss at time t with respect to η can be computed as the product of the derivative of ℓ_t with respect to θ (the usual input of SG) and the derivative of θ_t with respect to η:

$$\frac{\partial}{\partial \log \eta} \ell_t(T_t(\eta)) = \partial_\theta \ell_t(T_t(\eta)) \cdot A_t(\eta) \tag{17.13}$$

where

$$A_t(\eta) := \frac{\partial T_t(\eta)}{\partial \log \eta}. \tag{17.14}$$

Computation of the quantity A_t and its approximation h_t to be introduced later, are the main focus of this text.

Lemma 1. *The derivative $A_t(\eta)$ may be computed through the following recursion equation. $A_0(\eta) = 0$ and, for $t \geq 0$,*

$$A_{t+1}(\eta) = A_t(\eta) + \frac{\eta}{f(t)} \partial_\theta \ell_t(T_t(\eta)) + \frac{\eta}{f(t)} \partial_\theta^2 \ell_t(T_t(\eta)) \cdot A_t(\eta). \tag{17.15}$$

The proof lies in Section .3.1. This update of A involves the Hessian of the loss function with respect to θ, evaluated in the direction of A_t. Often this quantity is unavailable or too costly. Therefore we will use a finite difference approximation instead:

$$\partial_\theta^2 \ell_t(T_t(\eta)) \cdot A_t(\eta) \approx \partial_\theta \ell_t \left(T_t(\eta) + A_t(\eta)\right) - \partial_\theta \ell_t(T_t(\eta)). \tag{17.16}$$

This design ensures that the resulting update on $A_t(\eta)$ uses the gradient of ℓ_t only once:

$$A_{t+1}(\eta) \approx A_t(\eta) + \frac{\eta}{f(t)} \partial_\theta \ell_t \left(T_t(\eta) + A_t(\eta)\right). \tag{17.17}$$

An alternative approach would be to compute the Hessian in the direction A_t by numerical differentiation.

17.2.2 LLR on SG: preliminary version with simplified expressions (SG/SG)

Even with the approximation above, computing the quantities A_t would have a quadratic cost in t: each time we update η thanks to (17.4), we would need to compute anew all the $A_s(\eta)$, $s \leq t$, as well as the whole trajectory $\theta_t = T_t(\eta)$, at each iteration t. We therefore replace the $A_t(\eta)$'s by online approximations, the quantities h_t, which implement the same evolution equation (17.17) as A_t, disregarding the fact that η may have changed in the meantime. These quantities will be interpreted more properly in Section 17.4 as derivatives taken along the effective trajectory of the ascent. This yields the SG/SG algorithm.

14. This is an abuse of notation as T_t is not a function of $\log \eta$ but of η. Formally, we would need to replace T_t with $T_t \circ \exp$, which we refrain from doing to avoid burdensome notation.

Algorithm 2 (SG/SG). *We maintain $\theta_t \in \Theta$ (current parameter), $\eta_t \in \mathbb{R}$ (current step size) and $h_t \in T_{\theta_t}\Theta$ (approximation of the derivative of θ_t with respect to $\log(\eta)$).*

The first two are initialised arbitrarily, and h_0 is set to 0.

The update equations read:

$$\begin{cases} \log \eta_{t+1} = \log \eta_t + \dfrac{1}{\mu_t}\, \partial_\theta \ell_t(\theta_t) \cdot h_t \\ h_{t+1} = h_t + \dfrac{\eta_{t+1}}{f(t)}\, \partial_\theta \ell_t\left(\theta_t + h_t\right) \\ \theta_{t+1} = \theta_t + \dfrac{\eta_{t+1}}{f(t)} \partial_\theta \ell_t(\theta_t)\,, \end{cases} \tag{17.18}$$

where μ_t is some learning rate on $\log \eta$, such as $\mu_t = \sqrt{t+2}\log(t+3)$.

17.2.3 LLR on SG: efficient version (SG/AG)

To obtain better performances, we actually use an adagrad-inspired scheme to update the logarithm of the step size.

Algorithm 3 (SG/AG). *We maintain $\theta_t \in \Theta$ (current parameter), $\eta_t \in \mathbb{R}$ (current step size), $h_t \in T_{\theta_t}\Theta$ (approximation of the derivative of θ_t with respect to $\log(\eta)$), $n_t \in \mathbb{R}$ (average of the squared norms of $\partial \ell_t \circ T_t / \partial \log \eta$), and $d_t \in \mathbb{R}$ (renormalising factor for the computation of n_t).*

θ et η are initially set to θ_0 and η_0, the other variables are set to 0.

At each time t, we compute $\mu_t \in \mathbb{R}$ (a rate used in several updates), and $\lambda_t \in \mathbb{R}$ (the approximate derivative of $\ell_t \circ \theta_t$ with respect to $\log(\eta)$ at η_t).

The update equations read:

$$\begin{cases} \mu_t = \sqrt{t+2}\log(t+3) \\ \lambda_t = l(\theta_t) \cdot h_t \\ d_{t+1} = \left(1 - \dfrac{1}{\mu_t}\right) d_t + \dfrac{1}{\mu_t} \\ n_{t+1}^2 = \left(\left(1 - \dfrac{1}{\mu_t}\right) n_t^2 + \dfrac{1}{\mu_t}\lambda_t^2\right) d_{t+1}^{-1} \\ \log \eta_{t+1} = \log \eta_t + \dfrac{1}{\mu_t}\, \dfrac{\lambda_t}{n_{t+1}} \\ h_{t+1} = h_t + \dfrac{\eta_{t+1}}{f(t)}\, \partial_\theta \ell_t\left(\theta_t + h_t\right) \\ \theta_{t+1} = \theta_t + \dfrac{\eta_{t+1}}{f(t)}\, \partial_\theta \ell_t(\theta_t). \end{cases} \tag{17.19}$$

17.2.4 LLR on other Stochastic Gradient algorithms

The LLR procedure may be applied to any stochastic gradient algorithm of the form

$$\theta_{t+1} = F(\theta_t, \eta_t) \tag{17.20}$$

where θ_t may store all the information maintained by the algorithm, not necessarily just a parameter value. Appendix .2 presents the algorithm in this case. Appendix .1 presents SVRG/AG, which is the particular case of this procedure applied to SVRG with an AdaGrad scheme for the update of η_t.

17.3 Experiments on SG and SVRG

We now present the experiments conducted to test our procedure. We first describe the experimental set up, then discuss the results.

17.3.1 Presentation of the experiments

We conducted ascents on synthetic data generated by three different probabilistic models: a one-dimensional Gaussian model, a Bernoulli model and a 50-dimensional linear regression model. Each model has two components: a generative distribution, and a set of distributions used to approximate the former.

One Dimensional Gaussian Model. The mean and value of the Gaussian generative distribution were set to 5 and 2 respectively. Let us note p_θ the density of a standard Gaussian random variable. The function to maximise we used is:

$$\ell_t(\theta) = \log p_\theta(x_t) = -\frac{1}{2}(x_t - \theta)^2. \tag{17.21}$$

Bernoulli model. The parameter in the standard parameterisation for the Bernoulli model was set to $p = 0.3$, but we worked with a logit parameterisation $\theta = \log(p/(1-p))$ for both the generative distribution and the discriminative function. The latter is then:

$$\ell_t(\theta) = \theta \cdot x_t - \log\left(1 + e^\theta\right). \tag{17.22}$$

Fifty-dimensional Linear Regression model. In the last model, we compute a fixed random matrix M. We then draw samples Z from a standard 50-dimensional Gaussian distribution. We then use M to make random linear combinations $X = MZ$ of the coordinates of the Z vectors. Then we observe X and try to recover first coordinate of the sample Z. The solution θ^* is the first row of the inverse of M. Note Y the first coordinate of Z so that the regression pair is (X, Y). We want to maximise:

$$\ell_t(\theta) = -\frac{1}{2}\left(y_t - \theta \cdot x_t\right)^2, \tag{17.23}$$

For each model, we drew 2500 samples from the data (7500 for the 50-dimensional model), then conducted ascents on those with on the one hand the SG and SVRG algorithms, and on the other hand their LLR counterparts, SG/SG and SVRG/SG, respectively.

17.3.2 Description and analysis of the results

For each model, we present four different types of results. We start with the trajectories of the ascents for several initial values of η (in the 50-dimensional case, we plot the first entry of $\theta^T \cdot M$). Then we present the cumulated regrets. Next we show the evolution of the logarithm of η_t along the ascents for the LLR algorithms. Finally, we compare this to trajectories of the non-adaptive algorithms with good initial values of η. Each time, we present three figures, one for each model.

Each figure of Figures 17.1 to 17.3 is made of four graphs: the upper ones are those of SG and SVRG, the lower ones are those of SG/SG and SVRG/SG. Figures 17.1 to 17.3 present the trajectories of the ascents for several orders of

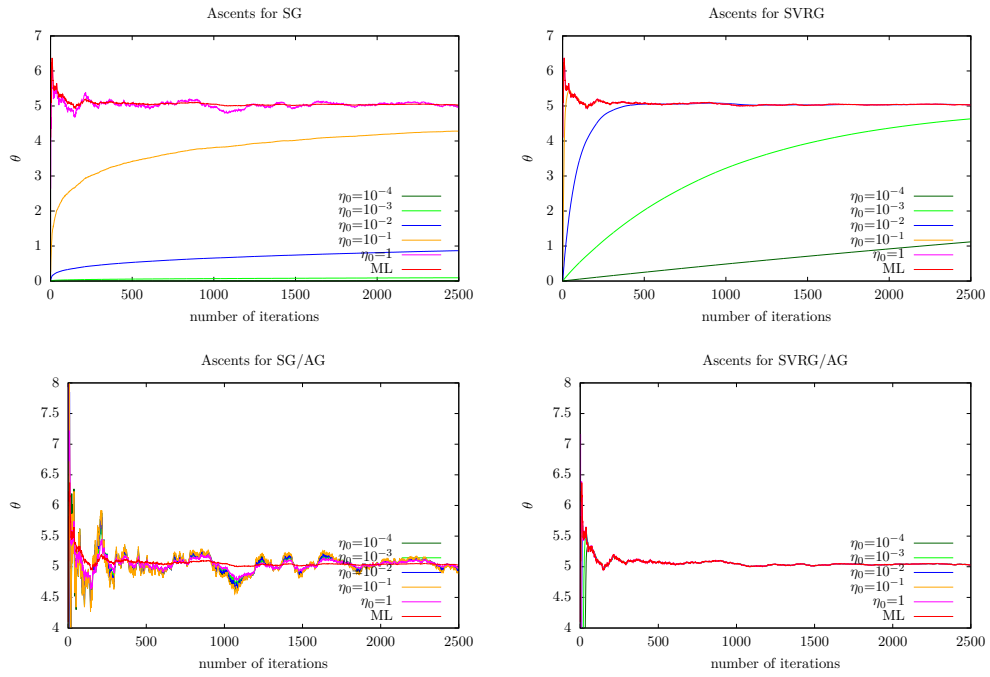

Figure 17.1: Trajectories of the ascents for a Gaussian model in one dimension for several algorithms and several η_0's

magnitude of η_0, while Figures 17.4 to 17.6 present the cumulated regrets for the same η_0's. The trajectory of the running maximum likelihood (ML) is displayed in red in each plot.

Trajectories of θ

Each figure for the ascent looks the same: there are several well distinguishable trajectories in the graphs of the standard algorithms, the upper ones, while trajectories are much closer to each other in those of the LLR algorithms, the lower ones.

Indeed, for many values of η, the standard algorithms will perform poorly. For instance, low values of η will result in dramatically low convergence towards the ML, as may be seen in some trajectories of the SG graphs. The SVRG algorithm performs noticeably better, but may start to oscillate, as in Figures 17.2 and 17.3.

These inconveniences are significantly improved by the use of LLR procedures. Indeed, in each model, almost every trajectory gets close to that of the ML in the SG/AG graphs. In the SVRG/AG graphs, the oscillations are overwhelmingly damped. Improvements for SG, though significant, are not as decisive in the linear regression model as in the other two, probably due to its greater complexity.

Cumulated regrets

Each curve of Figures 17.4 to 17.6 represents the difference between the cumulated regret of the algorithm used and that of the ML, for the η_0 chosen. The curves of SG and SVRG all go upwards, which means that the difference increases with time, whereas those of SG/AG and SVRG/SG tend to stagnate strikingly quickly. Actually, the trajectories for the linear regression model do not stagnate, but they

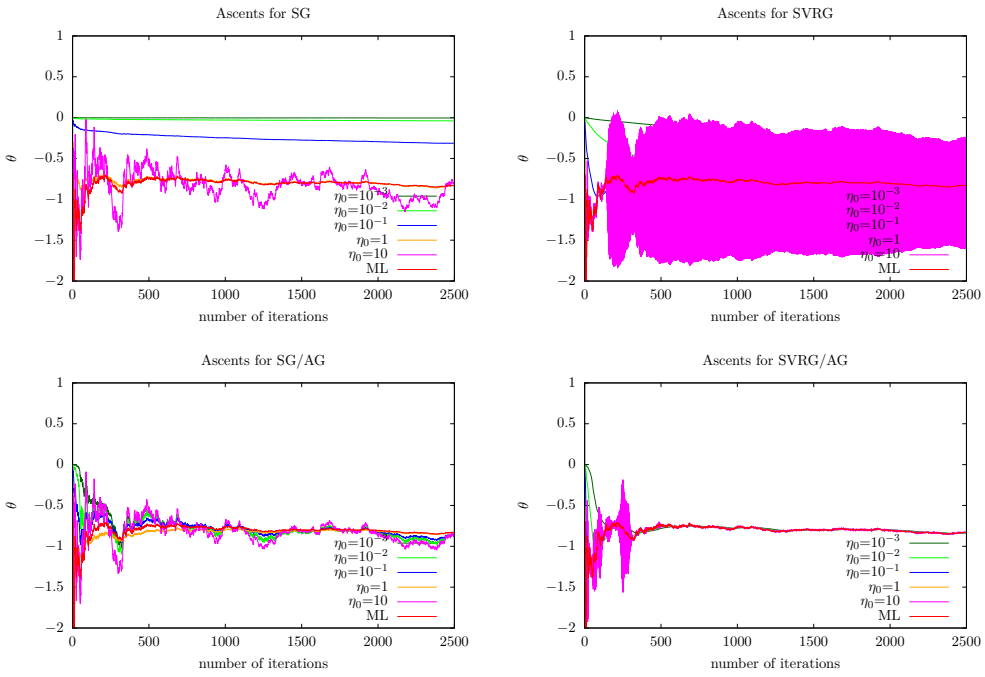

Figure 17.2: Trajectories of the ascents for a Bernoulli model for several algorithms and several η_0's

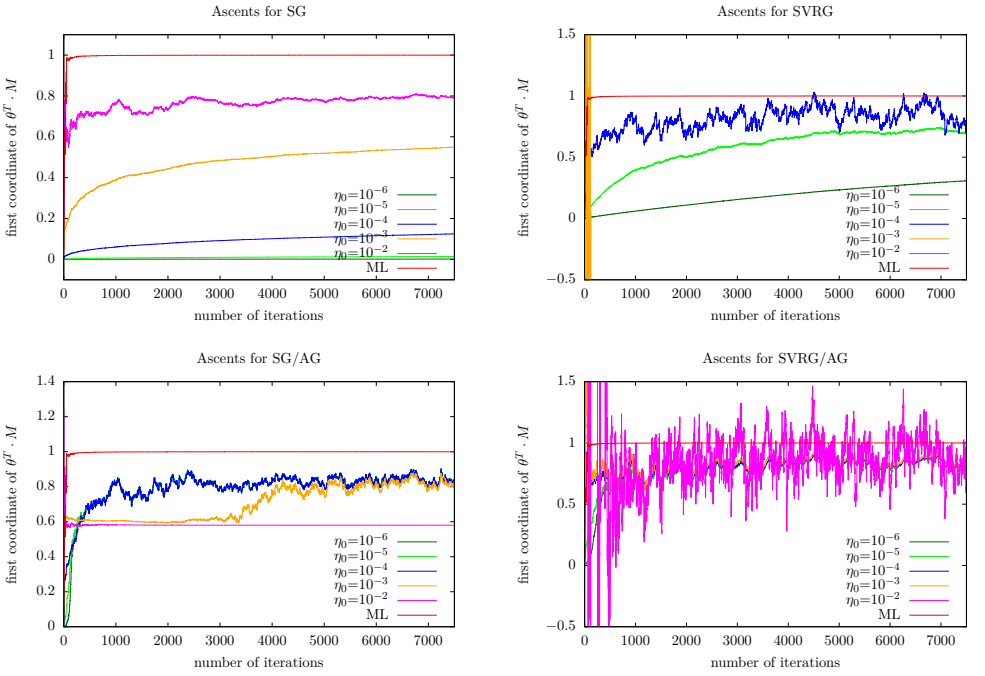

Figure 17.3: Trajectories of the ascents for a 50-dimensional linear regression model for several algorithms and several η_0's

are still significantly better for the LLR algorithms than for the original ones. The stagnation means that the values of the parameter found by these algorithms are very quickly as good as the Maximum Likelihood for the prediction task. Arguably, the

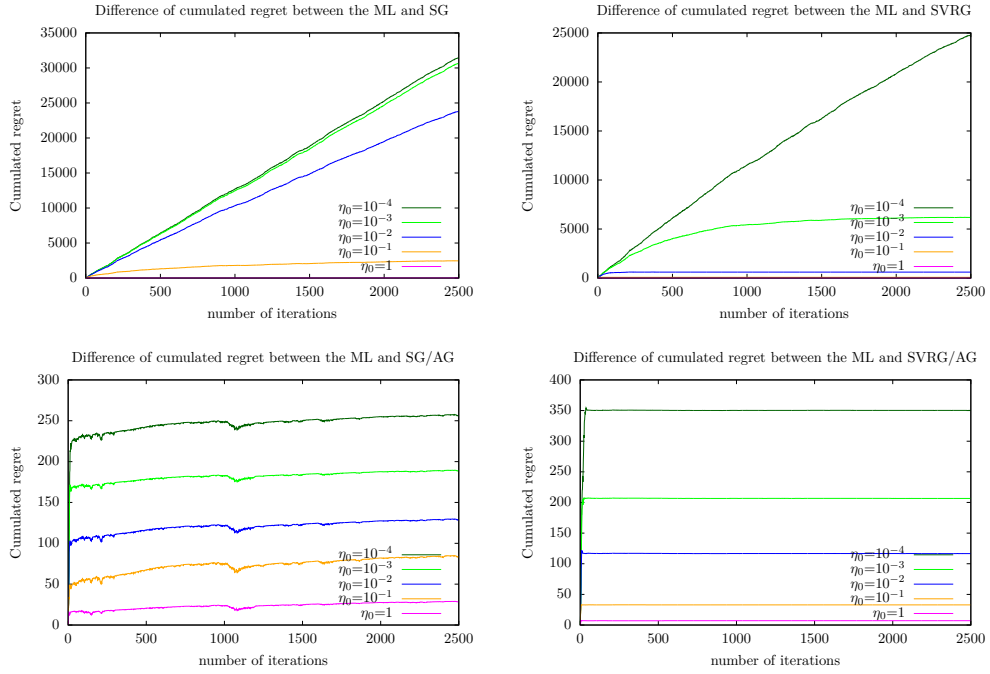

Figure 17.4: Difference between the cumulated regrets of the algorithm and of the ML for a Gaussian model in one dimension for several algorithms and several η_0's

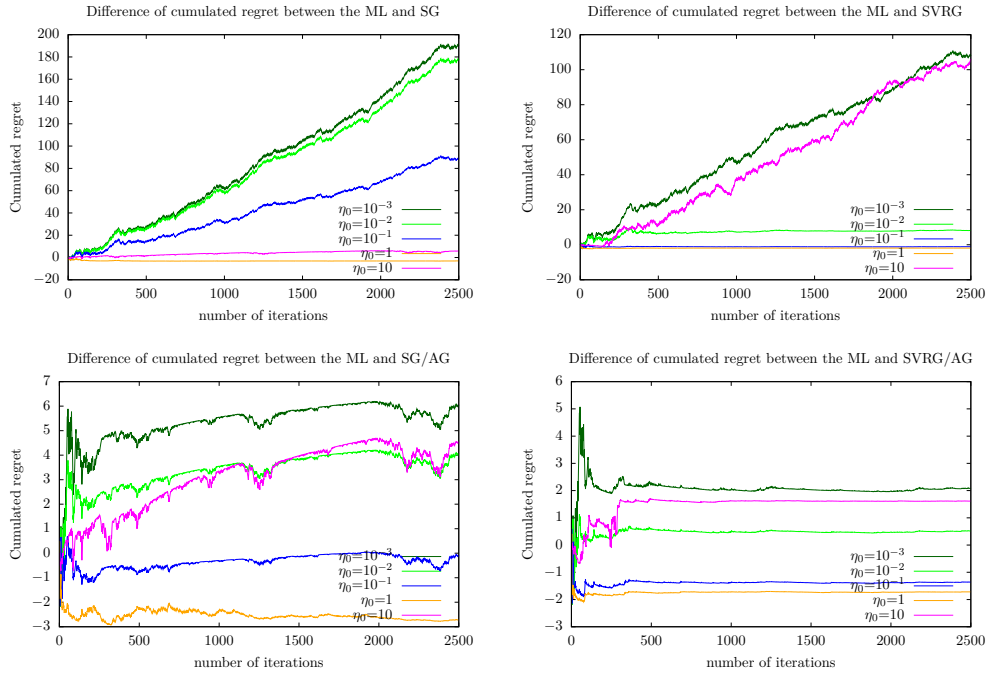

Figure 17.5: Difference between the cumulated regrets of the algorithm and of the ML for a Bernoulli model for several algorithms and several η_0's

fluctuations of the ascents around the later are therefore not a defect of the model: the cumulated regret graphs show that they are irrelevant for the minimisation at hand.

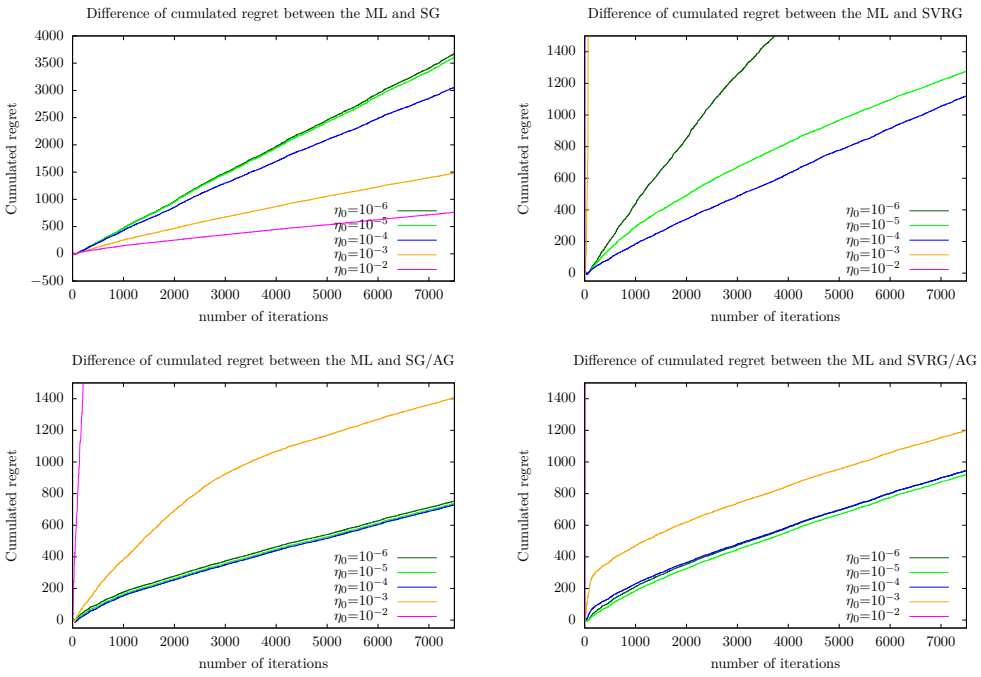

Figure 17.6: Difference between the cumulated regrets of the algorithm and of the ML for a 50-dimensional linear regression model for several algorithms and several η_0's

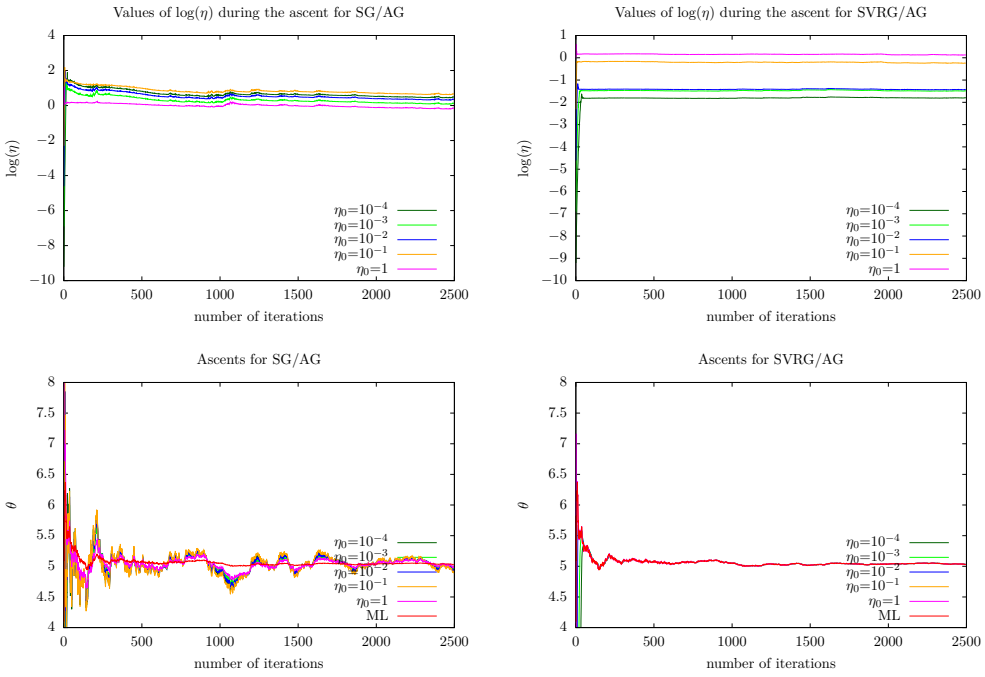

Figure 17.7: Evolution of $\log(\eta_t)$ in regard of the corresponding ascents for a Gaussian model in one dimension for SG and $SVRG$ with LLR and several η_0's

Evolution of the step size of the LLR algorithms during the ascents

Figures 17.7 to 17.9 show the evolution of the value of the logarithm of η_t in the LLR procedures for the three models, in regard of the trajectories of the corre-

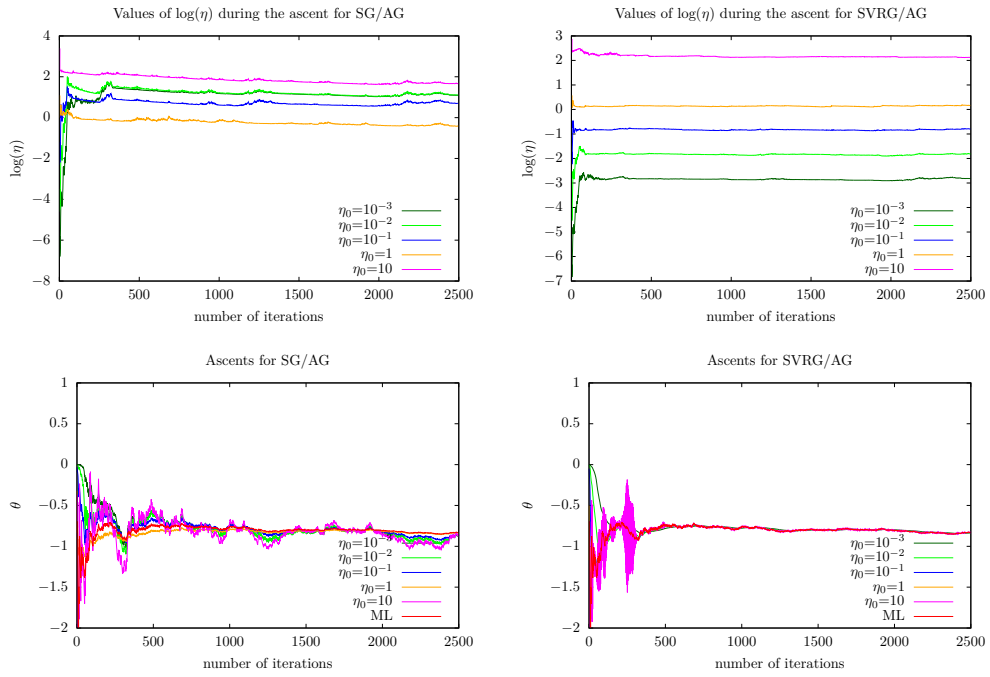

Figure 17.8: Evolution of $\log(\eta_t)$ in regard of the corresponding ascents for a Bernoulli model in one dimension for SG and SVRG with LLR and several η_0's

sponding ascents. For the Gaussian and Bernoulli models, in Figures 17.7 and 17.8, $\log(\eta_t)$ tends to stagnate quite quickly. This may seem a desirable behaviour : the algorithms have reached good values for η_t, and the ascent may accordingly proceed with those. However, this analysis may seem somewhat unsatisfactory due to the $1/f(t)$ dampening term in the parameter update, which remains unaltered by our procedure. For the linear regression model, in Figure 17.9, the convergence takes longer in the SG/SG case, and even in the SVRG/SG one, which may be explained again by the complexity of the model.

LLR versus hand-crafted learning rates

Figures 17.10 to 17.12 show the trajectories of the ascents for LLR algorithms with poor initial values of the step size, compared to the trajectories of the original algorithms with hand-crafted optimal values of η. The trajectories of the original algorithms appear in red. They possess only two graphs each, where all the trajectories are pretty much undistinguishable from another. This shows that the LLR algorithms show acceptable behaviour even with poor initial values of η, proving the procedure is able to rescue very badly initialised algorithms. However, one caveat is that the LLR procedure encounters difficulties dealing with too large values of η_0, and is much more efficient at dealing with small values of η_0. We have no satisfying explanation of this phenomenon yet. We thus suggest, in practice, to underestimate rather than overestimate the initial value η_0.

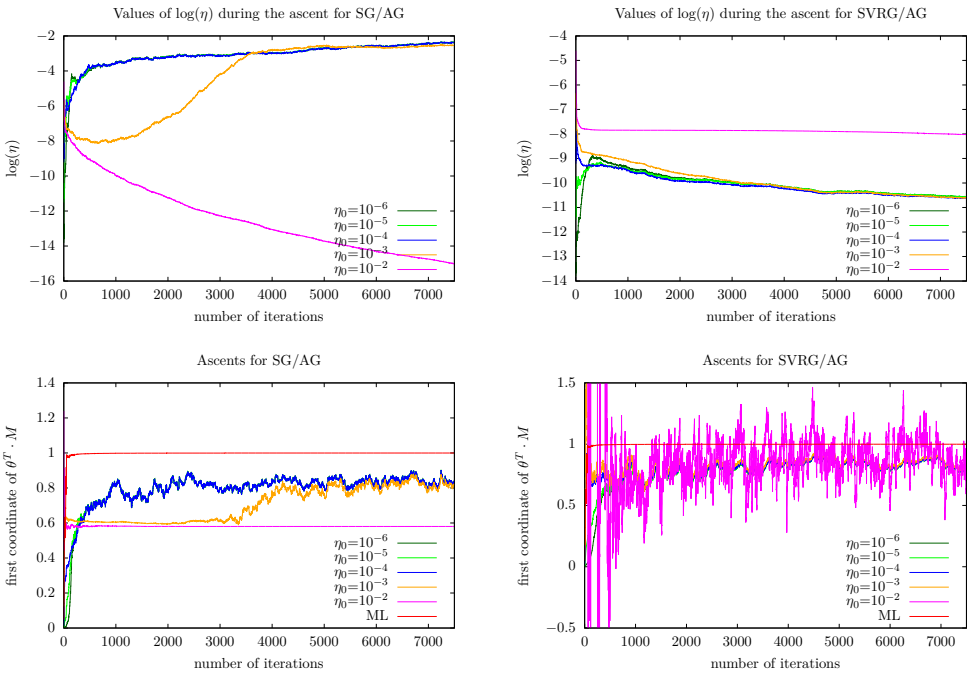

Figure 17.9: Evolution of $\log(\eta_t)$ in regard of the corresponding ascents for a 50-dimensional linear regression model for SG and SVRG with LLR and several η_0's

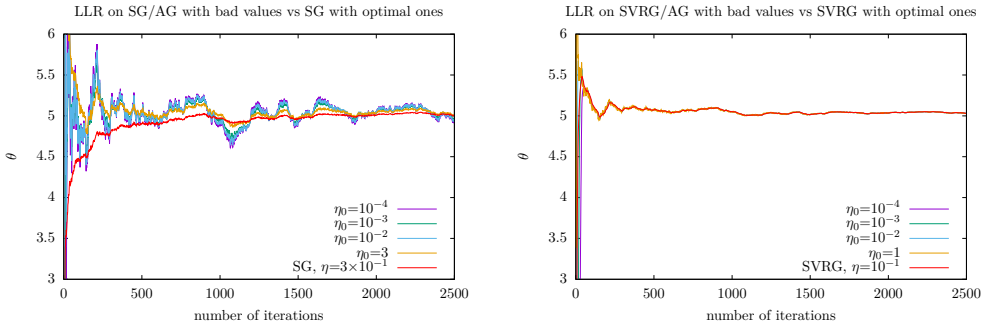

Figure 17.10: Trajectories of the ascents for a Gaussian model in one dimension for LLR algorithms with poor η_0's and original algorithms with empirically optimal η's

17.3.3 η_t in a quadratic model

In a quadratic deterministic one-dimensional model, where we want to maximise:

$$f(\theta) = -\alpha \frac{x^2}{2}, \tag{17.24}$$

SG is numerically stable if, and only if,

$$\left| 1 - \frac{\alpha\eta}{f(t)} \right| < 1, \tag{17.25}$$

that is

$$\frac{\eta}{2f(t)} < \alpha^{-1}. \tag{17.26}$$

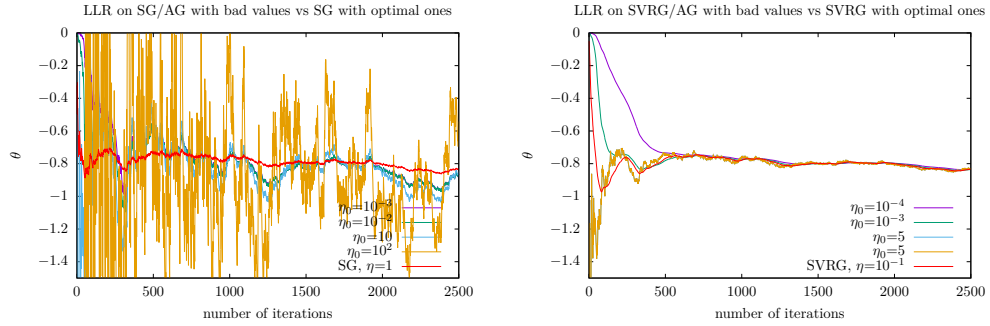

Figure 17.11: Trajectories of the ascents for a Bernoulli model for LLR algorithms with poor η_0's and original algorithms with empirically optimal η's

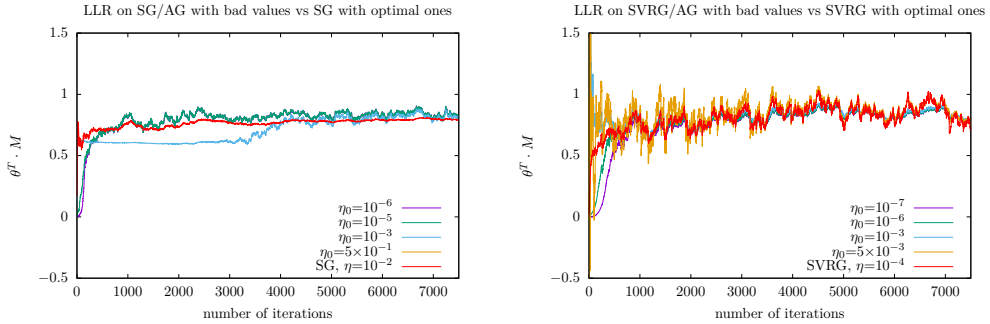

Figure 17.12: Trajectories of the ascents for a 50-dimensional linear regression model for LLR algorithms with poor η_0's and original algorithms with empirically optimal η's

Each graph of Figure 17.13 has two curves, one for the original algorithm, the other for its LLR version. The curve of the LLR version goes down quickly, then much more slowly, while the other curve goes down slowly all the time. This shows that, for $\alpha = 10^8$, the ratio above converges quickly towards α^{-1} for SG/AG and SVRG/AG, showing the ascent on η is indeed efficient. Then, the algorithm has converged, and η_t stays nearly constant, so much so that the LLR curve behaves like the other one. However, the convergence of η_t happens too slowly: θ_t takes very large values before η_t reaches this value, and even though it eventually converges to 0, such behaviour is unacceptable in practise.

17.4 A pathwise interpretation of the derivatives

Until now, we have tried to optimise the step size for a stochastic gradient ascent. This may be interpreted as conducting a gradient ascent on the subspace of the ascent algorithms which gathers the stochastic gradient algorithms, parametrised by $\eta \in \mathbb{R}$. However, we had to replace the $A_t(\eta)$'s by the h_t's because computing the former gave our algorithm a quadratic complexity in time. Indeed, adhesion to Equation 1 entails using $A_0(\eta_1)$ to compute $A_1(\eta_1)$, for instance. Likewise, $A_0(\eta_2)$ and $A_1(\eta_2)$ would be necessary to compute $A_2(\eta_2)$, and this scheme would repeat itself for every iteration.

We now introduce a formalism which shows the approximations we use are ac-

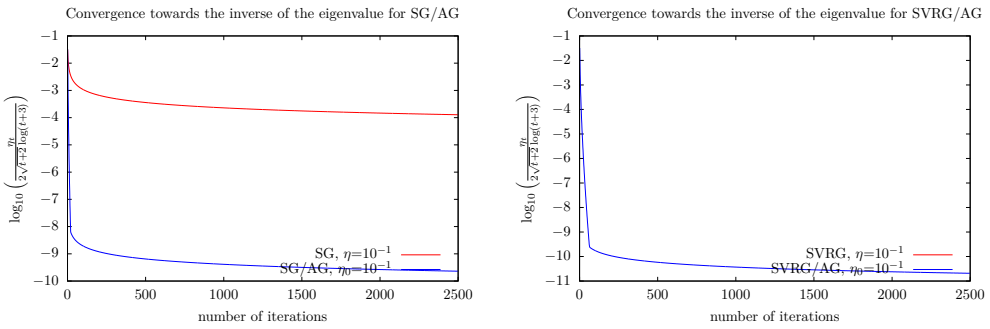

Figure 17.13: Evolution of $\log_{10}\left(\frac{\eta_t}{2\sqrt{t+2}\log(t+3)}\right)$ for a quadratic deterministic one-dimensional model for SG, SVRG and their LLR versions

tually derivatives taken alongside the effective trajectory of the ascent. It will also allow us to devise a new algorithm. It will, however, not account for the approximation of the Hessian.

To this avail, let us parameterise stochastic gradient algorithms by a sequence of step sizes

$$\boldsymbol{\eta} = (\eta_0, \eta_1, \ldots) \tag{17.27}$$

such that at iteration t, the update equation for θ_t becomes:

$$\theta_{t+1} = \theta_t + \frac{\eta_{t+1}}{f(t)} \partial_\theta \ell_t(\theta_t). \tag{17.28}$$

17.4.1 The loss as a function of step size: extension of the formalism

Consider the space $\mathcal{S}$ of infinite real sequences

$$\boldsymbol{\eta} = (\eta_0, \eta_1, \eta_2, \ldots) \tag{17.29}$$

We expand the T_t operators defined in Section 17.2 to similar ones defined on $\mathcal{S}$, with the same notation. Namely, define $T_0(\boldsymbol{\eta}) = \theta_0$ and, for $t > 0$,

$$T_t : \boldsymbol{\eta} \in \mathcal{S} \mapsto T_t(\boldsymbol{\eta}) \in \mathbb{R} \tag{17.30}$$

where θ_t has been obtained thanks to t iterations of (17.28). T_t is a regular function of $\boldsymbol{\eta}$, as the computations only involve

$$\eta_0, \eta_1, \ldots, \eta_t, \tag{17.31}$$

and so take place in finite-dimensional spaces. This will apply in all the computations below. As before, we work on a space we call $\log(\mathcal{S})$, the image of $\mathcal{S}$ by the mapping

$$\boldsymbol{\eta} = (\eta_t)_{t\geq 0} \in \mathcal{S} \mapsto \log(\boldsymbol{\eta}) = (\log \eta_t)_{t\geq 0}, \tag{17.32}$$

but we do not change notation for the functions $\boldsymbol{\eta} \mapsto T_t(\boldsymbol{\eta})$, as in Section 17.2.

17.4.2 The update of the step size in the SG/SG algorithm as a gradient ascent

We now prove that in SG/SG, when the Hessian is used without approximations, the step size η_t indeed follows a gradient ascent scheme.

Proposition 17.1. *Let*

$$(\theta_t)_{t\geq 0}\,, \quad \boldsymbol{\eta} = (\eta_t)_{t\geq 0} \tag{17.33}$$

be the sequences of parameters and step-sizes obtained with the SG/SG algorithm, where the Hessian is not approximated: this is Algorithm 2 where the update on h_t is replaced with

$$h_{t+1} = h_t + \frac{\eta}{\mu_t}\partial_\theta \ell_t(\theta_t) + \frac{\eta}{\mu_t}\,\partial_\theta^2 \ell_t(\theta_t)\cdot h_t. \tag{17.34}$$

Define e in the tangent plane of $\log(\mathcal{S})$ at $\log(\boldsymbol{\eta})$ by

$$e_t = 1, \quad t \geq 0. \tag{17.35}$$

Then, for all $t \geq 0$,

$$\log \eta_{t+1} = \log \eta_t + \frac{1}{\mu_t}\frac{\partial}{\partial e}\ell_t(T_t(\boldsymbol{\eta})). \tag{17.36}$$

The proof lies in Appendix .3.2.

17.4.3 A new algorithm, using a notion of "memory" borrowed from "No More Pesky Learning Rates"

We would now like to compute the change in η implied by a small modification of all previous coordinates η_s for s less than the current time t, but to compute the modification differently according to whether the coordinate s is "outdated" or not. To do it, we use the quantity τ_t defined in Section 4.2 of "No More Pesky Learning Rates" [15] as the "number of samples in recent memory". We want to discard the old η's and keep the recent ones. Therefore, at each time t, we compute

$$\gamma_t = \exp(-1/\tau_t). \tag{17.37}$$

Choose $\boldsymbol{\eta} \in \log(\mathcal{S})$, and consider the vector in the tangent plane to $\log(\mathcal{S})$ at $\boldsymbol{\eta}$:

$$e_t^j = \begin{cases} \displaystyle\prod_{k=j}^{t} \gamma_j, & j \leq t \\ 0, & j \geq t+1. \end{cases} \tag{17.38}$$

To run an algorithm using the e_t's instead of e as before, all we need to compute again is the formula for the update of the derivative below:

$$\mathcal{H}_t := \frac{\partial}{\partial e_t} T_t(\boldsymbol{\eta}). \tag{17.39}$$

$\mathcal{H}_t$ may indeed be computed, thanks to the following result.

Proposition 17.2. *The update equation of $\mathcal{H}_t$ is:*

$$\mathcal{H}_{t+1} = \gamma_{t+1}\mathcal{H}_t + \gamma_{t+1}\,\frac{\eta_{t+1}}{f(t)}\partial_\theta \ell_t(T_t(\eta^t)) + \gamma_{t+1}\frac{\eta_{t+1}}{f(t)}\partial_T^2 \ell_t(T_t(\boldsymbol{\eta}))\cdot \mathcal{H}_t. \tag{17.40}$$

The proof lies in Section .3.2.

Acknowledgements

The first author would like to thank Gaetan Marceau-Caron for his advice on programming, and Jérémy Bensadon for crucial help with LaTeX.

15. Schaul, Zhang, and LeCun, "No More Pesky Learning Rates", op. cit.

.1 LLR applied to the Stochastic Variance Reduced Gradient

The Stochastic Variance Reduced Gradient (SVRG) was introduced by Johnson and Zhang in "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction" [16]. We define here a version intended for online use.

Algorithm 4 (SVRG online)**.** *We maintain $\theta_t, \theta^b \in \Theta$ (current parameter and base parameter) and $s^b_t \in T_{\simeq\theta_t}\Theta$ (sum of the gradients of the ℓ_s computed at θ_s up to time t).*

θ is set to θ_0 and θ^b along s^b to 0.

The update equations read:

$$\begin{cases} s^b_{t+1} = s^b_t + \partial_\theta \ell_t(\theta^b) \\ \theta_{t+1} = \theta_t + \eta \left(\partial_\theta \ell_t(\theta_t) - \partial_\theta \ell_t(\theta^b) + \dfrac{s^b_{t+1}}{t+1} \right). \end{cases} \tag{41}$$

We now present the LLR version, obtained by updating the η of SVRG thanks to an SG ascent. We call this algorithm "SVRG/SG".

Algorithm 5 (SVRG/AG)**.** *We maintain $\theta_t, \theta^b \in \Theta$ (current parameter and base parameter), $\eta_t \in \mathbb{R}$ (current step size), $s^b_t \in T_{\simeq\theta_t}\Theta$ (sum of the gradients of the ℓ_s computed at θ_s up to time t), $h_t \in T_{\theta_t}\Theta$ (approximation of the derivative of T_t with respect to $\log(\eta)$ at η_t) and the real numbers n_t (average of the squared norms of the λ_s defined below) and d_t (renormalising factor for the computation of n_t).*

θ is set to θ_0, the other variables are set to 0.

At each time t, we compute $\mu_t \in \mathbb{R}$ (a rate used in several updates), and $\lambda_t \in \mathbb{R}$ (the approximate derivative of $\ell_t \circ \theta_t$ with respect to $\log(\eta)$ at η_t).

The update equations read:

$$\begin{cases} \mu_t = \sqrt{t+2}\log(t+3) \\ \lambda_t = \partial_\theta \ell_t(\theta_t) \cdot h_t \\ d_{t+1} = \left(1 - \dfrac{1}{\mu_t}\right) d_t + \dfrac{1}{\mu_t} \\ n^2_{t+1} = \left(\left(1 - \dfrac{1}{\mu_t}\right) n^2_t + \dfrac{1}{\mu_t}\lambda^2_t \right) d^{-1}_{t+1} \\ \eta_{t+1} = \eta_t \exp\left(\dfrac{1}{\mu_t} \dfrac{\lambda_t}{n_{t+1}} \right) \\ s^b_{t+1} = s^b_t + \partial_\theta \ell_t(\theta^b) \\ h_{t+1} = h_t + \eta_{t+1} \left(\partial_\theta \ell_t(\theta_t + h_t) - \partial_\theta \ell_t(\theta^b) + \dfrac{s^b_{t+1}}{t+1} \right) \\ \theta_{t+1} = \theta_t + \eta_{t+1} \left(\partial_\theta \ell_t(\theta_t) - \partial_\theta \ell_t(\theta^b) + \dfrac{s^b_{t+1}}{t+1} \right). \end{cases} \tag{42}$$

16. Johnson and Zhang, "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction", op. cit.

.2 LLR applied to a general stochastic gradient algorithm

Let Θ and H be two spaces. Θ is the space of parameters, H is that of hyperparameters. In this section, a parameter potentially means a tuple of parameters in the sense of other sections. For instance, in SVRG/SG online, we would call a parameter the couple

$$\left(\theta_t, \theta^b\right). \tag{43}$$

Likewise, in the same algorithm, we would call a hyperparameter the couple

$$(\eta_t, h_t). \tag{44}$$

Let

$$\begin{aligned} F: \quad & \Theta \times H \to \Theta \\ & (\theta, \eta) \mapsto F(\theta, \eta). \end{aligned} \tag{45}$$

be differentiable with respect to both variables. We consider the algorithm:

$$\theta_{t+1} = F(\theta_t, \eta_t). \tag{46}$$

Let us present its LLR version. We call it GEN/SG, GEN standing for "general".

Algorithm 6 (GEN/SG)**.** *We maintain $\theta_t \in \Theta$ (current parameter), $\eta_t \in H$ (current hyperparameter), $h_t \in T_{\theta_t}\Theta$ (approximation of the derivative of T_t in the direction of $e \in T_{\eta_t}H$).*

θ and η are set to user-defined values.

The update equations read:

$$\begin{cases} \eta_{t+1} = \eta_t + \alpha \partial_\theta \ell_t(\theta_t) \cdot h_t \\ h_{t+1} = \partial_\theta F(\theta_t, \eta_t) \cdot h_t + \partial_\eta F(\theta_t, \eta_t) \cdot \dfrac{\partial}{\partial e} \eta_t \\ \theta_{t+1} = F\left(\theta_t, \eta_{t+1}\right). \end{cases} \tag{47}$$

.3 Computations

.3.1 Computations for Section 17.2: proof of Fact 1

Proof. θ_0 is fixed, so $A_0(\eta) = 0$. Let $t \geq 0$. We differentiate (17.5) with respect to $\log(\eta)$, to obtain:

$$\frac{\partial}{\partial \log \eta} T_{t+1}(\eta) = \frac{\partial}{\partial \log \eta} T_t(\eta) + \frac{\eta}{f(t)} \partial_\theta \ell_t(\theta_t) + \frac{\eta}{f(t)} \partial_\theta^2 \ell_t(\theta_t(\eta)) \cdot \frac{\partial}{\partial \log \eta} T_t(\eta), \tag{48}$$

which concludes the proof. □

.3.2 Computations for Section 17.4

Computations for Section 17.4.2: proof of Proposition 17.1

To prove Proposition 17.1, we use the following three lemmas. The first two are technical, and are used in the proof of the third one, which provides an update formula for the derivative appearing in the statement of the proposition. We may have proceeded without these, as in the proof of Fact 1, but they allow the approach to be more generic.

Lemma 2. *Let*

$$\begin{array}{rccl} F_t: & \Theta \times \mathbb{R} & \to \Theta & \\ & (\theta, \eta) & \mapsto & F_t(\theta, \eta) = \theta + \frac{\eta}{f(t)} \partial_\theta \ell_t(\theta). \end{array} \tag{49}$$

Then,

$$\frac{\partial}{\partial \theta} F_t(\theta, \eta) = \mathrm{Id} + \frac{\eta}{f(t)} \partial_\theta^2 \ell_t(\theta) \tag{50}$$

and

$$\frac{\partial}{\partial \eta} F_t(\theta, \eta) = \frac{1}{f(t)} \partial_\theta \ell_t(\theta). \tag{51}$$

Id *is the identity on the tangent plane to* Θ *in* θ*.*

Lemma 3. *Let*

$$\begin{array}{rccl} V_t: & \mathcal{S} & \to \Theta \times \mathbb{R} & \\ & \boldsymbol{\eta} & \mapsto & V_t(\boldsymbol{\eta}) = (\theta_t(\eta), \eta_{t+1}). \end{array} \tag{52}$$

Consider $\log(\boldsymbol{\eta}) \in \log(\mathcal{S})$*, and any vector* e *tangent to* $\log(\mathcal{S})$ *at this point. Then the directional derivative of*

$$\begin{array}{rccl} F_t \circ V_t: & \mathcal{S} & \to \Theta & \\ & \boldsymbol{\eta} & \mapsto & F_t(V_t(\boldsymbol{\eta})) = T_t(\boldsymbol{\eta}) + \frac{\eta_{t+1}}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) \end{array} \tag{53}$$

at the point $\log(\boldsymbol{\eta})$ *and in the direction* e *is*

$$\frac{\partial}{\partial e} F_t \circ V_t(\boldsymbol{\eta}) = \frac{\partial}{\partial e} T_t(\eta) + \frac{\partial}{\partial e} \eta_{t+1} \frac{1}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) + \frac{\eta_{t+1}}{f(t)} \partial_\theta^2 \ell_t(T_t(\boldsymbol{\eta})) \cdot \frac{\partial}{\partial e} T_t(\eta). \tag{54}$$

We may then prove the following lemma.

Lemma 4. *Define*

$$\mathcal{H}_t = \frac{\partial}{\partial e} T_t(\boldsymbol{\eta}). \tag{55}$$

Then for all $t \geq 0$,

$$\mathcal{H}_{t+1} = \mathcal{H}_t + \frac{\eta_{t+1}}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) + \frac{\eta_{t+1}}{f(t)} \partial_\theta^2 \ell_t(T_t(\boldsymbol{\eta})) \cdot \mathcal{H}_t. \tag{56}$$

Proof. The update equation of $T_t(\boldsymbol{\eta})$, (17.28), is such that:

$$T_{t+1}(\boldsymbol{\eta}) = T_t(\boldsymbol{\eta}) + \frac{\eta_{t+1}}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) = F_t \circ V_t(\boldsymbol{\eta}). \tag{57}$$

From the above and Lemma 3,

$$\frac{\partial}{\partial e} T_{t+1}(\eta) = \frac{\partial}{\partial e} T_t(\boldsymbol{\eta}) + \frac{\partial}{\partial e} \eta_{t+1} \frac{1}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) + \frac{\eta_{t+1}}{f(t)} \partial_\theta^2 \ell_t(T_t(\boldsymbol{\eta})) \cdot \frac{\partial}{\partial e} T_t(\boldsymbol{\eta}). \tag{58}$$

Now,

$$\frac{\partial}{\partial e} \eta_{t+1} = \eta_{t+1}, \tag{59}$$

which concludes the proof. □

Finally, we prove Proposition 17.1.

Proof of Proposition 17.1. It is sufficient to prove that, for all $t \geq 0$,

$$\frac{\partial}{\partial e} \ell_t(T_t(\boldsymbol{\eta})) = \partial_\theta \ell_t(\theta_t) \cdot h_t, \tag{60}$$

that is,

$$\partial_\theta \ell_t(T_t(\boldsymbol{\eta})) \cdot \mathcal{H}_t = \partial_\theta \ell_t(\theta_t) \cdot h_t. \tag{61}$$

Therefore, it is sufficient to prove that, for all $t \geq 0$, $T_t(\boldsymbol{\eta}) = \theta_t$ and $\mathcal{H}_t = h_t$. $T_0(\eta) = \theta_0$ by construction and, since θ_0 does not depend on $\boldsymbol{\eta}$, $\mathcal{H}_0 = 0 = h_0$. Assuming the results hold up to iteration t, it is straighforward that $T_{t+1}(\boldsymbol{\eta}) = \theta_{t+1}$, since for all $s \leq t$, $T_s(\boldsymbol{\eta}) = \theta_s$. Therefore, thanks to Lemma 4, $\mathcal{H}_t$ and h_t have the same update, so that $\mathcal{H}_{t+1} = h_{t+1}$, which concludes the proof. □

Computations for Section 17.4.3: proof of Proposition 17.2

Proof. Thanks to (58) in Lemma 3,

$$\begin{aligned} \frac{\partial}{\partial e_{t+1}} T_{t+1}(\eta) = \frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}) + \frac{\partial}{\partial e_{t+1}} \eta_{t+1} \frac{1}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) \\ + \frac{\eta_{t+1}}{f(t)} \partial_\theta^2 \ell_t(T_t(\boldsymbol{\eta})) \cdot \frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}), \end{aligned} \tag{62}$$

that is:

$$\mathcal{H}_{t+1} = \frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}) + \frac{\partial}{\partial e_{t+1}} \eta_{t+1} \frac{1}{f(t)} \partial_\theta \ell_t(T_t(\boldsymbol{\eta})) + \frac{\eta_{t+1}}{f(t)} \partial_\theta^2 \ell_t(T_t(\boldsymbol{\eta})) \cdot \frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}). \tag{63}$$

We first prove:

$$\frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}) = \gamma_{t+1} \frac{\partial}{\partial e_t} T_t(\eta). \tag{64}$$

Define $(f_j)_{j\geq 0}$ the canonical basis of the tangent plane to $\log(\mathcal{S})$ at $\boldsymbol{\eta}$. Then,

$$e_{t+1} = \gamma_{t+1}(e_t + f_{t+1}). \tag{65}$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}) &= \frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}) \\ &= \gamma_{t+1} \frac{\partial}{\partial e_t} T_t(\eta^t) + \frac{\partial}{\partial f_{t+1}} T_t(\eta^t) \\ &= \gamma_{t+1} \frac{\partial}{\partial e_t} T_t(\eta^t)\end{aligned} \tag{66}$$

because the last term is 0. Therefore,

$$\frac{\partial}{\partial e_{t+1}} T_t(\boldsymbol{\eta}) = \gamma_{t+1} \mathcal{H}_t. \tag{67}$$

Then, thanks to (58),

$$\frac{\partial}{\partial e_{t+1}} \eta_{t+1} = \gamma_{t+1} \eta_{t+1}, \tag{68}$$

which is true since

$$\frac{\partial}{\partial e_{t+1}} \eta_{t+1} = \gamma_{t+1} \frac{\partial}{\partial f_{t+1}} \eta_{t+1} = \gamma_{t+1} \eta_{t+1}, \tag{69}$$

and concludes the proof. □

Index pour la partie RTRL

e_t^*	État de la trajectoire stable	58
$\mathrm{Id}_\Theta^\infty$	Envoie un vecteur sur la suite constante égale à celui-ci	56
$\mathcal{E}_t$	Espace des états	51
$\mathcal{L}(\Theta, \mathcal{E}_t)$	Espace des différentielles	55
Θ	Espace du paramètre	51
$T\mathcal{E}_t$	Espace tangent à $\mathcal{E}_t$	51
$T\Theta$	Espace tangent à Θ	51

F

g_t	Fonction couple état-paramètre	75
	Fonction d'échelle	79
L	Fonction d'échelle pour le calcul des longueurs d'intervalle	90
$\boldsymbol{T}_t$	Fonction de transition du système dynamique	54
$\boldsymbol{e}_t$	Fonction donnant l'état à l'instant t	55
$\boldsymbol{J}_t$	Fonction donnant la différentielle à l'instant t	56
$\boldsymbol{\theta}_t^{\mathrm{o}}$	Fonction donnant le paramètre de l'algorithme RTRL en boucle ouverte	121
ℓ^1, ℓ^2 et M_p	Fonctions d'échelle utilisées	89, 159

H

$T_{t_0}^{\Lambda^*}$	Horizon de contrôle de la plus grande valeur propre	127
$T_{t_0}^{r_\Theta^*}$	Horizon de maintien dans B_Θ^*	118

I

k_2	Indice pour la convergence	140
$T_{t_0}^\infty$	Infimum de $T_{t_0}^{r_\Theta^*}$ et $T_{t_0}^{\Lambda^*}$	128
$k_1(\mu)$	Infimum pour le maintien de $\boldsymbol{\theta}$ dans B_Θ^*	139
I_k	Intervalles de temps $[T_k, T_{k+1}[$	91

M

M_2	Majorant	63
M_4	Majorant	104
m_t	Majorant des différentielles des pertes	74
M_1	Majorant des différentielles secondes par rapport au paramètre	61
M_7	Majorant des termes en grand O	131
Λ^*	Majorant des valeurs propres des pertes sur le paramètre	77
M_ξ	Majorant du bruit	111
$\lambda_{\min}$	Minorant des valeurs propres des différentielles secondes sur les pertes en θ^*	137

N

	Norme d'opérateur	52
	Norme d'une application bilinéaire	52
	Norme d'une application trilinéaire	53
(T_k)	Nouvelle échelle de temps	90

O

Φ	Opérateur de déplacement sur Θ	101
ϕ	Opérateur de mise à jour sur le paramètre	102
$\Omega_{T\Theta\times\mathbb{R}_+}$	Ouvert pour le contrôle de ϕ	102

P

$\theta_t(\mu)$	Paramètre obtenu avec RTRL et la suite de pas $\mu\,\boldsymbol{\eta}$	140

θ^*	Paramètre optimal	96
θ^*	Paramètre stable	58
p_t	Perte sur les états	38, 73
$\Lambda_{t_0,\,t}(\theta)$	Plus grande valeur propre en θ	127
$\lambda_{t_0,\,t}(\theta)$	Plus petite valeur propre en θ	127
$\Pi_{s,\,t}^{t_0}$	Produit des différentielles	114

R

r_ξ	Rayon d'une boule contenant les différentielles bruitées	111
r_Θ^*	Rayon de la boule stable pour le paramètre	65
$r_\mathcal{E}$	Rayon des boules de contrôle pour les états	59
$r_{\mathcal{L}(\Theta,\,\mathcal{E})}$	Rayon des boules de contrôle pour les différentielles	59
$r_\mathcal{E}^*$	Rayon des boules stables pour les états	65
$r_{\mathcal{L}(\Theta,\,\mathcal{E})}^*$	Rayon des boules stables pour les différentielles	65

S

$\boldsymbol{\eta} = (\eta_t)$	Suite de pas de descente utilisée	89, 160
$\tilde{\boldsymbol{\eta}}$	Suite des pas de descente renormalisée	94
	Suite engendrée par une fonction d'échelle	79

T

$b_{t_0,\,t}(\boldsymbol{\eta})$	Terme additif dans la propriété centrale	131
$b^k(\mu)$	Terme additif le long de la suite (T_k)	139

Index pour la partie « NoBack-Track »

O

$\odot$	Opérateur d'égalisation des normes	165
$\mathcal{R}$	Opérateur de réduction	167
$\mathcal{R}_t$	Opérateur de réduction à l'instant t	171

P

$\theta_t(\mu)$	Paramètre obtenu avec NBT et la suite de pas $\mu\,\boldsymbol{\eta}$	185
A_k	Produit cartésien des boules optimales	189
$\otimes$	Produit tensoriel des espaces	170
$\otimes$	Produit tensoriel des vecteurs	170

R

r_ξ	Rayon de la boule contenant les estimées NBT	179
$r_{T\mathcal{E}\times T\Theta}$	Rayon des $B_{T\mathcal{E}_t\times T\Theta}$	174
$r(\gamma_1,\,\gamma_2)$	Rayon pour la zone sous l'hyperbole stable	169

T

u_k	Terme général du produit infini	161
$\mathcal{F}_t$	Tribu engendrée par les $\varepsilon(s)$, pour $1 \leq s \leq t$	171

V

$\varepsilon(t)$	Vecteur des variables de Bernoulli à l'instant t	171

Z

$\mathcal{Z}_t$	Zone sous une hyperbole stable par les opérateurs de réduction	173

Bibliographie

Articles scientifiques

[AM79] Brian D. O. Anderson et John B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, 1979.

[Ama98] Shun-ichi Amari. "Natural gradient works efficiently in learning". In : *Neural Comput.* 10 (2 fév. 1998), p. 251–276. issn : 0899-7667. doi : 10.1162/089976698300017746. url : http://portal.acm.org/citation.cfm?id=287476.287477.

[BM17] Borja Balle et Odalric-Ambrym Maillard. "Spectral Learning from a Single Trajectory under Finite-State Policies". In : *Proceedings of the 34th International Conference on Machine Learning*. Sous la dir. de Doina Precup et Yee Whye Teh. T. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia : PMLR, juin 2017, p. 361–370.

[Bot10] Léon Bottou. "Large-scale machine learning with stochastic gradient descent". In : *Proceedings of COMPSTAT'2010*. Springer, 2010, p. 177–186.

[Bot99] Léon Bottou. "On-line learning and stochastic approximations". In : *On-line learning in neural networks*. Sous la dir. de David Saad. Cambridge University Press New york, NY, USA, 1999. Chap. 2, p. 9–42.

[BSF94] Yoshua Bengio, Patrice Simard et Paolo Frasconi. "Learning Long-Term Dependencies with Gradient Descent is Difficult". In : *IEEE Transactions on Neural Networks* 5 (2 mar. 1994), p. 157–166.

[CMR05] Olivier Cappé, Eric Moulines et Tobias Ryden. *Inference in Hidden Markov Models*. Springer-Verlag New York, 2005. doi : 10.1007/0-387-28982-8.

[Dau+14] Yann N. Dauphin et al. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In : *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Sous la dir. de Zoubin Ghahramani et al. 2014, p. 2933–2941.

[DBL14] Aaron Defazio, Francis Bach et Simon Lacoste-Julien. “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives”. In : *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Sous la dir. de Zoubin Ghahramani et al. 2014.

[Del96] Bernard Delyon. “General results on the convergence of stochastic algorithms”. In : *IEEE Transactions on Automatic Control* 41 (9 1996), p. 1245–1255.

[DFB16] Aymeric Dieuleveut, Nicolas Flammarion et Francis Bach. *Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression*. Rapp. tech. 2016.

[DHS11] John Duchi, Elad Hazan et Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization”. In : *The Journal of Machine Learning Research* 12 (2011), p. 2121–2159.

[DLM99] Bernard Delyon, Marc Lavielle et Eric Moulines. “Convergence of a stochastic approximation version of the EM algorithm”. In : *The Annals of Statistics* 27 (1 1999), p. 94–128.

[FP96] Jean-Claude Fort et Gilles Pagès. “Convergence of Stochastic Algorithms : From the Kushner-Clark Theorem to the Lyapunov Functional Method”. In : *Advances in Applied Probability* 4 (28 1996), p. 1072–1094.

[Gha+14] Zoubin Ghahramani et al., éds. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. 2014.

[HSW89] Kurt Hornik, Maxwell Stinchcombe et Halbert White. “Multilayer Feedforward Networks are Universal Approximators”. In : *Neural Networks* 2 (1989), p. 359–366.

[Jae02] Herbert Jaeger. *A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the “echo state network” approach*. Rapp. tech. 159. German National Research Center for Information Technology, GMD, 2002.

[JZ13] Rie Johnson et Tong Zhang. “Accelerating Stochastic Gradient Descent using Predictive Variance Reduction”. In : *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Sous la dir. de Christopher J. C. Burges et al. 2013.

[KC78] Harold J. Kushner et Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. T. 26. Applied Mathematical Sciences. Springer-Verlag New York, 1978.

[KY03] Harold J. Kushner et George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag New York, 2003.

[KY95] Harold J. Kushner et J. Yang. “Analysis of adaptive step-size SA algorithms for parameter tracking”. In : *IEEE Transactions on Automatic Control* 40 (8 1995), p. 1403–1410.

[Lju77] Lennart Ljung. “Analysis of recursive stochastic algorithms”. In : *IEEE Transactions on Automatic Control* 22 (4 1977), p. 551–575.

[Löc15] Eva Löcherbach. “Ergodicity and speed of convergence to equilibrium for diffusion processes”. Cours disponible sur la page web de l’auteur, à l’adresse `https://eloecherbach.u-cergy.fr/cours.pdf`. 2015.

[Mah10] Ashique Mahmood. “Automatic step-size adaptation in incremental supervised learning”. Mém.de mast. Edmonton, Alberta, United States of America : University of Alberta, 2010.

[Mal12] Stéphane Mallat. “Group Invariant Scattering”. In : *Communications in Pure and Applied Mathematics* 65 (10 oct. 2012), p. 1331–1398.

[MDA15] Douglas Maclaurin, David Duvenaud et Ryan Adams. “Gradient-based Hypermarameter Optimization through Reversible Learning”. In : *Proceedings of The 32nd International Conference on Machine Learning*. Sous la dir. de Francis Bach et David Blei. 2015.

[MO15] Pierre-Yves Massé et Yann Ollivier. “Speed learning on the fly”. In : *preprint* (2015).

[OCT16] Yann Ollivier, Guillaume Charpiat et Corentin Tallec. “Training recurrent networks online without backtracking”. In : *preprint* (2016).

[Oll15a] Yann Ollivier. “Riemannian metrics for neural networks I: feedforward networks”. In : *Information and Inference: A journal of the IMA* 4 (2 2015), p. 108–153.

[Oll15b] Yann Ollivier. “Riemannian metrics for neural networks II: recurrent networks and learning symbolic data sequences”. In : *Information and Inference: A journal of the IMA* 4 (2 2015), p. 153–193.

[Oll17] Yann Ollivier. “Online natural gradient as a Kalman filter”. In : *preprint* (2017).

[OTC15] Yann Ollivier, Corentin Tallec et Guillaume Charpiat. “Training recurrent networks online without backtracking”. In : *preprint* (2015).

[Pea95] B.A. Pearlmutter. “Gradient calculations for dynamic recurrent neural networks: a survey”. In : *IEEE Transactions on Neural Networks* 6 (5 sept. 1995), p. 1212–1228. doi : 10.1109/72.410363.

[RM51] Herbert Robbins et Sutton Monro. “A Stochastic Approximation Method”. In : *The Annals of Mathematical Statistics* 22.3 (1951), p. 400–407.

[Sag+15] Levent Sagun et al. “Explorations on high dimensional landscapes”. Article accepté pour un atelier à ICLR 2015, disponible sur arxiv à l’adresse https://arxiv.org/pdf/1412.6615.pdf. 2015.

[SLB13] Mark Schmidt, Nicolas Le Roux et Francis Bach. *Minimizing finite sums with the Stochastic Average Gradient*. Rapp. tech. 00860051. HAL, 2013.

[SZL13] Tom Schaul, Sixin Zhang et Yann LeCun. “No More Pesky Learning Rates”. In : *Proceedings of The 30th International Conference on Machine Learning*. Sous la dir. de Sanjoy Dasgupta et David McAllester. JMLR, 2013, p. 343–351.

[TO17a] Corentin Tallec et Yann Ollivier. “Unbiased Online Recurrent Optimization”. In : *preprint* (2017).

[TO17b] Corentin Tallec et Yann Ollivier. “Unbiasing Truncated Backpropagation Through Time”. In : *preprint* (2017).

Articles de presse et sources multimédia

[Cho16] Tanguy Chouard. “The Go Files: AI computer wraps up 4-1 victory against human champion”. In : *Nature* (mar. 2016).

[LeC16] Yann LeCun. “L’apprentissage profond : une révolution en intelligence artificielle”. Leçon inaugurale au Collège de France, disponible à l’adresse `https://www.college-de-france.fr/site/yann-lecun/inaugural-lecture-2016-02-04-18h00.htm`. 2016.

[Sol17] Olivia Solon. “Oh the humanity! Poker computer trounces humans in big step for AI”. In : *The Guardian* (jan. 2017).

Table des figures

Table des matières

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Autour De L'Usage des gradients en apprentissage statistique

Mots-clefs : apprentissage statistique, optimisation stochastique, systèmes dynamiques

Résumé : Nous établissons un théorème de convergence locale de l'algorithme classique d'optimisation de système dynamique RTRL, appliqué à un système non linéaire. L'algorithme RTRL est un algorithme en ligne, mais il doit maintenir une grande quantité d'informations, ce qui le rend impropre à entraîner des systèmes d'apprentissage de taille moyenne. L'algorithme « NoBackTrack » y remédie en maintenant une approximation aléatoire non biaisée de faible taille de ces informations. Nous prouvons également la convergence avec probabilité arbitrairement proche de un, de celui-ci vers l'optimum local atteint par l'algorithme RTRL.

Nous formalisons également l'algorithme LLR et en effectuons une étude expérimentale, sur des données synthétiques. Cet algorithme met à jour de manière adaptive le pas d'une descente de gradient, par descente de gradient sur celui-ci. Il apporte ainsi une réponse partielle au problème de la fixation numérique du pas de descente, dont le choix influence fortement la procédure de descente et qui doit sinon faire l'objet d'une recherche empirique potentiellement longue par le praticien.

Title: About the Use of Gradients in Machine Learning

Keywords: machine learning, stochastic optimisation, dynamical systems

Abstract: We prove a local convergence theorem for the classical dynamical system optimisation algorithm called RTRL, in a non linear setting. The RTRL works on line, but maintains a huge amount of information, which makes it unfit to train even moderately large learning models. The "NoBackTrack"algorithm turns it by replacing these informations by a non biased, low dimension, random approximation. We also prove the convergence with arbitrarily close to one probability, of this algorithm to the local optimum reached by the RTRL algorithm.

We also formalise the LLR algorithm and conduct experiments on it, on synthetic data. This algorithm updates in an adaptive fashion the step size of a gradient descent, by conducting a gradient descent on this very step size. It therefore partially solves the issue of the numerical choice of a step size in a gradient descent. This choice influences strongly the descent and must otherwise be hand-picked by the user, following a potentially long research.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France