Random monotone operators and application to stochastic optimization

Adil Salim

▶ **To cite this version:**

Adil Salim. Random monotone operators and application to stochastic optimization. Optimization and Control [math.OC]. Université Paris Saclay (COmUE), 2018. English. NNT : 2018SACLT021 . tel-01960496

HAL Id: tel-01960496
https://pastel.hal.science/tel-01960496v1

Submitted on 19 Dec 2018

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

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

université PARIS-SACLAY

TELECOM ParisTech

NNT : 2018SACLT021

Thèse de doctorat

Opérateurs monotones aléatoires et application à l'optimisation stochastique

Thèse de doctorat de l'Université Paris-Saclay
préparée à Télécom ParisTech

Ecole doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Automatique, Traitement du Signal, Traitement des Images, Robotique

Thèse présentée et soutenue à Paris, le 26/11/2018, par

ADIL SALIM

Composition du Jury :

Antonin Chambolle Directeur de recherche CNRS, Ecole Polytechnique	Président
Jérôme Bolte Professeur, Université Toulouse 1 Capitole	Rapporteur
Bruno Gaujal Directeur de recherche INRIA, Laboratoire d'Informatique de Grenoble	Rapporteur
Panayotis Mertikopoulos Chargé de recherches CNRS, Laboratoire d'Informatique de Grenoble	Examinateur
Walid Hachem Directeur de recherche CNRS, Université Paris-Est Marne-la-Vallée	Directeur de thèse
Pascal Bianchi Professeur, Télécom ParisTech	Co-directeur de thèse
Jérémie Jakubowicz Chief Data Officer, Vente-privee	Invité
Eric Moulines Professeur, Ecole Polytechnique	Invité

Remerciements

Tout d'abord, je remercie mes directeurs de thèse, Pascal et Walid pour m'avoir proposé ce beau sujet et pour m'avoir encadré pendant ces trois années. J'ai eu de la chance de tomber sur vous. On aura passé de bons moments à parler approximation stochastique à "l'heure où on démontre les théorèmes facilement" ou en déplacements à Troyes, Bordeaux, Nice... Merci pour votre bienveillance, votre patience et votre aide pour ma recherche de postdoc. J'ai énormément appris à vos côtés. Je te remercie également Jérémie, pour m'avoir permis d'effectuer cette thèse et ainsi que pour ta gentillesse et le soutien matériel que tu m'as apporté.

Je remercie vivement Jérôme Bolte et Bruno Gaujal d'avoir accepté de rapporter cette thèse. Merci également à Antonin Chambolle, Panayotis Mertikopoulos et Eric Moulines d'avoir accepté de faire partie du jury. Présenter mes travaux devant vous tous est un grand honneur.

I would also like to thank Volkan Cevher and Peter Richtarik for accepting to host me at EPFL and at KAUST.

Enfin, je te remercie Olivier Fercoq pour l'aide précieuse que tu m'as apporté.

J'ai croisé beaucoup de monde à Télécom. Alors je dédie cette thèse à mes amis et collègues de Comelec, Achraf, Akram (see u in Jeddah), Marwa, Mohamed, Mehdi, Julien, Xavier, Alaa, Hussein, Samet, Yvonne, Chantal, Hamidou et tous ceux que j'oublie. J'ai quitté Comelec à la suite d'une... disons restructuration pour venir travailler à TSI (hum IDS, pardon). Aussi, je tiens à remercier l'équipe que j'y ai trouvé. D'abord mes amis de l'ENSAE, Anna et Moussab, qui sont depuis six ans dans ma promotion. Il sera difficile d'énumérer tout ce qu'on a partagé ici (musique, maths, business, gossip, voyages, mariage...). Mais essentiellement, on fait ce qu'on sait faire, on monte des coups. Guillaume, nos discussions autour de l'optimisation, du rap et du soulevé de terre m'ont beaucoup apporté. Je te souhaite une belle carrière au sein de la franc-maçonnerie. Une spéciale pour le bureau des quatre. Huge, my BAI, RDV au Ghana. Ceux qui doivent entrer à Télécom par la fenêtre, Pierre A. (Calgary Yeah !) et Mathurin (Pouloulou). Et enfin Pierre L., l'homme de la situation, pour les fous rires et pour m'avoir installé Ubuntu. Tu as changé le cours de ma thèse ;). Salutations à Mastane tah les numbers one et Massil de Montréal Rive Sud 94230 t'entends? Dédicasse à Gabriela et Robin pour sa sagesse en termes d'altérophilie et de ski. Big up aux anciens aussi, Nico et Mael (je veux faire de l'oseiiillleeeee), Ray Bro (l'homme qui perd 2 fois son passeport en 1 voyage). J'en place une pour Albert aussi, merci de m'attendre haha. Sans oublier les nouveaux qui vont poursuivre dans les sillons de l'approximation stochastique, Anas (soit solide !) et Sholom, le thésard de nuit. Profitez bien ! Enfin, je ne peux pas terminer ce paragraphe sans passer par la start-up (Alexandre : merci pour les conseils, Alex : préviens moi quand tu retournes à Tahiti, ça m'intéresse, Hamid : on se fera un voyage aux US un jour, ça va être drôle) et par les contrées plus reculées de Télécom, Tom (qui devrait devenir docteur quelques heures après moi, normalement), Valentin et Kevin (cesse de raconter des bêtises).

Je dédie également cette thèse à la communauté scientifique que j'ai cotoyé, les profs de Télécom (Robert, Umut, François, Ons, Joseph, Slim, Alexandre, Stéphan, Olivier), les personnes que j'ai rencontré en conférence (Guillaume, Gabor, ...), ceux qui sont venu squatter (Noufel, Loic), les anciens de

l'ENSAE (Mehdi, Alexander, Badr, Vincent, Pierre A.), d'Orsay (Henri) et Florian, qui m'a donné le goût de la recherche. Je remercie également les profs (avec une mention spéciale pour M. Patte) qui m'ont fait aimé cette discipline, les mathématiques.

Ces trois années m'ont aussi permis de m'initier à differents sports, tels que l'escalade avec Arnaud ou le JJB avec–My nigga My nigga–Ams Warr Sow Pastore, Aket dit le Jardinier et tous les membres du club, Hos !

Un grand merci à mes amis ! Du Hood à l'école en passant par la prépa et Stralmi, on a fait du chemin ! J'ai plein de souvenirs qui me viennent en tête à cet instant. On en aura des choses à raconter en vieillissant. Big up à la team Very Bad Trip, Rich Gang (Arnold, tu es le prochain sur la liste), Niggaz in Paris, Revna, Prémices et les Expats. Madjer, je n'ai pas osé mettre ta citation au début, mais j'y ai pensé fort. J'ai également une pensée pour Marcel, et ceux qui nous ont quitté. Enfin, une mention spéciale pour mon compagnon d'infortune Sami, et le physicien fou Quentin.

Enfin je souhaite dédier cette thèse à ma famille. Ma famille au sens large, le groupe Famille et ma belle-famille. Merci à ma belle-mère et ma belle-famille de nous soutenir au quotidien, vous êtes d'une aide précieuse et on a de la chance de vous avoir. Je rejoins Abi et Abdullah au rang de docteur. Zaki et Houssam, je vous souhaite toute la réussite pour la suite. Je dédie cette thèse à mes oncles Schubert et Zaidou, mon cousin Aouad et mes nièces Naïma et Aya. Enfin, à mon frère Irfane qui m'a montré le chemin, mon père, et ma mère qui m'a toujours soutenu et couvert pour que je puisse étudier sans me soucier du reste. Voilà la récompense pour tes sacrifices. Finalement j'embrasse mon épouse Kawtar qui me supporte au quotidien :). Je suis très heureux d'avoir partagé ces dernières années à tes côtés et je garde d'excellents souvenirs de cette période riche en voyages, délires et émotions. Que cela dure ! Mais attention : c'est fini les t-shirts à fleurs haha ! La vie nous a fait un magnifique cadeau qu'on a appelé Imrane et que j'embrasse également. Merci de m'aider à me lever le matin.

Contents

Chapter 1

Introduction

1.1 Theoretical context : Stochastic Approximation

In the fields of machine learning, statistics or signal processing, many methods rely on an underlying optimization algorithm. In modern applications of data science, it is often not possible to run these algorithms on a single computer. Indeed, when a large amount of data has to be processed, or when streams of data arrive online, either classical algorithms need to be simplified or several computers have to be used. These modifications of classical algorithms can often be formalized by the introduction of randomness in the iterations. To see this, first consider the case of big data problems. Since each iteration of classical algorithms would process the whole dataset, simplified versions of these algorithms will rather process a small randomly chosen amount of data at each iteration. Then, when this task is tackled by a connected network of computing agents, there must be communications inside the network to solve the problem. These communications are often required to happen randomly in the network if it is large. Moreover, in practical settings, the agents compute and communicate only at random instants. Finally, online learning problems need a full knowledge of the distribution of the data to be solved completely. Since streams of data arrive online, the distribution of the data is revealed across time to the user through random realizations. In other words, solving online learning problems requires to be able to work in noisy environments. The algorithms used in the contexts mentioned above can be formalized as optimization algorithms for which the function to minimize is unknown but revealed across the iterations. The literature of stochastic optimization, which studies these algorithms and which this thesis belongs, lies at the intersection of the mathematical optimization and the literature of stochastic approximation. Stochastic optimization algorithms find numerous applications in signal processing and machine learning [34]. Since the seminal work of Robbins and Monro [99] in 1951, stochastic optimization algorithms are analyzed through the prism of the stochastic approximation literature. We start by briefly recalling the goal of stochastic approximation algorithms.

1.1.1 Robbins-Monro algorithm

The stochastic approximation literature studies algorithms that take the form

$$x_{n+1} = x_n + \gamma_{n+1} h(\xi_{n+1}, x_n) \tag{1.1}$$

where x_n are random vectors valued in some Euclidean space X, (ξ_n) is a sequence of random variables (r.v) valued in a measure space Ξ, $(\gamma_n)_n$ is a sequence of positive step sizes and $h : \Xi \times \mathsf{X} \to \mathsf{X}$ is measurable. It is often assumed that the sequence (ξ_n) is independent and identically distributed (i.i.d).

The aim of the algorithm is to find a zero of the expectation function $\mathsf{H}(x) = \mathbb{E}_{\xi_1}(h(\xi_1, x))$ assumed to exist, *i.e* an element $x \in \mathsf{X}$ such that $\mathsf{H}(x) = 0$. Denote as $Z(\mathsf{H})$ the set of zeros of H. Under some assumptions on h and the step sizes (γ_n), it is known that (x_n) converges to $Z(\mathsf{H})$. The strength of stochastic approximation algorithm (1.1) is to be able to find a zero of H without evaluating the expectation $\mathsf{H}(x)$. Indeed, in many applications, the computation of H is intractable. There is a Law of Large Number effect that allows to smooth the randomness. A widely studied example of Robbins-Monro algorithm (1.1) is the stochastic gradient algorithm.

Example 1. The stochastic gradient algorithm aims at finding a minimizer of a differentiable function $\mathsf{F} : \mathsf{X} \to \mathbb{R}$. The function F is itself written as an expectation with respect to (w.r.t.) some r.v ξ

$$\mathsf{F}(x) = \mathbb{E}_{\xi}(f(\xi, x)), \tag{1.2}$$

where $f(\xi, \cdot)$ is a differentiable. The stochastic gradient algorithm update is written

$$x_{n+1} = x_n - \gamma_{n+1} \nabla f(\xi_{n+1}, x_n) \tag{1.3}$$

where the gradient ∇f is taken w.r.t. the second variable (x), (ξ_n) is a sequence of i.i.d copies of ξ and (γ_n) is a sequence of positive step sizes. Algorithm (1.3) can be cast as an instance of (1.1) by setting $h \equiv \nabla f$. If $f(s, \cdot)$ is convex, the following interchange property holds $\mathsf{H}(x) = \mathbb{E}_{\xi}(\nabla f(\xi, x)) = \nabla \mathsf{F}(x)$ for every $x \in \mathsf{X}$. Since F is convex, $Z(\mathsf{H}) = \arg\min \mathsf{F}$. Under mild assumptions, the sequence (x_n) converges to an element in $\arg\min \mathsf{F}$.

Two regimes that require different tools can be considered to analyze stochastic approximation algorithms : the case where $\gamma_n \to_{n \to +\infty} 0$ and the case where $\gamma_n \equiv \gamma > 0$. Typically, in the so-called decreasing step sizes case (first case), the sequence (x_n) of iterates converges almost surely (a.s.) to a zero of H. In the constant step size case (second case), the iterates quickly reach a small neighborhood of the set of solutions $Z(\mathsf{H})$ in a burn-in phase, and then fluctuate around the set of zeros. The main advantage of the decreasing step sizes case is to exhibit the a.s. convergence of the iterates. Although the constant step size case lacks the a.s convergence in general, the use of a constant step size is often more suitable in online learning settings.

A standard method to study evolution equations like (1.1) is the Ordinary Differential Equation (ODE) method, which was introduced in the 70's by Ljung [79] and extensively studied by Kushner and coworkers (see *e.g.* the book [73]). This method allows to study the dynamical behavior of stochastic approximation algorithms and to prove their convergence. Assume that H is a Lipschitz continuous function over X and consider the unique differentiable function $\mathsf{x} : \mathbb{R}_+ \to \mathsf{X}$ such that $\mathsf{x}'(t) = \mathsf{H}(\mathsf{x}(t))$ (where x' denotes the derivative of x) starting at some prescribed value $a \in \mathsf{X} : \mathsf{x}(0) = a$. The ODE method relies on relating the iterates x_n and the function x. More precisely, (x_n) is seen as a noisy Euler discretization of the function x.

1.1.2 A general framework

In this thesis, we develop a more general framework for stochastic approximation because the framework (1.1) fails to cover some important applications, see Sec. 1.2. Consider the following evolution equation

$$x_{n+1} = x_n + \gamma h_{\gamma}(\xi_{n+1}, x_n) \tag{1.4}$$

where the step size $\gamma > 0$ is taken constant, (ξ_n) is i.i.d with distribution μ and h_γ is a measurable function indexed by γ. Let us assume in most generality that

$$\mathbb{E}_{\xi}(h_{\gamma}(\xi, x)) \to_{\gamma \to 0} \mathsf{H}(x) \tag{1.5}$$

where $H : X \to X$ is some function. If H is a Lipschitz continuous function and the convergence (1.5) holds uniformly for x in the compact sets of X, then the ODE method can be applied and we are let back to the situation of the previous paragraph. However, this kind of assumption is too restrictive in many contexts, especially those mentioned in Sec. 1.2 below. We shall focus on the case where the convergence does not hold for every x, is not uniform over compact sets, and above all, H is a *set valued mapping* instead of being single valued. We are therefore led to study more general stochastic approximation algorithms.

1.2 Motivations

1.2.1 Stochastic Proximal Point algorithm

A first motivation for studying the general framework (1.4) comes from non smooth stochastic optimization. Consider a convex function $G : X \to (-\infty, +\infty]$ which is lower semicontinuous (lsc) and proper (we shall write $G \in \Gamma_0(X)$). Denote $\partial G(x)$ the set of all subgradients of G at x. The subdifferential $\partial G : X \rightrightarrows X$ is a set valued function. Consider $x \in X$. The proximity operator [87] of G at x is *the* minimizer of the (strongly convex) objective function:

$$\mathrm{prox}_{\gamma G}(x) = \arg\min_{y \in X} G(y) + \frac{1}{2\gamma}\|x - y\|^2, \tag{1.6}$$

where $\gamma > 0$, and the Moreau envelope [130] of G at x is the associated minimum value

$$G_\gamma(x) = \min_{y \in X} G(y) + \frac{1}{2\gamma}\|x - y\|^2. \tag{1.7}$$

Moreover, the Moreau envelope is differentiable and its gradient is a $1/\gamma$-Lipschitz continuous function that satisfies (see [12])

$$\mathrm{prox}_{\gamma G} = I - \gamma \nabla G_\gamma \tag{1.8}$$

where I is the identity of X. The goal of the proximal point algorithm [82] is to find a minimizer of G (equivalently a point x such that $0 \in \partial G(x)$, called hereafter a zero of ∂G) by iterating

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n), \tag{1.9}$$

where $\gamma > 0$. It is known that the sequence (x_n) converges to a minimizer of G. The proximal point algorithm enjoys good stability properties, among which exhibiting convergence for any $\gamma > 0$. The main drawback of this algorithm is that each iterate is implicitly defined, *i.e* one has to solve an optimization problem to find x_{n+1}. This operation can often be costly. Although the proximity operator of some classical functions has a closed form[1], the proximal point algorithm is not practical in many cases.

A way to simplify the iterations is to represent $G(x) = \mathbb{E}_\xi(g(\xi, x))$ where $g(\xi, \cdot)$ is a convex function and to apply the constant step stochastic proximal point algorithm [22, 121]:

$$x_{n+1} = \mathrm{prox}_{\gamma g(\xi_{n+1}, \cdot)}(x_n), \tag{1.10}$$

where $\gamma > 0$ and (ξ_n) are i.i.d copies of ξ. This algorithm can be seen as a generalization of the splitting algorithm of Passty [94] to infinitely many functions. Many loss functions used in machine learning can be

[1] See the website www.proximity-operator.net

written as expectations $\mathsf{G}(x) = \mathbb{E}_\xi(g(\xi, x))$ for which $\mathrm{prox}_{\gamma g(\xi,\cdot)}$ is easily computable whereas neither G nor $\mathrm{prox}_{\gamma \mathsf{G}}$ can be evaluated. A typical situation is the case where these loss functions boil down to finite sums $\mathsf{G}(x) = N^{-1}\sum_{i=1}^N g(i, x)$ and $\mathrm{prox}_{\gamma g(i,\cdot)}$ can be easily computed but $\mathrm{prox}_{\gamma \mathsf{G}}$ is intractable. This is *e.g.* the case for the classification problems like Support Vector Machine (SVM) or logistic regression. Another example comes from distributed optimization in the context where a network of computing agents is required to minimize a "global" cost function $\mathsf{G} = N^{-1}\sum_i g(i, \cdot)$, under the restriction that the "local" cost function $g(i, \cdot)$ is only known by the agent i. Hence, the network can only perform local computations involving each agent i and their respective cost function $g(i, \cdot)$ separately. In all these situations, the proposed algorithm is an instance of (1.10) where ξ is a uniform r.v. over $\{1, \ldots, N\}$.

Note that (1.10) can be cast in the form (1.4) by setting $h_\gamma(s, x) = -\nabla g_\gamma(s, x)$, where $g_\gamma(s, \cdot)$ is the Moreau envelope of $g(s, \cdot)$. In this case, $\mathsf{H} = \partial \mathsf{G}$ is set valued.

1.2.2 Stochastic Proximal Gradient algorithm

In optimization algorithms, proximity operators are often used to handle regularizations or constraints. In these cases, the problem to be solved is

$$\min_{x\in \mathsf{X}} \mathsf{F}(x) + \mathsf{G}(x),$$

where $\mathsf{G} \in \Gamma_0(\mathsf{X})$ is a convex function and F is assumed differentiable. The proximal gradient algorithm generalizes the proximal point algorithm (1.9) and is written

$$x_{n+1} = \mathrm{prox}_{\gamma \mathsf{G}}(x_n - \gamma \nabla \mathsf{F}(x_n)), \tag{1.11}$$

where $\gamma > 0$. If $\nabla \mathsf{F}$ is Lipschitz continuous (we shall say that F is smooth), if F is convex and if γ is enough small, then it is known that (x_n) converges to a minimizer of $\mathsf{F} + \mathsf{G}$. A first instance of this algorithm is the projected gradient algorithm to solve $\min_{\mathcal{C}} \mathsf{F}$. This algorithm can be seen as an application of (1.11) by setting $\mathsf{G} = \iota_{\mathcal{C}}$, where $\iota_{\mathcal{C}}$ the convex indicator function of the convex set $\mathcal{C}$. In this case, $\mathrm{prox}_{\mathsf{G}} = \Pi_{\mathcal{C}}$ is the projector onto $\mathcal{C}$. Each iteration of this algorithm requires to evaluate the projection $\Pi_{\mathcal{C}}$, which is sometimes intractable. However, the set $\mathcal{C}$ can often be represented as an intersection of simpler convex sets $\mathcal{C}_s$, *i.e* $\mathcal{C} = \bigcap_{s\in\Xi} \mathcal{C}_s$ where projections onto $\mathcal{C}_s$ can be easily computed. Another instance of the proximal gradient algorithm comes from structured problem in which G is a regularization term. The function G is represented as $\mathsf{G} = \sum_i g(i, \cdot)$, where $\mathrm{prox}_{\gamma \mathsf{G}}$ is hard to compute but $\mathrm{prox}_{\gamma g(i,\cdot)}$ can be evaluated. This is *e.g* the case for the overlapping group lasso : $\mathsf{G}(x) = \sum_i \|x_{S_i}\|$ where $\mathsf{X} = \mathbb{R}^N$, the S_i are subsets of $\{1, \ldots, N\}$ and x_{S_i} is the restriction of x to S_i. This is also the case for the total variation regularization : $\mathsf{G}(x) = \sum_{\{i,j\}\in E} \|x(i) - x(j)\|$ where $G = (V, E)$ is a graph, with V the set of nodes and E the set of edges, and where $x \in \mathbb{R}^V$. In all these examples, the proximal gradient algorithm cannot be implemented because it involves the computation of $\mathrm{prox}_{\gamma \mathsf{G}}$. However, in all these examples G can be seen as an expectation w.r.t. some r.v. ξ, $\mathsf{G}(x) = \mathbb{E}_\xi(g(\xi, x))$ (where the expectation sometimes boils down to a finite sum). In general, the stochastic proximal gradient algorithm aims at minimizing $\mathsf{F}(x) + \mathsf{G}(x) = \mathbb{E}_\xi(f(\xi, x)) + \mathbb{E}_\xi(g(\xi, x))$ by iterating [24, 2, 3]

$$x_{n+1} = \mathrm{prox}_{\gamma g(\xi_{n+1},\cdot)}(x_n - \gamma \nabla f(\xi_{n+1}, x_n)). \tag{1.12}$$

This algorithm can be cast in the form (1.4) by setting

$$h_\gamma(s, x) = \frac{1}{\gamma}(\mathrm{prox}_{\gamma g(s,\cdot)}(x - \gamma \nabla f(s, x)) - x).$$

Using (1.8), $h_\gamma(s, x) = -\nabla f(s, x) - \nabla g_\gamma(s, x - \gamma \nabla f(s, x))$. Moreover, $\mathsf{H} = \nabla \mathsf{F} + \partial \mathsf{G}$ in this case.

1.2.3 Stochastic Douglas Rachford algorithm

When minimizing a sum of two convex functions $\mathsf{F}+\mathsf{G}$, the Douglas Rachford algorithm [78] enjoys more numerical stability than the proximal gradient algorithm at the cost of implementing a proximity operator for F instead of a gradient. Moreover, any positive constant step size can be used in Douglas Rachford iterations to converge to a minimizer. In order to design an algorithm that enjoys the good features of Douglas Rachford algorithm without the iteration complexity, we are interested in the stochastic Douglas Rachford algorithm with constant step size, in which the proximity operator of F (resp. G) is randomized. To this end, F and G are represented as expectations, as in Sec. 1.2.2 and the resulting stochastic Douglas Rachford algorithm is also covered by our general framework (1.4).

1.2.4 Monotone operators and Stochastic Forward Backward algorithm

Maximal monotone operators are set valued functions that generalize the subdifferentials [12, 36]. Many optimization problems can be reformulated as finding zeros of a monotone operator (which is not necessarily a subdifferential). In this respect, the Forward Backward (FB) algorithm is a further generalization of the proximal gradient algorithm. The goal of this algorithm is to find a zero of a sum of two maximal monotone operators.

In this thesis, we refer to an operator as a set valued function $\mathsf{A}:\mathsf{X}\rightrightarrows\mathsf{X}$. The inverse operator A^{-1} is defined by the relation $y\in\mathsf{A}(x)\Leftrightarrow x\in\mathsf{A}^{-1}(y)$. An operator A is said monotone if $\langle y-y',x-x'\rangle\geq 0$ as soon as $y\in\mathsf{A}(x)$ and $y'\in\mathsf{A}(x')$. Under a maximality condition [85] of A, the resolvent of A, $\mathsf{J}_\gamma=(I+\gamma\mathsf{A})^{-1}$ is a single valued function. In this case, A is called a maximal monotone operator and J_γ is a contraction defined on X. Maximal monotone operators generalize subdifferentials of convex functions and resolvents generalize proximity operators. Indeed, $\mathsf{A}=\partial\mathsf{G}$ is a maximal monotone operator if $\mathsf{G}\in\Gamma_0(\mathsf{X})$, and its resolvent is $\mathrm{prox}_{\gamma\mathsf{G}}$. For every $\gamma>0$, the Yosida approximation of A is defined by $\mathsf{A}_\gamma=\frac{1}{\gamma}(I-\mathsf{J}_\gamma)$. Using (1.8), it is immediately seen that $\mathsf{A}_\gamma=\nabla\mathsf{G}_\gamma$ if $\mathsf{A}=\partial\mathsf{G}$. The set of zeros of A is defined to be $Z(\mathsf{A})=\mathsf{A}^{-1}(0)$. Many problems in optimization can be reformulated as finding a zero of a maximal monotone operator. For example, in the subdifferential case, $Z(\partial\mathsf{G})=\arg\min\mathsf{G}$. Given another maximal monotone operator which is single valued B, the Forward Backward algorithm aims at finding an element in $Z(\mathsf{A}+\mathsf{B})$ by iterating

$$x_{n+1}=\mathsf{J}_\gamma(x_n-\gamma\mathsf{B}(x)). \tag{1.13}$$

If A and B are subdifferentials, the Forward Backward algorithm boils down to the proximal gradient algorithm. Under a so called cocoercivity assumption of B, this algorithm is known to converge to a zero of $\mathsf{A}+\mathsf{B}$ if γ is small enough.

Beyond minimization problems, saddle points problems arise naturally in optimization and machine learning (see *e.g* [83]). We the saddle points problems are convex-concave, they can be reformulated as finding a zero of a sum of two monotone operators $\mathsf{A}+\mathsf{B}$ [101]. In optimization, primal dual algorithms like Douglas-Rachford [78], ADMM [62], Chambolle Pock [42] or Vu Condat [51, 124] can be seen as (skillful) instances of the FB algorithm. This FB algorithm is applied to the convex concave saddle point problem of finding so called primal dual optimal points of the initial optimization problem.

In order to develop a stochastic version of these primal dual algorithms, we are interested in a stochastic version of the Forward Backward algorithm. To this end, we consider a new tool called a random monotone operator $A(\xi,\cdot)$, *i.e* ξ is a r.v. and $A(\xi,\cdot)$ is a maximal monotone operator. Measurability issues due to the fact that A is set valued will be treated in the next chapter. Denote $J_\gamma(s,\cdot)$ the resolvent of $A(s,\cdot)$ and $A_\gamma(s,\cdot)$ its Yosida approximation. Consider an other random monotone

operator $B(\xi,\cdot)$ which is single valued. The constant step stochastic FB algorithm is written

$$x_{n+1} = J_\gamma(\xi_{n+1}, x_n - \gamma B(\xi_{n+1}, x_n)). \tag{1.14}$$

The aim of the stochastic FB algorithm is to find a zero of the so called mean operator $\mathsf{A}(x) + \mathsf{B}(x) = \mathbb{E}_\xi(A(\xi,x)) + \mathbb{E}_\xi(B(\xi,x))$. Integrability issues due to the fact that $A(\xi,x)$ is a set-valued r.v. will be treated in the next chapter, along with the definition of the expectation of a set-valued r.v. We just mention here the fact that in the subdifferential case $A(s,x) = \partial g(s,x)$, under mild assumptions [102], $\mathbb{E}_\xi(A(\xi,x)) = \partial \mathsf{G}(x)$ where $\mathsf{G}(x) = \mathbb{E}_\xi(g(\xi,x))$ (we shall say that the interchange property holds). The stochastic FB (1.14) can be cast into the framework (1.4) by setting $h_\gamma(s,x) = -B(s,x) - A_\gamma(s, x - \gamma B(s,x))$ and $\mathsf{H} = \mathsf{A} + \mathsf{B}$.

1.2.5 Fluid limit of parallel queues

Beyond stochastic optimization algorithms, the framework (1.4) can be used to study general Markov chains. For example the framework (1.4) is considered in [61] to study Markov chains with discontinuous drift. Using the notation of (1.4), this means that even $h_\gamma(s,\cdot)$ is discontinuous. We give an application example that comes from queueing theory. We are interested in establishing the long-run behavior of the number of users in a model of parallel queues. Users arrive at random instant in the queues and the queues are served following a prioritizing rule. After scaling the problem in order to study the so called fluid scaled process [61], the evolution of the number of users in the queues fits our framework (1.4).

1.3 Dynamics of Robbins-Monro algorithm

To better understand the methods used in this thesis, we get back to the Robbins-Monro algorithm of Sec. 1.1.1. We provide the main arguments behind the ODE method. We shall focus on the constant step case that will be of interest in the first part of this thesis. More precisely, we study the evolution equation (1.1) with $\gamma_n \equiv \gamma > 0$.

1.3.1 Known facts related with dynamical systems

Consider a Lipschitz continuous function $\mathsf{H} : \mathsf{X} \to \mathsf{X}$. Then, it is well known that for every $a \in \mathsf{X}$, the ODE $x' = \mathsf{H}(x)$ with initial condition $x(0) = a$ admits an unique solution over $\mathbb{R}_+$ [73]. We denote by $\Phi(a)$ this solution and abusively denote $\Phi(a)(t) = \Phi(a,t)$ for every $t \geq 0$. It is known that Φ satisfies the property of being a semiflow over X, *i.e.* $\Phi(\cdot, s+t) = \Phi(\cdot,t) \circ \Phi(\cdot,s)$ for every $t, s \geq 0$. The essence of the ODE method is to study the behavior of the interpolated process obtained from the iterates (x_n) of the algorithm (1.1) as being an approximation of the ODE solution. To perform this analysis, some important notions related to the dynamics of the semiflow Φ need to be introduced. A probability measure π over X is called an invariant measure for Φ if $\pi = \pi\Phi(\cdot,t)^{-1}$ for every $t > 0$. The set of invariant measures for Φ is denoted $\mathcal{I}(\Phi)$. A point $x \in \mathsf{X}$ is said recurrent for Φ if $x = \lim_{k\to+\infty} \Phi(x, t_k)$ for some sequence $t_k \to +\infty$. The Birkhoff center BC_Φ of Φ is the closure of the set of recurrent points. The celebrated Poincaré's recurrence theorem [53, Th. II.6.4 and Cor. II.6.5] says that the support of any $\pi \in \mathcal{I}(\Phi)$ is a subset of BC_Φ. The goal of the two next sections is to prove that the sequence of iterates (x_n) defined by (1.1) with a constant step size $\gamma_n \equiv \gamma > 0$ converges in probability to the set BC_Φ as $n \to +\infty$ and $\gamma \to 0$. Indeed, BC_Φ is often a set of interest while looking for zeros of H. For example, if $\Phi(a,t)$ converges to $Z(\mathsf{H})$ as $t \to +\infty$ for every $a \in \mathsf{X}$, then $Z(\mathsf{H}) = \mathsf{BC}_\Phi$.

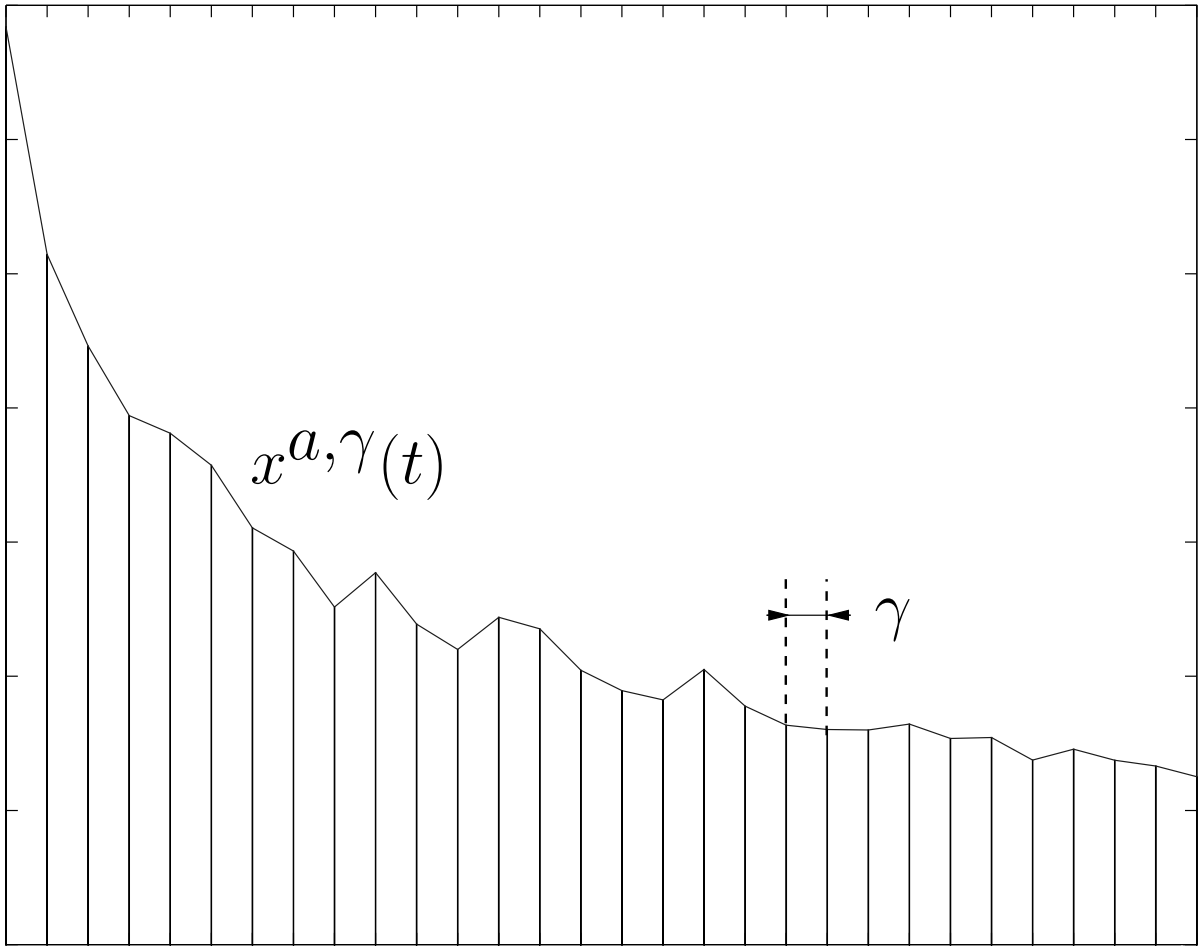

Figure 1.1: The linearly interpolated process of the iterates with step size $\gamma > 0$.

1.3.2 Convergence of stochastic processes

Consider a sequence (x_n) satisfying (1.1) with a constant step size $\gamma_n \equiv \gamma > 0$ starting from $x_0 = a$. Consider the linearly interpolated process of the sequence of iterates (see Fig. 1.1) $\mathsf{x}^{a,\gamma}$ over $\mathbb{R}_+$, piecewise defined for every $t \geq 0$ by

$$\mathsf{x}^{a,\gamma}(t) = x_n + (t - n\gamma)\frac{x_{n+1} - x_n}{\gamma}, \qquad t \in [n\gamma, (n+1)\gamma), \qquad n \in \mathbb{N}. \tag{1.15}$$

As a continuous time stochastic process, $\mathsf{x}^{a,\gamma}$ can be seen as a r.v. in the space $C(\mathbb{R}_+, \mathsf{X})$ of continuous functions endowed with the topology of the uniform convergence over bounded intervals. The ODE method first consists in showing that $\mathsf{x}^{a,\gamma} \longrightarrow_{\gamma \to 0} \Phi(a)$ in distribution in $C(\mathbb{R}_+, \mathsf{X})$ (*i.e* narrowly, see [14]). In other words, one can show that for every $T > 0$, $\sup_{t \in [0,T]} \|\mathsf{x}^{a,\gamma}(t) - \Phi(a,t)\| \longrightarrow 0$ in probability as $\gamma \to 0$ under some prescribed assumptions.

This important result does not suffice to characterize the long run behavior of the iterates *i.e* the case $T = +\infty$. What is ultimately needed is the long-run behavior of the process $\mathsf{x}^{a,\gamma}$ in terms of the Birkhoff center of the semiflow Φ. A stability result is needed.

1.3.3 Stability result

To characterize the long-run behavior of $\mathsf{x}^{a,\gamma}$, the sequence (x_n) is viewed as a Markov chain with transition kernel P_γ. The advocated stability result typically ensures that the set of invariant measures for $P_\gamma, \gamma \in (0, \gamma_0)$ is relatively compact for some $\gamma_0 > 0$. Under such a condition, the first result on the narrow convergence of $\mathsf{x}^{a,\gamma}$ can be used to show that every cluster point of the invariant measures of the Markov chain as $\gamma \to 0$ is an invariant measure for the semiflow Φ [60]. Using Poincaré's recurrence theorem, such cluster points are supported by the Birkhoff center BC_Φ. A reformulation of this result is the following :

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}(d(x_n, \mathsf{BC}_\Phi) > \varepsilon) \longrightarrow_{\gamma\to 0} 0. \tag{1.16}$$

1.4 From ODE to Differential Inclusions

Motivated by the examples of 1.2, we shall relax the classical assumptions used in the ODE method and study the framework (1.4) where H is allowed to be set valued. In this situation, the classical ODE is replaced with a *Differential Inclusion* (DI): $\dot{\mathsf{x}} \in \mathsf{H}(\mathsf{x})$ defined on the set of absolutely continuous functions over $\mathbb{R}_+$, where $\dot{\mathsf{x}}$ denotes the derivative of x defined almost everywhere. Stochastic approximation algorithms built on DI have recently aroused an important research effort to which this thesis belongs [17, 80]. In this work, two kinds of DI with different behaviors are of interest.

1. First, the case where $\mathsf{H}(x)$ is convex compact and not empty for every $x \in \mathsf{X}$ and H is upper semicontinuous [6] (usc) *i.e* for every $a \in \mathsf{X}$, and for every open set U such that $\mathsf{H}(a) \subset U$, there exists a neighborhood V of a such that $x \in V \Rightarrow \mathsf{H}(x) \subset U$. Assuming that for every $a \in \mathsf{X}$ there exists a solution to the DI starting at a (this holds under a linear growth assumption on H), the solution is in general not unique and the semiflow associated to the DI is hence set valued [6]. This kind of DI is of interest in many applications including game theory, or queueing systems.

2. Second, the case where $-\mathsf{H}$ is a maximal monotone operator [36]. In this case, we considered in particular the situation where the domain of H is strictly included in X, which is of obvious interest for many stochastic optimization algorithms.

1.5 Contributions

1.5.1 Convergence analysis with a constant step size

We first focus on the analysis of constant step stochastic approximation algorithms having a DI as a limit. We shall study the case where $(h_{\gamma_n}(s, x_n))_{n\in\mathbb{N}}$ converges to the set $H(s, x)$ as $n \to +\infty$ if $x_n \to x$ and $\gamma_n \to 0$. The function H is represented as a set valued expectation $\mathsf{H}(x) = \mathbb{E}_\xi(H(\xi, x))$. The set valued expectation is formally defined as a *selection integral* and generalizes the Lebesgue integral to set valued mappings, see Chap. 2. To study the dynamics of the iterates (x_n) given by (1.4) we adapt the general approach of Sec. 1.3 to DI $\dot{\mathsf{x}} \in \mathsf{H}(\mathsf{x})$ in the usc case and the monotone case of Sec. 1.4, each case requiring a specific treatment and exhibiting a specific convergence result.

The upper semicontinuous case

In this case, $H(s,\cdot)$ is assumed to be a proper ($\exists x \in \mathsf{X}, H(s,x) \neq \emptyset$) usc operator. We denote as $\Phi(a)$ the *set* of solutions to the DI $\dot{x} \in \mathsf{H}(x)$ starting at a. We assume that $\Phi(a)$ is not empty, and Φ can be seen as set valued flow. Set valued analogues to classical dynamical systems results 1.3.1 will be considered. This framework is introduced in the paper [107] under the additional assumption that X is a compact space. In our work, we relaxed the compactness assumption, which extends the scope of the algorithm (1.4) to *e.g.*, proximal non convex stochastic gradient algorithm 1.2.2, or queuing algorithms such as 1.2.5. Denote $\mathcal{I}(\Phi)$ the set of invariant measures for the set valued flow Φ, a notion introduced in [107]. Denoting d a distance that metricizes the topology over $C(\mathbb{R}_+, \mathsf{X})$ of uniform convergence over compact intervals, we first prove the dynamical result

$$\sup_{a\in\mathcal{K}} \mathbb{P}(\mathsf{d}(\mathsf{x}^{a,\gamma}, \Phi(a)) > \varepsilon) \longrightarrow_{\gamma\to 0} 0, \tag{1.17}$$

for every compact set $\mathcal{K} \subset \mathsf{X}$ and every $\varepsilon > 0$, where $\mathsf{d}(\mathsf{x}^{a,\gamma}, \Phi(a))$ denotes the distance from the function $\mathsf{x}^{a,\gamma}$ to the set $\Phi(a)$. Under a stability assumption of the Markov chain (x_n) this dynamical result is used to characterize the long-run behavior of the iterates :

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}(d(x_n, \mathsf{BC}_\Phi) > \varepsilon) \longrightarrow_{\gamma\to 0} 0, \tag{1.18}$$

where (x_n) is the process satisfying (1.4) with step size $\gamma > 0$. Similar results involving the empirical means $\overline{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ are obtained. Finally, stability conditions based on the so-called Pakes-Has'minskii criterion are provided in the context of the stochastic proximal non convex gradient algorithm (under a Łojasiewicz assumption [5, 30]) and in a model of parallel queues [61].

The monotone case

In this case, $-H(s,\cdot)$ is assumed to be a maximal monotone operator with domain $D(s) = \{x \in \mathsf{X}, H(s,x) \neq \emptyset\}$. If $D(s) = \mathsf{X}$, then $H(s,\cdot)$ is usc [97] and a dynamical result can be obtained from (1.17). We shall allow the domains $D(s)$ to vary with s. This covers the contexts of the stochastic proximal point algorithm 1.2.1, the stochastic proximal gradient algorithm in the convex case 1.2.2, the Douglas-Rachford algorithm 1.2.3 and the stochastic Forward Backward algorithm 1.2.4. With a proof that explicitly leverages the maximal monotonicity of $-H(s,\cdot)$ and allows the domains to be random, we first show that

$$\sup_{a\in\mathcal{K}\cap\mathcal{D}} \mathbb{P}(\mathsf{d}(\mathsf{x}^{a,\gamma}, \Phi(a)) > \varepsilon) \longrightarrow_{\gamma\to 0} 0, \tag{1.19}$$

where $\mathcal{D}$ is the domain of the mean operator H. Then, under a stability assumption of the Markov chain (x_n), it is shown that if H satisfies the so called demipositivity assumption (see Chap. 2), then

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}\left(d(x_k, Z(\mathsf{H})) > \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,. \tag{1.20}$$

Similar results that hold whether H is demipositive or not are obtained for the empirical means of the iterates. Finally, practical criteria ensuring the stability of the Markov chain (x_n) are provided in various instances of the stochastic Forward-Backward algorithm, including the stochastic proximal gradient algorithm of Sec. 1.2.2, the case where $H(s,\cdot)$ is linear and monotone, etc.

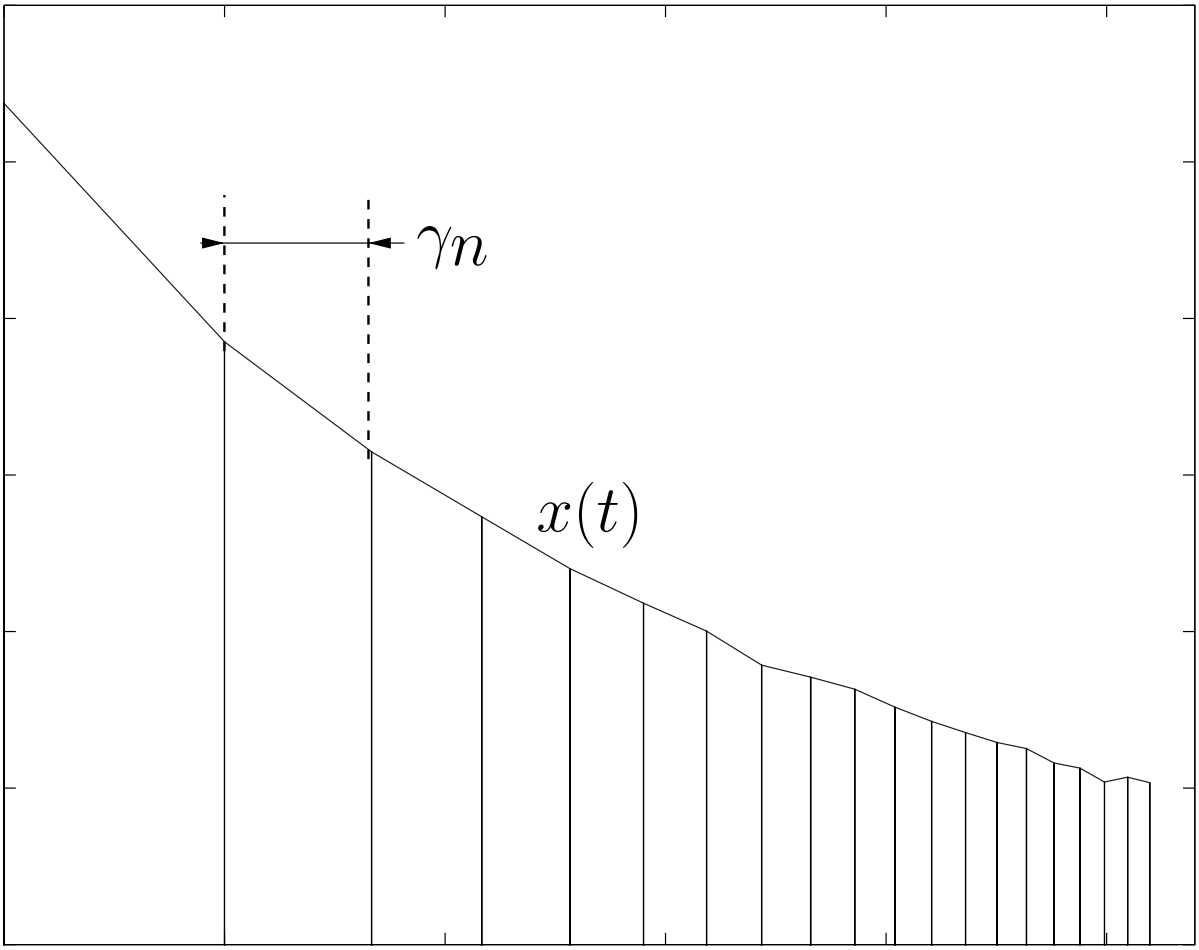

Figure 1.2: The linearly interpolated process of the iterates with step sizes γ_n.

1.5.2 Applicative contexts using decreasing step sizes

Stochastic approximation with decreasing step sizes

The ODE method can also be used to study decreasing step sizes algorithm, *i.e* the evolution equation (1.1) with $\gamma_n \to 0$. In this case, the linearly interpolated process x of (x_n) over $\mathbb{R}_+$ with timeframe γ_n is considered, see Fig. 1.2. The general idea is to prove the almost sure convergence of the interpolated process x to the solution of the ODE $\mathsf{x}' = \mathsf{H}(\mathsf{x})$. More precisely, the interpolated process is proven to be an almost sure Asymptotic Pseudo Trajectory (APT) of the ODE, a concept introduced by Benaïm and Hirsch in the field of dynamical systems [15]. It is shown that $\mathsf{d}(\mathsf{x}(t+\cdot), \Phi(\mathsf{x}(t))) \to_{t\to+\infty} 0$ a.s., where we recall that d metricizes the topology of the uniform convergence over compact sets and where Φ is the semiflow associated with the ODE. Then, the asymptotic convergence of the sequence of iterates (x_n) of the algorithm (1.1) can be obtained from the APT property. This notion has been generalized to monotone DI in [24]. More precisely, the paper [24] studies the decreasing step size analogue of the stochastic Forward Backward algorithm (1.14)

$$x_{n+1} = J_{\gamma_{n+1}}(\xi_{n+1}, x_n - \gamma_{n+1} B(\xi_{n+1}, x_n)), \tag{1.21}$$

where $\gamma_n \to 0$. It is proven that the interpolated process of the iterates (x_n) is an almost sure APT of the DI $\dot{\mathsf{x}} \in \mathsf{H}(\mathsf{x})$ where $\mathsf{H} = -(\mathsf{A} + \mathsf{B})$ is the mean monotone operator. Then, it is deduced that (x_n)

converges to an element of $Z(\mathsf{H})$ as $n \to +\infty$ if $-\mathsf{H}$ is monotone and demipositive, and the sequence of empirical means $(\overline{x}_n)_n$ converges to a solution as $n \to +\infty$ whether $-\mathsf{H}$ is demipositive or not. In this thesis, we proceed with Algorithm (1.21) in two directions. First, we apply this algorithm to solve a general composite optimization problem under linear constrains. The functions defining the objective function and the matrices defining the constraints are allowed to be represented as expectations, see Sec. 1.5.2 below. Second, we generalize the stochastic proximal gradient algorithm with decreasing step sizes to solve a regularized optimization problem over a large and general graph, see Sec. 1.5.2 below.

A fully stochastic primal dual algorithm

A first example comes from primal dual optimization algorithms. Consider four convex functions $\mathsf{F}, \mathsf{G} \in \Gamma_0(\mathsf{X})$ and $\mathsf{P}, \mathsf{Q} \in \Gamma_0(\mathsf{Z})$ where Z is an Euclidean space. Consider the following optimization problem:

$$\min_{(x,z)\in \mathsf{X}\times\mathsf{Z}} \mathsf{F}(x) + \mathsf{G}(x) + \mathsf{P}(z) + \mathsf{Q}(z) \quad \text{subject to} \quad \mathsf{A}x + \mathsf{B}z = c \tag{1.22}$$

where $\mathsf{A} : \mathsf{X} \to \mathsf{V}$ and $\mathsf{B} : \mathsf{Z} \to \mathsf{V}$ are matrices with values in the Euclidean space V, and $\mathsf{c} \in \mathsf{V}$ is a vector. In order to identify a minimizer of (1.22), primal dual methods typically generate a sequence of primal estimates $(x_n, z_n)_{n\in\mathbb{N}}$ and a sequence of dual estimates $(\lambda_n)_{n\in\mathbb{N}}$ jointly converging to a saddle point $((x, z), \lambda)$ of the Lagrangian function associated with (1.22). Under some qualification condition, (x, z) is a solution of Problem (1.22) and λ is a solution of a dual formulation of (1.22). The formulation (1.22) encompasses the formulation of classical primal dual algorithms [62, 42, 51, 124]. In these algorithms, F, P are treated explicitly (*i.e* through their gradient) and G, Q are treated implicitly (*i.e* through their proximity operator). We shall focus on the case where all functions to be minimized are given as statistical expectations, as well as the matrices and the vector defining the linear constraints. In other words, $\mathsf{F}(x) = \mathbb{E}_\xi(f(\xi, x))$ where $f(\xi, \cdot)$ is a convex function. A similar representation is allowed for G, P and Q. Besides, $\mathsf{A} = \mathbb{E}(A)$ where A is a random matrix. A similar representation is allowed for B and c. These expectations are unknown but revealed across the time through i.i.d realizations. Only stochastic (sub)gradients or stochastic proximity operators are available to the user. To solve this problem, we first remark that saddle points of the Lagrangian can be seen as zeros of a sum of two well chosen maximal monotone operators which are given as a set valued expectations. Hence, the stochastic FB algorithm can be applied and leads to a converging algorithm. To our knowledge, the proposed algorithm is the first fully stochastic primal dual algorithm. Application to distributed and asynchronous optimization will be considered.

Online regularization over large graphs

Consider a graph $G = (V, E)$ where V is the set of vertices and E is the set of edges. We first consider the resolution of the following programming problem

$$\min_{x\in\mathbb{R}^V} F(x) + \mathrm{TV}(x, G) \tag{1.23}$$

where $F \in \Gamma_0(\mathbb{R}^V)$ and $\mathrm{TV}(x, G) = \sum_{\{i,j\}\in E} |x(i) - x(j)|$ is the Total Variation regularization over the graph G. When applying the proximal gradient algorithm to solve this problem, there exist quite affordable methods to implement the proximity step in the special case where the graph is a simple path without loops. However, when the graph is large and unstructured, the computation of the proximity operator is more difficult. To overcome this difficulty, we introduced an algorithm that we called "Snake" and that consists in randomizing the proximity operator in such a way that it becomes computable. More

precisely, Snake selects random simple paths in the graph and performs the proximal gradient algorithm over these simple paths. Hence, only proximity operators over simple paths are computed and Snake take benefits of existing fast methods. Then, Snake is generalized to any regularization term tied to the graph geometry for which there exists fast methods to compute the proximity operator over a simple path. Snake is an instance of a generalization of the stochastic proximal gradient algorithm, whose convergence is proven. Numerical experiments are conducted over large graphs.

1.6 Outline of the Thesis

The next chapter is an introduction to some important notions used in the thesis. Then, the first part of the thesis studies the stochastic approximation framework (1.4) with a constant step size, mainly from a theoretical point of view. It consists in three chapters. Chapter 3 is related to Differential Inclusion involving an upper semicontinuous operator, and is based on the publication [28]. In Chapter 4, an analysis of the stochastic Forward Backward algorithm is performed, based on [25, 26, 27]. In Chapter 5, the stochastic Douglas Rachford algorithm is studied and applications to structured regularization and distributed optimization is considered. This chapter is based on the papers [89, 111] and the technical report [110]. Applications of stochastic approximation algorithms with decreasing step size are considered in the second part of the thesis. After recalling the main ideas behind the proof techniques in Chapter 6, we first consider a fully stochastic primal dual algorithm in Chapter 7, based on the work [112]. Finally, we provide an application to solve regularized problems over graphs in Chapter 8 ([109, 113, 114]). Chapter 9 is devoted to a conclusion. The technical report [110] can be found in the Appendix A.

Chapter 2

Preliminaries

2.1 General notations

If E is a topological space, the Borel σ-field of E is denoted as $\mathscr{B}(E)$ and the set of probability measures on E endowed with its Borel field is denoted $\mathcal{M}(E)$. If $(\Xi, \mathscr{G}, \mu)$ is a probability space and X and Euclidean space endowed with its Borel σ-field, the Banach space of measurable functions $\varphi : \Xi \to \mathsf{X}$ such that $\|\varphi\|^p$ is μ-integrable (for $p \geq 1$) is denoted $\mathcal{L}^p(\Xi, \mathscr{G}, \mu; \mathsf{X})$. The notation $C(E, F)$ is used to denote the set of continuous functions from E to the topological space F. The notation $C_b(E)$ stands for the set of bounded functions in $C(E, \mathbb{R})$.

We use the conventions $\sup \emptyset = -\infty$ and $\inf \emptyset = +\infty$. Notation $\lfloor x \rfloor$ stands for the floor value of x. If (E, d) is a metric space, for every $x \in E$ and $S \subset E$, we define $d(x, S) = \inf\{d(x, y) : y \in S\}$. We say that a sequence $(x_n, n \in \mathbb{N})$ on E converges to S, noted $x_n \to_n S$ or simply $x_n \to S$, if $d(x_n, S)$ tends to zero as n tends to infinity. For $\varepsilon > 0$, we define the ε-neighborhood of the set S as $S_\varepsilon := \{x \in E : d(x, S) < \varepsilon\}$. The closure of S is denoted by $\mathrm{cl}(S)$, and its complementary set by S^c. The characteristic function of S is the function $\mathbb{1}_S : E \to \{0, 1\}$ equal to one on S and to zero elsewhere. If E is an Euclidean space, the convex hull of S is denoted by $\mathrm{co}(S)$.

2.2 Set valued mappings and monotone operators

Consider an Euclidean space X. We recall some basic facts related with set valued mappings and their associated differential inclusions with emphasis on maximal monotone operators over X. These facts will be used in the proofs without mention. For more details, the reader is referred to the treatises [40], [12], [6], [36], or to the tutorial paper [96].

2.2.1 Basic facts on set valued mappings

Consider a set valued mapping (or operator) $\mathsf{H} : \mathsf{X} \rightrightarrows \mathsf{X}$, *i.e.*, for each $x \in \mathsf{X}$, $\mathsf{H}(x)$ is a subset of X. The domain and the graph of H are the respective subsets of X and $\mathsf{X} \times \mathsf{X}$ defined as $\mathrm{dom}(\mathsf{H}) := \{x \in \mathsf{X} : \mathsf{H}(x) \neq \emptyset\}$, and $\mathrm{gr}(\mathsf{H}) := \{(x, y) \in \mathsf{X} \times \mathsf{X} : y \in \mathsf{H}(x)\}$. The operator H is proper if $\mathrm{dom}(\mathsf{H}) \neq \emptyset$. Besides, H is said *upper semi continuous* (usc) at a point $a \in \mathsf{X}$ if for every open set U containing $\mathsf{H}(a)$, there exists $\eta > 0$, such that for every $x \in \mathsf{H}$, $\|x - a\| < \eta$ implies $\mathsf{H}(x) \subset U$. It is said usc if it is usc at every point [6, Chap. 1.4].

An operator $\mathsf{A} : \mathsf{X} \rightrightarrows \mathsf{X}$ is monotone if $\forall x, x' \in \mathrm{dom}(\mathsf{A})$, $\forall y \in \mathsf{A}(x), \forall y' \in \mathsf{A}(x')$, it holds that $\langle y - y', x - x' \rangle \geq 0$. A proper monotone operator A is said maximal if its graph $\mathrm{gr}(\mathsf{A})$ is a maximal

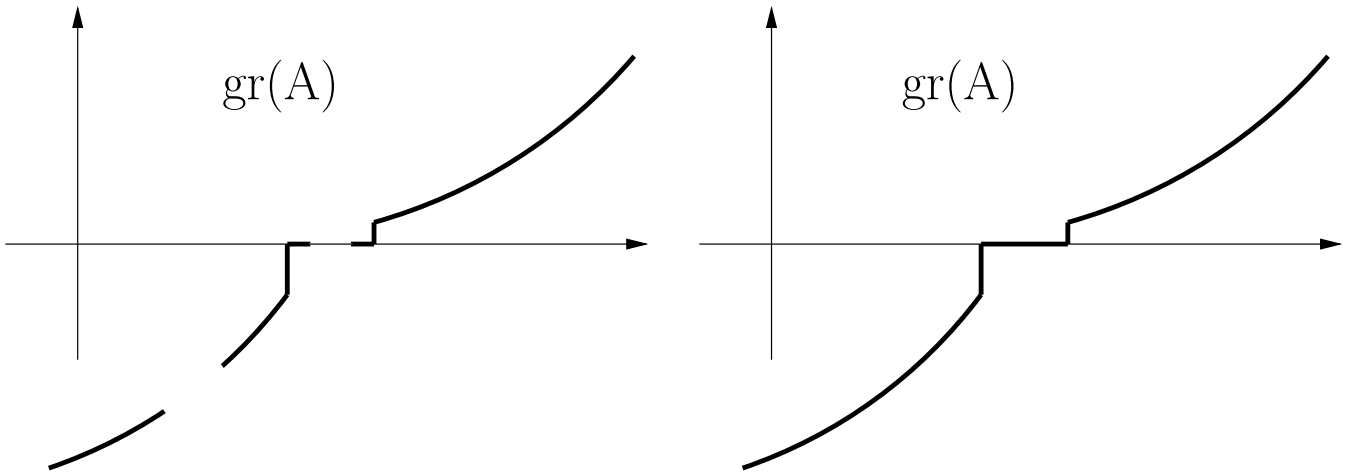

Figure 2.1: Left: A monotone operator over $\mathbb{R}$ which is not maximal. Right: A maximal monotone extension of the monotone operator

element in the inclusion ordering over $\mathsf{X} \times \mathsf{X}$ among graphs of monotone operators (see Fig. 2.1). Equivalently, the monotone operator A is maximal if the following property holds :

$$\forall (x,y) \in \mathsf{X}^2, \quad \forall (x',y') \in \mathrm{gr}(\mathsf{A}), \quad \langle y-y', x-x' \rangle \geq 0 \quad \Longrightarrow \quad (x,y) \in \mathrm{gr}(\mathsf{A}).$$

This maximality condition implies that $\mathrm{gr}(\mathsf{A})$ is a closed subset of $\mathsf{X} \times \mathsf{X}$. Denote by I the identity operator, and by A^{-1} the inverse of the operator A, defined by the fact that $y \in \mathsf{A}(x) \Leftrightarrow x \in \mathsf{A}^{-1}(y)$. Using the monotonicity of A, one can check that $\forall (x,y),(x',y') \in \mathrm{gr}(\mathsf{A}), \forall \gamma > 0, \|x-x'\| \leq \|(x+\gamma y) - (x'+\gamma y')\|$. In other words, if A is a monotone operator, then the resolvent operator defined for every $\gamma > 0$ by $\mathsf{J}_\gamma := (I+\gamma \mathsf{A})^{-1}$ can be identified with a 1-Lipschitz continuous function (*i.e* a contraction). It is well known that A belongs to the set $\mathscr{M}(\mathsf{X})$ of the maximal monotone operators on X if and only if, $\mathrm{dom}(\mathsf{J}_\gamma) = \mathsf{X}$ ([85]). If $\mathrm{dom}(\mathsf{A}) = \mathsf{X}$, then A is usc [97]. We also know that when $\mathsf{A} \in \mathscr{M}(\mathsf{X})$, the closure $\mathrm{cl}(\mathrm{dom}(\mathsf{A}))$ of $\mathrm{dom}(\mathsf{A})$ is convex, and $\lim_{\gamma \to 0} \mathsf{J}_\gamma(x) = \Pi_{\mathrm{cl}(\mathrm{dom}(\mathsf{A}))}(x)$, where Π_S is the projector on the closed convex set S. It holds that $\mathsf{A}(x)$ is closed and convex for all $x \in \mathrm{dom}(\mathsf{A})$. We can therefore put $\mathsf{A}_0(x) = \Pi_{\mathsf{A}(x)}(0)$, in other words, $\mathsf{A}_0(x)$ is the minimum norm element of $\mathsf{A}(x)$. Of importance is the so called Yosida regularization of A for $\gamma > 0$, defined as the single-valued operator $\mathsf{A}_\gamma = (I - \mathsf{J}_\gamma)/\gamma$. This is a $1/\gamma$-Lipschitz operator on X that satisfies $\mathsf{A}_\gamma(x) \to \mathsf{A}_0(x)$ and $\|\mathsf{A}_\gamma(x)\| \uparrow \|\mathsf{A}_0(x)\|$ for all $x \in \mathrm{dom}(\mathsf{A})$. One can also check that $\mathsf{A}_\gamma(x) \in \mathsf{A}(\mathsf{J}_\gamma(x))$ for all $x \in \mathsf{X}$.

A typical maximal monotone operator is the subdifferential ∂f of a function $f \in \Gamma_0(\mathsf{X})$, the set of proper, convex, and lower semicontinuous (lsc) functions on X. In this case, the resolvent $(I+\gamma \partial f)^{-1}$ for $\gamma > 0$ is the proximity operator of γf. The Yosida regularization of ∂f for $\gamma > 0$ coincides with the gradient of the so called Moreau's envelope $f_\gamma(x) := \min_y f(y) + \|y-x\|^2/(2\gamma)$ of f.

2.2.2 Differential Inclusions (DI)

We now turn to the differential inclusions induced by operators. First consider a set valued mapping $\mathsf{H} : \mathsf{X} \rightrightarrows \mathsf{X}$ and $a \in \mathsf{X}$, a solution to the Differential Inclusion (DI) $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$ on $\mathbb{R}_+$ starting at a is an absolutely continuous mapping $\mathsf{x} : \mathbb{R}_+ \to \mathsf{X}$ such that $\mathsf{x}(0) = a$, and $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$, where $\dot{\mathsf{x}}$ denotes the derivative of x defined almost everywhere. Consider the set valued mapping $\Phi : \mathsf{X} \rightrightarrows C(\mathbb{R}_+, \mathsf{X})$,

such that for every $a \in \mathsf{X}$, $\Phi(a)$ is the set of solutions to the DI starting at a. We refer to Φ as the evolution system induced by H. For every subset $S \subset \mathsf{X}$, we define $\Phi(S) = \bigcup_{x \in S} \Phi(x)$.

In the case where H is usc with nonempty compact convex values and satisfies the linear growth condition

$$\exists c > 0,\ \forall x \in \mathsf{X},\ \sup\{\|y\| \ : y \in \mathsf{H}(x)\} \leq c(1 + \|x\|)\ , \tag{2.1}$$

then, $\mathrm{dom}(\Phi) = \mathsf{X}$, see *e.g.* [6], and moreover, $\Phi(\mathsf{X})$ is closed in the space $C(\mathbb{R}_+, \mathsf{X})$ endowed with the topology of uniform convergence over compact sets of $\mathbb{R}_+$.

Assume from now on that $\mathsf{H} = -\mathsf{A} \in \mathscr{M}(\mathsf{X})$. Then for every $a \in \mathrm{dom}(\mathsf{A})$, $\Phi(a)$ contains exactly one function, still denoted $\Phi(a)$ [36]. Note that in the case where $\mathsf{A} = \nabla F, F \in \Gamma_0(\mathsf{X})$, the DI boils down to the so-called gradient flow. In the sequel, we denote $\Phi(a,t) = \Phi(a)(t)$. For every $t \geq 0$, the map $\Phi(\cdot, t) : \mathrm{dom}(\mathsf{A}) \to \mathrm{dom}(\mathsf{A})$ is a contraction and can be uniquely extended to a contraction from $\mathrm{cl}(\mathrm{dom}(\mathsf{A}))$ to $\mathrm{cl}(\mathrm{dom}(\mathsf{A}))$. Denoting $\Phi(\cdot, t)$ this extension, Φ becomes a semiflow on $\mathrm{cl}(\mathrm{dom}(\mathsf{A})) \times \mathbb{R}_+$, being a continuous $\mathrm{cl}(\mathrm{dom}(\mathsf{A})) \times \mathbb{R}_+ \to \mathrm{cl}(\mathrm{dom}(\mathsf{A}))$ function satisfying

$$\Phi(\cdot, 0) = I \quad \text{and} \quad \Phi(x, t+s) = \Phi(\Phi(x,s), t) \tag{2.2}$$

for each $x \in \mathrm{cl}(\mathrm{dom}(\mathsf{A}))$, and $t, s \geq 0$.

The set of zeros $Z(\mathsf{A}) := \{x \in \mathrm{dom}(\mathsf{A}) \ : \ 0 \in \mathsf{A}(x)\}$ of A is a closed convex set which coincides with the set of equilibrium points $\{x \in \mathrm{cl}(\mathrm{dom}(\mathsf{A})) \ : \ \forall t \geq 0, \Phi(x,t) = x\}$ of Φ. The trajectories $\Phi(x, \cdot)$ of the semiflow do not necessarily converge to $Z(\mathsf{A})$. A counterexample is given by the linear maximal monotone operator A defined on $\mathbb{R}^2$ by $A(y,z) = (z, -y)$ whose set of zeros is $Z(A) = (0,0)$. The DI associated to A boils down to a linear differential equation in $\mathbb{R}^2$, $(\dot{\mathsf{y}}(t), \dot{\mathsf{z}}(t)) = (-\mathsf{z}(t), \mathsf{y}(t))$ for every $t \geq 0$. To solve this equation, consider $i \in \mathbb{C}$ and denote $\mathsf{x}(t) = \mathsf{y}(t) + i\mathsf{z}(t) \in \mathbb{C}$. The DI is equivalent to $\dot{\mathsf{x}}(t) = i\mathsf{x}(t)$ whose solutions can be written $\mathsf{x}(t) = a\exp(it), a \in \mathbb{C}$. Finally, x does not converge to zero as $t \to +\infty$ in general. However, the ergodic theorem for the semiflows generated by the elements of $\mathscr{M}(\mathsf{X})$ states that if $Z(\mathsf{A}) \neq \emptyset$, then for each $x \in \mathrm{cl}(\mathrm{dom}(\mathsf{A}))$, the averaged function

$$\begin{array}{rccc} \overline{\Phi} & : & \mathrm{cl}(\mathrm{dom}(\mathsf{A})) \times \mathbb{R}_+ & \longrightarrow & \mathrm{cl}(\mathrm{dom}(\mathsf{A})) \\ & & (x,t) & \longmapsto & \dfrac{1}{t}\displaystyle\int_0^t \Phi(x,s)\,ds \end{array}$$

(with $\overline{\Phi(\cdot,0)} = \Phi(\cdot,0)$), converges to an element of $Z(\mathsf{A})$ as $t \to \infty$. The convergence of the trajectories of the semiflow itself to an element of $Z(\mathsf{A})$ is ensured when A is demipositive [38]. An operator $\mathsf{A} \in \mathscr{M}(\mathsf{X})$ is said demipositive if there exists $w \in Z(\mathsf{A})$ such that for every sequence $((u_n, v_n) \in \mathrm{gr}(\mathsf{A}))$ such that (u_n) converges to u, and such that (v_n) is bounded,

$$\langle u_n - w, v_n \rangle \xrightarrow[n\to\infty]{} 0 \quad \Rightarrow \quad u \in Z(\mathsf{A}).$$

Under this condition and if $Z(\mathsf{A}) \neq \emptyset$, then for all $x \in \mathrm{cl}(\mathrm{dom}(\mathsf{A}))$, $\Phi(x,t)$ converges as $t \to \infty$ to an element of $Z(\mathsf{A})$.

2.3 Random monotone operators

A sequence of elements $(\mathsf{A}_n)_{n\in\mathbb{N}}$ in $\mathscr{M}(\mathsf{X})$ is said to converge to an element $\mathsf{A} \in \mathscr{M}(\mathsf{X})$ if for every $\gamma > 0, x \in \mathsf{X}$, $(I + \gamma\mathsf{A}_n)^{-1}(x) \to (I + \gamma\mathsf{A})^{-1}(x)$ as $n \to +\infty$. Endowed with this topology, $\mathscr{M}(\mathsf{X})$ is a Polish space [4]. Moreover, the subset of all subdifferentials over X is closed. A random monotone operator is defined to be a random variable A from a probability space $(\Xi, \mathscr{G}, \mu)$ to $(\mathscr{M}(\mathsf{X}), \mathscr{B}(\mathscr{M}(\mathsf{X})))$.

Let $F : \Xi \rightrightarrows \mathsf{X}$ be a set valued function such that $F(s)$ is a closed set for each $s \in \Xi$. The function F is said *measurable* if $\{s \,:\, F(s) \cap H \neq \emptyset\} \in \mathscr{G}$ for any set $H \in \mathscr{B}(\mathsf{X})$. An equivalent definition for the mesurability of F requires that the domain $\mathrm{dom}(F) := \{s \in \Xi \,:\, F(s) \neq \emptyset\}$ of F belongs to $\mathscr{G}$, and that there exists a sequence of measurable functions $\varphi_n : \mathrm{dom}(F) \to \mathsf{X}$ such that $F(s) = \mathrm{cl}\,\{\varphi_n(s)\}_n$ for all $s \in \mathrm{dom}(F)$ [40, Chap. 3][65]. These functions are called measurable selections of F. Consider a function $A : (\Xi, \mathscr{G}, \mu) \to (\mathscr{M}(\mathsf{X}), \mathscr{B}(\mathscr{M}(\mathsf{X})))$. For every $s \in \Xi, \gamma > 0, x \in \mathsf{X}$, $(I + \gamma A(s))^{-1}(x)$ is denoted $J_\gamma(s, x)$. There is equivalence between [4]

1. A is a random monotone operator
2. $s \mapsto \mathrm{gr}(A(s))$ is measurable as a closed set-valued $\Xi \rightrightarrows \mathsf{X} \times \mathsf{X}$ function
3. For every $\gamma > 0, x \in \mathsf{X}$, the function $s \mapsto J_\gamma(s, x)$ from $(\Xi, \mathscr{G})$ to $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$ is measurable.

Example 2 (Random subdifferential)**.** Consider a function $g : \Xi \times \mathsf{X} \to (-\infty, \infty]$. The function g is said a convex normal integrand [125] if $g(s, \cdot)$ is convex, and if the set-valued mapping $s \mapsto \mathrm{epi}\, g(s, \cdot)$ is closed-valued and measurable, where epi is the epigraph of a function. Then, the function $s \mapsto \partial g(s, \cdot)$ is an example of random monotone operator [4].

Assume now that $F : \Xi \rightrightarrows \mathsf{X}$ is measurable and that $\mu(\mathrm{dom}(F)) = 1$. Consider the set

$$\mathfrak{S}_F^p := \{\varphi \in \mathcal{L}^p(\Xi, \mathscr{G}, \mu; \mathsf{X}) \,:\, \varphi(s) \in F(s)\ \mu - \text{a.e.}\}\,. \tag{2.3}$$

of $\mathcal{L}^p$ integrable selection of F. If $\mathfrak{S}_F^1 \neq \emptyset$, the function F is said integrable. The *selection integral* [86] of F is the set

$$\int F d\mu := \mathrm{cl}\left\{\int_\Xi \varphi d\mu \;:\; \varphi \in \mathfrak{S}_F^1\right\}. \tag{2.4}$$

For a random monotone operator $A : \Xi \to \mathscr{M}(\mathsf{X})$, since $J_\gamma(s, x)$ is measurable in s and continuous in x (being non expansive), $J_\gamma : \Xi \times \mathsf{X} \to \mathsf{X}$ is $\mathscr{G} \otimes \mathscr{B}(\mathsf{X})/\mathscr{B}(\mathsf{X})$ measurable by Carathéodory's theorem. Denoting $A_\gamma(s, x)$ the image of x by the Yosida regularization of the operator $A(s)$, this implies the measurability of $A_\gamma : \Xi \times \mathsf{X} \to \mathsf{X}$ for every $\gamma > 0$. Moreover, denoting by $D(s)$ the domain of $A(s)$, $s \mapsto \mathrm{cl}(D(s))$ is measurable, which implies that the function $s \mapsto d(x, D(s))$ is measurable for each $x \in \mathsf{X}$. Denoting as $A(s, x)$ the image of x by the operator $A(s)$, the set valued function $s \mapsto A(s, x)$ is also measurable. In particular, the function $s \mapsto A_0(s, x)$ is measurable for each $x \in \mathsf{X}$, where $A_0(s, x) := \Pi_{A(s,x)}(0)$. The essential intersection $\mathcal{D}$ of the domains $D(s)$ is defined as [67]

$$\mathcal{D} := \bigcup_{G \in \mathscr{G} : \mu(G) = 0} \bigcap_{s \in \Xi \setminus G} D(s)\,, \tag{2.5}$$

in other words, $x \in \mathcal{D} \;\Leftrightarrow\; \mu(\{s \,:\, x \in D(s)\}) = 1$. Let us assume that $\mathcal{D} \neq \emptyset$, and that the set-valued mapping $A(\cdot, x)$ is integrable for each $x \in \mathcal{D}$. For all $x \in \mathcal{D}$, we can define

$$\mathcal{A}(x) := \int_\Xi A(s, x)\, \mu(ds)\,.$$

We shall sometimes use the notation $\mathcal{A}(x) = \mathbb{E}_\xi(A(\xi, x))$ where ξ is a random variable with distribution μ. One can immediately see that the operator $\mathcal{A} : \mathcal{D} \rightrightarrows \mathsf{X}$ so defined is a monotone operator.

Example 3 (Interchange property)**.** Let $g : \Xi \times \mathsf{X} \to (-\infty, \infty]$ be a convex normal integrand, and let $\mathsf{G}(x) = \int g(s, x) \mu(ds)$, where the integral is defined as the sum

$$\int_{\{s\,:\,g(s,x) \in [0,\infty)\}} g(s, x)\, \mu(ds) + \int_{\{s\,:\,g(s,x) \in]-\infty, 0[\}} g(s, x)\, \mu(ds) + I(x)\,,$$

and

$$I(x) = \begin{cases} +\infty, & \text{if } \mu(\{s : g(s,x) = \infty\}) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

and where the convention $(+\infty) + (-\infty) = +\infty$ is used. The function G is a lower semi continuous convex function if $\mathsf{G}(x) > -\infty$ for all x [125], which we assume. Assume also that G is proper. Note that this implies that $g(s,\cdot) \in \Gamma_0(\mathsf{X})$ for μ-almost all s. We provide conditions under which the selection integral (set to $\emptyset$ for the values of x for which $\mathfrak{S}^1_{\partial g(\cdot,x)} = \emptyset$) $\int \partial g(s,x)\mu(ds)$ boils down to $\partial\mathsf{G}(x)$. We shall write that the interchange property holds since in this case,

$$\partial\mathsf{G}(x) = \int \partial g(s,x)\mu(ds).$$

First, this will be the case if $\int |g(s,x)|\,\mu(ds) < \infty$ for all $x \in \mathsf{X}$. By [102, page 179], this will also be the case if the following conditions hold: *i)* the set-valued mapping $s \mapsto \operatorname{cl}\operatorname{dom} g(s,\cdot)$ is constant μ-a.e., where $\operatorname{dom} g(s,\cdot)$ is the domain of $g(s,\cdot)$, *ii)* $\mathsf{G}(x) < \infty$ whenever $x \in \operatorname{dom} g(s,\cdot)$ μ-a.e., *iii)* there exists $x_0 \in \mathsf{X}$ at which G is finite and continuous. Another case where this interchange is permitted is the following. Let m be a positive integer, and let $\mathcal{C}_1, \ldots \mathcal{C}_m$ be a collection of closed and convex subsets of X. Let $\mathcal{C} = \cap_{i=k}^m \mathcal{C}_k \neq \emptyset$, and assume that the normal cone $N_{\mathcal{C}}(x)$ of $\mathcal{C}$ at x satisfies the identity $N_{\mathcal{C}}(x) = \sum_{k=1}^m N_{\mathcal{C}_k}(x)$ for each $x \in \mathsf{X}$, where the summation is the usual set summation. As is well known, this identity holds true under a qualification condition of the type $\cap_{k=1}^m \operatorname{ri}\mathcal{C}_k \neq \emptyset$ (see also [11] for other conditions). Now, assume that $\Xi = \{1, \ldots, m\}$ and that μ is an arbitrary probability measure putting a positive weight on each $\{k\} \subset \Xi$. Let $g(s,x)$ be the indicator function

$$g(s,x) = \iota_{\mathcal{C}_s}(x) \text{ for } (s,x) \in \Xi \times \mathsf{X}. \tag{2.6}$$

Then it is obvious that g is a convex normal integrand, $\mathsf{G} = \iota_{\mathcal{C}}$, and $\partial\mathsf{G}(x) = \int \partial g(s,x)\mu(ds)$. We can also combine these two types of conditions: let $(\Sigma, \mathscr{T}, \nu)$ be a probability space, where $\mathscr{T}$ is ν-complete, and let $h : \Sigma \times \mathsf{X} \to (-\infty, \infty]$ be a convex normal integrand satisfying the conditions *i)–iii)* above. Consider the closed and convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_m$ introduced above, and let α be a probability measure on the set $\{0, \ldots, m\}$ such that $\alpha(\{k\}) > 0$ for each $k \in \{0, \ldots, m\}$. Now, set $\Xi = \Sigma \times \{0, \ldots, m\}$, $\mu = \nu \otimes \alpha$, and define $g : \Xi \times \mathsf{X} \to (-\infty, \infty]$ as

$$g(s,x) = \begin{cases} \alpha(0)^{-1} h(u,x) & \text{if } k = 0, \\ \iota_{\mathcal{C}_k}(x) & \text{otherwise}, \end{cases}$$

where $s = (u,k) \in \Sigma \times \{0, \ldots, m\}$. Then it is clear that

$$\mathsf{G}(x) = \frac{1}{\alpha(0)} \int_\Sigma h(u,x)\nu(du) + \iota_{\mathcal{C}}(x)\,,$$

and

$$\partial\mathsf{G}(x) = \int \partial g(s,x)\mu(ds) = \frac{1}{\alpha(0)} \mathbb{E}_\nu \partial h(\cdot,x) + \sum_{k=1}^m N_{\mathcal{C}_k}(x)\,.$$

Part I

Stochastic approximation with a constant step size

Chapter 3

Constant Step Stochastic Approximations Involving Differential Inclusions: Stability, Long-Run Convergence and Applications

The purpose of this chapter is to study the constant step stochastic approximation framework (1.4) in the case where the underlying Differential Inclusion is induced by an upper semicontinuous operator, see Sec. 1.4. We consider a Markov chain (x_n) whose kernel is indexed by a scaling parameter $\gamma > 0$, referred to as the step size. The aim is to analyze the behavior of the Markov chain in the doubly asymptotic regime where $n \to \infty$ then $\gamma \to 0$. First, under mild assumptions on the so-called drift of the Markov chain, we show that the interpolated process converges narrowly to the solutions of a DI involving an usc set-valued map with closed and convex values. Second, we provide verifiable conditions which ensure the stability of the iterates. Third, by putting the above results together, we establish the long run convergence of the iterates (x_n) as $\gamma \to 0$, to the Birkhoff center of the DI. The ergodic behavior of the iterates is also provided. Our findings are applied to the problem of nonconvex proximal stochastic optimization and a fluid model of parallel queues.

3.1 Introduction

In this chapter, we consider a Markov chain $(x_n, n \in \mathbb{N})$ with values in an Euclidean space X. We assume that the probability transition kernel P_γ is indexed by a scaling factor γ, which belongs to some interval $(0, \gamma_0)$. The aim of the chapter is to analyze the long term behavior of the Markov chain in the regime where γ is small. The map

$$g_\gamma(x) := \int \frac{y - x}{\gamma} P_\gamma(x, dy)\,, \tag{3.1}$$

assumed well defined for all $x \in \mathsf{X}$, is called the *drift* or the *mean field*. The Markov chain admits the representation

$$x_{n+1} = x_n + \gamma\, g_\gamma(x_n) + \gamma\, U_{n+1}\,, \tag{3.2}$$

where U_{n+1} is a zero-mean martingale increment noise *i.e.,* the conditional expectation of U_{n+1} given the past samples is equal to zero. A case of interest in the chapter is given by iterative models of the form:

$$x_{n+1} = x_n + \gamma\, h_\gamma(\xi_{n+1}, x_n)\,, \tag{3.3}$$

where $(\xi_n, n \in \mathbb{N}^*)$ is a sequence of i.i.d random variables indexed by the set $\mathbb{N}^*$ of positive integers and defined on a probability space Ξ with probability law μ, and $\{h_\gamma\}_{\gamma\in(0,\gamma_0)}$ is a family of maps on $\Xi \times \mathsf{X} \to \mathsf{X}$. In this case, the drift g_γ has the form:

$$g_\gamma(x) = \int h_\gamma(s,x)\,\mu(ds)\,. \tag{3.4}$$

Our results are as follows.

1. **Dynamical behavior.** Assume that the drift g_γ has the form (3.4). Assume that for μ-almost all s and for every sequence $((\gamma_k, z_k) \in (0,\gamma_0) \times \mathsf{X}, k \in \mathbb{N})$ converging to $(0,z)$,

$$h_{\gamma_k}(s,z_k) \to H(s,z)$$

where $H(s,z)$ is a subset of X (the Euclidean distance between $h_{\gamma_k}(s,z_k)$ and the set $H(s,z)$ tends to zero as $k \to \infty$). Denote by $\mathsf{x}^\gamma(t)$ the continuous-time stochastic process obtained by a piecewise linear interpolation of the sequence x_n, where the points x_n are spaced by a fixed time step γ on the positive real axis. As $\gamma \to 0$, and assuming that $H(s,\cdot)$ is a proper and upper semicontinuous (usc) (see Sec. 2.2.1) map with closed convex values, we prove that x^γ converges narrowly (in the topology of uniform convergence on compact sets) to the set of solutions of the differential inclusion (DI)

$$\dot{\mathsf{x}}(t) \in \int H(s,\mathsf{x}(t))\mu(ds)\,, \tag{3.5}$$

where for every $x \in \mathsf{X}$, $\int H(s,x)\mu(ds)$ is the *selection integral* of $H(\cdot,x)$, see Sec. 2.3.

2. **Tightness.** As the iterates are not *a priori* supposed to be in a compact subset of X, we investigate the issue of stability. We posit a verifiable *Pakes-Has'minskii* condition on the Markov chain (x_n). The condition ensures that the iterates are stable in the sense that the random occupation measures

$$\Lambda_n := \frac{1}{n+1}\sum_{k=0}^{n} \delta_{x_k} \qquad (n \in \mathbb{N})$$

(where δ_a stands for the Dirac measure at point a), form a tight family of random variables on the Polish space of probability measures equipped with the Lévy-Prokhorov distance. The same criterion allows to establish the existence of invariant measures of the kernels P_γ, and the tightness of the family of all invariant measures, for all $\gamma \in (0,\gamma_0)$. As a consequence of Prokhorov's theorem, these invariant measures admit cluster points as $\gamma \to 0$. Under a Feller assumption on the kernel P_γ, we prove that every such cluster point is an invariant measure for the DI (3.5). Here, since the flow generated by the DI is in general set-valued, the notion of invariant measure is borrowed from [59].

3. **Long-run convergence.** Using the above results, we investigate the behavior of the iterates in the asymptotic regime where $n \to \infty$ and, next, $\gamma \to 0$. Denoting by $d(a,B)$ the distance between a point $a \in \mathsf{X}$ and a subset $B \subset \mathsf{X}$, we prove that for all $\varepsilon > 0$,

$$\lim_{\gamma\to 0} \limsup_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{P}\left(d(x_k, \mathsf{BC}_\Phi) > \varepsilon\right) = 0\,, \tag{3.6}$$

where BC is the Birkhoff center of the flow Φ induced by the DI (3.5), and $\mathbb{P}$ stands for the probability. We also characterize the ergodic behavior of these iterates. Setting $\overline{x}_n = \frac{1}{n+1}\sum_{k=0}^{n} x_k$, we prove that

$$\lim_{\gamma\to 0} \limsup_{n\to\infty} \mathbb{P}\left(d(\overline{x}_n, \mathrm{co}(L_{\mathrm{av}})) > \varepsilon\right) = 0\,, \tag{3.7}$$

where $\mathrm{co}(L_{\mathrm{av}})$ is the convex hull of the limit set of the averaged flow associated with (3.5) (see Sec. 3.4).

4. **Applications.** We investigate several application scenarios. First, we consider the problem of non-convex stochastic optimization, and analyze the convergence of a constant step size proximal stochastic gradient algorithm. The latter finds application in the optimization of deep neural networks [77]. We show that the interpolated process converges narrowly to a DI, which we characterize. We also provide sufficient conditions allowing to characterize the long-run behavior of the algorithm leading to a convergence proof of the proximal stochastic non-convex gradient algorithm ([98]). Second, we explain that our results apply to the characterization of the fluid limit of a system of parallel queues. The model is introduced in [8, 61]. Whereas the narrow convergence of the interpolated process was studied in [61], less is known about the stability and the long-run convergence of the iterates. We show how our results can be used to address this problem.

Chapter organization. In Sec. 3.2, we introduce the application examples. In Sec. 3.3, we briefly discuss the literature. Sec. 3.4 is devoted to the mathematical background and to the notations. The main results are given in Sec. 3.5. The tightness of the interpolated process as well as its narrow convergence towards the solution set of the DI (Th. 3.5.1) are proven in Sec. 3.6. Turning to the Markov chain characterization, Prop. 3.5.2, who explores the relations between the cluster points of the Markov chains invariant measures and the invariant measures of the flow induced by the DI, is proven in Sec. 3.7. A general result describing the asymptotic behavior of a functional of the iterates with a prescribed growth is provided by Th. 3.5.3, and proven in Sec. 3.8. Finally, in Sec. 3.9, we show how the results pertaining to the ergodic convergence and to the convergence of the iterates (Th. 3.5.4 and 3.5.5 respectively) can be deduced from Th. 3.5.3. Finally, Sec. 3.10 is devoted to the application examples. We prove that our hypotheses are satisfied.

3.2 Examples

Example 4. *Non-convex stochastic optimization.* Consider the problem

$$\text{minimize } \mathbb{E}_\xi(\ell(\xi, x)) + r(x) \text{ w.r.t } x \in \mathsf{X}, \tag{3.8}$$

where $\ell(\xi, \cdot)$ is a (possibly non-convex) differentiable function on $\mathsf{X} \to \mathbb{R}$ indexed by a random variable (r.v.) ξ, $\mathbb{E}_\xi$ represents the expectation w.r.t. ξ, and $r : \mathsf{X} \to \mathbb{R}$ is a convex function. The problem typically arises in deep neural networks [129, 115]. In the latter case, x represents the collection of weights of the network, ξ represents a random training example of the database, and $\ell(\xi, x)$ is a risk function which quantifies the inadequacy between the sample response and the network response. Here, $r(x)$ is a regularization term which prevents the occurence of undesired solutions. A typical regularizer used in machine learning is the ℓ_1-norm $\|x\|_1$ that promotes sparsity or generalizations like $\|Dx\|_1$, where

D is a matrix, that promote structured sparsity. A popular algorithm used to find an approximate solution to Problem (3.8) is the proximal stochastic gradient algorithm, which reads

$$x_{n+1} = \mathrm{prox}_{\gamma r}(x_n - \gamma \nabla \ell(\xi_{n+1}, x_n)) \,, \tag{3.9}$$

where $(\xi_n, n \in \mathbb{N}^*)$ are i.i.d. copies of the r.v. ξ, where ∇ represents the gradient w.r.t. parameter x, and where the proximity operator of r is the mapping on $\mathsf{X} \to \mathsf{X}$ defined by

$$\mathrm{prox}_{\gamma r} : x \mapsto \arg\min_{y \in \mathsf{X}} \left(\gamma\, r(y) + \frac{\|y - x\|^2}{2} \right) .$$

The drift g_γ has the form (3.4) where $h_\gamma(\xi, x) = \gamma^{-1}(\mathrm{prox}_{\gamma r}(x - \gamma \nabla \ell(\xi, x)) - x)$ and μ represents the distribution of the r.v. ξ. Under adequate hypotheses, we prove that the interpolated process converges narrowly to the solutions to the DI

$$\dot{\mathsf{x}}(t) \in -\nabla_x \mathbb{E}_\xi(\ell(\xi, \mathsf{x}(t))) - \partial r(\mathsf{x}(t)) \,,$$

where ∂r represents the subdifferential of a function r, defined by

$$\partial r(x) := \{u \in \mathsf{X} \,:\, \forall y \in \mathsf{X},\, r(y) \geq r(x) + \langle u, y - x \rangle\}$$

at every point $x \in \mathsf{X}$. We provide a sufficient condition under which the iterates (3.9) satisfy the Pakes-Has'minskii criterion, which in turn, allows to characterize the long-run behavior of the iterates.

Example 5. *Fluid limit of a system of parallel queues with priority.* We consider a time slotted queuing system composed of N queues. The following model is inspired from [8, 61]. We denote by y_n^k the number of users in the queue k at time n. We assume that a random number of $A_{n+1}^k \in \mathbb{N}$ users arrive in the queue k at time $n+1$. The queues are prioritized: the users of Queue k can only be served if all users of Queues ℓ for $\ell < k$ have been served. Whenever the queue k is non-empty and the queues ℓ are empty for all $\ell < k$, one user leaves Queue k with probability $\eta_k > 0$. Starting with $y_0^k \in \mathbb{N}$, we thus have

$$y_{n+1}^k = y_n^k + A_{n+1}^k - B_{n+1}^k \mathbb{1}_{\{y_n^k > 0,\, y_n^{k-1} = \cdots = y_n^1 = 0\}} \,,$$

where B_{n+1}^k is a Bernoulli r.v. with parameter η_k, and where $\mathbb{1}_S$ denotes the indicator of an event S, equal to one on that set and to zero otherwise. We assume that the process $((A_n^1, \ldots, A_n^N, B_n^1, \ldots, B_n^N), n \in \mathbb{N}^*)$ is iid, and that the random variables A_n^k have finite second moments. We denote by $\lambda_k := \mathbb{E}(A_n^k) > 0$ the arrival rate in Queue k. Given a scaling parameter $\gamma > 0$ which is assumed to be small, we are interested in the *fluid-scaled process*, defined as $x_n^k = \gamma y_n^k$. This process is subject to the dynamics:

$$x_{n+1}^k = x_n^k + \gamma\, A_{n+1}^k - \gamma\, B_{n+1}^k \mathbb{1}_{\{x_n^k > 0,\, x_n^{k-1} = \cdots = x_n^1 = 0\}} \,. \tag{3.10}$$

The Markov chain $x_n = (x_n^1, \ldots, x_n^N)$ admits the representation (3.2), where the drift g_γ is defined on $\gamma \mathbb{N}^N$, and is such that its k-th component $g_\gamma^k(x)$ is

$$g_\gamma^k(x) = \lambda_k - \eta_k \mathbb{1}_{\{x^k > 0,\, x^{k-1} = \cdots = x^1 = 0\}} \,, \tag{3.11}$$

for every $k \in \{1, \ldots, N\}$ and every $x = (x^1, \ldots, x^N) \in \gamma \mathbb{N}^N$. Introduce the vector

$$\boldsymbol{u}_k := (\lambda_1, \cdots, \lambda_{k-1}, \lambda_k - \eta_k, \lambda_{k+1}, \ldots, \lambda_N)$$

for all k. Let $\mathbb{R}_+ := [0, +\infty)$, and define the set-valued map on $\mathbb{R}_+^N$

$$\mathsf{H}(x) := \begin{cases} \boldsymbol{u}_1 & \text{if } x^{(1)} > 0 \\ \mathrm{co}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) & \text{if } x^1 = \cdots = x^{k-1} = 0 \text{ and } x^k > 0\,, \end{cases} \tag{3.12}$$

where co is the convex hull. Clearly, $g_\gamma(x) \in \mathsf{H}(x)$ for every $x \in \gamma\mathbb{N}^N$. In [61, § 3.2], it is shown that the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$ has a unique solution. Our results imply the narrow convergence of the interpolated process to this solution, hence recovering a result of [61]. More importantly, if the following stability condition

$$\sum_{k=1}^{N} \frac{\lambda_k}{\eta_k} < 1 \tag{3.13}$$

holds, our approach allows to establish the tightness of the occupation measure of the iterates x_n, and to characterize the long-run behavior of these iterates. We prove that in the long-run, the sequence (x_n) converges to zero in the sense of (3.6). The ergodic convergence in the sense of (3.7) can be also established with a small extra effort.

3.3 About the Literature

When the drift g_γ does not depend on γ and is supposed to be a Lispchitz continuous map, the long term behavior of the iterates x_n in the small step size regime has been studied in the treatises [20, 14, 73, 31, 16] among others. In particular, narrow convergence of the interpolated process to the solution of an Ordinary Differential Eq. (ODE) is established. The authors of [60] introduce a Pakes-Has'minskii criterion to study the long-run behavior of the iterates.

The recent interest in the stochastic approximation when the ODE is replaced with a differential inclusion dates back to [17], where decreasing steps were considered. A similar setting is considered in [58]. A Markov noise was considered in the recent manuscript [128]. We also mention [59], where the ergodic convergence is studied when the so called weak asymptotic pseudo trajectory property is satisfied. The case where the DI is built from maximal monotone operators is studied in [22] and [24].

Differential inclusions arise in many applications, which include game theory (see [17, 18], [107] and the references therein), convex optimization [24], queuing theory or wireless communications, where stochastic approximation algorithms with non continuous drifts are frequently used, and can be modelled by differential inclusions [61].

Differential inclusions with a constant step were studied in [107]. The paper [107] extends previous results of [19] to the case of a DI. The key result established in [107] is that the cluster points of the collection of invariant measures of the Markov chain are invariant for the flow associated with the DI. Prop. 3.5.2 of the present chapter restates this result in a more general setting and using a shorter proof, which we believe to have its own interest. Moreover, the so-called GASP model studied by [107] does not cover certain applications, such as the ones provided in Sec. 3.2, for instance. In addition, [107] focusses on the case where the space is compact, which circumvents the issue of stability and simplifies the mathematical arguments. However, in many situations, the compactness assumption does not hold, and sufficient conditions for stability need to be formulated. Finally, we characterize the asymptotic behavior of the iterates (x_n) (as well as their Cesarò means) in the doubly asymptotic regime where $n \to \infty$ then $\gamma \to 0$. Such results are not present in [107].

3.4 Background

The space $C(\mathbb{R}_+, \mathsf{X})$ is endowed with the topology of uniform convergence on compact sets which is metrized by the distance d defined for every $\mathsf{x}, \mathsf{y} \in C(\mathbb{R}_+, \mathsf{X})$ by

$$\mathsf{d}(\mathsf{x}, \mathsf{y}) := \sum_{n \in \mathbb{N}} 2^{-n} \left(1 \wedge \sup_{t \in [0,n]} \|\mathsf{x}(t) - \mathsf{y}(t)\| \right), \tag{3.14}$$

where $\|\cdot\|$ denotes the Euclidean norm in X.

3.4.1 Random Probability Measures

The support $\mathrm{supp}(\mu)$ of a probability measure $\mu \in \mathcal{M}(\mathsf{X})$ is the smallest closed set G such that $\mu(G) = 1$. The set $\mathcal{M}(\mathsf{X})$ is endowed with the topology of narrow convergence: a sequence $(\mu_n)_{n \in \mathbb{N}}$ on $\mathcal{M}(\mathsf{X})$ converges to a measure $\mu \in \mathcal{M}(\mathsf{X})$ (denoted $\mu_n \Rightarrow \mu$) if for every $f \in C_b(\mathsf{X})$, $\mu_n(f) \to \mu(f)$, where $\mu(f)$ is a shorthand for $\int f(x)\mu(dx)$. Endowed with this topology, $\mathcal{M}(\mathsf{X})$ is metrizable by the Lévy-Prokhorov distance and is a Polish space. Moreover, for every nonnegative measurable (resp. bounded measurable) function $f : (\mathsf{X}, \mathscr{B}(\mathsf{X})) \to (\mathsf{X}, \mathscr{B}(\mathsf{X}))$, $\mu \mapsto \mu(f)$ is measurable from $(\mathcal{M}(\mathsf{X}), \mathscr{B}(\mathcal{M}(\mathsf{X})))$ to $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$. A subset $\mathcal{G}$ of $\mathcal{M}(\mathsf{X})$ is said tight if for every $\varepsilon > 0$, there exists a compact subset K of X such that for all $\mu \in \mathcal{G}$, $\mu(K) > 1 - \varepsilon$. We shall often say that a family of random variable is tight instead of saying that the family of their distributions is tight. Prokhorov's theorem gives a practical criterion for relative compactness of probability measures : $\mathcal{G}$ is tight iff it is a relatively compact subset of $\mathcal{M}(\mathsf{X})$.

We denote by δ_a the Dirac measure at the point $a \in \mathsf{X}$. If X is a random variable on some measurable space $(\Omega, \mathscr{F})$ into $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$, we denote by $\delta_X : \Omega \to \mathcal{M}(\mathsf{X})$ the measurable mapping defined by $\delta_X(\omega) = \delta_{X(\omega)}$. If $\Lambda : (\Omega, \mathscr{F}) \to (\mathcal{M}(\mathsf{X}), \mathscr{B}(\mathcal{M}(\mathsf{X})))$ is a random variable on the set of probability measures, we denote by $\mathbb{E}\Lambda$ the probability measure defined by $(\mathbb{E}\Lambda)(f) := \mathbb{E}(\Lambda(f))$, for every $f \in C_b(\mathsf{X})$.

3.4.2 Invariant Measures of Set-Valued Evolution Systems

The shift operator $\Theta : C(\mathbb{R}_+, \mathsf{X}) \to C(\mathbb{R}_+, C(\mathbb{R}_+, \mathsf{X}))$ is defined by : for every $\mathsf{x} \in C(\mathbb{R}_+, \mathsf{X})$, $\Theta(\mathsf{x}) : t \mapsto \mathsf{x}(t + \cdot)$. Consider a set-valued mapping $\Phi : \mathsf{X} \rightrightarrows C(\mathbb{R}_+, \mathsf{X})$. When Φ is single valued (*i.e.*, for all $a \in \mathsf{X}$, $\Phi(a)$ is a continuous function), a measure $\pi \in \mathcal{M}(\mathsf{X})$ is called an *invariant measure* for Φ, or Φ-invariant, if for all $t > 0$, $\pi = \pi\Phi(\cdot, t)^{-1}$, where $\Phi(a, t)$ denotes $\Phi(a)(t)$. For all $t \geq 0$, we define the projection $p_t : C(\mathbb{R}_+, \mathsf{X}) \to \mathsf{X}$ by $p_t(\mathsf{x}) = \mathsf{x}(t)$.

The definition can be extended as follows to the case where Φ is set-valued.

Definition 3.4.1. A probability measure $\pi \in \mathcal{M}(\mathsf{X})$ is said invariant for Φ if there exists $\upsilon \in \mathcal{M}(C(\mathbb{R}_+, \mathsf{X}))$ s.t.

(i) $\mathrm{supp}(\upsilon) \subset \mathrm{cl}(\Phi(\mathsf{X}))$;

(ii) υ is Θ-invariant ;

(iii) $\upsilon p_0^{-1} = \pi$.

When Φ is single valued, both definitions coincide. The above definition is borrowed from [59] (see also [84]). Note that $\mathrm{cl}(\Phi(\mathsf{X}))$ can be replaced by $\Phi(\mathsf{X})$ whenever the latter set is closed (sufficient conditions for this have been provided above).

The limit set of a function $\mathsf{x} \in C(\mathbb{R}_+, \mathsf{X})$ is defined as

$$L_{\mathsf{x}} := \bigcap_{t \geq 0} \mathrm{cl}(\mathsf{x}([t, +\infty))) \,.$$

It coincides with the set of points of the form $\lim_n \mathsf{x}(t_n)$ for some sequence $t_n \to \infty$. Consider now a set valued mapping $\Phi : \mathsf{X} \rightrightarrows C(\mathbb{R}_+, \mathsf{X})$. The limit set $L_{\Phi(a)}$ of a point $a \in \mathsf{X}$ for Φ is

$$L_{\Phi(a)} := \bigcup_{\mathsf{x} \in \Phi(a)} L_{\mathsf{x}} \,,$$

and $L_\Phi := \bigcup_{a \in \mathsf{X}} L_{\Phi(a)}$. A point a is said recurrent for Φ if $a \in L_{\Phi(a)}$. The Birkhoff center of Φ is the closure of the set of recurrent points

$$\mathsf{BC}_\Phi := \mathrm{cl}\{a \in \mathsf{X} \,:\, a \in L_{\Phi(a)}\} \,.$$

The following result, established in [59] (see also [7]), is a consequence of the celebrated recurrence theorem of Poincaré.

Proposition 3.4.1. Let $\Phi : \mathsf{X} \rightrightarrows C(\mathbb{R}_+, \mathsf{X})$. Assume that $\Phi(\mathsf{X})$ is closed. Let $\pi \in \mathcal{M}(\mathsf{X})$ be an invariant measure for Φ. Then, $\pi(\mathsf{BC}_\Phi) = 1$.

We denote by $\mathcal{I}(\Phi)$ the subset of $\mathcal{M}(\mathsf{X})$ formed by all invariant measures for Φ. We define

$$\mathscr{I}(\Phi) := \{\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X})) \,:\, \forall A \in \mathscr{B}(\mathcal{M}(\mathsf{X})), \mathcal{I}(\Phi) \subset A \,\Rightarrow\, \mathfrak{m}(A) = 1\} \,.$$

We define the mapping $\mathrm{av} : C(\mathbb{R}_+, \mathsf{X}) \to C(\mathbb{R}_+, \mathsf{X})$ by

$$\mathrm{av}(\mathsf{x}) : t \mapsto \frac{1}{t} \int_0^t \mathsf{x}(s)\, ds \,,$$

and $\mathrm{av}(\mathsf{x})(0) = \mathsf{x}(0)$. Finally, we define the average flow $\mathrm{av}(\Phi) : \mathsf{X} \rightrightarrows C(\mathbb{R}_+, \mathsf{X})$ by $\mathrm{av}(\Phi)(a) = \{\mathrm{av}(\mathsf{x}) : \mathsf{x} \in \Phi(a)\}$ for each $a \in \mathsf{X}$.

3.5 Main Results

3.5.1 Dynamical Behavior

Choose $\gamma_0 > 0$. For every $\gamma \in (0, \gamma_0)$, we introduce a probability transition kernel P_γ on $\mathsf{X} \times \mathscr{B}(\mathsf{X}) \to [0, 1]$.

Let $(\Xi, \mathscr{G}, \mu)$ be an arbitrary probability space.

Assumption (RM). There exist a $\mathscr{G} \otimes \mathscr{B}(\mathsf{X})/\mathscr{B}(\mathsf{X})$-measurable map $h_\gamma : \Xi \times \mathsf{X} \to \mathsf{X}$ and $H : \Xi \times \mathsf{X} \rightrightarrows \mathsf{X}$ such that:

i) For every $x \in \mathsf{X}$,

$$\int \frac{y - x}{\gamma} P_\gamma(x, dy) = \int h_\gamma(s, x) \mu(ds) \,.$$

ii) For every s μ-a.e. and for every converging sequence $(u_n, \gamma_n) \to (u^\star, 0)$ on $\mathsf{X} \times (0, \gamma_0)$,

$$h_{\gamma_n}(s, u_n) \to H(s, u^\star)\,.$$

iii) For all s μ-a.e., $H(s, \cdot)$ is proper, usc, with closed convex values.

iv) For every $x \in \mathsf{X}$, $H(\cdot, x)$ is μ-integrable. We set $\mathsf{H}(x) := \int H(s, x)\, \mu(ds)$.

v) For every $T > 0$ and every compact set $K \subset \mathsf{X}$,

$$\sup\{\|\mathsf{x}(t)\| : t \in [0, T], \mathsf{x} \in \Phi(a), a \in K\} < \infty\,.$$

where Φ is the evolution system induced by H.

vi) For every compact set $K \subset \mathsf{X}$, there exists $\varepsilon_K > 0$ such that

$$\sup_{x \in K} \sup_{0 < \gamma < \gamma_0} \int \left\| \frac{y - x}{\gamma} \right\|^{1+\varepsilon_K} P_\gamma(x, dy) < \infty\,, \tag{3.15}$$

$$\sup_{x \in K} \sup_{0 < \gamma < \gamma_0} \int \|h_\gamma(s, x)\|^{1+\varepsilon_K}\, \mu(ds) < \infty\,. \tag{3.16}$$

Assumption i) implies that the drift has the form (3.1). As mentioned in the introduction, this is for instance useful in the case of iterative Markov models such as (3.3). Assumption v) requires implicitly that the set of solutions $\Phi(a)$ is non-empty for any value of a. It holds true if, *e.g.*, the linear growth condition (2.1) on H is satisfied.

On the canonical space $\Omega := \mathsf{X}^{\mathbb{N}}$ equipped with the σ-algebra $\mathscr{F} := \mathscr{B}(\mathsf{X})^{\otimes \mathbb{N}}$, we denote by $X : \Omega \to \mathsf{X}^{\mathbb{N}}$ the canonical process defined by $X_n(\omega) = \omega_n$ for every $\omega = (\omega_k, k \in \mathbb{N})$ and every $n \in \mathbb{N}$, where $X_n(\omega)$ is the n-th coordinate of $X(\omega)$. For every $\nu \in \mathcal{M}(\mathsf{X})$ and $\gamma \in (0, \gamma_0)$, we denote by $\mathbb{P}^{\nu,\gamma}$ the unique probability measure on $(\Omega, \mathscr{F})$ such that X is an homogeneous Markov chain with initial distribution ν and transition kernel P_γ. We denote by $\mathbb{E}^{\nu,\gamma}$ the corresponding expectation. When $\nu = \delta_a$ for some $a \in \mathsf{X}$, we shall prefer the notation $\mathbb{P}^{a,\gamma}$ to $\mathbb{P}^{\delta_a,\gamma}$.

For every $\gamma > 0$, we introduce the measurable map on $(\Omega, \mathscr{F}) \to (C(\mathbb{R}_+, \mathsf{X}), \mathscr{B}(C(\mathbb{R}_+, \mathsf{X})))$, such that for every $x = (x_n, n \in \mathbb{N})$ in Ω,

$$\mathsf{X}_\gamma(x)\ : t \mapsto x_{\lfloor \frac{t}{\gamma} \rfloor} + (t/\gamma - \lfloor t/\gamma \rfloor)(x_{\lfloor \frac{t}{\gamma} \rfloor + 1} - x_{\lfloor \frac{t}{\gamma} \rfloor})\,.$$

The random variable X_γ will be referred to as the linearly *interpolated process*. On the space $C(\mathbb{R}_+, \mathsf{X})$ endowed with $\mathscr{B}(C(\mathbb{R}_+, \mathsf{X})))$, the distribution of the r.v. X_γ is $\mathbb{P}^{\nu,\gamma}\mathsf{X}_\gamma^{-1}$.

Theorem 3.5.1. Suppose that Assumption (RM) is satisfied. Then, for every compact set $K \subset \mathsf{X}$, the family $\{\mathbb{P}^{a,\gamma}\mathsf{X}_\gamma^{-1} : a \in K, 0 < \gamma < \gamma_0\}$ is tight. Moreover, for every $\varepsilon > 0$,

$$\sup_{a \in K} \mathbb{P}^{a,\gamma}\left(\mathsf{d}(\mathsf{X}_\gamma, \Phi(K)) > \varepsilon\right) \xrightarrow[\gamma \to 0]{} 0\,,$$

where Φ is the evolution system induced by H.

3.5.2 Convergence Analysis

For each $\gamma \in (0, \gamma_0)$, we denote by

$$\mathcal{I}(P_\gamma) := \{\pi \in \mathcal{M}(\mathsf{X}) \,:\, \pi = \pi P_\gamma\}$$

the set of invariant probability measures of P_γ. Letting $\mathcal{P} = \{P_\gamma, 0 < \gamma < \gamma_0\}$, we define $\mathcal{I}(\mathcal{P}) = \bigcup_{\gamma \in (0,\gamma_0)} \mathcal{I}(P_\gamma)$. We say that a measure $\nu \in \mathcal{M}(\mathsf{X})$ is a cluster point of $\mathcal{I}(\mathcal{P})$ as $\gamma \to 0$, if there exists a sequence $\gamma_j \to 0$ and a sequence of measures $(\pi_j, j \in \mathbb{N})$ s.t. $\pi_j \in \mathcal{I}(P_{\gamma_j})$ for all j, and $\pi_j \Rightarrow \nu$.

We define

$$\mathscr{I}(P_\gamma) := \{\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X})) \,:\, \mathrm{supp}(\mathfrak{m}) \subset \mathcal{I}(P_\gamma)\}\,,$$

and $\mathscr{I}(\mathcal{P}) = \bigcup_{\gamma \in (0,\gamma_0)} \mathscr{I}(P_\gamma)$. We say that a measure $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X}))$ is a cluster point of $\mathscr{I}(\mathcal{P})$ as $\gamma \to 0$, if there exists a sequence $\gamma_j \to 0$ and a sequence of measures $(\mathfrak{m}_j, j \in \mathbb{N})$ s.t. $\mathfrak{m}_j \in \mathscr{I}(P_{\gamma_j})$ for all j, and $\mathfrak{m}_j \Rightarrow \mathfrak{m}$.

Proposition 3.5.2. Suppose that Assumption (RM) is satisfied. Then,

i) As $\gamma \to 0$, any cluster point of $\mathcal{I}(\mathcal{P})$ is an element of $\mathcal{I}(\Phi)$;

ii) As $\gamma \to 0$, any cluster point of $\mathscr{I}(\mathcal{P})$ is an element of $\mathscr{I}(\Phi)$.

In order to explore the consequences of this Prop., we introduce two supplementary assumptions. The first is the so-called Pakes-Has'minskii tightness criterion, who reads as follows [60]:

Assumption (PH). There exists measurable mappings $V : \mathsf{X} \to [0, +\infty)$, $\psi : \mathsf{X} \to [0, +\infty)$ and two functions $\alpha : (0, \gamma_0) \to (0, +\infty)$, $\beta : (0, \gamma_0) \to \mathbb{R}$, such that

$$\sup_{\gamma \in (0,\gamma_0)} \frac{\beta(\gamma)}{\alpha(\gamma)} < \infty \qquad \text{and} \qquad \lim_{\|x\| \to +\infty} \psi(x) = +\infty\,,$$

and for every $\gamma \in (0, \gamma_0)$,

$$P_\gamma V \leq V - \alpha(\gamma)\psi + \beta(\gamma)\,.$$

We recall that a transition kernel P on $\mathsf{X} \times \mathscr{B}(\mathsf{X}) \to [0, 1]$ is said *Feller* if the mapping $Pf : x \mapsto \int f(y)P(x, dy)$ is continuous for any $f \in C_b(\mathsf{X})$. If P is Feller, then the set of invariant measures of P is a closed subset of $\mathcal{M}(\mathsf{X})$. The following assumption ensures that for all $\gamma \in (0, \gamma_0)$, P_γ is Feller.

Assumption (FL). For every $s \in \Xi$, $\gamma \in (0, \gamma_0)$, the function $h_\gamma(s, \cdot)$ is continuous.

Theorem 3.5.3. Let Assumptions (RM), (PH) and (FL) be satisfied. Let ψ and V be the functions specified in (PH). Let $\nu \in \mathcal{M}(\mathsf{X})$ s.t. $\nu(V) < \infty$. Let $\mathcal{U} := \bigcup_{\pi \in \mathcal{I}(\Phi)} \mathrm{supp}(\pi)$. Then, for all $\varepsilon > 0$,

$$\limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}(d(X_k, \mathcal{U}) > \varepsilon) \xrightarrow[\gamma \to 0]{} 0\,. \tag{3.17}$$

Let Y an Euclidean space and $f \in C(\mathsf{X}, \mathsf{Y})$. Assume that there exists $M \geq 0$ and $\varphi : \mathsf{Y} \to \mathbb{R}_+$ such that $\lim_{\|a\| \to \infty} \varphi(a)/\|a\| = +\infty$ and

$$\forall a \in \mathsf{X},\ \varphi(f(a)) \leq M(1 + \psi(a))\,. \tag{3.18}$$

Then, the set $\mathcal{S}_f := \{\pi(f) \,:\, \pi \in \mathcal{I}(\Phi) \text{ and } \pi(\|f(\cdot)\|) < \infty\}$ is nonempty. For all $n \in \mathbb{N}$, $\gamma \in (0, \gamma_0)$, the r.v.

$$F_n := \frac{1}{n+1} \sum_{k=0}^{n} f(X_k)$$

is $\mathbb{P}^{\nu,\gamma}$-integrable, and satisfies for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \; d\left(\mathbb{E}^{\nu,\gamma}(F_n)\,, \mathcal{S}_f\right) \xrightarrow[\gamma\to 0]{} 0\,, \tag{3.19}$$

$$\limsup_{n\to\infty} \; \mathbb{P}^{\nu,\gamma}\left(d\left(F_n\,, \mathcal{S}_f\right) \geq \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,. \tag{3.20}$$

Theorem 3.5.4. Let Assumptions (RM), (PH) and (FL) be satisfied. Assume that $\Phi(\mathsf{X})$ is closed. Let ψ and V be the functions specified in (PH). Let $\nu \in \mathcal{M}(\mathsf{X})$ s.t. $\nu(V) < \infty$. Assume that

$$\lim_{\|a\|\to\infty} \frac{\psi(a)}{\|a\|} = +\infty\,.$$

For all $n \in \mathbb{N}$, define $\overline{X}_n := \frac{1}{n+1} \sum_{k=0}^{n} X_k$. Then, for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \; d\left(\mathbb{E}^{\nu,\gamma}(\overline{X}_n)\,, \mathrm{co}(L_{\mathrm{av}(\Phi)})\right) \xrightarrow[\gamma\to 0]{} 0\,,$$

$$\limsup_{n\to\infty} \; \mathbb{P}^{\nu,\gamma}\Big(d\Big(\overline{X}_n\,, \mathrm{co}(L_{\mathrm{av}(\Phi)})\Big) \geq \varepsilon\Big) \xrightarrow[\gamma\to 0]{} 0\,,$$

Theorem 3.5.5. Let Assumptions (RM), (PH) and (FL) be satisfied. Assume that $\Phi(\mathsf{X})$ is closed. Let ψ and V be the functions specified in (PH). Let $\nu \in \mathcal{M}(\mathsf{X})$ s.t. $\nu(V) < \infty$. Then, for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \; \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}\left(d\left(X_k\,, \mathsf{BC}_\Phi\right) \geq \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,.$$

3.6 Proof of Th. 3.5.1

The first lemma is a straightforward adaptation of the *convergence theorem* [6, Chap. 1.4, Th. 1, pp. 60]. Hence, the proof is omitted. We denote by λ_T the Lebesgue measure on $[0, T]$.

Lemma 3.6.1. Let $\{F_\xi : \xi \in \Xi\}$ be a family of mappings on $\mathsf{X} \rightrightarrows \mathsf{X}$. Let $T > 0$ and for all $n \in \mathbb{N}$, let $u_n : [0, T] \to \mathsf{X}$, $v_n : \Xi \times [0, T] \to \mathsf{X}$ be measurable maps w.r.t $\mathscr{B}([0, T])$ and $\mathscr{G} \otimes \mathscr{B}([0, T])$ respectively. Note for simplicity $\mathcal{L}^1 := \mathcal{L}^1(\Xi \times [0, T], \mathscr{G} \otimes \mathscr{B}([0, T]), \mu \otimes \lambda_T; \mathbb{R})$. Assume the following.

i) For all (ξ, t) $\mu \otimes \lambda_T$-a.e., $(u_n(t), v_n(\xi, t)) \to_n \mathrm{gr}(F_\xi)$.

ii) (u_n) converges λ_T-a.e. to a function $u : [0, T] \to \mathsf{X}$.

iii) For all n, $v_n \in \mathcal{L}^1$ and converges weakly in $\mathcal{L}^1$ to a function $v : \Xi \times [0, T] \to \mathsf{X}$.

iv) For all ξ μ-a.e., F_ξ is proper upper semi continuous with closed convex values.

Then, for all (ξ, t) $\mu \otimes \lambda_T$-a.e., $v(\xi, t) \in F_\xi(u(t))$.

Given $T > 0$ and $0 < \delta \leq T$, we denote by

$$w_{\mathsf{x}}^T(\delta) := \sup\{\|\mathsf{x}(t) - \mathsf{x}(s)\| : |t-s| \leq \delta, (t,s) \in [0,T]^2\}$$

the modulus of continuity on $[0,T]$ of any $\mathsf{x} \in C(\mathbb{R}_+, \mathsf{X})$.

Lemma 3.6.2. For all $n \in \mathbb{N}$, denote by $\mathscr{F}_n \subset \mathscr{F}$ the σ-field generated by the r.v. $\{X_k, 0 \leq k \leq n\}$. For all $\gamma \in (0, \gamma_0)$, define $Z_{n+1}^\gamma := \gamma^{-1}(X_{n+1} - X_n)$. Let $K \subset \mathsf{X}$ be compact. Let $\{\bar{\mathbb{P}}^{a,\gamma}, a \in K, 0 < \gamma < \gamma_0\}$ be a family of probability measures on $(\Omega, \mathscr{F})$ satisfying the following uniform integrability condition:

$$\sup_{n \in \mathbb{N}^*, a \in K, \gamma \in (0,\gamma_0)} \bar{\mathbb{E}}^{a,\gamma}(\|Z_n^\gamma\| \mathbb{1}_{\|Z_n^\gamma\| > A}) \xrightarrow{A \to +\infty} 0\,. \tag{3.21}$$

Then, $\{\bar{\mathbb{P}}^{a,\gamma}\mathsf{X}_\gamma^{-1} : a \in K, 0 < \gamma < \gamma_0\}$ is tight. Moreover, for any $T > 0, \varepsilon > 0$,

$$\sup_{a \in K} \bar{\mathbb{P}}^{a,\gamma}\left(\max_{0 \leq n \leq \lfloor \frac{T}{\gamma} \rfloor} \gamma \left\| \sum_{k=0}^n \left(Z_{k+1}^\gamma - \bar{\mathbb{E}}^{a,\gamma}(Z_{k+1}^\gamma | \mathscr{F}_k)\right)\right\| > \varepsilon\right) \xrightarrow{\gamma \to 0} 0\,. \tag{3.22}$$

Proof. We prove the first point. Set $T > 0$, let $0 < \delta \leq T$, and choose $0 \leq s \leq t \leq T$ s.t. $t - s \leq \delta$. Let $\gamma \in (0, \gamma_0)$ and set $n := \lfloor \frac{t}{\gamma} \rfloor$, $m := \lfloor \frac{s}{\gamma} \rfloor$. For any $R > 0$,

$$\begin{aligned}\|\mathsf{X}_\gamma(t) - \mathsf{X}_\gamma(s)\| &\leq \gamma \sum_{k=m+2}^n \|Z_k^\gamma\| + \gamma(t/\gamma - n)\|Z_{n+1}^\gamma\| + \gamma((m+1) - s/\gamma)\|Z_{m+1}^\gamma\| \\ &\leq \gamma(t/\gamma - s/\gamma)R + \gamma \sum_{k=m+1}^{n+1} \|Z_k^\gamma\| \mathbb{1}_{\|Z_k^\gamma\| > R}\,.\end{aligned}$$

Recalling that $t - s \leq \delta$ and using Markov inequality, we obtain

$$\begin{aligned}\bar{\mathbb{P}}^{a,\gamma}\mathsf{X}_\gamma^{-1}(\{x : w_x^T(\delta) > \varepsilon\}) &\leq \bar{\mathbb{P}}^{a,\gamma}\left(\gamma \sum_{k=1}^{\lfloor \frac{T}{\gamma} \rfloor + 1} \|Z_k^\gamma\| \mathbb{1}_{\|Z_k^\gamma\| > R} > \varepsilon - \delta R\right) \\ &\leq (T + \gamma_0) \frac{\sup_{k \in \mathbb{N}^*} \bar{\mathbb{E}}^{a,\gamma}\left(\|Z_k^\gamma\| \mathbb{1}_{\|Z_k^\gamma\| > R}\right)}{\varepsilon - \delta R}\,,\end{aligned}$$

provided that $R\delta < \varepsilon$. Choosing $R = \varepsilon/(2\delta)$ and using the uniform integrability,

$$\sup_{a \in K, 0 < \gamma < \gamma_0} \bar{\mathbb{P}}^{a,\gamma}\mathsf{X}_\gamma^{-1}(\{x : w_x^T(\delta) > \varepsilon\}) \xrightarrow{\delta \to 0} 0\,.$$

As $\{\bar{\mathbb{P}}^{a,\gamma}\mathsf{X}_\gamma^{-1} p_0^{-1}, 0 < \gamma < \gamma_0, a \in K\}$ is obviously tight, the tightness of $\{\bar{P}^{a,\gamma}\mathsf{X}_\gamma^{-1}, a \in K, 0 < \gamma < \gamma_0\}$ follows from [29, Th. 7.3]

We prove the second point. We define $M_{n+1}^{a,\gamma} := \sum_{k=0}^n \left(Z_{k+1}^\gamma - \bar{\mathbb{E}}^{a,\gamma}(Z_{k+1}^\gamma | \mathscr{F}_k)\right)$. We introduce

$$\eta_{n+1}^{a,\gamma,\leq} := Z_{n+1}^\gamma \mathbb{1}_{\|Z_{n+1}^\gamma\| \leq R} - \bar{\mathbb{E}}^{a,\gamma}\left(Z_{n+1}^\gamma \mathbb{1}_{\|Z_{n+1}^\gamma\| \leq R} | \mathscr{F}_n\right)$$

and we define $\eta_{n+1}^{a,\gamma,>}$ in a similar way, by replacing $\leq$ with $>$ in the right hand side of the above equation. Clearly, for all $a \in K$, $\gamma M_{n+1}^{a,\gamma} = S_{n+1}^{a,\gamma,\leq} + S_{n+1}^{a,\gamma,>}$ where $S_{n+1}^{a,\gamma,\leq} := \gamma \sum_{k=0}^n \eta_{k+1}^{a,\gamma,\leq}$ and $S_{n+1}^{a,\gamma,>}$ is defined similarly. Thus,

$$\gamma \|M_{n+1}^{a,\gamma}\| \leq \|S_{n+1}^{a,\gamma,\leq}\| + \|S_{n+1}^{a,\gamma,>}\|.$$

Under $\bar{\mathbb{P}}^{a,\gamma}$, the random processes $S^{a,\gamma,\leq}$ and $S^{a,\gamma,>}$ are $\mathscr{F}_n$-adapted martingales. Defining $q_\gamma := \lfloor \frac{T}{\gamma} \rfloor + 1$, we obtain by Doob's martingale inequality and by the boundedness of the increments of $S_n^{a,\gamma,\leq}$ that

$$\bar{\mathbb{P}}^{a,\gamma}\left(\max_{1\leq n\leq q_\gamma} \|S_n^{a,\gamma,\leq}\| > \varepsilon\right) \leq \frac{\bar{\mathbb{E}}^{a,\gamma}(\|S_{q_\gamma}^{a,\gamma,\leq}\|)}{\varepsilon} \leq \frac{\bar{\mathbb{E}}^{a,\gamma}(\|S_{q_\gamma}^{a,\gamma,\leq}\|^2)^{1/2}}{\varepsilon} \leq \frac{2}{\varepsilon}\gamma R\sqrt{q_\gamma}\,,$$

and the right hand side tends to zero uniformly in $a \in K$ as $\gamma \to 0$. By the same inequality,

$$\bar{\mathbb{P}}^{a,\gamma}\left(\max_{1\leq n\leq q_\gamma} \|S_n^{a,\gamma,>}\| > \varepsilon\right) \leq \frac{2}{\varepsilon} q_\gamma \gamma \sup_{k\in\mathbb{N}^*} \bar{\mathbb{E}}^{a,\gamma}\left(\|Z_k^\gamma\| \mathbb{1}_{\|Z_k^\gamma\|>R}\right)\,.$$

Choose an arbitrarily small $\delta > 0$ and select R as large as need in order that the supremum in the right hand side is no larger than $\varepsilon\delta/(2T + 2\gamma_0)$. Then the left hand side is no larger than δ. Hence, the proof is concluded. □

For any $R > 0$, define $h_{\gamma,R}(s,x) := h_\gamma(s,x)\mathbb{1}_{\|x\|\leq R}$. Let $H_R(s,x) := H(s,x)$ if $\|x\| < R$, $\{0\}$ if $\|x\| > R$, and X otherwise. Denote the corresponding selection integral as $\mathsf{H}_R(x) = \int H_R(s,x)\,\mu(ds)$. Define $\tau_R(x) := \inf\{n \in \mathbb{N} : \|x_n\| > R\}$ for all $x \in \Omega$. We also introduce the measurable mapping $B_R : \Omega \to \Omega$, given by

$$B_R(x) : n \mapsto x_n \mathbb{1}_{n<\tau_R(x)} + x_{\tau_R(x)}\mathbb{1}_{n\geq\tau_R(x)}$$

for all $x \in \Omega$ and all $n \in \mathbb{N}$.

Lemma 3.6.3. Suppose that Assumption (RM) is satisfied. Then, for every compact set $K \subset \mathsf{X}$, the family $\{\mathbb{P}^{a,\gamma}B_R^{-1}\mathsf{X}_\gamma^{-1}, \gamma \in (0,\gamma_0), a \in K\}$ is tight. Moreover, for every $\varepsilon > 0$,

$$\sup_{a\in K} \mathbb{P}^{a,\gamma}B_R^{-1}\left[d(\mathsf{X}_\gamma, \Phi_{\mathsf{H}_R}(K)) > \varepsilon\right] \xrightarrow[\gamma\to 0]{} 0\,.$$

Proof. We introduce the measurable mapping $M_{\gamma,R} : \Omega \to \mathsf{X}^{\mathbb{N}}$ s.t. for all $x \in \Omega$, $M_{\gamma,R}(x)(0) := 0$ and

$$M_{\gamma,R}(x)(n) := (x_n - x_0) - \gamma\sum_{k=0}^{n-1}\int h_{\gamma,R}(s,x_k)\mu(ds)$$

for all $n \in \mathbb{N}^*$. We also introduce the measurable mapping $\mathsf{G}_{\gamma,R} : C(\mathbb{R}_+,\mathsf{X}) \to C(\mathbb{R}_+,\mathsf{X})$ s.t. for all $\mathsf{x} \in C(\mathbb{R}_+,\mathsf{X})$,

$$\mathsf{G}_{\gamma,R}(\mathsf{x})(t) := \int_0^t \int h_{\gamma,R}(s,\mathsf{x}(\gamma\lfloor u/\gamma\rfloor))\,\mu(ds)\,du\,.$$

We first express the interpolated process in integral form. For every $x \in \mathsf{X}^{\mathbb{N}}$ and $t \geq 0$,

$$\mathsf{X}_\gamma(x)(t) = x_0 + \int_0^t \gamma^{-1}(x_{\lfloor\frac{u}{\gamma}\rfloor+1} - x_{\lfloor\frac{u}{\gamma}\rfloor})\,du\,.$$

We have the decomposition

$$x_n = x_0 + \gamma\sum_{k=0}^{n-1}\int h_{\gamma,R}(s,x_k)\mu(ds) + M_{\gamma,R}(x)(n).$$

Then, interpolating,

$$\mathsf{X}_\gamma(x) = x_0 + \mathsf{G}_{\gamma,R}\circ\mathsf{X}_\gamma(x) + \mathsf{X}_\gamma\circ M_{\gamma,R}(x)\,. \tag{3.23}$$

The uniform integrability condition (3.21) is satisfied when letting $\bar{\mathbb{P}}^{a,\gamma} := \mathbb{P}^{a,\gamma} B_R^{-1}$. First, note that $B_R(x)(n+1) = B_R(x)(n) + (x_{n+1} - x_n)\mathbb{1}_{\tau_R(x)>n}$ and that $\tau_R(x) > n \Leftrightarrow \|B_R(x)(n)\| \leq R \Rightarrow x_n = B_R(x)(n)$. Note also that, w.r.t $(\mathscr{F}_n)$, $\tau_R(X)$ is a stopping time and $B_R(X)$ is adapted. Then, using (RM)–i),

$$\begin{aligned}
\mathbb{E}^{a,\gamma}(\gamma^{-1}(B_R(X)(n+1) - B_R(X)(n))|\mathscr{F}_n) &= \mathbb{E}^{a,\gamma}\left(\frac{X_{n+1} - X_n}{\gamma}\mathbb{1}_{\tau_R(X)>n}|\mathscr{F}_n\right) \\
&= \mathbb{1}_{\tau_R(X)>n}\int \frac{y - X_n}{\gamma} P_\gamma(X_n, dy) \\
&= \mathbb{1}_{\tau_R(X)>n}\int h_\gamma(s, X_n)\mu(ds) \\
&= \mathbb{1}_{\|B_R(x)(n)\|\leq R}\int h_\gamma(s, B_R(X)(n))\mu(ds) \\
&= \int h_{\gamma,R}(s, B_R(X)(n))\mu(ds).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}^{a,\gamma}\left(\left\|\frac{B_R(X)(n+1) - B_R(X)(n)}{\gamma}\right\|^{1+\epsilon_K}\right) &= \mathbb{E}^{a,\gamma}\left(\left\|\frac{X_{n+1} - X_n}{\gamma}\right\|^{1+\epsilon_K}\mathbb{1}_{\tau_R(X)>n}\right) \\
&= \mathbb{E}^{a,\gamma}\left(\int\left\|\frac{y - X_n}{\gamma}\right\|^{1+\epsilon_K} P_\gamma(X_n, dy)\mathbb{1}_{\tau_R(X)>n}\right) \\
&\leq \mathbb{E}^{a,\gamma}\left(\int\left\|\frac{y - X_n}{\gamma}\right\|^{1+\epsilon_K} P_\gamma(X_n, dy)\mathbb{1}_{\|X_n\|\leq R}\right) \\
&\leq \sup_{\|x\|\leq R}\int\left\|\frac{y - x}{\gamma}\right\|^{1+\epsilon_K} P_\gamma(x, dy).
\end{aligned}$$

The condition (3.21) follows from hypothesis (3.15). Thus, Lem. 3.6.2 implies that for all $\varepsilon > 0$ and $T > 0$,

$$\sup_{a\in K}\bar{\mathbb{P}}^{a,\gamma}\left(\max_{0\leq n\leq\lfloor\frac{T}{\gamma}\rfloor}\|M_{\gamma,R}(x)(n+1)\| > \varepsilon\right) \xrightarrow{\gamma\to 0} 0\,.$$

It is easy to see that for all $x \in \Omega$, the function $\mathsf{X}_\gamma \circ M_{\gamma,R}(x)$ is bounded on every compact interval $[0,T]$ by $\max_{0\leq n\leq\lfloor\frac{T}{\gamma}\rfloor}\|M_{\gamma,R}(x)(n+1)\|$. This in turns leads to:

$$\sup_{a\in K}\bar{\mathbb{P}}^{a,\gamma}(\|\mathsf{X}_\gamma \circ M_{\gamma,R}\|_{\infty,T} > \varepsilon) \xrightarrow{\gamma\to 0} 0\,, \tag{3.24}$$

where the notation $\|\mathsf{x}\|_{\infty,T}$ stands for the uniform norm of x on $[0,T]$.

As a second consequence of Lem. 3.6.2, the family $\{\bar{\mathbb{P}}^{a,\gamma}\mathsf{X}_\gamma^{-1}, 0 < \gamma < \gamma_0, a \in K\}$ is tight. Choose any subsequence (a_n, γ_n) s.t. $\gamma_n \to 0$ and $a_n \in K$. Using Prokhorov's theorem and the compactness of K, there exists a subsequence (which we still denote by (a_n, γ_n)) and there exist some $a^\star \in K$ and some $\upsilon \in \mathcal{M}(C(\mathbb{R}_+, \mathsf{X}))$ such that $a_n \to a^\star$ and $\bar{\mathbb{P}}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1}$ converges narrowly to υ. By Skorokhod's representation theorem, we introduce some r.v. z, $\{\mathsf{x}_n, n \in \mathbb{N}\}$ on $C(\mathbb{R}_+, \mathsf{X})$ with respective distributions υ and $\bar{\mathbb{P}}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1}$, defined on some other probability space $(\Omega', \mathscr{F}', \mathbb{P}')$ and such that $d(\mathsf{x}_n(\omega), \mathsf{z}(\omega)) \to 0$ for all $\omega \in \Omega'$. By (3.23) and (3.24), the sequence of r.v.

$$\mathsf{x}_n - \mathsf{x}_n(0) - \mathsf{G}_{\gamma_n,R}(\mathsf{x}_n)$$

converges in probability to zero in $(\Omega', \mathscr{F}', \mathbb{P}')$, as $n \to \infty$. One can extract a subsequence under which this convergence holds in the almost sure sense. Therefore, there exists an event of probability one s.t., everywhere on this event,

$$\mathsf{z}(t) = \mathsf{z}(0) + \lim_{n\to\infty} \int_0^t \int_\Xi h_{\gamma_n,R}(s, \mathsf{x}_n(\gamma_n \lfloor u/\gamma_n \rfloor))\, \mu(ds)\, du \qquad (\forall t \geq 0)\,,$$

where the limit is taken along the former subsequence. We now select an ω s.t. the above convergence holds, and omit the dependence on ω in the sequel (otherwise stated, z and x_n are treated as elements of $C(\mathbb{R}_+, \mathsf{X})$ and no longer as random variables). Set $T > 0$. As (x_n) converges uniformly on $[0,T]$, there exists a compact set K' (which depends on ω) such that $\mathsf{x}_n(\gamma_n \lfloor t/\gamma_n \rfloor) \in K'$ for all $t \in [0,T]$, $n \in \mathbb{N}$. Define

$$v_n(s,t) := h_{\gamma_n,R}(s, \mathsf{x}_n(\gamma_n \lfloor t/\gamma_n \rfloor))\,.$$

By Eq. (3.16), the sequence $(v_n, n \in \mathbb{N})$ forms a bounded subset of $\mathcal{L}^{1+\varepsilon_{K'}} := \mathcal{L}^{1+\varepsilon_{K'}}(\Xi \times [0,T], \mathscr{G} \otimes \mathscr{B}([0,T]), \mu \otimes \lambda_T; \mathsf{X})$. By the Banach-Alaoglu theorem, the sequence converges weakly to some mapping $v \in \mathcal{L}^{1+\varepsilon_{K'}}$ along some subsequence. This has two consequences. First,

$$\mathsf{z}(t) = \mathsf{z}(0) + \int_0^t \int_\Xi v(s,u)\, \mu(ds)\, du\,, \quad (\forall t \in [0,T])\,. \tag{3.25}$$

Second, for $\mu \otimes \lambda_T$-almost all (s,t), $v(s,t) \in H_R(s, \mathsf{z}(t))$. In order to prove this point, remark that, by Assumption (RM),

$$v_n(s,t) \to H_R(s, \mathsf{z}(t))$$

for almost all (s,t). This implies that the couple $(\mathsf{x}_n(\gamma_n \lfloor t/\gamma_n \rfloor), v_n(s,t))$ converges to $\mathrm{gr}(H_R(s,\cdot))$ and the second point thus follows from Lem. 3.6.1. By Fubini's theorem, there exists a negligible set of $[0,T]$ s.t. for all t outside this set, $v(\cdot,t)$ is an integrable selection of $H_R(\cdot, \mathsf{z}(t))$. As $H(\cdot,x)$ is integrable for every $x \in \mathsf{X}$, the same holds for H_R. Denoting by H_R the selection integral of H_R and Φ_{H_R} the evolution system induced by H_R, Eq. (3.25) implies that $\mathsf{z} \in \Phi_{\mathsf{H}_R}(K)$. We have shown that for any sequence $((a_n, \gamma_n), n \in \mathbb{N})$ on $K \times (0, \gamma_0)$ s.t. $\gamma_n \to 0$, there exists a subsequence along which, for every $\varepsilon > 0$, $\mathbb{P}^{a_n,\gamma_n} B_R^{-1}(d(\mathsf{X}_{\gamma_n}, \Phi_{\mathsf{H}_R}(K)) > \varepsilon) \to 0$. This proves the lemma. □

End of the proof of Th. 3.5.1.

We first prove the second statement. Set an arbitrary $T > 0$. Define $d_T(\mathsf{x},\mathsf{y}) := \|x - y\|_{\infty,T}$. It is sufficient to prove that for any sequence $((a_n, \gamma_n), n \in \mathbb{N})$ s.t. $\gamma_n \to 0$, there exists a subsequence along which $\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n}, \Phi(K)) > \varepsilon) \to 0$. Choose $R > R_0(T)$, where $R_0(T) := \sup\{\|\mathsf{x}(t)\| : t \in [0,T], \mathsf{x} \in \Phi_{\mathsf{H}}(a), a \in K\}$ is finite by Assumption (RM). It is easy to show that any $\mathsf{x} \in \Phi_{\mathsf{H}_R}(K)$ must satisfy $\|\mathsf{x}\|_{\infty,T} < R$. Thus, when $R > R_0(T)$, any $\mathsf{x} \in \Phi_{\mathsf{H}_R}(K)$ is such that there exists $\mathsf{y} \in \Phi(K)$ with $d_T(\mathsf{x},\mathsf{y}) = 0$ *i.e.*, the restrictions of x and y to $[0,T]$ coincide. As a consequence of the Lem. 3.6.3, each sequence (a_n, γ_n) chosen as above admits a subsequence along which, for all $\varepsilon > 0$,

$$\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \Phi(K)) > \varepsilon) \to 0\,. \tag{3.26}$$

The event $d_T(\mathsf{X}_\gamma \circ B_R, \mathsf{X}_\gamma) > 0$ implies the event $\|\mathsf{X}_\gamma \circ B_R\|_{\infty,T} \geq R$, which in turn implies by the triangular inequality that $d_T(\mathsf{X}_\gamma \circ B_R, \Phi(K)) \geq R - R_0(T)$. Therefore,

$$\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \mathsf{X}_{\gamma_n}) > \varepsilon) \leq \mathbb{P}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \Phi(K)) \geq R - R_0(T))\,. \tag{3.27}$$

By (3.26), the right hand side converges to zero. Using (3.26) again along with the triangular inequality, it follows that $\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n}, \Phi(K)) > \varepsilon) \to 0$, which proves the second statement of the theorem.

We prove the first statement (tightness). Using [29], this is equivalent to showing that for every $T > 0$, and for every sequence (a_n, γ_n) on $K \times (0, \gamma_0)$, the sequence $(\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1} p_0^{-1})$ is tight, and for each positive ε and η, there exists $\delta > 0$ such that $\limsup_n \mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\{x : w_x^T(\delta) > \varepsilon\}) < \eta$. Since $\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1} p_0^{-1} = \delta_{a_n}$, $(\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1} p_0^{-1})_n$ is tight.

First consider the case where $\gamma_n \to 0$. Fixing $T > 0$, letting $R > R_0(T)$ and using (3.27), it holds that for all $\varepsilon > 0$, $\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \mathsf{X}_{\gamma_n}) > \varepsilon) \to_n 0$. Moreover, we showed that $(\mathbb{P}^{a_n,\gamma_n} B_R^{-1} \mathsf{X}_{\gamma_n}^{-1})$ is tight. In addition, for every $\mathsf{x}, \mathsf{y} \in C(\mathbb{R}_+, \mathsf{X})$, it holds by the triangle inequality that $w_{\mathsf{x}}^T(\delta) \le w_{\mathsf{y}}^T(\delta) + 2 d_T(\mathsf{x}, \mathsf{y})$ for every $\delta > 0$. Thus,

$$\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\{x : w_x^T(\delta) > \varepsilon\}) \le \mathbb{P}^{a_n,\gamma_n} B_R^{-1} \mathsf{X}_{\gamma_n}^{-1}(\{x : w_x^T(\delta) > \varepsilon/2\}) + \mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \mathsf{X}_{\gamma_n}) > \varepsilon/4),$$

which leads to the tightness of $(\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1})$ when $\gamma_n \to 0$.

It remains to establish the tightness when $\liminf_n \gamma_n > \eta > 0$ for some $\eta > 0$. Note that for all $\gamma > \eta$,

$$w_{\mathsf{X}_\gamma^T(x)}(\delta) \le 2\delta \max_{k=0\ldots\lfloor T/\eta \rfloor + 1} \|x_k\| .$$

There exists n_0 such that for all $n \ge n_0$, $\gamma_n > \eta$ which implies by the union bound:

$$\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\{\mathsf{x} : w_{\mathsf{x}}^T(\delta) > \varepsilon\}) \le \sum_{k=0}^{\lfloor T/\eta \rfloor + 1} P_\gamma^k(a, B(0, (2\delta)^{-1}\varepsilon)^c) ,$$

where $B(0, r) \subset \mathsf{X}$ stands for the ball or radius r and where P_γ^k stands for the iterated kernel, recursively defined by

$$P_\gamma^k(a, \cdot) = \int P_\gamma(a, dy) P_\gamma^{k-1}(y, \cdot) \tag{3.28}$$

and $P_\gamma^0(a, \cdot) = \delta_a$. Using (3.15), it is an easy exercise to show, by induction, that for every $k \in \mathbb{N}$, $P_\gamma^k(a, B(0, r)^c) \to 0$ as $r \to \infty$. By letting $\delta \to 0$ in the above inequality, the tightness of $(\mathbb{P}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1})$ follows.

3.7 Proof of Prop. 3.5.2

To establish Prop. 3.5.2–i), we consider a sequence $((\pi_n, \gamma_n), n \in \mathbb{N})$ such that $\pi_n \in \mathcal{I}(P_{\gamma_n})$, $\gamma_n \to 0$, and (π_n) is tight. We first show that the sequence $(\upsilon_n := \mathbb{P}^{\pi_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}, n \in \mathbb{N})$ is tight, then we show that every cluster point of (υ_n) satisfies the conditions of Def. 3.4.1.

Given $\varepsilon > 0$, there exists a compact set $K \subset \mathsf{X}$ such that $\inf_n \pi_n(K) > 1 - \varepsilon/2$. By Th. 3.5.1, the family $\{\mathbb{P}^{a,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}, a \in K, n \in \mathbb{N}\}$ is tight. Let $\mathcal{C}$ be a compact set of $C(\mathbb{R}_+, \mathsf{X})$ such that $\inf_{a \in K, n \in \mathbb{N}} \mathbb{P}^{a,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\mathcal{C}) > 1 - \varepsilon/2$. By construction of the probability measure $\mathbb{P}^{\pi_n,\gamma_n}$, it holds that $\mathbb{P}^{\pi_n,\gamma_n}(\cdot) = \int_{\mathsf{X}} \mathbb{P}^{a,\gamma_n}(\cdot)\, \pi_n(da)$. Thus,

$$\upsilon_n(\mathcal{C}) \ge \int_K \mathbb{P}^{a,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\mathcal{C})\, \pi_n(da) > (1 - \varepsilon/2)^2 > 1 - \varepsilon ,$$

which shows that (υ_n) is tight.

Since $\pi_n = \upsilon_n p_0^{-1}$, and since the projection p_0 is continuous, it is clear that every cluster point π of $\mathcal{I}(\mathcal{P})$ as $\gamma \to 0$ can be written as $\pi = \upsilon p_0^{-1}$, where υ is a cluster point of a sequence (υ_n). Thus, Def. 3.4.1–(iii) is satisfied by π and υ. To establish Prop. 3.5.2–i), we need to verify the conditions (i)

and (ii) of Def. 3.4.1. In the remainder of the proof, we denote with a small abuse as (n) a subsequence along which (υ_n) converges narrowly to υ.

To establish the validity of Def. 3.4.1–(i), we prove that for every $\eta > 0$, $\upsilon_n((\Phi_{\mathsf{H}}(\mathsf{X}))_\eta) \to 1$ as $n \to \infty$; the result will follow from the convergence of (υ_n). Fix $\varepsilon > 0$, and let $K \subset \mathsf{X}$ be a compact set such that $\inf_n \pi_n(K) > 1 - \varepsilon$. We have

$$\begin{aligned}
\upsilon_n((\Phi_{\mathsf{H}}(\mathsf{X}))_\eta) &= \mathbb{P}^{\pi_n,\gamma_n}(d(\mathsf{X}_{\gamma_n}, \Phi(\mathsf{X})) < \eta) \\
&\geq \mathbb{P}^{\pi_n,\gamma_n}(d(\mathsf{X}_{\gamma_n}, \Phi(K)) < \eta) \\
&\geq \int_K \mathbb{P}^{a,\gamma_n}(d(\mathsf{X}_{\gamma_n}, \Phi(K)) < \eta)\, \pi_n(da) \\
&\geq (1-\varepsilon) \inf_{a\in K} \mathbb{P}^{a,\gamma_n}(d(\mathsf{X}_{\gamma_n}, \Phi(K)) < \eta)\,.
\end{aligned}$$

By Th. 3.5.1, the infimum at the right hand side converges to 1. Since $\varepsilon > 0$ is arbitrary, we obtain the result.

It remains to establish the Θ-invariance of υ (Condition (ii)). Equivalently, we need to show that

$$\int f(\mathsf{x})\, \upsilon(d\mathsf{x}) = \int f(\Theta(\mathsf{x})(t))\, \upsilon(d\mathsf{x}) \tag{3.29}$$

for all $f \in C_b(C(\mathbb{R}_+, \mathsf{X}))$ and all $t > 0$. We shall work on (υ_n) and make $n \to \infty$. Write $\eta_n := t - \gamma_n\lfloor t/\gamma_n \rfloor$. Thanks to the P_{γ_n}–invariance of π_n, $\Theta(\mathsf{x}(\gamma_n\lfloor t/\gamma_n\rfloor + \cdot))(\eta_n)$ and $\Theta(\mathsf{x})(t)$ are equal in law under $\upsilon_n(d\mathsf{x})$. Thus,

$$\begin{aligned}
\int f(\Theta(\mathsf{x})(t))\, \upsilon_n(d\mathsf{x}) &= \int f(\Theta(\mathsf{x}(\gamma_n\lfloor t/\gamma_n\rfloor + \cdot))(\eta_n))\, \upsilon_n(d\mathsf{x}) \\
&= \int f(\Theta(\mathsf{x})(\eta_n))\, \upsilon_n(d\mathsf{x}).
\end{aligned} \tag{3.30}$$

Using Skorokhod's representation theorem, there exists a probability space $(\Omega', \mathscr{F}', \mathbb{P}')$ and random variables $(\bar{\mathsf{x}}_n, n \in \mathbb{N})$ and $\bar{\mathsf{x}}$ over this probability space, with values in $C(\mathbb{R}_+, \mathsf{X})$, such that for every $n \in \mathbb{N}$, the distribution of $\bar{\mathsf{x}}_n$ is υ_n, the distribution of $\bar{\mathsf{x}}$ is υ and $\mathbb{P}'$-a.s,

$$d(\bar{\mathsf{x}}_n, \bar{\mathsf{x}}) \longrightarrow_{n\to+\infty} 0,$$

i.e, $(\bar{\mathsf{x}}_n)$ converges to $\bar{\mathsf{x}}$ as $n \to +\infty$ uniformly over compact sets of $\mathbb{R}_+$. Since $\eta_n \to_{n\to+\infty} 0$, $\mathbb{P}'$-a.s, $d(\Theta(\bar{\mathsf{x}}_n)(\eta_n), \bar{\mathsf{x}}) \longrightarrow_{n\to+\infty} 0$. Hence,

$$\int f(\Theta(\mathsf{x})(\eta_n))\, \upsilon_n(d\mathsf{x}) \xrightarrow[n\to\infty]{} \int f(\mathsf{x})\, \upsilon(d\mathsf{x})\,.$$

Recalling Eq. (3.30), we have shown that $\int f(\Theta(\mathsf{x})(t))\, \upsilon_n(d\mathsf{x}) \xrightarrow[n\to\infty]{} \int f(\Theta(\mathsf{x})(t))\, \upsilon(d\mathsf{x})$. Since

$$\int f(\mathsf{x})\, \upsilon_n(d\mathsf{x}) \xrightarrow[n\to\infty]{} \int f(\mathsf{x})\, \upsilon(d\mathsf{x}),$$

the identity (3.29) holds true. Prop. 3.5.2–i) is proven.

We now prove Prop. 3.5.2–ii). Consider a sequence $((\mathfrak{m}_n, \gamma_n), n \in \mathbb{N})$ such that $\mathfrak{m}_n \in \mathscr{I}(P_{\gamma_n})$, $\gamma_n \to 0$, and $\mathfrak{m}_n \Rightarrow \mathfrak{m}$ for some $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X}))$. Since the space $\mathcal{M}(\mathsf{X})$ is separable, Skorokhod's representation theorem shows that there exists a probability space $(\Omega', \mathscr{F}', \mathbb{P}')$, a sequence of $\Omega' \to \mathcal{M}(\mathsf{X})$ random variables (Λ_n) with distributions $\mathfrak{m}_n$, and a $\Omega' \to \mathcal{M}(\mathsf{X})$ random variable Λ with distribution $\mathfrak{m}$ such that $\Lambda_n(\omega) \Rightarrow \Lambda(\omega)$ for each $\omega \in \Omega'$. Moreover, there is a probability one subset of Ω' such that $\Lambda_n(\omega)$ is a P_{γ_n}–invariant probability measure for all n and for every ω in this set. For each of these ω, we can construct on the space $(\mathsf{X}^{\mathbb{N}}, \mathscr{F})$ a probability measure $\mathbb{P}^{\Lambda_n(\omega),\gamma_n}$ as we did in Sec. 3.5.1. By the same argument as in the proof of Prop. 3.5.2–i), the sequence $(\mathbb{P}^{\Lambda_n(\omega),\gamma_n}\mathsf{X}_{\gamma_n}^{-1}, n \in \mathbb{N})$ is tight, and any cluster point υ satisfies the conditions of Def. 3.4.1 with $\Lambda(\omega) = \upsilon p_0^{-1}$. Prop. 3.5.2 is proven.

3.8 Proof of Th. 3.5.3

3.8.1 Technical lemmas

Using Prokhorov's theorem, some compact sets of $\mathcal{M}(\mathsf{X})$ are given by the following.

Lemma 3.8.1. Given a family $\{K_j, j \in \mathbb{N}\}$ of compact sets of X, the set

$$U := \{\mu \in \mathcal{M}(\mathsf{X}) \,:\, \forall j \in \mathbb{N}, \mu(K_j) \geq 1 - 2^{-j}\}$$

is a compact set of $\mathcal{M}(\mathsf{X})$.

Proof. The set U is tight hence relatively compact by Prokhorov's theorem. It is moreover closed. Indeed, let $(\mu_n, n \in \mathbb{N})$ represent a sequence of U s.t. $\mu_n \Rightarrow \mu$. Then, for all $j \in \mathbb{N}$, $\mu(K_j) \geq \limsup_n \mu_n(K_j) \geq 1 - 2^{-j}$ since K_j is closed. □

A tightness criterion of probability measures *over* the space $E = \mathcal{M}(\mathsf{X})$ i.e *in* the space $\mathcal{M}(\mathcal{M}(\mathsf{X}))$ can be given.

For any $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X}))$, we denote by $e(\mathfrak{m})$ the probability measure in $\mathcal{M}(\mathsf{X})$ such that for every $f \in C_b(\mathsf{X})$,

$$e(\mathfrak{m}) : f \mapsto \int \mu(f)\mathfrak{m}(d\mu)\,.$$

Otherwise stated, $e(\mathfrak{m})(f) = \mathfrak{m}(\mathcal{T}_f)$ where $\mathcal{T}_f : \mu \mapsto \mu(f)$.

Lemma 3.8.2. Let X be a real random variable such that $X \leq 1$ with probability one, and $\mathbb{E}X \geq 1 - \varepsilon$ for some $\varepsilon \geq 0$. Then $\mathbb{P}[X \geq 1 - \sqrt{\varepsilon}] \geq 1 - \sqrt{\varepsilon}$.

Proof. $1 - \varepsilon \leq \mathbb{E}X \leq \mathbb{E}X\mathbb{1}_{X<1-\sqrt{\varepsilon}} + \mathbb{E}X\mathbb{1}_{X\geq 1-\sqrt{\varepsilon}} \leq (1 - \sqrt{\varepsilon})(1 - \mathbb{P}[X \geq 1 - \sqrt{\varepsilon}]) + \mathbb{P}[X \geq 1 - \sqrt{\varepsilon}]$. The result is obtained by rearranging. □

Lemma 3.8.3. Let $\mathcal{L}$ be a family on $\mathcal{M}(\mathcal{M}(\mathsf{X}))$. If $\{e(\mathfrak{m}) \,:\, \mathfrak{m} \in \mathcal{L}\}$ is tight, then $\mathcal{L}$ is tight.

Proof. Let $\varepsilon > 0$ and choose any integer k s.t. $2^{-k+1} \leq \varepsilon$. For all $j \in \mathbb{N}$, choose a compact set $K_j \subset \mathsf{X}$ s.t. for all $\mathfrak{m} \in \mathcal{L}$, $e(\mathfrak{m})(K_j) > 1 - 2^{-2j}$. Define U as the set of measures $\nu \in \mathcal{M}(\mathsf{X})$ s.t. for all $j \geq k$, $\nu(K_j) \geq 1 - 2^{-j}$. By Lem. 3.8.1, U is compact. For all $\mathfrak{m} \in \mathcal{L}$, the union bound implies that

$$\mathfrak{m}(\mathcal{M}(\mathsf{X})\backslash U) \leq \sum_{j=k}^{\infty} \mathfrak{m}\{\nu : \nu(K_j) < 1 - 2^{-j}\}$$

By Lem. 3.8.2, $\mathfrak{m}\{\nu : \nu(K_j) \geq 1 - 2^{-j}\} \geq 1 - 2^{-j}$. Therefore, $\mathfrak{m}(\mathcal{M}(\mathsf{X})\backslash U) \leq \sum_{j=k}^{\infty} 2^{-j} = 2^{-k+1} \leq \varepsilon$. This proves that $\mathcal{L}$ is tight. □

Moreover, $e : \mathcal{M}(\mathcal{M}(\mathsf{X})) \to \mathcal{M}(\mathsf{X})$ is continuous.

Lemma 3.8.4. Let $(\mathfrak{m}_n, n \in \mathbb{N})$ be a sequence on $\mathcal{M}(\mathcal{M}(\mathsf{X}))$, and consider $\bar{\mathfrak{m}} \in \mathcal{M}(\mathcal{M}(\mathsf{X}))$. If $\mathfrak{m}_n \Rightarrow \bar{\mathfrak{m}}$, then $e(\mathfrak{m}_n) \Rightarrow e(\bar{\mathfrak{m}})$.

Proof. For any $f \in C_b(\mathsf{X})$, $\mathcal{T}_f \in C_b(\mathcal{M}(\mathsf{X}))$. Thus, $\mathfrak{m}_n(\mathcal{T}_f) \to \bar{\mathfrak{m}}(\mathcal{T}_f)$. □

When a sequence $(\mathfrak{m}_n, n \in \mathbb{N})$ of $\mathcal{M}(\mathcal{M}(\mathsf{X}))$ converges narrowly to $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X}))$, it follows from the above proof that $\mathfrak{m}_n\mathcal{T}_f^{-1} \Rightarrow \mathfrak{m}\mathcal{T}_f^{-1}$ for all bounded continuous f. The purpose of the next lemma is to extend this result to the case where f is not necessarily bounded, but instead, satisfies some uniform integrability condition. For any vector-valued function f, we use the notation $\|f\| := \|f(\cdot)\|$.

Lemma 3.8.5. Let $f \in C(\mathsf{X},\mathsf{Y})$ where Y is an Euclidean space. Define by $\mathcal{T}_f : \mathcal{M}(\mathsf{X}) \to \mathbb{R}$ the mapping s.t. $\mathcal{T}_f(\nu) := \nu(f)$ if $\nu(\|f\|) < \infty$ and equal to zero otherwise. Let $(\mathfrak{m}_n, n \in \mathbb{N})$ be a sequence on $\mathcal{M}(\mathcal{M}(\mathsf{X}))$ and let $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(\mathsf{X}))$. Assume that $\mathfrak{m}_n \Rightarrow \mathfrak{m}$ and

$$\lim_{K\to\infty} \sup_n e(\mathfrak{m}_n)(\|f\|\mathbb{1}_{\|f\|>K}) = 0\,. \tag{3.31}$$

Then, $\nu(\|f\|) < \infty$ for all ν $\mathfrak{m}$-a.e. and $\mathfrak{m}_n\mathcal{T}_f^{-1} \Rightarrow \mathfrak{m}\mathcal{T}_f^{-1}$.

Proof. By Eq. (3.31), $e(\mathfrak{m})(\|f\|) < \infty$. This implies that for all ν $\mathfrak{m}$-a.e., $\nu(\|f\|) < \infty$. Choose $h \in C_b(\mathsf{Y})$ s.t. h is L-Lipschitz continuous. We must prove that $\mathfrak{m}_n\mathcal{T}_f^{-1}(h) \to \mathfrak{m}\mathcal{T}_f^{-1}(h)$. By the above remark, $\mathfrak{m}\mathcal{T}_f^{-1}(h) = \int h(\nu(f))d\mathfrak{m}(\nu)$, and by Eq (3.31), $\mathfrak{m}_n\mathcal{T}_f^{-1}(h) = \int h(\nu(f))d\mathfrak{m}_n(\nu)$. Choose $\varepsilon > 0$. By Eq. (3.31), there exists $K_0 > 0$ s.t. for all $K > K_0$, $\sup_n e(\mathfrak{m}_n)(\|f\|\mathbb{1}_{\|f\|>K}) < \varepsilon$. For every such K, define the bounded function $f_K \in C(\mathsf{X},\mathsf{Y})$ by $f_K(x) = f(x)(1 \wedge K/\|f(x)\|)$. For all $K > K_0$, and for all $n \in \mathbb{N}$,

$$\begin{aligned}
|\mathfrak{m}_n\mathcal{T}_f^{-1}(h) - \mathfrak{m}_n\mathcal{T}_{f_K}^{-1}(h)| &\le \int |h(\nu(f)) - h(\nu(f_K))|d\mathfrak{m}_n(\nu) \\
&\le L\int \nu(\|f - f_K\|)d\mathfrak{m}_n(\nu) \\
&\le L\int \nu(\|f\|\mathbb{1}_{\|f\|>K})d\mathfrak{m}_n(\nu) \le L\varepsilon\,.
\end{aligned}$$

By continuity of $\mathcal{T}_{f_K}$, it holds that $\mathfrak{m}_n\mathcal{T}_{f_K}^{-1}(h) \to \mathfrak{m}\mathcal{T}_{f_K}^{-1}(h)$. Therefore, for every $K > K_0$,

$$\limsup_n |\mathfrak{m}_n\mathcal{T}_f^{-1}(h) - \mathfrak{m}\mathcal{T}_{f_K}^{-1}(h)| \le L\varepsilon\,.$$

As $\nu(\|f\|) < \infty$ for all ν $\mathfrak{m}$-a.e., the dominated convergence theorem implies that $\nu(f_K) \to \nu(f)$ as $K \to \infty$, $\mathfrak{m}$-a.e. As h is bounded and continuous, a second application of the dominated convergence theorem implies that $\int h(\nu(f_K))d\mathfrak{m}(\nu) \to \int h(\nu(f))d\mathfrak{m}(\nu)$, which reads $\mathfrak{m}\mathcal{T}_{f_K}^{-1}(h) \to \mathfrak{m}\mathcal{T}_f^{-1}(h)$. Thus, $\limsup_n |\mathfrak{m}_n\mathcal{T}_f^{-1}(h) - \mathfrak{m}\mathcal{T}_f^{-1}(h)| \le L\varepsilon$. As a consequence, $\mathfrak{m}_n\mathcal{T}_f^{-1}(h) \to \mathfrak{m}\mathcal{T}_f^{-1}(h)$ as $n \to \infty$, which completes the proof. □

3.8.2 Narrow Cluster Points of the Empirical Measures

Let $P : \mathsf{X} \times \mathscr{B}(\mathsf{X}) \to [0,1]$ be a probability transition kernel. For $\nu \in \mathcal{M}(\mathsf{X})$, we denote by $\mathbb{P}^{\nu,P}$ the probability on $(\Omega, \mathscr{F})$ such that X is an homogeneous Markov chain with initial distribution ν and transition kernel P.

For every $n \in \mathbb{N}$, we define the measurable mapping $\Lambda_n : \Omega \to \mathcal{M}(\mathsf{X})$ as

$$\Lambda_n(x) := \frac{1}{n+1}\sum_{k=0}^{n} \delta_{x_k} \tag{3.32}$$

for all $x = (x_k : k \in \mathbb{N})$. Note that

$$\mathbb{E}^{\nu,P}\Lambda_n = \frac{1}{n+1}\sum_{k=0}^{n} \nu P^k\,,$$

where $\mathbb{E}^{\nu,P}\Lambda_n = e(\mathbb{P}^{\nu,P}\Lambda_n^{-1})$, and P^k stands for the iterated kernel, recursively defined by $P^k(x,\cdot) = \int P(x,dy)P^{k-1}(y,\cdot)$ and $P^0(x,\cdot) = \delta_x$.

We recall that $\mathscr{I}(P)$ represents the subset of $\mathcal{M}(\mathcal{M}(\mathsf{X}))$ formed by the measures whose support is included in $\mathcal{I}(P)$.

Proposition 3.8.6. Let $P : \mathsf{X} \times \mathscr{B}(\mathsf{X}) \to [0,1]$ be a Feller probability transition kernel. Let $\nu \in \mathcal{M}(\mathsf{X})$.

1. Any cluster point of $\{\mathbb{E}^{\nu,P}\Lambda_n\,,\, n \in \mathbb{N}\}$ is an element of $\mathcal{I}(P)$.
2. Any cluster point of $\{\mathbb{P}^{\nu,P}\Lambda_n^{-1}\,,\, n \in \mathbb{N}\}$ is an element of $\mathscr{I}(P)$.

Proof. We omit the upper script $^{\nu,P}$. For all $f \in C_b(\mathsf{X})$, $\mathbb{E}\Lambda_n(Pf) - \mathbb{E}\Lambda_n(f) \to 0$. As P is Feller, any cluster point π of $\{\mathbb{E}\Lambda_n\,,\, n \in \mathbb{N}\}$ satisfies $\pi(Pf) = \pi(f)$. This proves the first point.

For every $f \in C_b(\mathsf{X})$ and $x \in \Omega$, consider the decomposition:

$$\Lambda_n(x)(Pf) - \Lambda_n(x)(f) = \frac{1}{n+1}\sum_{k=0}^{n-1}(Pf(x_k) - f(x_{k+1})) + \frac{Pf(x_n) - f(x_0)}{n+1}\,.$$

Using that f is bounded, Doob's martingale convergence theorem implies that the sequence

$$\Big(\sum_{k=0}^{n-1} k^{-1}(Pf(X_k) - f(X_{k+1}))\Big)_n$$

converges a.s. when n tends to infinity. By Kronecker's lemma, we deduce that

$$\frac{1}{n+1}\sum_{k=0}^{n-1}(Pf(X_k) - f(X_{k+1}))$$

tends a.s. to zero. Hence,

$$\Lambda_n(Pf) - \Lambda_n(f) \to 0 \text{ a.s.} \tag{3.33}$$

Now consider a subsequence (Λ_{φ_n}) which converges in distribution to some r.v. Λ as n tends to infinity. For a fixed $f \in C_b(\mathsf{X})$, the mapping $\nu \mapsto (\nu(f), \nu(Pf))$ on $\mathcal{M}(\mathsf{X}) \to \mathbb{R}^2$ is continuous. From the mapping theorem, $\Lambda_{\varphi_n}(f) - \Lambda_{\varphi_n}(Pf)$ converges in distribution to $\Lambda(f) - \Lambda(Pf)$. By (3.33), it follows that $\Lambda(f) - \Lambda(Pf) = 0$ on some event $\mathcal{E}_f \in \mathscr{F}$ of probability one. Denote by $C_\kappa(\mathsf{X}) \subset C_b(\mathsf{X})$ the set of continuous real-valued functions having a compact support, and let $C_\kappa(\mathsf{X})$ be equipped with the uniform norm $\|\cdot\|_\infty$. Introduce a dense denumerable subset S of $C_\kappa(\mathsf{X})$. On the probability-one event $\mathcal{E} = \cap_{f\in S}\mathcal{E}_f$, it holds that for all $f \in S$, $\Lambda(f) = \Lambda P(f)$. The same equality can be extended to any $f \in C_\kappa(\mathsf{X})$ by density of S and continuity of $f \mapsto (\nu(f), \nu(Pf))$ over $C_\kappa(\mathsf{X})$ for every $\nu \in \mathcal{M}(\mathsf{X})$. Hence, almost everywhere on $\mathcal{E}$, one has $\Lambda = \Lambda P$. □

3.8.3 Tightness of the Empirical Measures

Proposition 3.8.7. Let $\mathcal{P}$ be a family of transition kernels on X. Let $V : \mathsf{X} \to [0, +\infty)$, $\psi : \mathsf{X} \to [0, +\infty)$ be measurable. Let $\alpha : \mathcal{P} \to (0, +\infty)$ and $\beta : \mathcal{P} \to \mathbb{R}$. Assume that $\sup_{P\in\mathcal{P}} \frac{\beta(P)}{\alpha(P)} < \infty$ and $\psi(x) \to \infty$ as $\|x\| \to \infty$. Assume that for every $P \in \mathcal{P}$,

$$PV \le V - \alpha(P)\psi + \beta(P)\,.$$

Then, the following holds.

i) The family $\bigcup_{P\in\mathcal{P}} \mathcal{I}(P)$ is tight. Moreover, $\sup_{\pi\in\mathcal{I}(\mathcal{P})} \pi(\psi) < +\infty$.

ii) For every $\nu \in \mathcal{M}(\mathsf{X})$ s.t. $\nu(V) < \infty$, every $P \in \mathcal{P}$, $\{\mathbb{E}^{\nu,P}\Lambda_n\,,\, n \in \mathbb{N}\}$ is tight. Moreover, $\sup_{n\in\mathbb{N}} \mathbb{E}^{\nu,P}\Lambda_n(\psi) < \infty$.

Proof. For each $P \in \mathcal{P}$, PV is everywhere finite by assumption. Moreover,

$$\sum_{k=0}^{n} P^{k+1}V \leq \sum_{k=0}^{n} P^k V - \alpha(P) \sum_{k=0}^{n} P^k \psi + (n+1)\beta(P)\,.$$

Using that $V \geq 0$ and $\alpha(P) > 0$,

$$\frac{1}{n+1} \sum_{k=0}^{n} P^k \psi \leq \frac{V}{\alpha(P)(n+1)} + c\,,$$

where $c := \sup_{P \in \mathcal{P}} \beta(P)/\alpha(P)$ is finite. For any $M > 0$,

$$\begin{aligned} \frac{1}{n+1} \sum_{k=0}^{n} P^k(\psi \wedge M) &\leq \left(\frac{1}{n+1} \sum_{k=0}^{n} P^k \psi \right) \wedge M \\ &\leq \left(\frac{V}{\alpha(P)(n+1)} + c \right) \wedge M\,. \end{aligned} \tag{3.34}$$

Set $\pi \in \mathcal{I}(\mathcal{P})$, and consider $P \in \mathcal{P}$ such that $\pi = \pi P$. Inequality (3.34) implies that for every n,

$$\pi(\psi \wedge M) \leq \pi \left(\left(\frac{V}{\alpha(P)(n+1)} + c \right) \wedge M \right)\,.$$

By Lebesgue's dominated convergence theorem, $\pi(\psi \wedge M) \leq c$. Letting $M \to \infty$ yields $\pi(\psi) \leq c$. The tightness of $\mathcal{I}(\mathcal{P})$ follows from the convergence of $\psi(x)$ to ∞ as $\|x\| \to \infty$. Setting $M = +\infty$ in (3.34), and integrating w.r.t. ν, we obtain

$$\mathbb{E}^{\nu,P} \Lambda_n(\psi) \leq \frac{\nu(V)}{(n+1)\alpha(P)} + c\,,$$

which proves the second point. □

Proposition 3.8.8. We posit the assumptions of Prop. 3.8.7. Then,

1. The family $\mathscr{I}(\mathcal{P}) := \bigcup_{P \in \mathcal{P}} \mathscr{I}(P)$ is tight;
2. $\{\mathbb{P}^{\nu,P} \Lambda_n^{-1}\,,\ n \in \mathbb{N}\}$ is tight.

Proof. For every $\mathfrak{m} \in \mathscr{I}(\mathcal{P})$, it is easy to see that $e(\mathfrak{m}) \in \mathcal{I}(\mathcal{P})$. Thus, $\{e(\mathfrak{m}) : \mathfrak{m} \in \mathscr{I}(\mathcal{P})\}$ is tight by Prop. 3.8.7. By Lem. 3.8.3, $\mathscr{I}(\mathcal{P})$ is tight. The second point follows from the equality $\mathbb{E}^{\nu,P} \Lambda_n = e(\mathbb{P}^{\nu,P} \Lambda_n^{-1})$ along with Prop. 3.8.7 and Lem. 3.8.3. □

3.8.4 Main Proof

By continuity of $h_\gamma(s,\cdot)$ for every $s \in \Xi$, $\gamma \in (0, \gamma_0)$, the transition kernel P_γ is Feller. By Prop. 3.8.7 and Eq. (3.18), we have $\sup_n \mathbb{E}^{\nu,\gamma} \Lambda_n(\varphi \circ f) < \infty$ which, by de la Vallée-Poussin's criterion for uniform integrability, implies

$$\lim_{K \to \infty} \sup_n \mathbb{E}^{\nu,\gamma} \Lambda_n(\|f\| \mathbb{1}_{\|f\| > K}) = 0\,. \tag{3.35}$$

In particular, the quantity $\mathbb{E}^{\nu,\gamma} \Lambda_n(f) = \mathbb{E}^{\nu,\gamma}(F_n)$ is well-defined.

We now prove the statement (3.19). By contradiction, assume that for some $\delta > 0$, there exists a positive sequence $\gamma_j \to 0$, such that for all $j \in \mathbb{N}$, $\limsup_{n\to\infty} d\left(\mathbb{E}^{\nu,\gamma_j}\Lambda_n(f), \mathcal{S}_f\right) > \delta$. For every j, there exists an increasing sequence of integers $(\varphi_n^j, n \in \mathbb{N})$ converging to $+\infty$ s.t.

$$\forall n,\ d\left(\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(f), \mathcal{S}_f\right) > \delta\,. \tag{3.36}$$

By Prop. 3.8.7, the sequence $(\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}, n \in \mathbb{N})$ is tight. By Prokhorov's theorem and Prop. 3.8.6, there exists $\pi_j \in \mathcal{I}(P_{\gamma_j})$ such that, as n tends to infinity, $\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j} \Rightarrow \pi_j$ along some subsequence. By the uniform integrability condition (3.35), $\pi_j(\|f\|) < \infty$ and $\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(f) \to \pi_j(f)$ as n tends to infinity, along the latter subsequence. By Eq. (3.36), for all $j \in \mathbb{N}$, $d(\pi_j(f), \mathcal{S}_f) \geq \delta$. By Prop. 3.8.7, $\sup_{\pi\in\mathcal{I}(\mathcal{P})} \pi(\psi) < +\infty$. Since $\varphi \circ f \leq M(1+\psi)$, de la Vallée-Poussin's criterion again implies that

$$\sup_{\pi\in\mathcal{I}(\mathcal{P})} \pi(\|f\|\mathbb{1}_{\|f\|>K}) < \infty\,. \tag{3.37}$$

Also by Prop. 3.8.7, the sequence (π_j) is tight. Thus $\pi_j \Rightarrow \pi$ along some subsequence, for some measure π which, by Prop. 3.5.2, is invariant for Φ. The uniform integrability condition (3.37) implies that $\pi(\|f\|) < \infty$ (hence, the set $\mathcal{S}_f$ is non-empty) and $\pi_j(f) \to \pi(f)$ as j tends to infinity, along the above subsequence. This shows that $d(\pi(f), \mathcal{S}_f) > \delta$, which is absurd. The statement (3.19) holds true (and in particular, $\mathcal{S}_f$ must be non-empty).

The proof of the statement (3.17) follows the same line, by replacing f with the function $\mathbb{1}_{\overline{\mathcal{U}_\varepsilon}}$. We briefly explain how the proof adapts, without repeating all the arguments. In this case, $\mathcal{S}_{\mathbb{1}_{\mathcal{U}_\varepsilon^c}}$ is the singleton $\{0\}$, and Eq. (3.36) reads $\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(\mathcal{U}_\varepsilon^c) > \delta$. By the Portmanteau theorem, $\limsup_n \mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(\mathcal{U}_\varepsilon^c) \leq \pi_j(\mathcal{U}_\varepsilon^c)$ where the $\limsup$ is taken along some subsequence. The contradiction follows from the fact that $\limsup \pi_j(\mathcal{U}_\varepsilon^c) \leq \pi(\overline{\mathcal{U}_\varepsilon^c}) = 0$ (where the $\limsup$ is again taken along the relevant subsequence).

We prove the statement (3.20). Assume by contradiction that for some (other) sequence $\gamma_j \to 0$, $\limsup_{n\to\infty} \mathbb{P}^{\nu,\gamma_j}\left(d\left(\Lambda_n(f), \mathcal{S}_f\right) \geq \varepsilon\right) > \delta$. For every j, there exists a sequence $(\varphi_n^j, n \in \mathbb{N})$ s.t.

$$\forall n,\ \mathbb{P}^{\nu,\gamma_j}\left(d\left(\Lambda_{\varphi_n^j}(f), \mathcal{S}_f\right) \geq \varepsilon\right) > \delta\,. \tag{3.38}$$

By Prop. 3.8.8, $(\mathbb{P}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}^{-1}, n \in \mathbb{N})$ is tight, one can extract a further subsequence (which we still denote by (φ_n^j) for simplicity) s.t. $\mathbb{P}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}^{-1}$ converges narrowly to a measure $\mathfrak{m}_j$ as n tends to infinity, which, by Prop. 3.8.6, satisfies $\mathfrak{m}_j \in \mathscr{I}(P_{\gamma_j})$. Noting that $e(\mathbb{P}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}^{-1}) = \mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}$ and recalling Eq. (3.35), Lem. 3.8.5 implies that $\nu'(\|f\|) < \infty$ for all ν' $\mathfrak{m}_j$-a.e., and $\mathbb{P}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}^{-1}\mathcal{T}_f^{-1} \Rightarrow \mathfrak{m}_j\mathcal{T}_f^{-1}$, where we recall that $\mathcal{T}_f(\nu') := \nu'(f)$ for all ν' s.t. $\nu'(\|f\|) < \infty$. As $(\mathcal{S}_f)_\varepsilon^c$ is a closed set,

$$\begin{aligned}\mathfrak{m}_j\mathcal{T}_f^{-1}((\mathcal{S}_f)_\varepsilon^c) &\geq \limsup_n \mathbb{P}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}^{-1}\mathcal{T}_f^{-1}((\mathcal{S}_f)_\varepsilon^c)\\ &= \limsup_n \mathbb{P}^{\nu,\gamma_j}\left(d\left(\Lambda_{\varphi_n^j}(f), \mathcal{S}_f\right) \geq \varepsilon\right) > \delta\,.\end{aligned}$$

By Prop. 3.8.7, $(\mathfrak{m}_j)$ is tight, and one can extract a subsequence (still denoted by $(\mathfrak{m}_j)$) along which $\mathfrak{m}_j \Rightarrow \mathfrak{m}$ for some measure $\mathfrak{m}$ which, by Prop. 3.5.2, belongs to $\mathscr{I}(\Phi)$. For every j, $e(\mathfrak{m}_j) \in \mathcal{I}(P_{\gamma_j})$. By the uniform integrability condition (3.37), one can apply Lem. 3.8.5 to the sequence $(\mathfrak{m}_j)$. We deduce that $\nu'(\|f\|) < \infty$ for all ν' $\mathfrak{m}$-a.e. and $\mathfrak{m}_j\mathcal{T}_f^{-1} \Rightarrow \mathfrak{m}\mathcal{T}_f^{-1}$. In particular,

$$\mathfrak{m}\mathcal{T}_f^{-1}((\mathcal{S}_f)_\varepsilon^c) \geq \limsup_j \mathfrak{m}_j\mathcal{T}_f^{-1}((\mathcal{S}_f)_\varepsilon^c) > \delta\,.$$

Since $\mathfrak{m} \in \mathscr{I}(\Phi)$, it holds that $\mathfrak{m}\mathcal{T}_f^{-1}((\mathcal{S}_f)_\varepsilon^c) = 0$, hence a contradiction.

3.9 Proofs of Th. 3.5.4 and 3.5.5

3.9.1 Proof of Th. 3.5.4

In this proof, we set $L = L_{\mathrm{av}(\Phi)}$ to simplify the notations. It is straightforward to show that the identity mapping $f(x) = x$ satisfies the hypotheses of Th. 3.5.3 with $\varphi = \psi$. Hence, it is sufficient to prove that $\mathcal{S}_f$ is a subset of $\overline{\mathrm{co}}(L)$, the closed convex hull of L. Choose $q \in \mathcal{S}_I$ and let $q = \int x d\pi(x)$ for some $\pi \in \mathcal{I}(\Phi)$ admitting a first order moment. There exists a Θ-invariant measure $\upsilon \in \mathcal{M}(C(\mathbb{R}_+, \mathsf{X}))$ s.t. $\mathrm{supp}(\upsilon) \subset \Phi(\mathsf{X})$ and $\upsilon p_0^{-1} = \pi$. We remark that for all $t > 0$,

$$q = \upsilon(p_0) = \upsilon(p_t) = \upsilon(p_t \circ \mathrm{av})\,, \tag{3.39}$$

where the second identity is due to the shift-invariance of υ, and the last one uses Fubini's theorem. Again by the shift-invariance of υ, the family $\{p_t, t > 0\}$ is uniformly integrable w.r.t. υ. By Tonelli's theorem, $\sup_{t>0} \upsilon(\|p_t \circ \mathrm{av}\, \|\mathbb{1}_S) \leq \sup_{t>0} \upsilon(\|p_t\| \mathbb{1}_S)$ for every $S \in \mathscr{B}(C(\mathbb{R}_+, \mathsf{X}))$. Hence, the family $\{p_t \circ \mathrm{av}, t > 0\}$ is υ-uniformly integrable as well. In particular, $\{p_t \circ \mathrm{av}, t > 0\}$ is tight in $(C(\mathbb{R}_+, \mathsf{X}), \mathscr{B}(C(\mathbb{R}_+, \mathsf{X})), \upsilon)$. By Prokhorov's theorem, there exists a sequence $t_n \to \infty$ and a measurable function $g : C(\mathbb{R}_+, \mathsf{X}) \to \mathsf{X}$ such that $p_{t_n} \circ \mathrm{av}$ converges in distribution to g as $n \to \infty$. By uniform integrability, $\upsilon(p_{t_n} \circ \mathrm{av}) \to \upsilon(g)$. Eq. (3.39) finally implies that

$$q = \upsilon(g)\,.$$

In order to complete the proof, it is sufficient to show that $g(\mathsf{x}) \in \overline{L}$ for every x υ-a.e., because $\overline{\mathrm{co}}(L) \subset \mathrm{co}(\overline{L})$. Set $\varepsilon > 0$ and $\delta > 0$. By the tightness of the r.v. $(p_{t_n} \circ \mathrm{av}, n \in \mathbb{N})$, choose a compact set K such that $\upsilon(p_{t_n} \circ \mathrm{av})^{-1}(K^c) \leq \delta$ for all n. As $\overline{L_\varepsilon}^c$ is an open set, one has

$$\upsilon g^{-1}(\overline{L_\varepsilon}^c) \leq \lim_n \upsilon(p_{t_n} \circ \mathrm{av})^{-1}(\overline{L_\varepsilon}^c) \leq \lim_n \upsilon(p_{t_n} \circ \mathrm{av})^{-1}(\overline{L_\varepsilon}^c \cap K) + \delta\,.$$

Let $\mathsf{x} \in \Phi(\mathsf{X})$ be fixed. By contradiction, suppose that $\mathbb{1}_{\overline{L_\varepsilon}^c \cap K}(p_{t_n}(\mathrm{av}(\mathsf{x})))$ does not converge to zero. Then, $p_{t_n}(\mathrm{av}(\mathsf{x})) \in \overline{L_\varepsilon}^c \cap K$ for every n along some subsequence. As K is compact, one extract a subsequence, still denoted by t_n, s.t. $p_{t_n}(\mathrm{av}(\mathsf{x}))$ converges. The corresponding limit must belong to the closed set L_ε^c, but must also belong to L by definition of x. This proves that $\mathbb{1}_{L_\varepsilon^c \cap K}(p_{t_n} \circ \mathrm{av}(\mathsf{x}))$ converges to zero for all $x \in \Phi(\mathsf{X})$. As $\mathrm{supp}(\upsilon) \subset \Phi(\mathsf{X})$, $\mathbb{1}_{\overline{L_\varepsilon}^c \cap K}(p_{t_n} \circ \mathrm{av})$ converges to zero υ-a.s. By the dominated convergence theorem, we obtain that $\upsilon g^{-1}(\overline{L_\varepsilon}^c) \leq \delta$. Letting $\delta \to 0$ we obtain that $\upsilon g^{-1}(\overline{L_\varepsilon}^c) = 0$. Hence, $g(\mathsf{x}) \in \overline{L}$ for all x υ-a.e. The proof is complete.

3.9.2 Proof of Th. 3.5.5

Recall the definition $\mathcal{U} := \bigcup_{\pi \in \mathcal{I}(\Phi)} \mathrm{supp}(\pi)$. By Th. 3.5.3, for all $\varepsilon > 0$,

$$\limsup_{n \to \infty} \mathbb{E}^{\nu,\gamma} \Lambda_n(\mathcal{U}_\varepsilon^c) \xrightarrow[\gamma \to 0]{} 0,$$

where Λ_n is the random measure given by (3.32). By Th. 3.4.1, $\mathrm{supp}(\pi) \subset \mathsf{BC}_\Phi$ for each $\pi \in \mathcal{I}(\Phi)$. Thus, $\mathcal{U}_\varepsilon \subset (\mathsf{BC}_\Phi)_\varepsilon$. Hence, $\limsup_n \mathbb{E}^{\nu,\gamma} \Lambda_n(((\mathsf{BC}_\Phi)_\varepsilon)^c) \to 0$ as $\gamma \to 0$. This completes the proof.

3.10 Applications

In this section, we return to the Examples 4 and 5 of Sec. 3.2.

3.10.1 Non-Convex Optimization

Consider the algorithm (3.9) to solve problem (3.8) where $\ell : \Xi \times \mathsf{X} \to \mathbb{R}$, $r : \mathsf{X} \to \mathbb{R}$ and ξ is a random variable over a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with values in the measurable space $(\Xi, \mathscr{G})$ and with distribution μ. Assume that $\ell(s, \cdot)$ is continuously differentiable for every $s \in \Xi$, that $\ell(\cdot, x)$ is μ-integrable for every $x \in \mathsf{X}$ and that r is a convex and lower semicontinuous function. We assume that for every compact subset K of X, there exists $\varepsilon_K > 0$ s.t.

$$\sup_{x \in K} \int \|\nabla \ell(s, x)\|^{1+\varepsilon_K} \mu(ds) < \infty \,. \tag{3.40}$$

Define $L(x) := \mathbb{E}_\xi(\ell(\xi, x))$. Under Condition (3.40), it is easy to check that L is differentiable, and that $\nabla L(x) = \int \nabla \ell(s, x) \mu(ds)$. From now on, we assume moreover that ∇L is Lipschitz continuous. Condition (3.40) and the Lipschitz continuity of ∇L are satisfied under the following assumption : there exists $\varepsilon > 0$ such that $\nabla \ell(s, \cdot)$ is $C(s)$-Lipschitz continuous for μ-a.e s, where $C^{1+\varepsilon}$ is μ-integrable and there exists $x_\star \in \mathsf{X}$ such that $\|\nabla \ell(\cdot, x_\star)\|^{1+\varepsilon}$ is μ-integrable. Note that Lipschitz conditions of this type are usually unavoidable regarding the so-called explicit part (or forward part) of the proximal gradient algorithm (3.9). Letting $H(s, x) := -\nabla \ell(s, x) - \partial r(x)$, it holds that $H(\cdot, x)$ is proper, μ-integrable and usc [97], and that the corresponding selection integral $\mathsf{H}(x) := \int H(s, x) \mu(ds)$ is given by

$$\mathsf{H}(x) = -\nabla L(x) - \partial r(x) \,.$$

By [36, Th. 3.17, Remark 3.14], for every $a \in \mathsf{X}$, the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(x(t))$ admits a unique solution on $[0, +\infty)$ s.t. $\mathsf{x}(0) = a$.

The iterates x_n given by (3.9) satisfy (3.3) where $h_\gamma(s, x) := \gamma^{-1}(\mathrm{prox}_{\gamma r}(x - \gamma \nabla \ell(s, x)) - x)$. Moreover, the map h_γ satisfies Assumption (RM). Recall that

$$\begin{aligned} h_\gamma(s, x) &= -\nabla r_\gamma(x - \gamma \nabla \ell(s, x)) - \nabla \ell(s, x) && (3.41) \\ &\in -\partial r(\mathrm{prox}_{\gamma r}(x - \gamma \nabla \ell(s, x))) - \nabla \ell(s, x) \\ &\in -\partial r(x - \gamma h_\gamma(s, x)) - \nabla \ell(s, x) \,. && (3.42) \end{aligned}$$

In order to show that Assumption (RM)-ii) is satisfied, we need some estimate on $\|h_\gamma(s, x)\|$. Using Eq. (3.41) and the fact that ∇r_γ is γ^{-1}-Lipschitz continuous (see Sec. 2.2.1), we obtain that

$$\begin{aligned} \|h_\gamma(s, x)\| &\le \|\nabla r_\gamma(x)\| + 2\|\nabla \ell(s, x)\| \\ &\le \|\partial_0 r(x)\| + 2\|\nabla \ell(s, x)\| \,, && (3.43) \end{aligned}$$

where $\partial_0 r(x)$ the least norm element in $\partial r(x)$ for every $x \in \mathsf{X}$ (see Sec. 2.2.1). As $\partial_0 r$ is locally bounded and ∂r is usc, it follows from Eq. (3.42) that Assumption (RM)-ii) is satisfied. The estimate (3.43) also yields Assumption (RM)-vi). As a conclusion, Assumption (RM) is satisfied. In particular, the statement of Th. 3.5.1 holds.

To show that Assumption (PH) is satisfied, we first recall the Lojasiewicz (L) condition studied in [30]. We shall use the formulation of [72] which is a particular case of [30]. Assume that L is differentiable with a C-Lipschitz continuous gradient. We say that L and r satisfy the (L) condition with constant $\beta > 0$ if for every $x \in \mathsf{X}$,

$$\frac{1}{2} D_{L,r}(x, C) \ge \beta \left[(L + r)(x) - \min(L + r)\right]$$

where

$$D_{L,r}(x,C) := -2C \min_{y\in \mathsf{X}} \left[\langle \nabla L(x), y-x\rangle + \frac{C}{2}\|y-x\|^2 + r(y) - r(x) \right].$$

The (L) condition helps to prove the convergence of the (deterministic) proximal gradient algorithm applied to the (deterministic) problem of minimizing the sum $L+r$. We refer to [30] for practical cases where the (L) condition is satisfied. In our stochastic setting, we introduce the Stochastic Lojasiewicz condition (SL). We say that ℓ and r satisfy the (SL) condition if there exists $\beta > 0$ such that for every $x \in \mathsf{X}$,

$$\frac{1}{2}\int D_{\ell(s,\cdot),r}\left(x, \frac{1}{\gamma}\right)\mu(ds) \geq \beta\left[(L+r)(x) - \min(L+r)\right]$$

for all $\gamma \leq \frac{1}{2C}$. Note that (SL) is satisfied if for every $s \in \Xi$, $\ell(s,\cdot)$ and r satisfy the (L) condition with constant β. In the sequel, we assume that for every $x \in \mathsf{X}$, the random variable $\|\ell(x,\xi)\|$ is square integrable and denote by $W(x)$ its variance.

Proposition 3.10.1. Assume that the (SL) condition is satisfied, that $\gamma \leq \frac{1}{2C}$ and that

$$\beta(L(x)+r(x)) - W(x) \longrightarrow_{\|x\|\to+\infty} +\infty.$$

Then (PH) is satisfied.

Proof. Using (sub)differential calculus, it is easy to show that for every $n \in \mathbb{N}$,

$$x + \gamma h_\gamma(s,x) = \arg\min_{y\in\mathsf{X}} \left[\langle \nabla \ell(s,x), y-x\rangle + \frac{1}{2\gamma}\|y-x\|^2 + r(y) - r(x) \right].$$

Since ∇L is C-Lipschitz continuous, recalling that $\frac{\gamma^2 C}{2} - \frac{\gamma}{2} \leq -\frac{\gamma}{4}$,

$$\begin{aligned}
(L+r)(x+\gamma h_\gamma(s,x)) &= L(x+\gamma h_\gamma(s,x)) + r(x) + r(x+\gamma h_\gamma(s,x)) - r(x) \\
&\leq (L+r)(x) + \langle \nabla L(x), \gamma h_\gamma(s,x)\rangle + \frac{1}{2\gamma}\|\gamma h_\gamma(s,x)\|^2 \\
&\quad + \left(\frac{C}{2} - \frac{1}{2\gamma}\right)\|\gamma h_\gamma(s,x)\|^2 + r(x+\gamma h_\gamma(s,x)) - r(x) \\
&\leq (L+r)(x) + \langle \nabla \ell(s,x), \gamma h_\gamma(s,x)\rangle + \frac{1}{2\gamma}\|\gamma h_\gamma(s,x)\|^2 \\
&\quad + \langle \nabla L(x) - \nabla \ell(s,x), \gamma h_\gamma(s,x)\rangle + r(x+\gamma h_\gamma(s,x)) - r(x) \\
&\quad - \frac{\gamma}{4}\|h_\gamma(s,x)\|^2 \\
&\leq (L+r)(x) - \frac{\gamma}{2} D_{\ell(s,\cdot),r}(x,1/\gamma) - \frac{\gamma}{4}\|h_\gamma(s,x)\|^2 \\
&\quad - \gamma\langle \nabla \ell(s,x) - \nabla L(x), h_\gamma(s,x)\rangle. \qquad (3.44)
\end{aligned}$$

Using $|\langle a,b\rangle| \leq \|a\|^2 + \frac{1}{4}\|b\|^2$ in the last inner product, we finally have

$$(L+r)(x+\gamma h_\gamma(s,x)) \leq (L+r)(x) - \frac{\gamma}{2} D_{\ell(s,\cdot),r}(x,1/\gamma) + \gamma\|\nabla \ell(s,x) - \nabla L(x)\|^2. \qquad (3.45)$$

Integrating with respect to μ, we obtain

$$\begin{aligned}
\int (L+r)(x+\gamma h_\gamma(s,x))\mu(ds) &\leq (L+r)(x) + \gamma W(x) \\
&\quad - \gamma\beta\left((L+r)(x) - \min(L+r)\right).
\end{aligned}$$

Finally, the condition (PH) is satisfied with $\alpha(\gamma) = \gamma$, $\beta(\gamma) = 0$, $V = L + r - \min L + r$ and $\psi = \beta V - W$. □

Note that the assumptions of Prop. 3.10.1 are satisfied if the (SL) condition is satisfied, $L(x) + r(x) \to_{\|x\| \to +\infty} +\infty$ and the "variance" function W is bounded.

3.10.2 Fluid Limit of a System of Parallel Queues

Consider a positive integer N. We now apply the results of this chapter to the dynamical system described in Example 5 above. For a given $\gamma > 0$, the transition kernel P_γ of the Markov chain (x_n) whose entries are given by Eq. (3.10) is defined on $\gamma\mathbb{N}^N \times 2^{\gamma\mathbb{N}^N}$. This requires some small adaptations of the statements of the main results that we keep confined to this paragraph for the chapter readability. The limit behavior of the interpolated process (see Th. 3.5.1) is described by the following Prop., which has an analogue in [61]:

Proposition 3.10.2. For every compact set $K \subset \mathbb{R}^N$, the family $\{\mathbb{P}^{a,\gamma}\mathsf{X}_\gamma^{-1}, a \in K \cap \gamma\mathbb{N}^N, 0 < \gamma < \gamma_0\}$ is tight. Moreover, for every $\varepsilon > 0$,

$$\sup_{a \in K \cap \gamma\mathbb{N}^N} \mathbb{P}^{a,\gamma}\left(d(\mathsf{X}_\gamma, \Phi_{\mathsf{H}}(K)) > \varepsilon\right) \xrightarrow[\gamma \to 0]{} 0\,,$$

where the set-valued map H is given by (3.12).

Proof. To prove this Prop., we mainly need to check that Assumption (RM) is verified. We recall that the Markov chain (x_n) given by Eq. (3.10) admits the representation (3.2), where the function $g_\gamma = (g_\gamma^1, \ldots, g_\gamma^N)$ is given by (3.11). If we set $h_\gamma(s,x) = g_\gamma(x)$ (the fact that g_γ is defined on $\gamma\mathbb{N}^N$ instead of $\mathbb{R}_+^N$ is irrelevant), then for each sequence $(u_n, \gamma_n) \to (u^\star, 0)$ with $u_n \in \gamma_n\mathbb{N}^N$ and $x^\star \in \mathbb{R}_+^N$, it holds that $g_{\gamma_n}(u_n) \to \mathsf{H}(u^\star)$. Thus, Assumption (RM)–i) is verified with $H(s,x) = \mathsf{H}(x)$. Assumptions (RM)–ii) to (RM)–iv) are obviously verified. Since the set-valued map H satisfies the condition (2.1), Assumption (RM)–v) is verified. Finally, the finiteness assumption (3.15) with $\varepsilon_K = 2$ follows from the existence of second moments for the A_n^k, and (3.16) is immediate. The rest of the proof follows word for word the proof of Th. 3.5.1. □

The long run behavior of the iterates is provided by the following Prop.:

Proposition 3.10.3. Let $\nu \in \mathcal{M}(\mathbb{R}_+^N)$ be such that $\nu(\|\cdot\|^2) < \infty$. For each $\gamma > 0$, define the probability measure ν_γ on $\gamma\mathbb{N}^N$ as

$$\nu_\gamma(\{\gamma i_1, \gamma i_2, \ldots, \gamma i_N\}) = \nu(\gamma(i_1 - 1/2, i_1 + 1/2] \times \cdots \times \gamma(i_N - 1/2, i_N + 1/2])\,.$$

If Condition (3.13) is satisfied, then for all $\varepsilon > 0$,

$$\limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu_\gamma, \gamma}\left(d\left(X_k\,, 0\right) \geq \varepsilon\right) \xrightarrow[\gamma \to 0]{} 0\,.$$

To prove this Prop., we essentially show that the assumptions of Th. 3.5.5 are satisfied. In the course of the proof, we shall establish the existence of the (PH) criterion with a function ψ having a linear growth. With some more work, it is possible to obtain a (PH) criterion with a faster than linear growth for ψ, allowing to obtain the ergodic convergence as shown in Th. 3.5.4. This point will not be detailed here.

Proof. Considering the space $\gamma\mathbb{N}^N$ as a metric space equipped with the discrete topology, any probability transition kernel on $\gamma\mathbb{N}^N \times 2^{\gamma\mathbb{N}^N}$ is trivially Feller. Thus, Prop. 3.8.6 holds when letting $P = P_\gamma$ and $\nu \in \mathcal{M}(\gamma\mathbb{N}^N)$. Let us check that Assumption (PH) is verified if the stability condition (3.13) is satisfied. Let

$$V : \mathbb{R}_+^N \to \mathbb{R}_+, \quad x = (x^1, \dots, x^N) \mapsto \Big(\sum_{k=1}^N x^k/\eta^k\Big)^2 .$$

Given $1 \le k, \ell \le N$, define $f(x) = x^k x^\ell$ on $\gamma\mathbb{N}^2$. Using Eq. (3.10), the i.i.d property of the process $((A_n^1, \dots, A_n^N, B_n^1, \dots, B_n^N), n \in \mathbb{N})$ and the finiteness of the second moments of the A_n^k, we obtain

$$\begin{aligned}(P_\gamma f)(x) \le x^k x^\ell &- \gamma x^k \left(\eta^\ell \mathbb{1}_{\{x^\ell > 0,\, x^{\ell-1} = \cdots = x^1 = 0\}} - \lambda^\ell\right) \\ &- \gamma x^\ell \left(\eta^k \mathbb{1}_{\{x^k > 0,\, x^{k-1} = \cdots = x^1 = 0\}} - \lambda^k\right) + \gamma^2 C\,,\end{aligned}$$

where C is a positive constant. Thus, when $x \in \gamma\mathbb{N}^N$,

$$(P_\gamma V)(x) \le V(x) - 2\gamma \sum_{k=1}^N x^k/\eta^k \sum_{\ell=1}^N \left(\mathbb{1}_{\{x^\ell > 0,\, x^{\ell-1} = \cdots = x^1 = 0\}} - \lambda^\ell/\eta^\ell\right) + \gamma^2 C,$$

after modifying the constant C if necessary. If $x \neq 0$, then one and only one of the $\mathbb{1}_{\{x^\ell > 0,\, x^{\ell-1} = \cdots = x^1 = 0\}}$ is equal to one. Therefore, $(P_\gamma V)(x) \le V(x) - \gamma\psi(x) + \gamma^2 C$, where

$$\psi(x) = 2\Big(1 - \sum_{\ell=1}^N \lambda^\ell/\eta^\ell\Big) \sum_{k=1}^N x^k/\eta^k\,.$$

As a consequence, when Condition (3.13) is satisfied, the function ψ is coercive, and one can straightforwardly check that the statements of Prop. 3.8.7–i) and Prop. 3.8.7–ii) hold true under minor modifications, namely, $\bigcup_{P\in\mathcal{P}} \mathcal{I}(P)$ is tight in $\mathcal{M}(\mathbb{R}_+^N)$, since $\sup_{\pi\in\mathcal{I}(\mathcal{P})} \pi(\psi) < +\infty$, where $\mathcal{P} = \{P_\gamma\}_{\gamma\in(0,\gamma_0)}$. Moreover, for every $\nu \in \mathcal{M}(\mathbb{R}_+^N)$ s.t. $\nu(\|\cdot\|^2) < \infty$ and every $P \in \mathcal{P}$, $\{\mathbb{E}^{\nu,P}\Lambda_n\,, n \in \mathbb{N}\}$ is tight, since $\sup_{n\in\mathbb{N}} \mathbb{E}^{\nu,P}\Lambda_n(\psi) < \infty$. We can now follow the proof of Th. 3.5.5. Doing so, all it remains to show is that the Birkhoff center of the flow Φ_{H} is reduced to $\{0\}$. This follows from the fact that when Condition (3.13) is satisfied, all the trajectories of the flow Φ_{H} converge to zero, as shown in [61, § 3.2]. □

Chapter 4

A constant step Forward-Backward algorithm involving random maximal monotone operators

In this chapter, we continue the study of the stochastic approximation framework (1.4). We consider the case of a Differential Inclusion induced by a maximal monotone operator, see Sec. 1.4. The Forward-Backward algorithm is a classical method to find a zero of a monotone operator. We study a stochastic Forward-Backward algorithm with a constant step. At each time step, this algorithm involves an independent copy of a couple of random maximal monotone operators. As a first result, we show that the interpolated process obtained from the iterates converges narrowly in the small step regime to the solution of the DI induced by the sum of the mean operators. In order to control the long term behavior of the iterates, a stability result is needed in addition. To this end, the sequence of the iterates is seen as a homogeneous Feller Markov chain whose transition kernel is parameterized by the algorithm step size. We show that the cluster points of the Markov chains invariant measures in the small step regime are invariant for the semiflow induced by the DI. Conclusions regarding the long run behavior of the iterates for small steps follows from this fact. We also show that when the sum of the mean operators is demipositive, the probabilities that the iterates are away from the set of zeros of this sum are small in Cesàro mean. We study the ergodic behavior of these iterates as well. Finally, we consider applications of the proposed algorithm. In particular we perform a detailed analysis of the random proximal gradient algorithm with constant step.

4.1 Introduction

Given two maximal monotone operators A and B on the Euclidean space X, where B is single valued, the Forward-Backward splitting algorithm is an iterative algorithm for finding a zero of the sum operator $\mathsf{A}+\mathsf{B}$. It reads

$$x_{n+1} = (I+\gamma\mathsf{A})^{-1}(x_n - \gamma\mathsf{B}(x_n))\,, \tag{4.1}$$

where γ is a positive step. This algorithm consists in a forward step $(I-\gamma\mathsf{B})(x_n)$ followed by a backward step, where the resolvent $(I+\gamma\mathsf{A})^{-1}$ of A, known to be single valued as A is maximal monotone, is applied to the output of the former. When B satisfies a so called cocoercivity condition, and when the step γ is small enough, the convergence of the algorithm towards a zero of $\mathsf{A}+\mathsf{B}$ (provided it exists) is a well established fact [12, Ch. 25]. In the field of convex optimization, this algorithm can be used

to find a minimizer of the sum of two real functions $F+G$ on X, where F is a convex function which is defined on the whole X and which has a Lipschitz gradient, and where G is a convex, proper, and lower semi continuous (lsc) function ($G \in \Gamma_0(\mathsf{X})$). In this case, the Forward-Backward algorithm is known as the proximal gradient algorithm, and is written as $x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n))$, where $\mathrm{prox}_{\gamma G} := (I + \gamma \partial G)^{-1}$ is Moreau's proximity operator of γG.

In this chapter, we are interested in the situation where the operators A and B are replaced with random maximal monotone operators. Consider two random monotone operators (see Sec. 2.3) $A, B : \Xi \to \mathscr{M}(\mathsf{X})$ defined on a probability space $(\Xi, \mathscr{G}, \mu)$, and let $(\xi_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed (i.i.d) random variables from some probability space to $(\Xi, \mathscr{G})$ with the probability distribution μ. Assuming that for every $s \in \Xi$, $B(s)$ is a single-valued operator defined on the whole X, we examine the stochastic version of the Forward-Backward algorithm

$$x_{n+1} = (I + \gamma A(\xi_{n+1}))^{-1}(I - \gamma B(\xi_{n+1}))x_n\,, \quad \gamma > 0\,. \tag{4.2}$$

Our aim is to study the dynamical behavior of this algorithm in the limit of the small steps γ, where the effect of the noise due to the ξ_n will be smoothened.

To give an application example for this algorithm, let us consider again the minimization problem of the sum $F+G$, and let us assume that these functions are unknown to the observer (or difficult to compute), and are written as $F(x) = \mathbb{E}_{\xi_1} f(\xi_1, x)$ and $G(x) = \mathbb{E}_{\xi_1} g(\xi_1, x)$. When the functions f and g are known with $f(\xi_1, \cdot)$ being convex differentiable, and $g(\xi_1, \cdot) \in \Gamma_0(\mathsf{X})$, and when an i.i.d sequence (ξ_n) is available, we can approximatively solve the minimization problem of $F+G$ by resorting to the stochastic proximal gradient algorithm $x_{n+1} = \mathrm{prox}_{\gamma g(\xi_{n+1},\cdot)}(x_n - \gamma \nabla_x f(\xi_{n+1}, x_n))$. Similar algorithms has been studied in [24, 106] with the additional assumption that the step size γ vanishes as n tends to infinity. The main asset of such vanishing step size algorithms is that the iterates (with or without averaging) converge almost surely as the iteration index goes to infinity. This chapter focuses on the case where the step size γ is fixed w.r.t. n. As we shall see below, convergence holds in a weaker sense in this case. Loosely speaking, the iterates fluctuate in a small neighborhood of the set of sought solutions, but do not converge in an almost sure sense as $n \to \infty$. Yet, constant step size algorithms have raised a great deal of attention in the signal processing and machine learning literature ([55]). First, they are known to reach a neighborhood of the solution in a fewer number of iterations than the decreasing step algorithms. Second, they are in practice able to adapt to non stationary or slowly changing environments, and thus track a possible changing set of solutions. This is particularly helpful in adaptive signal processing for instance.

In order to study the dynamical behavior of (4.2), we introduce the operators defined for every $x \in \mathsf{X}$ by

$$\mathcal{A}(x) = \int A(s)(x)\,\mu(ds) \quad \text{and} \quad \mathcal{B}(x) = \int B(s)(x)\,\mu(ds)\,,$$

where the first integral is a selection integral (see Sec. 2.3, Eq. (2.4)). Assuming that the monotone operator $\mathcal{A}+\mathcal{B}$ is maximal, we consider a in the domain of $\mathcal{A}+\mathcal{B}$, and the DI (see Sec. 2.2.2)

$$\begin{cases} \dot{\mathsf{x}}(t) & \in & -(\mathcal{A}+\mathcal{B})(\mathsf{x}(t)) \\ \mathsf{x}(0) & = & a. \end{cases} \tag{4.3}$$

Let $\mathsf{x}^{a,\gamma}(t)$ be the continuous random process obtained by assuming that the iterates x_n are distant apart by the time step γ, and by interpolating linearly these iterates. Then, the first step of the approach undertaken in this chapter is to show that $\mathsf{x}^{a,\gamma}$ shadows the solution of the DI for small γ, in the sense that it converges narrowly to this solution as $\gamma \to 0$ in the topology of convergence on the compact sets

of $\mathbb{R}_+$. The same idea is behind the so-called ODE method which is frequently used in the stochastic approximation literature (see [14, 73] or Sec. 1.3).

The compact convergence alone is not enough to control the long term behavior of the iterates. A stability result is needed. To that end, the second step of the approach is to view the sequence (x_n) as a homogeneous Feller Markov chain whose transition kernel is parameterized by γ. In this context, the aim is to show that the set of invariant measures for this kernel is non empty, and that the family of invariant measures obtained for all γ belonging to some interval $(0, \gamma_0]$ is tight. We shall obtain a general tightness criterion which will be made more explicit in a number of situations of interest involving random maximal monotone operators.

The narrow convergence of $\mathrm{x}^{a,\gamma}$, together with the tightness of the Markov chain invariant measures, lead to the invariance of the small γ cluster points of these invariant measures with respect to the semiflow induced by the DI (4.3) (see [64, 60, 16] for similar contexts). Using these results, it becomes possible to characterize the long run behavior of the iterates (x_n). In particular, the proximity of these iterates to the set of zeros $Z(\mathcal{A}+\mathcal{B})$ of $\mathcal{A}+\mathcal{B}$ is of obvious interest. First, we show that when the operator $\mathcal{A}+\mathcal{B}$ is *demipositive* [38], the probabilities that the iterates are away from $Z(\mathcal{A}+\mathcal{B})$ are small in Cesàro mean. Whether $\mathcal{A}+\mathcal{B}$ is demipositive or not, we can also characterize the ergodic behavior of the algorithm, showing that when γ is small, the partial sums $n^{-1}\sum_1^n x_k$ stay close to $Z(\mathcal{A}+\mathcal{B})$ with a high probability.

Stochastic approximations with differential inclusions were considered in [17] and in [58] from the dynamical systems viewpoint. The case where the DI is defined by a maximal monotone operator was studied in [22], [24], and [106]. Instances of the random proximal gradient algorithm were treated in *e.g.*, [2] or [104]. All these references dealt with the decreasing step case, which requires quite different tools from the constant step case. This case is considered in [48] (see also [47]), which relies on a Robbins-Siegmund like approach requiring summability assumptions on the random errors. The constant step case is also dealt with in [107] and in Chap. 3 for generic differential inclusions. In the present work, we follow the line of reasoning of Chap. 3, noting that the case where the DI is defined by a maximal monotone operator has many specificities. For instance, a maximal monotone operator is not upper semi continuous in general, as it was assumed for the differential inclusions studied in [107] and Chap. 3. Another difference lies in the fact that we consider here the case where the domains of the operators $A(s)$ can be different. Finally, to be more practical, the tightness criterion for the Markov chain invariant measures requires a quite specific treatment in the context of the maximal monotone operators.

We close this paragraph by mentioning [21], where one of the studied stochastic proximal gradient algorithms can be cast in the general framework of (4.2).

Chapter organization. Sec. 4.2 introduces the main algorithm. Sec. 4.3 provides our assumptions and states our main result about the long run behavior of the iterates. A brief sketch of the proof is also provided for convenience, the detailed arguments being postponed to the end of the chapter. Sec. 4.4 provides some illustrations of our results in particular cases. The monotone operators involved are assumed to be subdifferentials, hence covering the context of numerical optimization. Our assumptions are discussed at length in this scenario. The case when the monotone operators are linear maps is addressed as well. Sec. 4.5 analyzes the dynamical behavior of the iterates. It is shown that the piecewise linear interpolation of the iterates converges narrowly, uniformly on compact sets, to a solution of the DI. The result, which has its own interest, is the first key argument to establish the main theorem of Sec. 4.3. The second argument is provided in Sec. 4.6, where we characterize the cluster points of the invariant measures (indexed by the step size) of the Markov chain formed by the iterates. The Appendices 4.7 and 4.8 are devoted to the proofs relative to Sec. 4.4 and 4.5 respectively.

4.2 Background and problem statement

Consider $\mathsf{A} \in \mathscr{M}(\mathsf{X})$ and the semiflow $\Phi : \mathrm{cl}(\mathrm{dom}(\mathsf{A})) \times \mathbb{R}_+ \to \mathrm{cl}(\mathrm{dom}(\mathsf{A}))$ associated to A (see Sec. 2.2.2).

We recall some of the most important notions related with the dynamical behavior of the semiflow Φ. Denote as $\mathcal{M}(\mathsf{X})$ the space of probability measures on X equipped with its Borel σ-field $\mathscr{B}(\mathsf{X})$. An element $\pi \in \mathcal{M}(\mathsf{X})$ is called an invariant measure for Φ if $\pi = \pi\Phi(\cdot, t)^{-1}$ for every $t > 0$. The set of invariant measures for Φ will be denoted $\mathcal{I}(\Phi)$. The limit set of the trajectory $\Phi(x, \cdot)$ of the semiflow Φ starting at x is the set

$$L_{\Phi(x,\cdot)} := \bigcap_{t \geq 0} \mathrm{cl}\left(\Phi(x, [t, \infty))\right)$$

of the limits of the convergent subsequences $(\Phi(x, t_k))_k$ as $t_k \to \infty$. A point $x \in \mathrm{cl}(\mathrm{dom}\,\mathsf{A})$ is said recurrent if $x \in L_{\Phi(x,\cdot)}$. The Birkhoff center BC_Φ of Φ is

$$\mathsf{BC}_\Phi := \mathrm{cl}\left\{x \in \mathrm{cl}(\mathrm{dom}\,\mathsf{A}) \,:\, x \in L_{\Phi(x,\cdot)}\right\},$$

i.e., the closure of the set of recurrent points of Φ. The celebrated Poincaré's recurrence theorem [53, Th. II.6.4 and Cor. II.6.5] says that the support of any $\pi \in \mathcal{I}(\Phi)$ is a subset of BC_Φ.

Proposition 4.2.1. Assume that $Z(\mathsf{A}) \neq \emptyset$, and let $\pi \in \mathcal{I}(\Phi)$. If A is demipositive, then $\mathrm{supp}(\pi) \subset Z(\mathsf{A})$. If π has a first moment, then, whether A is demipositive or not,

$$\int x\, \pi(dx) \in Z(\mathsf{A})\,.$$

Proof. When A is demipositive, $\Phi(x, t)$ converges to an element of $Z(\mathsf{A})$ as $t \to +\infty$ hence $Z(\mathsf{A})$ coincides straightforwardly with BC_Φ, and the first inclusion follows from Poincaré's recurrence theorem.

To show the second result, we start by proving that $\{\overline{\Phi}(\cdot, t) \,:\, t > 0\}$ is uniformly integrable as a family of random variables in $(\mathsf{X}, \mathscr{B}(\mathsf{X}), \pi)$. Let $\varepsilon > 0$. Since the family $\{\Phi(\cdot, t) \,:\, t \geq 0\}$ is identically distributed, it is uniformly integrable, thus, there exists $\eta_\varepsilon > 0$ such that $\sup_t \int_S \|\Phi(x, t)\|\, \pi(dx) \leq \varepsilon$ for all $S \in \mathscr{B}(\mathsf{X})$ satisfying $\pi(S) \leq \eta_\varepsilon$. By Tonelli's theorem,

$$\sup_{t>0} \int_S \|\overline{\Phi}(x, t)\|\, \pi(dx) \leq \sup_{t>0} \frac{1}{t} \int_0^t \int_S \|\Phi(x, s)\|\, \pi(dx)ds \leq \varepsilon\,,$$

which shows that, indeed, $\{\overline{\Phi}(\cdot, t) \,:\, t > 0\}$ is uniformly integrable [91, Prop. II-5-2]. By the ergodic theorem for semiflows generated by elements of $\mathscr{M}(\mathsf{X})$ (see Sec. 2.2.2), there exists a measurable function $f : \mathrm{cl}(\mathrm{dom}\,\mathsf{A}) \to Z(\mathsf{A})$ such that $\overline{\Phi}(\cdot, t) \to f$ as $t \to \infty$. Since

$$\int x\, \pi(dx) = \int \overline{\Phi}(x, t)\, \pi(dx) \quad \text{for all } t \geq 0\,,$$

we can make $t \to \infty$ and use the uniform integrability of $\{\overline{\Phi}(\cdot, t) : t > 0\}$ to obtain that $\int \|f\|\, d\pi < \infty$, and $\int x\, \pi(dx) = \int f(x)\, \pi(dx)$. The result follows from the closed convexity of $Z(\mathsf{A})$. □

4.2.1 Presentation of the stochastic Forward-Backward algorithm

Consider two random monotone operators $A, B : (\Xi, \mathscr{G}, \mu) \to \mathscr{M}(\mathsf{X})$ such that for every $s \in \Xi$, $B(s)$ is single-valued and continuous over X. Denoting $B(s, x)$ the image of x by the operator $B(s)$, recall that $s \mapsto B(s, x)$ is measurable. By Carathéodory's theorem, B is $\mathscr{G} \otimes \mathscr{B}(\mathsf{X})$-measurable seen as a function

defined on $\Xi \times \mathsf{X}$. Assuming that $B(\cdot, x)$ is μ-integrable for every $x \in \mathsf{X}$, we set $\mathcal{B}(x) := \int B(s,x)\mu(ds)$. Note that $\mathrm{dom}\,\mathcal{B} = \mathsf{X}$. Denote $A(s,x)$ the image of x by the operator $A(s)$, and $\mathcal{D}$ the essential intersection of the domains $D(s) = \mathrm{dom}(A(s))$ (see Eq. (2.5)). Assuming that $\mathcal{D} \neq \emptyset$ and that $A(\cdot, x)$ is integrable (see Sec. 2.3) for every $x \in \mathcal{D}$, we denote the selection integral $\mathcal{A}(x) := \int A(s,x)\mu(ds)$.

Let (ξ_n) be an i.i.d. sequence of random variables from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\Xi, \mathscr{G})$ with the distribution μ. Let x_0 be a X-valued random variable with probability law ν, and assume that x_0 and (ξ_n) are independent. Starting from x_0, our purpose is to study the behavior of the iterates

$$x_{n+1} = J_\gamma(\xi_{n+1}, x_n - \gamma B(\xi_{n+1}, x_n)), \quad n \in \mathbb{N}\,, \tag{4.4}$$

for a given $\gamma > 0$, where we recall the notation $J_\gamma(s,\cdot) := (I + \gamma A(s))^{-1}(\cdot)$ for every $s \in \Xi$.

In the deterministic case where the functions $A(s,\cdot)$ and $B(s,\cdot)$ are replaced with deterministic maximal monotone operators $\mathsf{A}(\cdot)$ and $\mathsf{B}(\cdot)$, with B still being assumed single-valued with $\mathrm{dom}(\mathsf{B}) = \mathsf{X}$, the algorithm coincides with the well-known Forward-Backward algorithm (4.1). Assuming that B is so-called cocoercive and that γ is not too large, the iterates given by (4.1) are known to converge to an element of $Z(\mathsf{A}+\mathsf{B})$, provided this set is not empty [12, Th. 25.8]. In the stochastic case who is of interest here, this convergence does not hold in general. Nonetheless, we shall show below that in the long run, the probability that the iterates or their empirical means stay away of $Z(\mathcal{A}+\mathcal{B})$ is small when γ is close to zero.

4.3 Assumptions and main results

We first observe that the process (x_n) described by Eq. (4.4) is a homogeneous Markov chain whose transition kernel P_γ is defined by the identity

$$P_\gamma(x,f) = \int f(J_\gamma(s, x - \gamma B(s,x)))\,\mu(ds)\,, \tag{4.5}$$

valid for each measurable and positive function f. The kernel P_γ and the initial measure ν determine completely the probability distribution of the process (x_n), seen as a $(\Omega, \mathscr{F}) \to (\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes\mathbb{N}})$ random variable. We shall denote this probability distribution on $(\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes\mathbb{N}})$ as $\mathbb{P}^{\nu,\gamma}$. We denote by $\mathbb{E}^{\nu,\gamma}$ the corresponding expectation. When $\nu = \delta_a$ for some $a \in \mathsf{X}$, we shall prefer the notations $\mathbb{P}^{a,\gamma}$ and $\mathbb{E}^{a,\gamma}$ to $\mathbb{P}^{\delta_a,\gamma}$ and $\mathbb{E}^{\delta_a,\gamma}$. From now on, (x_n) will denote the canonical process on the canonical space $(\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes\mathbb{N}})$.

We denote as $\mathscr{F}_n$ the sub-σ-field of $\mathscr{F}$ generated by the family $\{x_0, \{\xi_k^\gamma : 1 \le k \le n\}\}$, and we write $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$ for $n \in \mathbb{N}$.

In the remainder of the chapter, C will always denote a positive constant that does not depend on the time n nor on γ. This constant may change from a line of calculation to another. In all our derivations, γ will lie in the interval $(0, \gamma_0]$ where γ_0 is a fixed constant which is chosen as small as needed.

4.3.1 Assumptions

Assumption 4.3.1. For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K} \cap \mathcal{D}} \int \|A_0(s,x)\|^{1+\varepsilon}\,\mu(ds) < \infty.$$

Assumption 4.3.2. The monotone operator $\mathcal{A}$ is maximal.

Assumption 4.3.3. For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K}} \int \|B(s,x)\|^{1+\varepsilon} \, \mu(ds) < \infty \, .$$

The next assumption will mainly lead to the tightness of the invariant measures mentioned in the introduction.

We know that a point $x_\star$ is an element of $Z(\mathcal{A}+\mathcal{B})$ if there exists $\varphi \in \mathfrak{S}^1_{A(\cdot,x_\star)}$ such that $\int \varphi(s) \, \mu(ds) + \int B(s,x_\star) \, \mu(ds) = 0$. When $B(\cdot, x_\star) \in \mathcal{L}^2(\Xi, \mathscr{G}, \mu; \mathsf{X})$, and when the above function φ can be chosen in $\mathcal{L}^2(\Xi, \mathscr{G}, \mu; \mathsf{X})$, we say that such a zero admits a $\mathcal{L}^2$ representation (φ, B). In this case, we define

$$\begin{aligned} \psi_\gamma(x) := \int \Big\{ & \langle A_\gamma(s, x - \gamma B(s,x)) - \varphi(s), J_\gamma(s, x - \gamma B(s,x)) - x_\star \rangle \\ & + \langle B(s,x) - B(s,x_\star), x - x_\star \rangle \Big\} \, \mu(ds) \\ & + \gamma \int \|A_\gamma(s, x - \gamma B(s,x))\|^2 \mu(ds) - 6\gamma \int \|B(s,x) - B(s,x_\star)\|^2 \mu(ds) \, , \end{aligned} \tag{4.6}$$

where

$$A_\gamma(s,x) := \frac{x - J_\gamma(s,x)}{\gamma}$$

is the Yosida regularization of $A(s,x)$ for $\gamma > 0$.

Assumption 4.3.4. There exists $x_\star \in Z(\mathcal{A}+\mathcal{B})$ admitting a $\mathcal{L}^2$ representation (φ, B). The function $\Psi(x) := \inf_{\gamma \in (0, \gamma_0]} \psi_\gamma(x)$ satisfies one of the following properties:

(a) $\liminf_{\|x\| \to \infty} \dfrac{\Psi(x)}{\|x\|} > 0.$

(b) $\dfrac{\Psi(x)}{\|x\|} \xrightarrow[\|x\| \to \infty]{} \infty.$

(c) $\liminf_{\|x\| \to \infty} \dfrac{\Psi(x)}{\|x\|^2} > 0.$

Let us comment these assumptions.

Assumptions 4.3.1 and 4.3.3 are moment assumptions on $A_0(s,x)$ and $B(s,x)$ that are usually easy to check. Assumption 4.3.1 implies that for every $x \in \mathcal{D}$, $A_0(\,.\,,x)$ is integrable. Therefore, $A(\,.\,,x)$ is integrable. This implies that the domain of the selection integral $\mathcal{A}$ coincides with $\mathcal{D}$.

Conditions where Assumption 4.3.2 are satisfied can be found in [36, Chap. II.6] in the case where μ has a finite support, and in [24, Prop. 3.1] in other cases. When $A(s)$ is the subdifferential of a function $g(s,\cdot)$ belonging to $\Gamma_0(\mathsf{X})$, the maximality of $\mathcal{A}$ is established if we can exchange the expectation of $g(\xi_1, x)$ w.r.t. ξ_1 with the subdifferentiation w.r.t. x, in which case $\mathcal{A}$ would be equal to ∂G, where $G(x) = \int g(s,x) \, \mu(ds)$. This problem is dealt with in [125] (see also Sec. 4.4.1 below).

The first role of Assumption 4.3.4 is to ensure the tightness of the invariant measures of the kernels P_γ, as mentioned in the introduction. Beyond the tightness, this assumption controls the asymptotic behavior of functionals of the iterates with a prescribed growth condition at infinity. Assumption 4.3.4 will be specified and commented at length in Sec. 4.4.

Regarding the domains of the operators $A(s)$, two cases will be considered, according to whether these domains vary with s or not. We shall name these two cases the "common domain" case and the "different domains" case respectively. In the common domain case, our assumption is therefore:

Assumption 4.3.5 (Common domain case)**.** The set-valued function $s \mapsto D(s)$ is μ-almost everywhere constant.

In the common domain case, Assumptions 4.3.1–4.3.4 will be sufficient to state our results, whereas in the different domains case, three supplementary assumptions will be needed:

Assumption 4.3.6 (Different domains case)**.** $\forall x \in \mathsf{X}$, $\int d(x, D(s))^2 \, \mu(ds) \geq C\boldsymbol{d}(x)^2$, where $\boldsymbol{d}(\cdot)$ is the distance function to $\mathcal{D}$.

Assumption 4.3.7 (Different domains case)**.** For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{\gamma \in (0,\gamma_0], x \in \mathcal{K}} \frac{1}{\gamma^{1+\varepsilon}} \int \|J_\gamma(s,x) - \Pi_{\mathrm{cl}(D(s))}(x)\|^{1+\varepsilon} \, \mu(ds) < \infty \,.$$

Assumption 4.3.8 (Different domains case)**.** For all $\gamma \in (0, \gamma_0]$ and all $x \in \mathsf{X}$,

$$\int \left(\frac{\|J_\gamma(s,x) - \Pi_{\mathrm{cl}(D(s))}(x)\|}{\gamma} + \|B(s,x)\| \right) \mu(ds) \leq C(1 + \psi_\gamma(x)) \,.$$

Assumption 4.3.6 is rather mild, and is studied *e.g* in [90]. This assumption is easy to illustrate in the case where μ is a finite sum of Dirac measures. Following [11], we say that a finite collection of closed and convex subsets $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ over X is *linearly regular* if there exists $\kappa > 0$ such that for every x,

$$\max_{i=1\ldots m} d(x, \mathcal{C}_i) \geq \kappa d(x, \mathcal{C}), \quad \text{where } \mathcal{C} = \bigcap_{i=1}^m \mathcal{C}_i \,,$$

and where implicitly $\mathcal{C} \neq \emptyset$. Sufficient conditions for a collection of sets to satisfy the above condition can be found in [11] and the references therein. In the general case, Assumption 4.3.6 is studied in [90]

We know that when $\gamma \to 0$, $J_\gamma(s,x)$ converges to $\Pi_{\mathrm{cl}(D(s))}(x)$ for each (s,x). Assumptions 4.3.7 and 4.3.8 add controls on the convergence rate. The instantiations of these assumptions in the case of the stochastic proximal gradient algorithm will be provided in Sec. 4.4.1 below.

4.3.2 Main result

Lemma 4.3.1. Let Assumptions 4.3.2 and 4.3.3 hold true. Then, the monotone operator $\mathcal{A} + \mathcal{B}$ is maximal.

Proof. Assumption 4.3.3 implies that the monotone operator $\mathcal{B}$ is continuous on X. Therefore, $\mathcal{B}$ is maximal [36, Prop. 2.4]. The maximality of $\mathcal{A} + \mathcal{B}$ follows, since $\mathcal{A}$ is maximal by Assumption 4.3.2, and $\mathcal{B}$ has a full domain [36, Cor. 2.7]. □

Note that $\mathrm{dom}(\mathcal{A}+\mathcal{B}) = \mathcal{D}$. In the remainder of the chapter, we denote as $\Phi : \mathrm{cl}(\mathcal{D}) \times \mathbb{R}_+ \to \mathrm{cl}(\mathcal{D})$ the semiflow produced by the DI $\dot{\mathsf{x}}(t) \in -(\mathcal{A}+\mathcal{B})(\mathsf{x}(t))$. Recall that $\mathcal{I}(\Phi)$ is the set of invariant measures for the semiflow Φ.

We also write

$$\bar{x}_n := \frac{1}{n+1} \sum_{k=0}^n x_k \,.$$

We now state our main theorem.

Theorem 4.3.2. Let Assumptions 4.3.1, 4.3.2, 4.3.3, and 4.3.4–(a) be satisfied. Moreover, assume that either Assumption 4.3.5 or Assumptions 4.3.6–4.3.8 are satisfied.

Then, $\mathcal{I}(\Phi) \neq \emptyset$. Let $\nu \in \mathcal{M}(\mathsf{X})$ be with a finite second moment, and let $\mathcal{U} := \bigcup_{\pi \in \mathcal{I}(\Phi)} \mathrm{supp}(\pi)$. Then, for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}(d(x_k, \mathcal{U}) > \varepsilon) \xrightarrow[\gamma\to 0]{} 0\,. \tag{4.7}$$

In particular, if the operator $\mathcal{A} + \mathcal{B}$ is demipositive, then

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}\left(d(x_k, Z(\mathcal{A}+\mathcal{B})) > \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,. \tag{4.8}$$

Moreover, the set $\{\pi \in \mathcal{I}(\Phi) \,:\, \pi(\Psi) < \infty\}$ is not empty. Let Y an Euclidean space, and let $f : \mathsf{X} \to \mathsf{Y}$ be continuous. Assume that there exists $M \geq 0$ and $\varphi : \mathsf{Y} \to \mathbb{R}_+$ such that $\lim_{\|a\|\to\infty} \varphi(a)/\|a\| = \infty$, and

$$\forall a \in \mathsf{X},\ \ \varphi(f(a)) \leq M(1 + \Psi(a))\,.$$

Then, for all $n \in \mathbb{N}$, $\gamma \in (0, \gamma_0]$, the r.v.

$$F_n := \frac{1}{n+1} \sum_{k=0}^{n} f(x_k)$$

is $\mathbb{P}$-integrable, and satisfies for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \mathbb{P}^{\nu,\gamma}\left(d\left(F_n, \mathcal{S}_f\right) \geq \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,, \tag{4.9}$$

$$\limsup_{n\to\infty}\ d\left(\mathbb{E}^{\nu,\gamma}(F_n), \mathcal{S}_f\right) \xrightarrow[\gamma\to 0]{} 0\,. \tag{4.10}$$

where $\mathcal{S}_f := \{\pi(f) \,:\, \pi \in \mathcal{I}(\Phi)\}$. In particular, if $f(x) = x$, and if Assumption 4.3.4–(b) is satisfied, then

$$\limsup_{n\to\infty}\ \mathbb{P}^{\nu,\gamma}\left(d\left(\bar{x}_n, Z(\mathcal{A}+\mathcal{B})\right) \geq \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,, \tag{4.11}$$

$$\limsup_{n\to\infty}\ d\left(\mathbb{E}^{\nu,\gamma}(\bar{x}_n), Z(\mathcal{A}+\mathcal{B})\right) \xrightarrow[\gamma\to 0]{} 0\,. \tag{4.12}$$

By Lem. 4.3.1 and Prop. 4.2.1, the convergences (4.8), (4.11), and (4.12) are the consequences of (4.7), (4.9), and (4.10) respectively. We need to prove the latter.

4.3.3 Proof technique

We first observe that the Markov kernels P_γ are Feller, *i.e.*, they take the set $C_b(\mathsf{X})$ of the real, continuous, and bounded functions on X to $C_b(\mathsf{X})$. Indeed, for each $f \in C_b(\mathsf{X})$, Eq. (4.5) shows that $P_\gamma(\cdot, f) \in C_b(\mathsf{X})$ by the continuity of $J_\gamma(s,\cdot)$ and $B(s,\cdot)$, and by dominated convergence.

For each $\gamma > 0$, we denote as

$$\mathcal{I}(P_\gamma) := \{\pi \in \mathcal{M}(\mathsf{X}) \,:\, \pi = \pi P_\gamma\}$$

the set of invariant probability measures of P_γ. Define the family of kernels $\mathcal{P} := \{P_\gamma\}_{\gamma\in(0,\gamma_0]}$, and let

$$\mathcal{I}(\mathcal{P}) := \bigcup_{\gamma\in(0,\gamma_0]} \mathcal{I}(P_\gamma)$$

be the set of distributions π such that $\pi = \pi P_\gamma$ for at least one P_γ with $\gamma \in (0, \gamma_0]$.

The following proposition, which is valid for Feller Markov kernels, has been proven in Chap. 3.

Proposition 4.3.3. Let $V : \mathsf{X} \to [0, +\infty)$ and $Q : \mathsf{X} \to [0, +\infty)$ be measurable. Assume that $Q(x) \to \infty$ as $\|x\| \to \infty$. Assume that for each $\gamma \in (0, \gamma_0]$,

$$P_\gamma(x, V) \leq V(x) - \alpha(\gamma)Q(x) + \beta(\gamma)\,, \tag{4.13}$$

where $\alpha : (0, \gamma_0] \to (0, +\infty)$ and $\beta : (0, \gamma_0] \to \mathbb{R}$ satisfy $\sup_{\gamma \in (0,\gamma_0]} \frac{\beta(\gamma)}{\alpha(\gamma)} < \infty$. Then, the family $\mathcal{I}(\mathcal{P})$ is tight. Moreover, $\sup_{\pi \in \mathcal{I}(\mathcal{P})} \pi(Q) < \infty$.

Assume moreover that, as $\gamma \to 0$, any cluster point of $\mathcal{I}(\mathcal{P})$ is an element of $\mathcal{I}(\Phi)$. In particular, $\{\pi \in \mathcal{I}(\Phi) \,:\, \pi(Q) < \infty\}$ is not empty. Let $\nu \in \mathcal{M}(\mathsf{X})$ s.t. $\nu(V) < \infty$. Let $\mathcal{U} := \bigcup_{\pi \in \mathcal{I}(\Phi)} \mathrm{supp}(\pi)$. Then, for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}(d(x_k, \mathcal{U}) > \varepsilon) \underset{\gamma\to 0}{\longrightarrow} 0\,.$$

Let Y an Euclidean space and $f : \mathsf{X} \to \mathsf{Y}$ be continuous. Assume that there exists $M \geq 0$ and $\varphi : \mathsf{Y} \to \mathbb{R}_+$ such that $\lim_{\|a\|\to\infty} \varphi(a)/\|a\| = \infty$ and

$$\forall a \in \mathsf{X},\ \varphi(f(a)) \leq M(1 + Q(a))\,.$$

Then, for all $n \in \mathbb{N}$, $\gamma \in (0, \gamma_0]$, the r.v.

$$F_n := \frac{1}{n+1} \sum_{k=0}^{n} f(x_k)$$

is $\mathbb{P}^{\nu,\gamma}$-integrable, and satisfies for all $\varepsilon > 0$,

$$\limsup_{n\to\infty} d\left(\mathbb{E}^{\nu,\gamma}(F_n)\,, \mathcal{S}_f\right) \underset{\gamma\to 0}{\longrightarrow} 0\,, \quad \text{and} \quad \limsup_{n\to\infty} \mathbb{P}^{\nu,\gamma}\left(d\left(F_n\,, \mathcal{S}_f\right) \geq \varepsilon\right) \underset{\gamma\to 0}{\longrightarrow} 0\,,$$

where $\mathcal{S}_f := \{\pi(f) \,:\, \pi \in \mathcal{I}(\Phi)\}$.

Proof. Assume that Eq. (4.13) holds. By Prop. 3.8.7, $\mathcal{I}(\mathcal{P})$ is tight and $\sup_{\pi \in \mathcal{I}(\mathcal{P})} \pi(Q) < \infty$, which proves the first point. Assume moreover that, as $\gamma \to 0$, any cluster point of $\mathcal{I}(\mathcal{P})$ is an element of $\mathcal{I}(\Phi)$. By the tightness of $\mathcal{I}(\mathcal{P})$ and the Prokhorov theorem, such a cluster point π exists, and satisfies $\pi(Q) < \infty$ by the first point just shown. The rest of the proof follows Sec. 3.8.4 word-for-word. □

In order to prove Th. 4.3.2, it is enough to show that the assumptions of Prop. 4.3.3 are satisfied. Namely, we need to establish (4.13) and to show that the cluster points of $\mathcal{I}(\mathcal{P})$ as $\gamma \to 0$ are elements of $\mathcal{I}(\Phi)$.

In Sec. 4.5, we show that the linearly interpolated process constructed from the sequence (x_n) converges narrowly as $\gamma \to 0$ to a DI solution in the topology of uniform convergence on compact sets. The main result of this section is Th. 4.5.1, which has its own interest. To prove this theorem, we establish the tightness of the linearly interpolated process (Lem. 4.5.3), then we show that the limit points coincide with the DI solution (Lem. 4.5.4–4.5.6). In Sec. 4.6, we start by establishing the inequality (4.13), which is shown in Lem. 4.6.1 with $Q(x) = \Psi(x)$. Using the tightness of $\mathcal{I}(\mathcal{P})$ in conjunction with Th. 4.5.1, Lem 4.6.2 shows that the cluster points of $\mathcal{I}(\mathcal{P})$ are elements of $\mathcal{I}(\Phi)$. In the different domains case, this lemma requires that the invariant measures of P_γ put most of their weights in a thickening of the domain $\mathcal{D}$ of order γ. This fact is established by Lem. 4.6.3.

4.4 Case studies - Tightness of the invariant measures

Before proving the main results, we first address three important cases: the case of the random proximal gradient algorithm, the case where $A(s)$ is an affine monotone operator and $B(s) = 0$, and the case where $\mathcal{D}$ is bounded. The main problem is to ensure that one of the cases of Assumption 4.3.4 is verified. We close the section with a general condition ensuring that Assumption 4.3.4–(a) is verified. The proofs are postponed to Appendix 4.7.

4.4.1 A random proximal gradient algorithm

Let $(\Sigma, \mathscr{A}, \zeta)$ be a probability space, where $\mathscr{A}$ is ζ-complete. Let $h : \Sigma \times \mathsf{X} \to (-\infty, \infty]$ a convex normal integrand (see Sec. 2.3). To simplify the presentation, we furthermore assume that h is finite everywhere, noting that the results can be extended to the case where h can take the value ∞. Recall that $s \mapsto \partial h(s, \cdot)$ is a random monotone operator (in all the following, the subdifferential or the gradient of a function in (s, x) will be meant to be taken w.r.t. x). Assume that $\int |h(s, x)| \zeta(ds) < \infty$ for all $x \in \mathsf{X}$, and consider the convex function $H(x) := \int h(s, x)\, \zeta(ds)$ defined on X. By *e.g.*, [102, page 179], $\partial H(x) = \int \partial h(s, x)\, \zeta(ds)$.

Let $f : \Sigma \times \mathsf{X} \to \mathbb{R}$ be such that $f(\cdot, x)$ is $\mathscr{A}$-measurable for all $x \in \mathsf{X}$, and $f(s, \cdot)$ is convex and continuously differentiable for all $s \in \Sigma$. Moreover, assume that $\int |f(s, x)|\, \zeta(ds) < \infty$ for all $x \in \mathsf{X}$, and define the function $F(x) := \int f(s, x)\, \zeta(ds)$ on X. This function is differentiable with $\nabla F(x) = \int \nabla f(s, x)\, \zeta(ds)$.

Finally, given $m \in \mathbb{N}^*$, let $\{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ be a collection of closed and convex subsets of X. We assume that $\bigcap_{i=1}^m \mathrm{ri}(\mathcal{C}_i) \neq \emptyset$, where ri is the relative interior of a set.

Our purpose is to approximatively solve the optimization problem

$$\min_{x \in \mathcal{C}} F(x) + H(x), \quad \mathcal{C} := \bigcap_{i=1}^m \mathcal{C}_i \tag{4.14}$$

whether the minimum is attained. Let (u_n) be an iid sequence on Σ with the probability measure ζ. Let (I_n) be an iid sequence on $\{0, 1, \dots, m\}$ with the probability measure α such that $\alpha(k) = \mathbb{P}(I_1 = k) > 0$ for each k. Assume that (I_n) and (u_n) are independent. In order to solve the problem (4.14), we consider the iterates

$$x_{n+1} = \begin{cases} \mathrm{prox}_{\alpha(0)^{-1}\gamma h(u_{n+1}, \cdot)}(x_n - \gamma \nabla f(u_{n+1}, x_n)) & \text{if } I_{n+1} = 0, \\ \Pi_{\mathcal{C}_{I_{n+1}}}(x_n - \gamma \nabla f(u_{n+1}, x_n)) & \text{otherwise,} \end{cases} \tag{4.15}$$

for $\gamma > 0$. This problem can be cast in the general framework of the stochastic proximal gradient algorithm presented in the introduction. On the space $\Xi := \Sigma \times \{0, \dots, m\}$, define the iid random variables $\xi_n := (u_n, I_n)$ with the measure $\mu := \zeta \otimes \alpha$. Denoting as ι_S the indicator function of the set S, let $g : \Xi \times \mathsf{X} \to (-\infty, \infty]$ be defined as

$$g(s, x) := \begin{cases} \alpha(0)^{-1} h(u, x) & \text{if } i = 0, \\ \iota_{\mathcal{C}_i}(x) & \text{otherwise,} \end{cases}$$

where $s = (u, i)$. Then, Problem (4.14) is equivalent to minimizing the sum $F(x) + G(x)$, where

$$G(x) := \int g(s, x)\, \mu(ds) = \sum_{k=1}^m \iota_{\mathcal{C}_k}(x) + H(x)\,.$$

It is furthermore clear that the algorithm (4.15) is the instance of the general algorithm (4.4) that corresponds to $A(s) = \partial g(s,\cdot)$ and $B(s) = \nabla f(u,\cdot)$ for $s = (u,i)$. With our assumptions, the qualification conditions hold, and the three sets $\arg\min(F+G)$, $Z(\partial G + \nabla F)$, and $Z(\mathcal{A}+\mathcal{B})$ coincide.

Before going further, we recall some well known facts regarding the coercive functions belonging to $\Gamma_0(\mathsf{X})$. A function $q \in \Gamma_0(\mathsf{X})$ is said coercive if $\lim_{\|x\|\to\infty} q(x) = \infty$. It is said supercoercive if $\lim_{\|x\|\to\infty} q(x)/\|x\| = \infty$. The three following conditions are equivalent: i) q is coercive, ii) there exists $a \in \mathbb{R}$ such that the level set $\mathrm{lev}_{\leq a}\, q$ is non empty and compact, iii) $\liminf_{\|x\|\to\infty} q(x)/\|x\| > 0$ (see *e.g.*, [12, Prop. 11.11 and 11.12] and [32, Prop. 1.1.5]).

The main result of this paragraph is the following:

Proposition 4.4.1. Let the following hypotheses hold true:

H1 There exists $x_\star \in Z(\partial G + \nabla F)$ admitting a $\mathcal{L}^2$ representation $(\varphi((u,i)), \nabla f(u,x_\star))$.

H2 There exists $c > 0$ s.t. for every $x \in \mathsf{X}$,

$$\int \langle \nabla f(s,x) - \nabla f(s,x_\star), x - x_\star\rangle\, \zeta(ds) \geq c \int \|f(s,x) - f(s,x_\star)\|^2\, \zeta(ds).$$

H3 The function $F+G$ satisfies one of the following properties:

(a) $F+G$ is coercive.

(b) $F+G$ is supercoercive.

Then, Assumption 4.3.4–(a) (resp., Assumption 4.3.4–(b)) holds true if Hypothesis H3–(a) (resp., Hypothesis H3–(b)) holds true.

Let us comment these hypotheses. A light condition ensuring the truth of Hypothesis H1 is provided by the following lemma.

Lemma 4.4.2. Assume that there exists $x_\star \in Z(\partial G + \nabla F)$ satisfying the two following conditions: $\int \|\nabla f(u,x_\star)\|^2\, \zeta(du) < \infty$, and there exists an open neighborhood $\mathcal{N}$ of $x_\star$ such that $\int h(u,x)^2\, \zeta(du) < \infty$ for all $x \in \mathcal{N}$. Then, Hypothesis H1 is verified.

We now turn to Hypothesis H2. When studying the deterministic Forward-Backward algorithm (4.1), it is standard to assume that B is cocoercive, in other words, that there exists a constant $L > 0$ such that $\langle \mathsf{B}(x) - \mathsf{B}(y), x - y\rangle \geq L\|\mathsf{B}(x) - \mathsf{B}(y)\|^2$ [12, Th. 25.8]. A classical case where this is satisfied is the case where B is the gradient of a convex differentiable function having a $1/L$-Lipschitz continuous gradient, as is shown by the Baillon-Haddad Th. [12, Cor. 18.16]. In our case, if we assume that there exists a nonnegative measurable function $\beta(s)$ such that $\|\nabla f(s,x) - \nabla f(s,x')\| \leq \beta(s)\|x - x'\|$, then by the Baillon-Haddad theorem,

$$\langle \nabla f(s,x) - \nabla f(s,x'), x - x'\rangle \geq \frac{1}{\beta(s)} \|\nabla f(s,x) - \nabla f(s,x')\|^2 .$$

Thus, one obvious case where Hypothesis H2 is satisfied is the case where $\beta(s)$ is bounded.

Using proposition 4.4.1, we can now obtain the following corollary to Th. 4.3.2.

Corollary 4.4.3. Let Hypotheses H1–H3 hold true. Assume in addition the following hypotheses:

C1 For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K} \cap \mathcal{C}} \int \|\partial_0 h(u,x)\|^{1+\varepsilon} \, \zeta(du) < \infty,$$

where $\partial_0 h(u,\cdot)$ is the least norm element of $\partial h(u,\cdot)$.

C2 For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K}} \int \|\nabla f(u,x)\|^{1+\varepsilon} \, \zeta(du) < \infty \, .$$

C3 The sets $\mathcal{C}_1, \ldots, \mathcal{C}_m$ are linearly regular.

C4 For all $\gamma \in (0, \gamma_0]$ and all $x \in \mathsf{X}$,

$$\int \left(\|\nabla h_\gamma(u,x)\| + \|\nabla f(u,x)\| \right) \zeta(du) \leq C(1 + |F(x) + H_\gamma(x)|) \, ,$$

where $h_\gamma(u,\cdot)$ is the Moreau envelope of $h(u,\cdot)$.

Then, for each probability measure ν having a finite second moment,

$$\limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma} \left(d(x_k, \arg\min(F+G)) > \varepsilon \right) \xrightarrow[\gamma \to 0]{} 0 \, .$$

Moreover, if Hypothesis H3–(b) is satisfied, then

$$\limsup_{n \to \infty} \; \mathbb{P}^{\nu,\gamma} \left(d\left(\bar{x}_n, \arg\min(F+G)\right) \geq \varepsilon \right) \xrightarrow[\gamma \to 0]{} 0, \text{ and}$$
$$\limsup_{n \to \infty} \; d\left(\mathbb{E}^{\nu,\gamma}(\bar{x}_n), \arg\min(F+G)\right) \xrightarrow[\gamma \to 0]{} 0 \, .$$

Proof. With the hypotheses H1–H3 and C1–C4, one can check that the assumptions 4.3.1–4.3.8 are verified. Note that $\partial G + \nabla F$ is a demipositive operator, being the subdifferential of a $\Gamma_0(\mathsf{X})$ function having a minimizer [38]. The results of the corollary follow from those of Th. 4.3.2. □

4.4.2 The case where $A(s)$ is affine

In all the remainder of this section, we shall focus on the validity of Assumption 4.3.4. We assume that $B = 0$, and that

$$A(s,x) = H(s)x + d(s),$$

where $H : \Xi \to \mathscr{L}(\mathsf{X})$ where $\mathscr{L}(\mathsf{X})$ is the space of linear operator over X and $d : \Xi \to \mathsf{X}$ are two $\mathscr{G}$-measurable functions. It is easily seen that the affine operator $A(s)$ is monotone if and only if $H(s) + H(s)^\star$ is a positive semidefinite operator (we shall write $H(s) + H(s)^\star \geq 0$), a condition that we shall assume in this subsection. Moreover, assuming that

$$\int (\|H(s)\|^2 + \|d(s)\|^2) \, \mu(ds) < \infty,$$

the operator

$$\mathcal{A}(x) = \Big(\int H(s) \, \mu(ds) \Big) x + \int d(s) \, \mu(ds) := \boldsymbol{H} x + \boldsymbol{d}$$

exists and is a maximal monotone operator with the domain X. When $\boldsymbol{d}$ belongs to the image of $\boldsymbol{H}$, $Z(\mathcal{A}) \neq \emptyset$, and every $x_\star \in Z(\mathcal{A})$ has a unique $\mathcal{L}^2$ representation $(\varphi(s) = H(s)x_\star + d(s), 0)$. We have the following proposition:

Proposition 4.4.4. If $\boldsymbol{H} + \boldsymbol{H}^T > 0$, then $\boldsymbol{H}$ is invertible, $Z(\mathcal{A}) = \{x_\star\}$ with $x_\star = -\boldsymbol{H}^{-1}\boldsymbol{d}$, and Assumption 4.3.4–(c) is verified.

4.4.3 The case where the domain $\mathcal{D}$ is bounded

Proposition 4.4.5. Let the following hypotheses hold true:

H1 The domain $\mathcal{D}$ is bounded.

H2 There exists a constant $C > 0$ such that

$$\forall x \in \mathsf{X},\ \int d(s,x)^2\,\mu(ds) \geq C\boldsymbol{d}(x)^2.$$

H3 There exists $x_\star \in Z(\mathcal{A}+\mathcal{B})$ admitting a $\mathcal{L}^2$ representation.

H4 There exists $c > 0$ s.t. for every $x \in \mathsf{X}$, For all γ small enough,

$$\int \langle B(s,x) - B(s,x_\star), x - x_\star\rangle\,\mu(ds) \geq c\int \|B(s,x) - B(s,x_\star)\|^2\,\mu(ds)\,.$$

Then, Assumption 4.3.4–(c) is satisfied.

4.4.4 A case where Assumption 4.3.4–(a) is valid

We close this section by providing a general condition that guarantees the validity of Assumption 4.3.4–(a). For simplicity, we focus on the case where $B(s) = 0$, noting that the result can be easily extended to the case where $B(s) \neq 0$ when a cocoercivity hypothesis of the type of Prop. 4.4.5–H4 is satisfied.

We denote by $\boldsymbol{S}(\rho, d)$ the sphere of X with center ρ and radius d. We also denote by $\operatorname{int} S$ the interior of a set S.

Proposition 4.4.6. Assume that $B(s) = 0$, and that there exists $x_\star \in Z(\mathcal{A}) \cap \operatorname{int}\mathcal{D}$ admitting a $\mathcal{L}^2$ representation $\varphi \in \mathfrak{S}^2_{A(\cdot,x_\star)}$. Assume that there exists a set $\Sigma \in \mathscr{G}$ such that $\mathcal{D} \subset \cap_{s\in\Sigma} D(s)$, $\mu(\Sigma) > 0$, and such that for all $s \in \Sigma$, there exists $\delta(s) > 0$ satisfying $\boldsymbol{S}(\varphi(s), \delta(s)) \subset \operatorname{int}\mathcal{D}$, and

$$\forall x \in \boldsymbol{S}(\varphi(s), \delta(s)),\quad \inf_{y\in A(s,x)} \langle y - \varphi(s), x - x_\star\rangle > 0.$$

Then, Assumption 4.3.4–(a) is satisfied.

Note that the $\inf$ in the statement of this proposition is attained, as is revealed by the proof.

4.5 Narrow convergence towards the DI solutions

4.5.1 Main result

The set $C(\mathbb{R}_+, \mathsf{X})$ of continuous functions from $\mathbb{R}_+$ to X is equipped with the topology of uniform convergence on the compact intervals, who is known to be compatible with the distance d defined as

$$\mathsf{d}(\mathsf{x},\mathsf{y}) := \sum_{n\in\mathbb{N}^*} 2^{-n}\left(1 \wedge \sup_{t\in[0,n]} \|\mathsf{x}(t) - \mathsf{y}(t)\|\right).$$

For every $\gamma > 0$, we introduce the measurable map $\mathsf{X}_\gamma : (\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes\mathbb{N}}) \to (C(\mathbb{R}_+,\mathsf{X}), \mathscr{B}(C(\mathbb{R}_+,\mathsf{X})))$, defined for every $x = (x_n : n \in \mathbb{N})$ in $\mathsf{X}^{\mathbb{N}}$ as

$$\mathsf{X}_\gamma(x)\ : t \mapsto x_{\lfloor \frac{t}{\gamma}\rfloor} + (t/\gamma - \lfloor t/\gamma\rfloor)(x_{\lfloor \frac{t}{\gamma}\rfloor+1} - x_{\lfloor \frac{t}{\gamma}\rfloor})\,.$$

This map will be referred to as the linearly interpolated process. When $x = (x_n)$ is the process with the probability measure $\mathbb{P}^{\nu,\gamma}$ defined above, the distribution of the r.v. X_γ is $\mathbb{P}^{\nu,\gamma}\mathsf{X}_\gamma^{-1}$. If S is a subset of X and $\varepsilon > 0$, we denote by $S_\varepsilon := \{a \in \mathsf{X} : d(a, S) < \varepsilon\}$ the ε-neighborhood of S. The aim of the present section is to establish the following result:

Theorem 4.5.1. Let Assumptions 4.3.1–4.3.3 hold true. Let either Assumption 4.3.5 or Assumptions 4.3.6-4.3.7 hold true. Then, for every $\eta > 0$, for every compact set $\mathcal{K} \subset \mathsf{X}$ s.t. $\mathcal{K} \cap \mathcal{D} \neq \emptyset$,

$$\forall M \geq 0, \quad \sup_{a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}} \mathbb{P}^{a,\gamma} \left(\mathsf{d}(\mathsf{X}_\gamma, \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), \cdot)) > \eta \right) \xrightarrow[\gamma \to 0]{} 0. \tag{4.16}$$

Using the Yosida regularization $A_\gamma(s, x)$ of $A(s, x)$, the iterates (4.4) can be rewritten as $x_0 = a \in \mathcal{D}_{\gamma M}$ and

$$x_{n+1} = x_n - \gamma B(\xi_{n+1}, x_n) - \gamma A_\gamma(\xi_{n+1}, x_n - \gamma B(\xi_{n+1}, x_n)). \tag{4.17}$$

Setting $h_\gamma(s, x) := -B(s, x) - A_\gamma(s, x - \gamma B(s, x))$, the iterates (4.4) can be cast into the same form as the one studied in Chap. 3 (i.e Eq. 1.4). The following result, which we state here mainly for the ease of the reading, is a straightforward consequence of Th. 3.5.1, Chap. 3.

Proposition 4.5.2. Let Assumptions 4.3.1–4.3.3 hold true. Assume moreover that for every $s \in \Xi$, $D(s) = \mathsf{X}$. Then, Eq. (4.16) holds true.

Proof. It is sufficient to check that the mapping h_γ satisfies the Assumption (RM) of Th. 3.5.1, Chap. 3. Assumption i) is satisfied by definition of h_γ. As $D(\cdot)$ is a constant equal to X, the operator $A(s, \cdot)$ is upper semi continuous as a set-valued operator [97]. Thus, $H(s, \cdot) := -A(s, \cdot) - B(s, \cdot)$ is proper, upper semi continuous with closed convex values, and μ-integrable. Hence, the assumptions iv) and iii) are satisfied. Assumption v) is satisfied by the natural properties of the semiflow induced by the maximal monotone map $\mathcal{A} + \mathcal{B}$, whereas Assumption vi) directly follows from the present Assumptions 4.3.1 and 4.3.3 and the definition of h_γ. One should finally verify Assumption ii), which states that for every converging sequence $(u_n, \gamma_n) \to (u^\star, 0)$, $h_{\gamma_n}(s, u_n) \to H(s, u^\star)$, for every $s \in \Xi$. To this end, it is sufficient to prove that

$$A_{\gamma_n}(s, u_n - \gamma_n B(s, u_n)) \to A(s, u^\star)\,. \tag{4.18}$$

Choose $\varepsilon > 0$. As $A(s, \cdot)$ is upper semi continuous, there exists $\eta > 0$ s.t. $\forall u$, $\|u - u^\star\| < \eta$ implies $A(s, u) \subset A(s, u^\star)_\varepsilon$. Let $v_n := J_{\gamma_n}(s, u_n - \gamma B(s, u_n))$. By the triangular inequality and the non-expansiveness of J_{γ_n},

$$\|v_n - u^\star\| \leq \|u_n - u^\star\| + \gamma_n \|B(s, u_n)\| + \|J_{\gamma_n}(u^\star) - u^\star\|\,,$$

where it is clear that each of the three terms in the right hand side tends to zero. Thus, there exists $N \in \mathbb{N}$ s.t. $\forall n \geq N$, $\|v_n - u^\star\| \leq \eta$, which in turn implies $A(s, v_n) \subset A(s, u^\star)_\varepsilon$. As $A_{\gamma_n}(s, u_n - \gamma_n B(s, u_n)) \in A(s, v_n)$, the convergence (4.18) is established. □

4.5.2 Proof of Th. 4.5.1

In the sequel, we prove Th. 4.5.1 under the set of Assumptions 4.3.6-4.3.7. The proof in the common domain case *i.e.*, when Assumption 4.3.5 holds, is somewhat easier and follows from the same arguments.

In order to prove Th. 4.5.1, we just have to weaken the assumptions of Prop. 4.5.2: for a given $s \in \Xi$, the domain $D(s)$ is not necessarily equal to X and the monotone operator $A(s, \,.\,)$ is not necessarily upper

semi continuous. Up to these changes, the proof is similar to the proof of Th. 3.5.1, Chap. 3, and the modifications are in fact confined to specific steps of the proof.

Choose a compact set $\mathcal{K} \subset \mathsf{X}$ s.t. $\mathcal{K} \cap \mathrm{cl}(\mathcal{D}) \neq \emptyset$. Choose $R > 0$ s.t. $\mathcal{K}$ is contained in the ball of radius R. For every $x = (x_n : n \in \mathbb{N})$ in $\mathsf{X}^{\mathbb{N}}$, define $\tau_R(x) := \inf\{n \in \mathbb{N} : x_n > R\}$ and introduce the measurable mapping $B_R : \mathsf{X}^{\mathbb{N}} \to \mathsf{X}^{\mathbb{N}}$, given by

$$B_R(x) : n \mapsto x_n \mathbb{1}_{n<\tau_R(x)} + x_{\tau_R(x)} \mathbb{1}_{n \geq \tau_R(x)} \, .$$

Consider the image measure $\bar{\mathbb{P}}^{a,\gamma} := \mathbb{P}^{a,\gamma} B_R^{-1}$, which corresponds to the law of the *truncated* process $B_R(x)$. The crux of the proof consists in showing that for every $\eta > 0$ and every $M > 0$,

$$\sup_{a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}} \bar{\mathbb{P}}^{a,\gamma} \left(\mathsf{d}(\mathsf{X}_\gamma, \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), \cdot)) > \eta \right) \xrightarrow[\gamma \to 0]{} 0. \tag{4.19}$$

Eq. (4.19) is the counterpart of Lem. 3.6.3. Once it has been proven, the conclusion follows verbatim from Sec. 3.6, End of the proof. Our aim is thus to establish Eq. (4.19). The proof follows the same steps as the proof of Lem. 3.6.3 up to some confined changes. Here, the steps of the proof which do not need any modification are recalled rather briefly (we refer the reader to Chap. 3 for the details). On the other hand, the parts which require an adaptation are explicitly stated as lemmas, whose detailed proofs are provided in Appendix 4.8.

Define $h_{\gamma,R}(s,a) := h_\gamma(s,a) \mathbb{1}_{\|a\| \leq R}$. First, we recall the following decomposition, established in Chap. 3:

$$\mathsf{X}_\gamma = \Pi_0 + \mathsf{G}_{\gamma,R} \circ \mathsf{X}_\gamma + \mathsf{X}_\gamma \circ M_{\gamma,R} \, , \tag{4.20}$$

$\bar{\mathbb{P}}^{a,\gamma}$ almost surely, where $\Pi_0 : \mathsf{X}^{\mathbb{N}} \to C(\mathbb{R}_+, \mathsf{X})$, $\mathsf{G}_{\gamma,R} : C(\mathbb{R}_+, \mathsf{X}) \to C(\mathbb{R}_+, \mathsf{X})$ and $M_{\gamma,R} : \mathsf{X}^{\mathbb{N}} \to \mathsf{X}^{\mathbb{N}}$ are the mappings respectively defined by

$$\begin{aligned}
&\Pi_0(x) : t \mapsto x_0 \\
&M_{\gamma,R}(x) : n \mapsto (x_n - x_0) - \gamma \sum_{k=0}^{n-1} \int h_{\gamma,R}(s, x_k) \mu(ds) \\
&\mathsf{G}_{\gamma,R}(\mathsf{x}) : t \mapsto \int_0^t \int h_{\gamma,R}(s, \mathsf{x}(\gamma \lfloor u/\gamma \rfloor)) \, \mu(ds) du \, ,
\end{aligned}$$

for every $x = (x_n : n \in \mathbb{N})$ and every $\mathsf{x} \in C(\mathbb{R}_+, \mathsf{X})$.

Lemma 4.5.3. For all $\gamma \in (0, \gamma_0]$ and all $x \in \mathsf{X}^{\mathbb{N}}$, define $Z_{n+1}^\gamma(x) := \gamma^{-1}(x_{n+1} - x_n)$. There exists $\varepsilon > 0$ such that:

$$\sup_{n \in \mathbb{N}, a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}, \gamma \in (0,\gamma_0]} \bar{\mathbb{E}}^{a,\gamma} \left(\left(\|Z_n^\gamma\| + \frac{\boldsymbol{d}(x_n)}{\gamma} \mathbb{1}_{\|x_n\| \leq R} \right)^{1+\varepsilon} \right) < +\infty \tag{4.21}$$

Using Lem. 3.6.2, the uniform integrability condition (4.21) implies[1] that $\{\bar{\mathbb{P}}^{a,\gamma} \mathsf{X}_\gamma^{-1} : a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}, \gamma \in (0, \gamma_0]\}$ is tight, and for any $T > 0$,

$$\sup_{a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}} \bar{\mathbb{P}}^{a,\gamma}(\|\mathsf{X}_\gamma \circ M_{\gamma,R}\|_{\infty,T} > \varepsilon) \xrightarrow{\gamma \to 0} 0 \, , \tag{4.22}$$

where the notation $\|\mathsf{x}\|_{\infty,T}$ stands for the uniform norm of x on $[0,T]$.

[1] Lem. 3.6.2 of Chap. 3 was actually shown with condition $[a \in \mathcal{K}]$ instead of $[a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}]$, but the proof can be easily adapted to the latter case.

Lemma 4.5.4. For an arbitrary sequence (a_n, γ_n) such that $a_n \in \mathcal{K} \cap \mathcal{D}_{\gamma_n M}$ and $\gamma_n \to 0$, there exists a subsequence (still denoted as (a_n, γ_n)) such that $(a_n, \gamma_n) \to (a^*, 0)$ for some $a^* \in \mathcal{K} \cap \mathrm{cl}(\mathcal{D})$, and there exists r.v. z and $(\mathsf{x}_n : n \in \mathbb{N})$ defined on some probability space $(\Omega', \mathscr{F}', \mathbb{P}')$ into $C(\mathbb{R}_+, \mathsf{X})$ s.t. x_n has the distribution $\bar{\mathbb{P}}^{a_n,\gamma_n} \mathsf{X}_{\gamma_n}^{-1}$ and $\mathsf{x}_n(\omega) \to \mathsf{z}(\omega)$ for all $\omega \in \Omega'$. Moreover, defining

$$u_n(t) := \mathsf{x}_n(\gamma_n \lfloor t/\gamma_n \rfloor) \ ,$$

the sequence (a_n, γ_n) and (x_n) can be chosen in such a way that the following holds $\mathbb{P}'$-a.e.

$$\sup_n \int_0^T \left(\frac{\boldsymbol{d}(u_n(t))}{\gamma_n} \mathbb{1}_{\|u_n(t)\| \le R} \right)^{1+\frac{\varepsilon}{2}} dt \ < +\infty \quad (\forall T > 0) \, , \tag{4.23}$$

where $\varepsilon > 0$ is the constant introduced in Lem. 4.5.3.

From now on, the proof of the convergence 4.19 will use the maximal monotonicity of the operators, hence the proof will differ from the proof of 3.6.3. Define

$$\begin{aligned} v_n(s,t) &:= -B(s, u_n(t)) \mathbb{1}_{\|u_n(t)\| \le R} \\ w_n(s,t) &:= -A_{\gamma_n}(s, u_n(t) - \gamma_n B(s, u_n(t))) \mathbb{1}_{\|u_n(t)\| \le R} \, . \end{aligned}$$

Then, $v_n(s,t) + w_n(s,t) = h_{\gamma_n, R}(s, u_n(t))$. Thanks to the convergence (4.22) and Lem. 4.5.4, Eq. (4.20) becomes : $\mathbb{P}'$-a.e.,

$$\mathsf{z}(t) = \mathsf{z}(0) + \lim_{n \to \infty} \int_0^t \int_\Xi v_n(s,u) + w_n(s,u) \, \mu(ds) \, du \qquad (\forall t \ge 0) \, . \tag{4.24}$$

We now select an $\omega \in \Omega'$ s.t. the events (4.23) and (4.24) are realized, and omit the dependence in ω in the sequel. Otherwise stated, u_n, v_n and w_n are handled from now on as deterministic functions, and no longer as random variables. The aim of the next lemmas is to analyze the integrand $v_n(s,u) + w_n(s,u)$.

Consider some $T > 0$ and let λ_T represent the Lebesgue measure on the interval $[0, T]$. To simplify notations, we set $\mathcal{L}^{1+\varepsilon}_{\mathsf{X}} := \mathcal{L}^{1+\varepsilon}(\Xi \times [0,T], \mathscr{G} \otimes \mathscr{B}([0,T]), \mu \otimes \lambda_T; \mathsf{X})$.

Lemma 4.5.5. The sequences $(v_n)_n$ and $(w_n)_n$ are bounded in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}}$.

The sequence of mappings $((s,t) \mapsto (v_n(s,t), w_n(s,t)))_n$ is bounded in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}\times\mathsf{X}}$ and therefore admits a weak cluster point in that space. We denote by (v, w) such a cluster point, where $v, w : \Xi \times [0,T] \to \mathsf{X}$. Let $H_R(s,x) := -A(s,x) - B(s,x)$ if $\|x\| < R$, $H_R(s,x) := \mathsf{X}$ if $\|x\| = R$ and $H_R(s,x) = \{0\}$ if $\|x\| > R$. Denote the corresponding selection integral as $\mathsf{H}_R(x) = \int H_R(s,x) \, \mu(ds)$.

Lemma 4.5.6. For every (s,t) $\mu \otimes \lambda_T$-a.e., $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(s,\cdot))$.

By Lem. 4.5.6 and Fubini's theorem, there is a λ_T-negligible set s.t. for every t outside this set, $v(\cdot,t) + w(\cdot,t)$ is an integrable selection of $H_R(\cdot, \mathsf{z}(t))$. Moreover, as (v,w) is a weak cluster point of (v_n, w_n) in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}\times\mathsf{X}}$, it holds that

$$\mathsf{z}(t) = \mathsf{z}(0) + \int_0^t \int_\Xi v(s,u) + w(s,u) \, \mu(ds) \, du \, , \qquad (\forall t \in [0,T]) \, .$$

By the above equality, z is a solution to the DI $\dot{\mathsf{x}} \in \mathsf{H}_R(\mathsf{x})$ with initial condition $\mathsf{z}(0) = a^*$. Denoting by $\Phi_R(a^*)$ the set of such solutions, this reads $\mathsf{z} \in \Phi_R(a^*)$. As $a^* \in \mathcal{K} \cap \mathrm{cl}(\mathcal{D})$, one has $\mathsf{z} \in \Phi_R(\mathcal{K} \cap \mathrm{cl}(\mathcal{D}))$ where we use the notation $\Phi_R(S) := \cup_{a \in S} \Phi_R(a)$ for every set $S \subset \mathsf{X}$. Extending the notation $\mathsf{d}(x, \mathsf{S}) :=$

$\inf_{y\in S} d(x,y)$, we obtain that $d(x_n, \Phi_R(\mathcal{K}\cap \mathrm{cl}(\mathcal{D}))) \to 0$. Thus, for every $\eta>0$, we have shown that $\bar{\mathbb{P}}^{a_n,\gamma_n}(d(X_{\gamma_n}, \Phi_R(\mathcal{K}\cap \mathrm{cl}(\mathcal{D}))) > \eta) \to 0$ as $n\to\infty$. We have thus proven the following result:

$$\forall \eta>0,\ \lim_{\gamma\to 0} \sup_{a\in\mathcal{K}\cap\mathcal{D}_{\gamma M}} \bar{\mathbb{P}}^{a,\gamma}(d(X_\gamma, \Phi_R(\mathcal{K}\cap \mathrm{cl}(\mathcal{D}))) > \eta) = 0\,.$$

Letting $T>0$ and choosing $R > \sup\{\|\Phi(a,t)\| : t\in[0,T], a\in\mathcal{K}\cap\mathrm{cl}(\mathcal{D})\}$ (the latter quantity being finite, see *e.g.* [36]), it is easy to show that any solution to the DI $\dot{x}\in H_R(x)$ with initial condition $a\in\mathcal{K}\cap\mathrm{cl}(\mathcal{D})$ coincides with $\Phi(a,\,.\,)$ on $[0,T]$. By the same arguments as in [28, Sec. 4 - End of the proof], Th. 4.5.1 follows.

4.6 Cluster points of the P_γ invariant measures. End of the proof of Th. 4.3.2

Lemma 4.6.1. Assume that there exists $x_\star\in Z(\mathcal{A}+\mathcal{B})$ that admits a $\mathcal{L}^2$ representation. Then,

$$P_\gamma(x, \|\cdot - x_\star\|^2) \le \|x-x_\star\|^2 - 0.5\gamma\psi_\gamma(x) + \gamma^2 C,$$

where ψ_γ is the function defined in (4.6).

Proof. By assumption, there exists a $\mathcal{L}^2$ representation (φ, B) of $x_\star$. By expanding

$$\|x_{n+1}-x_\star\|^2 = \|x_n-x_\star\|^2 + 2\langle x_{n+1}-x_n, x_n-x_\star\rangle + \|x_{n+1}-x_n\|^2\,,$$

and by using (4.17), we obtain

$$\begin{aligned}\|x_{n+1}-x_\star\|^2 = \|x_n-x_\star\|^2 &- 2\gamma\langle A_\gamma(\xi_{n+1}, x_n-\gamma B(\xi_{n+1},x_n)) + B(\xi_{n+1},x_n), x_n-x_\star\rangle \\ &+ \gamma^2\|A_\gamma(\xi_{n+1}, x_n-\gamma B(\xi_{n+1},x_n)) + B(\xi_{n+1},x_n)\|^2. \quad (4.25)\end{aligned}$$

Write $x=x_n$, $A_\gamma = A_\gamma(\xi_{n+1}, x_n-\gamma B(\xi_{n+1},x_n))$, $J_\gamma = J_\gamma(\xi_{n+1}, x_n-\gamma B(\xi_{n+1},x_n))$, $B=B(\xi_{n+1},x_n)$, $B_\star = (\xi_{n+1}, x_\star)$, and $\varphi=\varphi(\xi_{n+1})$ for conciseness. We write

$$\begin{aligned}\langle A_\gamma, x-x_\star\rangle &= \langle A_\gamma-\varphi, J_\gamma - x_\star\rangle + \langle A_\gamma-\varphi, x-\gamma B - J_\gamma\rangle + \gamma\langle A_\gamma-\varphi, B\rangle \\ &\quad + \langle\varphi, x-x_\star\rangle \\ &= \langle A_\gamma-\varphi, J_\gamma-x_\star\rangle + \gamma\|A_\gamma\|^2 - \gamma\langle A_\gamma,\varphi\rangle + \gamma\langle A_\gamma-\varphi, B\rangle + \langle\varphi, x-x_\star\rangle.\end{aligned}$$

We also write $\langle B, x-x_\star\rangle = \langle B-B_\star, x-x_\star\rangle + \langle B_\star, x-x_\star\rangle$ and $\gamma^2\|A_\gamma+B\|^2 = \gamma^2(\|A_\gamma\|^2+\|B\|^2 + 2\langle A_\gamma, B\rangle)$. Plugging these identities at the right hand side of (4.25), we obtain

$$\begin{aligned}\|x_{n+1}-x_\star\|^2 &= \|x-x_\star\|^2 - 2\gamma\left\{\langle A_\gamma-\varphi, J_\gamma-x_\star\rangle + \langle B-B_\star, x-x_\star\rangle\right\} - \gamma^2\|A_\gamma\|^2 \\ &\quad + 2\gamma^2\langle A_\gamma,\varphi\rangle + 2\gamma^2\langle\varphi, B\rangle + \gamma^2\|B\|^2 - 2\gamma\langle\varphi+B_\star, x-x_\star\rangle \\ &\le \|x-x_\star\|^2 - 2\gamma\left\{\langle A_\gamma-\varphi, J_\gamma-x_\star\rangle + \langle B-B_\star, x-x_\star\rangle\right\} - (\gamma^2/2)\|A_\gamma\|^2 \\ &\quad + (3\gamma^2/2)\|B\|^2 + 4\gamma^2\|\varphi\|^2 - 2\gamma\langle\varphi+B_\star, x-x_\star\rangle \\ &\le \|x-x_\star\|^2 - 2\gamma\left\{\langle A_\gamma-\varphi, J_\gamma-x_\star\rangle + \langle B-B_\star, x-x_\star\rangle\right\} - (\gamma^2/2)\|A_\gamma\|^2 \\ &\quad + 3\gamma^2\|B-B_\star\|^2 + 3\gamma^2\|B_\star\|^2 + 4\gamma^2\|\varphi\|^2 - 2\gamma\langle\varphi+B_\star, x-x_\star\rangle\end{aligned}$$

where the first inequality is due to the fact that $2\langle a,b\rangle \le \|a\|^2/2 + 2\|b\|^2$ and the second to the triangle inequality. Observe that the term between the braces at the right hand side of the last inequality is

nonnegative thanks to the monotonicity of $A(s,\cdot)$ and $B(s,\cdot)$. Taking the conditional expectation $\mathbb{E}_n$ at each side, the contribution of the last inner product at the right hand side disappears, and we obtain

$$P_\gamma(x, \|\cdot - x_\star\|^2) \le \|x - x_\star\|^2 - 0.5\gamma\psi_\gamma(x) + 4\gamma^2 \int \|\varphi(s)\|^2 \mu(ds) + 3\gamma^2 \int \|B(s, x_\star)\|^2 \mu(ds)$$

where ψ_γ is the function defined in (4.6). □

Given $k \in \mathbb{N}$, we denote by P_γ^k the kernel P_γ iterated k times. The iterated kernel is defined recursively as $P_\gamma^0(x, dy) = \delta_x(dy)$, and

$$P_\gamma^k(x, S) = \int P_\gamma^{k-1}(y, S)\, P_\gamma(x, dy)$$

for each $S \in \mathscr{B}(\mathsf{X})$.

Lemma 4.6.2. Let the assumptions of the statement of Th. 4.5.1 hold true. Assume that for all $\varepsilon > 0$, there exists $M > 0$ such that

$$\sup_{\gamma \in (0,\gamma_0]} \sup_{\pi \in \mathcal{I}(P_\gamma)} \pi((\mathcal{D}_{M\gamma})^c) \le \varepsilon. \tag{4.26}$$

Then, as $\gamma \to 0$, any cluster point of $\mathcal{I}(\mathcal{P})$ is an element of $\mathcal{I}(\Phi)$.

Note that in the common domain case, (4.26) is trivially satisfied, since the supports of all the invariant measures are included in $\mathrm{cl}(\mathcal{D})$.

Proof. Choose two sequences (γ_i) and (π_i) such that $\gamma_i \to 0$, $\pi_i \in \mathcal{I}(P_{\gamma_i})$ for all $i \in \mathbb{N}$, and π_i converges narrowly to some $\pi \in \mathcal{M}(\mathsf{X})$ as $i \to \infty$.

Let f be a real, bounded, and Lipschitz function on X with Lipschitz coefficient L. By definition, $\pi_i(f) = \pi_i(P_{\gamma_i}^k f)$ for all $k \in \mathbb{N}$. Set $t > 0$, and let $k_i = \lfloor t/\gamma_i \rfloor$. We have

$$\begin{aligned}
|\pi_i f - \pi_i(f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), t))| &= \left| \int (P_{\gamma_i}^{k_i}(a, f) - f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), t))) \pi_i(da) \right| \\
&\le \int \left| P_{\gamma_i}^{k_i}(a, f) - f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)) \right| \pi_i(da) \\
&\quad + \int \left| f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)) - f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), t)) \right| \pi_i(da) \\
&\le \int \mathbb{E}^{a,\gamma_i} \left| f(x_{k_i}) - f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)) \right| \pi_i(da) \\
&\quad + \int \left| f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)) - f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), t)) \right| \pi_i(da) \\
&:= U_i + V_i\,.
\end{aligned}$$

By the boundedness and the Lispchitz-continuity of f,

$$U_i \le \int \mathbb{E}^{a,\gamma_i} \left[2\|f\|_\infty \wedge L \|x_{k_i} - \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)\| \right] \pi_i(da)\,.$$

Fixing an arbitrarily small $\varepsilon > 0$, it holds by (4.26) that $\pi_i((\mathcal{D}_{M\gamma_i})^c) \le \varepsilon/2$ for a large enough M. By the tightness of (π_i), we can choose a compact $\mathcal{K} \subset \mathsf{X}$ s.t. for all i, $\pi_i(\mathcal{K}^c) \le \varepsilon/2$. With these choices, we obtain

$$U_i \le \sup_{a \in \mathcal{K} \cap \mathcal{D}_{M\gamma_i}} \mathbb{E}^{a,\gamma_i} \left[2\|f\|_\infty \wedge L \|x_{k_i} - \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)\| \right] + 2\|f\|_\infty \varepsilon\,.$$

Denoting as $(\cdot)_{[0,t]}$ the restriction of a function to the interval $[0,t]$, and observing that

$$\|x_{k_i} - \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)\| \leq \|(\mathsf{X}_\gamma(x) - \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), \cdot))_{[0,t]}\|_\infty,$$

we can now apply Th. 4.5.1 to obtain

$$\sup_{a\in\mathcal{K}\cap\mathcal{D}_{M\gamma_i}} \mathbb{E}^{a,\gamma_i}\left[2\|f\|_\infty \wedge L\|x_{k_i} - \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)\|\right] \xrightarrow[i\to\infty]{} 0\,.$$

As ε is arbitrary, we obtain that $U_i \to_i 0$. Turning to V_i, fix an arbitrary $\varepsilon > 0$, and choose a compact $\mathcal{K} \subset \mathsf{X}$ such that $\pi_i(\mathcal{K}^c) \leq \varepsilon$ for all i. We have

$$V_i \leq \sup_{a\in\mathcal{K}} \left|f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), k_i\gamma_i)) - f(\Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), t))\right| + 2\|f\|_\infty\varepsilon\,.$$

By the uniform continuity of the function $f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), \cdot)$ on the compact $\mathcal{K} \times [0,t]$, and by the convergence $k_i\gamma_i \uparrow t$, we obtain that $\limsup_i V_i \leq 2\|f\|_\infty\varepsilon$. As ε is arbitrary, $V_i \to_i 0$. In conclusion, $\pi_i f - \pi_i(f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), t)) \to_i 0$. Moreover, $\pi_i f - \pi_i(f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), t)) \to_i \pi f - \pi(f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), t))$ since $f(\cdot) - f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), t))$ is bounded continuous. Thus, $\pi f = \pi(f \circ \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\cdot), t))$. Since π_i converges narrowly to π, we obtain that for all $\eta > 0$, $\pi(\mathrm{cl}(\mathcal{D}_\eta)^c) \leq \liminf_i \pi_i(\mathrm{cl}(\mathcal{D}_\eta)^c) = 0$ by choosing ε arbitrarily small in (4.26) and making $\gamma_i \to 0$. Thus, $\mathrm{supp}(\pi) \subset \mathrm{cl}(\mathcal{D})$, and we obtain in conclusion that $\pi f = \pi(f \circ \Phi(\cdot, t))$ for an arbitrary real, bounded, and Lipschitz continuous function f. Thus, $\pi \in \mathcal{I}(\Phi)$. □

To establish (4.26) in the different domains case, we need the following lemma.

Lemma 4.6.3. Let Assumptions 4.3.6, 4.3.8, and 4.3.4–(a) hold true. Then, for all $\varepsilon > 0$, there exists $M > 0$ such that

$$\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi((\mathcal{D}_{M\gamma})^c) \leq \varepsilon.$$

Proof. We start by writing

$$\boldsymbol{d}(x_{n+1}) \leq \|x_{n+1} - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\| \leq \|x_{n+1} - \Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n)\| + \|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|.$$

On the one hand, we have by Assumption 4.3.8 and the nonexpansiveness of the resolvent that

$$\begin{aligned}\mathbb{E}_n^{a,\gamma}\|x_{n+1} - \Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n)\| &\leq \mathbb{E}_n^{a,\gamma}\|J_\gamma(\xi_{n+1}, x_n) - \Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n)\| + \gamma\mathbb{E}_n^{a,\gamma}\|B(\xi_{n+1}, x_n)\| \\ &\leq C\gamma(1 + \psi_\gamma(x_n))\,,\end{aligned}$$

on the other hand, since

$$\|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2 \leq \boldsymbol{d}(x_n)^2 - d(x_n, D(\xi_{n+1}))^2 \quad \text{(see (4.28))},$$

we can make use of Assumption 4.3.6 to obtain

$$\mathbb{E}_n^{a,\gamma}\|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\| \leq (\mathbb{E}_n^{a,\gamma}\|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2)^{1/2} \leq \rho\boldsymbol{d}(x_n)\,,$$

where $\rho \in [0,1)$. We therefore obtain that

$$\mathbb{E}_n^{a,\gamma}\boldsymbol{d}(x_{n+1}) \leq \rho\boldsymbol{d}(x_n) + C\gamma(1 + \psi_\gamma(x_n)).$$

By iterating, we end up with the inequality

$$P_\gamma^{n+1}(a,\boldsymbol{d}) \le \rho^{n+1}\boldsymbol{d}(a) + C\gamma \sum_{k=0}^{n} \rho^{n-k}(1 + P_\gamma^k(a,\psi_\gamma)). \tag{4.27}$$

From Assumption 4.3.4–(a) and Lem. 4.6.1, the inequality (4.13) in the statement of Prop. 4.3.3 is satisfied with $V(x) = \|x - x_\star\|^2$, $Q(x) = \Psi(x)$, $\alpha(\gamma) = \gamma/2$, and $\beta(\gamma) = C\gamma^2$. By the first part of this proposition, $\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi(\Psi) < \infty$. In particular, noting that $\boldsymbol{d}(x) \le \|x\| + \|\Pi_{\mathrm{cl}(\mathcal{D})}(0)\|$, we obtain that $\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi(\boldsymbol{d}) < \infty$. Moreover, with a small adaptation of the proof of Prop. 3.8.7 in Chap. 3 to the inequality of Lem. 4.6.1, we can show the slightly stronger result that $\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi(\psi_\gamma) < \infty$. Let $\gamma \in (0,\gamma_0]$ and $\pi \in \mathcal{I}(P_\gamma)$. We can integrate w.r.t π in (4.27) to obtain

$$\pi(\boldsymbol{d}) \le \rho^{n+1}\pi(\boldsymbol{d}) + C\gamma \sum_{k=0}^{n} \rho^{n-k}(1 + \pi(\psi_\gamma)).$$

Using Markov's inequality, we have for all $n \in \mathbb{N}$,

$$\pi((\mathcal{D}_{M\gamma})^c) \le \frac{\pi(\boldsymbol{d})}{M\gamma} \le \frac{\rho^{n+1}}{M\gamma}\pi(\boldsymbol{d}) + \frac{C}{M}\sum_{k=0}^{n} \rho^{n-k}(1 + \pi(\psi_\gamma)) \le \frac{\rho^{n+1}C}{M\gamma} + \frac{C}{M}.$$

By making $n \to \infty$, we obtain that $\pi((\mathcal{D}_{M\gamma})^c) \le C/M$, and the proof is concluded by taking M as large as required. □

Th. 4.3.2: proofs of the convergences (4.7), (4.9), and (4.10)

We need to check that the assumptions of Prop. 4.3.3 are satisfied. Lem. 4.6.1 shows that the inequality (4.13) is satisfied with $V(x) = \|x - x_\star\|^2$, $Q(x) = \Psi(x)$, $\alpha(\gamma) = \gamma/2$, and $\beta(\gamma) = C\gamma^2$, and Assumption 4.3.4–(a) ensures that $\Psi(x) \xrightarrow[\|x\|\to\infty]{} \infty$ as required.

When the assumptions of Th. 4.5.1, are satisfied, Lem. 4.6.2 shows with the help of Lem. 4.6.3 when needed that any cluster point of $\mathcal{I}(\mathcal{P})$ belongs to $\mathcal{I}(\Phi)$. The required convergences follow at once from Prop. 4.3.3. Th. 4.3.2 is proven.

4.7 Proofs relative to Sec. 4.4

4.7.1 Proof of Prop. 4.4.1

It is well known that the coercivity or the supercoercivity of a function $q \in \Gamma_0(\mathsf{X})$ can be characterized through the study of the recession function q^∞ of q, which is the function in $\Gamma_0(\mathsf{X})$ whose epigraph is the recession cone of the epigraph of q [101, §8], [76, § 6.8]. We recall the following fact.

Lemma 4.7.1. The function $q \in \Gamma_0(\mathsf{X})$ is coercive if and only if 0 is the only solution of the inequality $q^\infty(x) \le 0$. It is supercoercive if and only if $q^\infty = \iota_{\{0\}}$.

Proof. By [76, Prop. 6.8.4], $\mathrm{lev}_{\le 0}\, q^\infty$ is the recession cone of any level set $\mathrm{lev}_{\le a}\, q$ which is not empty [101, Th. 8.6]. Thus, q is coercive if and only if $\mathrm{lev}_{\le 0}\, q^\infty$ is the recession cone of a nonempty compact set, hence equal to $\{0\}$. The second point follows from [10, Prop. 2.16]. □

Lemma 4.7.2. For each $\gamma > 0$, $q^\infty = (q_\gamma)^\infty$.

Proof. By [76, Th. 6.8.5], the Legendre-Fenchel transform $(q^\infty)^*$ of q^∞ satisfies $(q^\infty)^* = \iota_{\mathrm{cl}\,\mathrm{dom}\,q^*}$. Since $q_\gamma = q \,\square\, ((2\gamma)^{-1}\|\cdot\|^2)$ where $\square$ is the infimal convolution operator, $(q_\gamma)^* = q^* + (\gamma/2)\|\cdot\|^2$. Therefore, $\mathrm{dom}\, q^* = \mathrm{dom}(q_\gamma)^*$, which implies that $(q^\infty)^* = ((q_\gamma)^\infty)^*$, and the result follows. □

Lemma 4.7.3 ([66, Th. II.2.1])**.** Assume that $q : \Xi \times \mathsf{X} \to (-\infty, \infty]$ is a normal integrand such that $q(s,\cdot) \in \Gamma_0(\mathsf{X})$ for almost every s. Assume that $Q(x) := \int q(s,x)\,\mu(ds)$ belongs to $\Gamma_0(\mathsf{X})$. Then, $Q^\infty(x) = \int q^\infty(s,x)\,\mu(ds)$, where $q^\infty(s,\cdot)$ is the recession function of $q(s,\cdot)$.

We now enter the proof of Prop. 4.4.1. Denote by $g_\gamma(s,\cdot)$ the Moreau envelope of the mapping $g(s,\cdot)$ defined above.

Lemma 4.7.4. Let Hypothesis H1 hold true. Then, for all $\gamma > 0$, the mapping

$$G^\gamma : x \mapsto \int g_\gamma(s,x)\,\mu(ds)\,,$$

is well defined on $\mathsf{X} \to \mathbb{R}$, and is convex (hence continuous) on X. Moreover, $G^\gamma \uparrow G$ as $\gamma \downarrow 0$.

Proof. Since $x_\star \in \mathrm{dom}\, G$ from Hypothesis H1, it holds from the definition of the function g that $\int |g(s,x_\star)|\,\mu(ds) < \infty$. Moreover, noting that $\varphi(s) \in \partial g(s,x_\star)$, the inequality $g(s,x) \geq \langle \varphi(s), x - x_\star\rangle + g(s,x_\star)$ holds. Thus,

$$\begin{aligned} g_\gamma(s,x) &= \inf_w \Big(g(s,w) + \frac{1}{2\gamma}\|w-x\|^2 \Big) \geq \inf_w \Big(\langle \varphi(s), w - x_\star\rangle + g(s,x_\star) + \frac{1}{2\gamma}\|w-x\|^2 \Big) \\ &= \langle \varphi(s), x - x_\star\rangle + g(s,x_\star) - \frac{\gamma}{2}\|\varphi(s)\|^2. \end{aligned}$$

Writing $x = x^+ - x^-$ where $x^+ = x \vee 0$, this inequality shows that $g_\gamma(\cdot,x_\star)^-$ is integrable. Moreover, since the Moreau envelope satisfies $g_\gamma(s,x) \leq g(s,x)$, we obtain that $g_\gamma(\cdot,x_\star)^+ \leq g(\cdot,x_\star)^+ \leq |g(\cdot,x_\star)|$ who is also integrable. Therefore, $|g_\gamma(\cdot,x_\star)|$ is integrable. For other values of x, we have

$$g_\gamma(s,x) = g_\gamma(s,x_\star) + \int_0^1 \langle x - x_\star, \nabla g_\gamma(s, x_\star + t(x-x_\star)) - \nabla g_\gamma(s,x_\star)\rangle\, dt + \langle x - x_\star, \nabla g_\gamma(s,x_\star)\rangle,$$

where $\nabla g_\gamma(s,x)$ is the gradient of $g_\gamma(s,x)$ w.r.t. x. Using the well know properties of the Yosida regularization (see Sec. 2.2.1), we obtain

$$|g_\gamma(s,x)| \leq |g_\gamma(s,x_\star)| + \frac{\|x-x_\star\|^2}{2\gamma} + \|x-x_\star\|\,\|\varphi(s)\|^2.$$

Consequently, $g_\gamma(\cdot,x)$ is integrable, thus, $G^\gamma(x)$ is defined for all $x \in \mathsf{X}$. The convexity and hence the continuity of G^γ follow trivially from the convexity of $g_\gamma(s,\cdot)$.

Since the integrand $g_\gamma(s,x)$ increases as γ decreases, so is the case of $G^\gamma(x)$. If $x \in \mathrm{dom}(G)$, it holds that $|g(\cdot,x)|$ is integrable. On the one hand, $g_\gamma(s,x)^+ \leq |g(s,x)|$, and on the other hand, $g_\gamma(s,x)^- \leq \|\varphi(s)\|\|x-x_\star\| + |g(s,x_\star)| + \|\varphi(s)\|^2$ for $\gamma \leq 2$. By the dominated convergence, $G^\gamma(x) \to G(x)$ as $\gamma \to 0$. If $x \notin \mathrm{dom}\, G$, then $\int g_\gamma(s,x)^+\mu(ds) \to \infty$ as $\gamma \to 0$ by monotone convergence, and $\int g_\gamma(s,x)^-\mu(ds)$ remains bound. Thus, $G^\gamma(x) \to \infty$. □

Lemma 4.7.5. Let Hypotheses H1 and H2 hold true. Then, for all γ small enough,

$$G^\gamma(x) + F(x) - G^\gamma(x_\star) - F(x_\star) \leq 2\psi_\gamma(x) + \gamma C,$$

where ψ_γ is given by (4.6).

Proof. By the convexity of $g_\gamma(s,\cdot)$ and $f(s,\cdot)$, we have

$$g_\gamma(s, x-\gamma\nabla f(s,x)) - g_\gamma(s,x_\star) \le \langle \nabla g_\gamma(s, x-\gamma\nabla f(s,x)), x-\gamma\nabla f(s,x) - x_\star\rangle, \text{ and}$$
$$f(s,x) - f(s,x_\star) - \langle \nabla f(s,x_\star), x-x_\star\rangle \le \langle \nabla f(s,x) - \nabla f(s,x_\star), x-x_\star\rangle.$$

Write $g_\gamma = g_\gamma(s, x-\gamma\nabla f(s,x))$, $\nabla f = \nabla f(s,x)$, $\mathrm{prox}_\gamma = \mathrm{prox}_{\gamma g(s,\cdot)}(x-\gamma\nabla f(s,x))$, $\varphi = \varphi(s)$, and $\nabla f_\star = \nabla f(s,x_\star)$. From these two inequalities, we obtain

$$\begin{aligned}
&g_\gamma(s, x-\gamma\nabla f(s,x)) - g_\gamma(s,x_\star) + f(s,x) - f(s,x_\star) - \langle \varphi(s) + \nabla f(s,x_\star), x-x_\star\rangle \\
&\le \langle \nabla g_\gamma, x - \gamma\nabla f - x_\star + \mathrm{prox}_\gamma - \mathrm{prox}_\gamma\rangle + \langle \nabla f - \nabla f_\star, x-x_\star\rangle - \langle \varphi, x - x_\star + \mathrm{prox}_\gamma - \mathrm{prox}_\gamma\rangle \\
&= \langle \nabla g_\gamma - \varphi, \mathrm{prox}_\gamma - x_\star\rangle + \langle \nabla f - \nabla f_\star, x-x_\star\rangle + \gamma\|\nabla g_\gamma\|^2 - \gamma\langle \varphi, \nabla g_\gamma + \nabla f\rangle.
\end{aligned}$$

Again, by the convexity of $g_\gamma(s,\cdot)$, we have

$$g_\gamma(s, x-\gamma\nabla f(s,x)) \ge g_\gamma(s,x) - \gamma\langle \nabla g_\gamma(s,x), \nabla f(s,x)\rangle.$$

Thus, we obtain

$$\begin{aligned}
&g_\gamma(s,x) - g_\gamma(s,x_\star) + f(s,x) - f(s,x_\star) - \langle \varphi(s) + \nabla f(s,x_\star), x-x_\star\rangle \\
&\le \langle \nabla g_\gamma - \varphi, \mathrm{prox}_\gamma - x_\star\rangle + \langle \nabla f - \nabla f_\star, x-x_\star\rangle + \gamma\|\nabla g_\gamma\|^2 - \gamma\langle \varphi, \nabla g_\gamma + \nabla f\rangle + \gamma\langle \nabla g_\gamma(s,x), \nabla f\rangle.
\end{aligned}$$

We now bound the sum of the last two terms at the right hand side. By the γ^{-1}-Lipschitz continuity of the Yosida regularization, $|\langle \nabla g_\gamma(s,x) - \nabla g_\gamma, \nabla f\rangle| \le \|\nabla f\|^2$. Using in addition the inequalities $|\langle a,b\rangle| \le \|a\|^2/2 + \|b\|^2/2$ and $\|\nabla f\|^2 \le 2\|\nabla f_\star\|^2 + 2\|\nabla f - \nabla f_\star\|^2$, we obtain

$$\begin{aligned}
\gamma\langle \nabla g_\gamma(s,x), \nabla f\rangle - \gamma\langle \varphi, \nabla g_\gamma + \nabla f\rangle &= \gamma\langle \nabla g_\gamma(s,x) - \nabla g_\gamma, \nabla f\rangle + \gamma\langle \nabla g_\gamma, \nabla f\rangle - \gamma\langle \varphi, \nabla g_\gamma + \nabla f\rangle \\
&\le 2\gamma\|\nabla f\|^2 + \gamma\|\nabla g_\gamma\|^2 + \gamma\|\varphi\|^2 \\
&\le 4\gamma\|\nabla f - \nabla f_\star\|^2 + 4\gamma\|\nabla f_\star\|^2 + \gamma\|\nabla g_\gamma\|^2 + \gamma\|\varphi\|^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&g_\gamma(s,x) - g_\gamma(s,x_\star) + f(s,x) - f(s,x_\star) - \langle \varphi(s) + \nabla f(s,x_\star), x-x_\star\rangle \\
&\le 2\left(\langle \nabla g_\gamma - \varphi, \mathrm{prox}_\gamma - x_\star\rangle + \langle \nabla f - \nabla f_\star, x-x_\star\rangle + \gamma\|\nabla g_\gamma\|^2 - 6\gamma\|\nabla f - \nabla f_\star\|^2\right) \\
&\quad + 16\gamma\|\nabla f - \nabla f_\star\|^2 - \langle \nabla f - \nabla f_\star, x-x_\star\rangle + \gamma\|\varphi\|^2 + 4\gamma\|\nabla f_\star\|^2.
\end{aligned}$$

Taking the integral with respect to $\mu(ds)$ at both sides, the contribution of the inner product $\langle \varphi + \nabla f_\star, x-x_\star\rangle$ vanishes. Recalling (4.6), we obtain

$$\begin{aligned}
&G_\gamma(x) + F(x) - G_\gamma(x_\star) - F(x_\star) \\
&\le 2\psi_\gamma(x) - \int \left(\langle \nabla f - \nabla f_\star, x-x_\star\rangle - 16\gamma\|\nabla f - \nabla f_\star\|^2\right) d\mu + \gamma\int \left(\|\varphi\|^2 + 4\|\nabla f_\star\|^2\right) d\mu.
\end{aligned}$$

Using Hypothesis H2, we obtain the desired result. □

End of the proof of Prop. 4.4.1. Let $\gamma_0 > 0$ be such that Lem. 4.7.5 holds true for all $\gamma \in (0, \gamma_0]$. Denoting as $q(s,\cdot)^\infty$ the recession function of $q(s,\cdot)$, we have

$$(G^{\gamma_0} + F)^\infty \overset{\text{(a)}}{=} \int \left((g_{\gamma_0}(s,\cdot))^\infty + f(s,\cdot)^\infty\right) \mu(ds) \overset{\text{(b)}}{=} \int \left(g(s,\cdot)^\infty + f(s,\cdot)^\infty\right) \mu(ds) \overset{\text{(c)}}{=} (G+F)^\infty,$$

where the equalities (a) and (c) are due to Lem. 4.7.3, and (b) is due to Lem. 4.7.2. Thus, by Lem. 4.7.1, $F+G$ is coercive (resp. supercoercive) if and only if $F+G^{\gamma_0}$ is coercive (resp. supercoercive). Consequently, since G^γ increases as γ decreases by Lem. 4.7.4, the hypotheses H1, H2, and H3–(a) (resp., H1, H2, H3–(b)) imply Assumption 4.3.4–(a) (resp. Assumption 4.3.4–(b)). Prop. 4.4.1 is proven.

4.7.2 Proof of Lem. 4.4.2

We first recall that $\partial G(\cdot) = \int \partial g(s,\cdot)\,\mu(ds)$, where

$$\partial g(s,\cdot) = \begin{cases} \alpha(0)^{-1}\partial h(u,\cdot) & \text{if } i=0,\\ \partial \iota_{\mathcal{C}_i} & \text{otherwise,}\end{cases}$$

for $s=(u,i)\in\Xi$. Let ψ be an arbitrary measurable $\Sigma\to\mathsf{X}$ function such that $\psi(u)\in\partial h(u,x_\star)$ for ζ-almost all $u\in\Sigma$ (such functions are called *measurable selections* of the set-valued function $\partial h(\cdot,x_\star)$). For each $d\in\mathsf{X}$, it holds by the convexity of $h(u,\cdot)$ that

$$\begin{aligned} h(u,x_\star+d) &\geq h(u,x_\star)+\langle\psi(u),d\rangle, \text{ and}\\ h(u,x_\star-d) &\geq h(u,x_\star)-\langle\psi(u),d\rangle,\end{aligned}$$

for ζ-almost all $u\in\Sigma$. Equivalently,

$$h(u,x_\star)-h(u,x_\star-d)\leq\langle\psi(u),d\rangle\leq h(u,x_\star+d)-h(u,x_\star).$$

Thus, if $\|d\|$ is small enough but otherwise d is arbitrary, we get from the second assumption of the statement that $\langle\psi(u),d\rangle$ is ζ-square-integrable. Thus, $\int\|\psi(u)\|^2\,\zeta(du)<\infty$ (see [66, Th. II.4.2] for a similar argument). Now, writing $s=(u,i)\in\Xi$, every measurable selection ϕ of $\partial g(\cdot,x_\star)$ is of the form

$$\phi(s)=\begin{cases}\alpha(0)^{-1}\psi(u) & \text{if } i=0,\\ \theta_i & \text{otherwise,}\end{cases}$$

where ψ is a measurable selection of $\partial h(\cdot,x_\star)$, and θ_i is an element of $\partial\iota_{\mathcal{C}_i}(x_\star)$. By what precedes, it is immediate that $\int\|\phi\|^2 d\mu<\infty$. By assumption, there exists a measurable selection φ of $\partial g(\cdot,x_\star)$ such that $\int(\varphi(s)+\nabla f(u,x_\star))\mu(ds)=0$. Using the first assumption, we get that the couple $(\varphi(s),\nabla f(u,x_\star))$ is a $\mathcal{L}^2$ representation of $x_\star$.

4.7.3 Proof of Prop. 4.4.4

The assertions about $Z(\mathcal{A})$ are straightforward. A small calculation shows that

$$\begin{aligned} J_\gamma(s,x) &= (I+\gamma H(s))^{-1}(x-\gamma d(s)), \quad \text{and}\\ A_\gamma(s,x) &= A(s,J_\gamma(s,x)) = (I+\gamma H(s))^{-1}(H(s)x+d(s)).\end{aligned}$$

Using these expressions, we obtain

$$\begin{aligned}\psi_\gamma(x) &= \int\Big\{\langle A(s,J_\gamma(s,x))-H(s)x_\star-d(s),J_\gamma(s,x)-x_\star\rangle+\gamma\|A(s,J_\gamma(s,x))\|^2\Big\}\,\mu(ds)\\ &= \int\Big\{(J_\gamma(s,x)-x_\star)^T\frac{H(s)+H^T(s)}{2}(J_\gamma(s,x)-x_\star)+\gamma\|A(s,J_\gamma(s,x))\|^2\Big\}\,\mu(ds).\end{aligned}$$

Since $(I+\gamma H(s))^{-1}$ and $H(s)(I+\gamma H(s))^{-1}$ are respectively the resolvent and the Yosida regularization of the linear, monotone and maximal operator $H(s)$, it holds that $\|(I+\gamma H(s))^{-1}\|\leq 1$, and $\|\gamma H(s)(I+\gamma H(s))^{-1}\|\leq 1$.

Denoting as $\|\cdot\|_S$ the semi norm associated with any semidefinite nonnegative matrix S, we write

$$\begin{aligned}\psi_\gamma(x) &\geq \int\|J_\gamma(s,x)-x_\star\|^2_{(H(s)+H^T(s))/2}\,\mu(ds)\\ &= \int\Big\|(I+\gamma H(s))^{-1}\Big((x-x_\star)-\gamma(H(s)x_\star+d(s))\Big)\Big\|^2_{(H(s)+H^T(s))/2}\,\mu(ds).\end{aligned}$$

Using the inequality $\|a-b\|^2 \geq 0.5\|a\|^2 - \|b\|^2$, we obtain that $\psi_\gamma(x) \geq 0.5 W_\gamma(x) - U_\gamma$, with

$$W_\gamma(x) = \int \left\|(I+\gamma H(s))^{-1}(x-x_\star)\right\|^2_{(H(s)+H^T(s))/2}\, \mu(ds), \quad \text{and}$$
$$U_\gamma = \gamma^2 \int \left\|(I+\gamma H(s))^{-1}(H(s)x_\star + d(s))\right\|^2_{(H(s)+H^T(s))/2}\, \mu(ds)$$
$$= \gamma \int \|H(s)x_\star + d(s))\|^2_{\gamma I_\gamma(s)}\, \mu(ds).$$

with

$$I_\gamma(s) = (I+\gamma H(s))^{-T} \frac{H(s)+H^T(s)}{2} (I+\gamma H(s))^{-1}.$$

From the inequalities shown above, we have

$$\left\|\gamma I_\gamma(s)\right\| \leq 1.$$

Therefore,

$$0 \leq U_\gamma \leq \gamma \int \|H(s)x_\star + d(s)\|^2\, \mu(ds) \leq \gamma C.$$

Turning to $W_\gamma(x)$, it holds that

$$W_\gamma(x) = (x-x_\star)^T \Big(\int I_\gamma(s)\, \mu(ds) \Big)(x-x_\star),$$

Since $\|I_\gamma(s)\| \leq \left\|\frac{H(s)+H^T(s)}{2}\right\|$ and $I_\gamma(s) \to_{\gamma\to 0} (H(s)+H^T(s))/2$, it holds by dominated convergence that $\int I_\gamma(s)\, \mu(ds) \to_{\gamma\to 0} \boldsymbol{H} + \boldsymbol{H}^T$. If $\boldsymbol{H}+\boldsymbol{H}^T > 0$, then there exists $\gamma_0 > 0$ such that

$$\inf_{\gamma \in (0,\gamma_0]} \lambda_{\min} \left(\int I_\gamma(s)\, \mu(ds) \right) > 0,$$

where $\lambda_{\min}$ is the smallest eigenvalue. Thus, Assumption 4.3.4–(c) is verified.

4.7.4 Proof of Prop. 4.4.5

Since $A_\gamma(s,\cdot)$ is $1/\gamma$-Lipschitz, $\|A_\gamma(s, x-\gamma B(s,x))\| \geq \|A_\gamma(s,x)\| - \|B(s,x)\| \geq \|A_\gamma(s,x)\| - \|B(s,x)-B(s,x_\star)\| - \|B(s,x_\star)\|$. Therefore,

$$\begin{aligned}
&\psi_\gamma(x) \\
&\geq \int \Big\{ \langle B(s,x)-B(s,x_\star), x-x_\star\rangle - 6\gamma \|B(s,x)-B(s,x_\star)\|^2 + \gamma \|A_\gamma(s, x-\gamma B(s,x))\|^2 \Big\}\, \mu(ds) \\
&\geq \int \Big\{ \langle B(s,x)-B(s,x_\star), x-x_\star\rangle - 8\gamma \|B(s,x)-B(s,x_\star)\|^2 + (\gamma/2)\|A_\gamma(s,x)\|^2 \\
&\qquad\qquad - 2\gamma \|B(s,x_\star)\|^2 \Big\}\, \mu(ds) \\
&\geq \frac{\gamma}{2} \int \|A_\gamma(s,x)\|^2\, \mu(ds) - \gamma C
\end{aligned}$$

for γ small enough, by Hypothesis H4. We now have

$$\gamma \int \|A_\gamma(s,x)\|^2\, \mu(ds) = \frac{1}{\gamma} \int \|x - J_\gamma(s,x)\|^2\, \mu(ds) \geq \frac{1}{\gamma} \int d(s,x)^2\, \mu(ds) \geq \frac{C}{\gamma} \boldsymbol{d}(x)^2$$

thanks to Hypothesis H2. The result follows from the boundedness of $\mathcal{D}$.

4.7.5 Proof of Prop. 4.4.6

To prove this proposition, we start with the following result.

Lemma 4.7.6. Let $\mathsf{A} \in \mathscr{M}(\mathsf{X})$ be such that

$$\exists (x_*, y_*) \in \mathrm{gr}(\mathsf{A}),\ \exists \delta > 0, \quad \boldsymbol{S}(x_*, \delta) \subset \mathrm{int}(\mathrm{dom}\,\mathsf{A}),\ \text{and}\ \forall x \in \boldsymbol{S}(x_*, \delta),\ \inf_{y \in \mathsf{A}(x)} \langle y - y_*, x - x_* \rangle > 0.$$

Then, assuming that $\mathrm{dom}\,\mathsf{A}$ is unbounded,

$$\liminf_{x \in \mathrm{dom}\,\mathsf{A}, \|x\| \to \infty} \frac{\inf_{y \in \mathsf{A}(x)} \langle y - y_*, x - x_* \rangle}{\|x\|} > 0.$$

Proof. Given a vector $u \in \mathsf{X}$, define the function

$$f_u(\lambda) = \inf_{y \in \mathsf{A}(x_* + \lambda u)} \langle y - y_*, u \rangle$$

for all $\lambda \geq 0$ such that $x_* + \lambda u \in \mathrm{dom}\,\mathsf{A}$. For all $\lambda_1 > \lambda_2$ in $\mathrm{dom}\, f_u$, and all $y_1 \in A(x_* + \lambda_1 u)$ and $y_2 \in A(x_* + \lambda_2 u)$, we have

$$\langle y_1 - y_*, u \rangle - \langle y_2 - y_*, u \rangle = \langle y_1 - y_2, u \rangle = \frac{1}{\lambda_1 - \lambda_2} \langle y_1 - y_2, x_* + \lambda_1 u - (x_* + \lambda_2 u) \rangle \geq 0.$$

Passing to the infima, we obtain that $f_u(\lambda_1) \geq f_u(\lambda_2)$, in other words, f_u is non decreasing.

For all $x \in \mathrm{dom}\,\mathsf{A}$ such that $\|x - x_*\| \geq \delta$, we have by setting $u = \delta(x - x_*)/\|x - x_*\|$

$$\inf_{y \in \mathsf{A}(x)} \langle y - y_*, x - x_* \rangle = \frac{\|x - x_*\|}{\delta} f_u(\delta^{-1} \|x - x_*\|) \geq \frac{\|x - x_*\|}{\delta} f_u(1).$$

For any $u \in \boldsymbol{S}(0, \delta)$, it holds by assumption that $f_u(1) = \inf_{y \in \mathsf{A}(x_* + u)} \langle y - y_*, u \rangle$ is positive. We shall show that $f_u(1)$ is lower semicontinuous (lsc) as a function of u on the sphere $\boldsymbol{S}(0, \delta)$. Since this sphere is compact, $f_u(1)$ attains its infimum on $\boldsymbol{S}(0, \delta)$, and the lemma will be proven.

It is well-known that A is locally bounded near any point in the interior if its domain [36, Prop. 2.9] [12, §21.4]. Thus, by the closedeness of $\mathrm{gr}(\mathsf{A})$, the $\inf$ in the expression of $f_u(1)$ is attained. Let $u_n \to u$, and write $f_{u_n}(1) = \langle y_n - y_*, u_n \rangle$. By the maximality of A, we obtain that for any accumulation point y of (y_n) (who exists by the local boundedness), it holds that $(u, y) \in \mathrm{gr}(\mathsf{A})$. Consequently, $\liminf_n f_{u_n}(1) \geq f_u(1)$, in other words, $f_u(1)$ is lsc. □

We now prove Prop. 4.4.6. Let us write

$$f(\gamma, s, x) = \frac{\langle A_\gamma(s, x) - \varphi(s), J_\gamma(s, x) - x_\star \rangle}{\|x\|} + \frac{\|x - J_\gamma(s, x)\|^2}{\gamma \|x\|}, \ \text{and}$$
$$g(s, x) = \inf_{\gamma \in (0, 1]} f(\gamma, s, x).$$

Note that $\psi_\gamma(x)/\|x\| = \int f(\gamma, s, x)\, \mu(ds)$. We shall show that $\liminf_{\|x\| \to \infty} g(s, x) > 0$ for all $s \in \Sigma$. Assume the contrary, namely, that there exist $s \in \Sigma$ and $\|x_k\| \to \infty$ such that $g(s, x_k) \to 0$. In these conditions, there exists a sequence (γ_k) in $(0, 1]$ such that $f(\gamma_k, s, x_k) \to 0$. By inspecting the second term in the expression of $f(\gamma_k, s, x_k)$, we obtain that $\|J_{\gamma_k}(s, x_k)\|/\|x_k\| \to 1$. Rewriting the first term as

$$\frac{\|J_{\gamma_k}(s, x_k)\|}{\|x_k\|} \frac{\langle A_{\gamma_k}(s, x_k) - \varphi(s), J_{\gamma_k}(s, x_k) - x_\star \rangle}{\|J_{\gamma_k}(s, x_k)\|},$$

and recalling that $A_{\gamma_k}(s,x_k) \in A(s, J_{\gamma_k}(s,x_k))$, Lem. 4.7.6 shows that the $\liminf$ of this term is positive, which raises a contradiction.

Note that

$$\inf_{\gamma\in(0,1]} \frac{\psi_\gamma(x)}{\|x\|} \geq \int g(s,x)\,\mu(ds).$$

Using Fatou's lemma, we obtain Assumption 4.3.4–(a).

4.8 Proofs relative to Sec. 4.5

4.8.1 Proof of Lem. 4.5.3

Let ε be the smallest of the three constants (also named ε) in Assumptions 4.3.1, 4.3.3 and 4.3.7 respectively where $\mathcal{K} = B(R)$. For every a, γ, the following holds for $\bar{\mathbb{P}}^{a,\gamma}$-almost all $x = (x_n : n \in \mathbb{N})$:

$$\begin{aligned}
\boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_{n+1}\|\leq R} &= \boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_{n+1}\|\leq R}(\mathbb{1}_{\|x_n\|\leq R} + \mathbb{1}_{\|x_n\|>R}) = \boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_{n+1}\|\leq R}\mathbb{1}_{\|x_n\|\leq R} \\
&\leq \boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_n\|\leq R} \\
&= \|x_{n+1} - \Pi_{\mathcal{D}}(x_{n+1})\|\mathbb{1}_{\|x_n\|\leq R} \\
&\leq \|x_{n+1} - \Pi_{\mathcal{D}}(x_n)\|\mathbb{1}_{\|x_n\|\leq R}\,.
\end{aligned}$$

Using the notation $\bar{\mathbb{E}}_n^{a,\gamma} = \bar{\mathbb{E}}^{a,\gamma}(\,.\,|x_0,\ldots,x_n)$, we thus obtain:

$$\bar{\mathbb{E}}_n^{a,\gamma}(\boldsymbol{d}(x_{n+1})^{1+\varepsilon}\mathbb{1}_{\|x_{n+1}\|\leq R}) \leq \int \|J_\gamma(s, x_n - \gamma B(s,x_n)) - \Pi_{\mathcal{D}}(x_n)\|^{1+\varepsilon}\mathbb{1}_{\|x_n\|\leq R}\,d\mu(s)\,.$$

By the convexity of $\|\cdot\|^{1+\varepsilon}$, for all $\alpha \in (0,1)$,

$$\|x+y\|^{1+\varepsilon} = \frac{1}{\alpha^{1+\varepsilon}}\left\|\alpha x + (1-\alpha)\frac{\alpha}{1-\alpha}y\right\|^{1+\varepsilon} \leq \alpha^{-\varepsilon}\|x\|^{1+\varepsilon} + (1-\alpha)^{-\varepsilon}\|y\|^{1+\varepsilon}\,.$$

Therefore, by setting $\delta_\gamma(s,a) := \|J_\gamma(s, a - \gamma B(s,a)) - \Pi_{D(s)}(a)\|$,

$$\begin{aligned}
\bar{\mathbb{E}}_n^{a,\gamma}(\boldsymbol{d}(x_{n+1})^{1+\varepsilon}\mathbb{1}_{\|x_{n+1}\|\leq R}) \leq \alpha^{-\varepsilon}\int \delta_\gamma(s,x_n)^{1+\varepsilon}\mathbb{1}_{\|x_n\|\leq R}\,d\mu(s) \\
+ (1-\alpha)^{-\varepsilon}\int \|\Pi_{D(s)}(x_n) - \Pi_{\mathcal{D}}(x_n)\|^{1+\varepsilon}\mathbb{1}_{\|x_n\|\leq R}\,d\mu(s)\,.
\end{aligned}$$

Note that for every $s \in \Xi$, $a \in \mathsf{X}$,

$$\|\delta_\gamma(s,a)\| \leq \|J_\gamma(s,a) - \Pi_{D(s)}(a)\| + \gamma\|B(s,a)\|\,.$$

Hence, by Assumptions 4.3.7 and 4.3.3, there exists a deterministic constant $C > 0$ s.t.

$$\sup_n \int \delta_\gamma(s,x_n)^{1+\varepsilon}\mathbb{1}_{\|x_n\|\leq R}\,d\mu(s) \leq C\gamma^{1+\varepsilon}\,.$$

Moreover, since $\Pi_{\mathrm{cl}(D(s))}$ is a firmly non expansive operator [12, Chap. 4], it holds that for all $u \in \mathrm{cl}(\mathcal{D})$, and for μ-almost all s,

$$\|\Pi_{\mathrm{cl}(D(s))}(x_n) - u\|^2 \leq \|x_n - u\|^2 - \|\Pi_{\mathrm{cl}(D(s))}(x_n) - x_n\|^2.$$

Taking $u = \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)$, we obtain that

$$\|\Pi_{\mathrm{cl}(D(s))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2 \leq \boldsymbol{d}(x_n)^2 - d(x_n, D(s))^2. \tag{4.28}$$

Making use of Assumption 4.3.6, and assuming without loss of generality that $\varepsilon \leq 1$, we obtain

$$\begin{aligned}\int \|\Pi_{\mathrm{cl}(D(s))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^{1+\varepsilon} \, d\mu(s) &\leq \left(\int \|\Pi_{\mathrm{cl}(D(s))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2 \, d\mu(s)\right)^{(1+\varepsilon)/2} \\ &\leq \alpha' \boldsymbol{d}(x_n)^{1+\varepsilon},\end{aligned}$$

for some $\alpha' \in [0,1)$. Choosing α close enough to zero, we obtain that there exists $\rho \in [0,1)$ such that

$$\bar{\mathbb{E}}_n^{a,\gamma}\left(\frac{\boldsymbol{d}(x_{n+1})^{1+\varepsilon}}{\gamma^{1+\varepsilon}} \mathbb{1}_{\|x_{n+1}\| \leq R}\right) \leq \rho \frac{\boldsymbol{d}(x_n)^{1+\varepsilon}}{\gamma^{1+\varepsilon}} \mathbb{1}_{\|x_n\| \leq R} + C.$$

Taking the expectation at both sides, iterating, and using the fact that $\boldsymbol{d}(x_0) = \boldsymbol{d}(a) < M\gamma$, we obtain that

$$\sup_{n \in \mathbb{N}, a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}, \gamma \in (0, \gamma_0]} \bar{\mathbb{E}}^{a,\gamma}\left(\left(\frac{\boldsymbol{d}(x_n)}{\gamma}\right)^{1+\varepsilon} \mathbb{1}_{\|x_n\| \leq R}\right) < +\infty\,. \tag{4.29}$$

Since $A_\gamma(s,\cdot)$ is γ^{-1}-Lipschitz continuous, $\|A_\gamma(s, x - \gamma B(s,x))\| \leq \|A_\gamma(s,x)\| + \|B(s,x)\|$. Moreover, choosing measurably $\tilde{x} \in \mathcal{D}$ in such a way that $\|x - \tilde{x}\| \leq 2\boldsymbol{d}(x)$, we obtain $\|A_\gamma(s,x)\| \leq \|A_0(s,\tilde{x})\| + 2\frac{\boldsymbol{d}(x)}{\gamma}$. Therefore, there exists R' depending only on R and $\mathcal{D}$ s.t.

$$\|A_\gamma(s,x)\| \mathbb{1}_{\|x\| \leq R} \leq \|A_0(s,\tilde{x})\| \mathbb{1}_{\|\tilde{x}\| \leq R'} + 2\frac{\boldsymbol{d}(x)}{\gamma} \mathbb{1}_{\|x\| \leq R}\,.$$

Thus,

$$\begin{aligned}\bar{\mathbb{E}}_n^{a,\gamma}(\|Z_{n+1}^\gamma\|^{1+\varepsilon}) &= \int \|h_{\gamma,R}(s,x_n)\|^{1+\varepsilon} d\mu(s) \\ &= \int \|B(s,x_n) + A_\gamma(s, x_n - \gamma B(s,x_n))\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \leq R} \, d\mu(s) \\ &\leq \int \left(2\|B(s,x_n)\| + \|A_0(s,\tilde{x}_n)\| + 2\frac{\boldsymbol{d}(x_n)}{\gamma}\right)^{1+\varepsilon} \mathbb{1}_{\|x_n\| \leq R'} \, d\mu(s)\,.\end{aligned} \tag{4.30}$$

By Assumption 4.3.3, $\int \|B(s,x_n)\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \leq R} \, d\mu(s) \leq C$ where the constant C depends only on ε and R. By Assumption 4.3.1, we also have $\int \|A_0(s,x_n)\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \leq R} \, d\mu(s) \leq C$ for some (other) constant C. The third term is controlled by Eq. (4.29). Taking expectations, the bound (4.21) is established.

4.8.2 Proof of Lem. 4.5.4

The first point can be obtained by straightforward application of Prokhorov and Skorokhod's theorems. However, to verify the second point, we need to construct the sequences more carefully. Choose $\varepsilon > 0$ as in Lem. 4.5.3. We define the process $Y^\gamma : \mathsf{X}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ s.t. for every $n \in \mathbb{N}$,

$$Y_n^\gamma(x) := \sum_{k=0}^{n-1} \frac{\boldsymbol{d}(x_k)^{1+\varepsilon/2}}{\gamma^{\varepsilon/2}} \mathbb{1}_{\|x_k\| \leq R}\,,$$

and we denote by $(X, Y^\gamma) : \mathsf{X}^{\mathbb{N}} \to (\mathsf{X} \times \mathbb{R})^{\mathbb{N}}$ the process given by $(X, Y^\gamma)_n(x) := (x_n, Y_n^\gamma(x))$. We define for every n, $\tilde{Z}_{n+1}^\gamma := \gamma^{-1}((X, Y^\gamma)_{n+1} - (X, Y^\gamma)_n)$. By Lem. 4.5.3, it is easily seen that

$$\sup_{n\in\mathbb{N}, a\in\mathcal{K}\cap\mathcal{D}_{\gamma M}, \gamma\in(0,\gamma_0]} \bar{\mathbb{E}}^{a,\gamma}\left(\|\tilde{Z}_n^\gamma\| \mathbb{1}_{\|\tilde{Z}_n^\gamma\|>A}\right) \xrightarrow{A\to+\infty} 0\,.$$

We now apply Lem. 3.6.2, only replacing X by $\mathsf{X}\times\mathbb{R}$ and $\bar{\mathbb{P}}^{a,\gamma}$ by $\bar{\mathbb{P}}^{a,\gamma}(X, Y^\gamma)^{-1}$. By this lemma, the family $\{\bar{\mathbb{P}}^{a,\gamma}(X, Y^\gamma)^{-1}\overline{\mathsf{X}}_\gamma^{-1} : a \in \mathcal{K}\cap\mathcal{D}_{\gamma M}, \gamma \in (0, \gamma_0]\}$ is tight, where $\overline{\mathsf{X}}_\gamma^{-1} : (\mathsf{X}\times\mathbb{R})^{\mathbb{N}} \to C(\mathbb{R}_+, \mathsf{X}\times\mathbb{R})$ is the piecewise linear interpolated process, defined in the same way as X_γ only substituting $\mathsf{X}\times\mathbb{R}$ with X in the definition. By Prokhorov's theorem, one can choose the subsequence (a_n, γ_n) s.t. $\bar{\mathbb{P}}^{a_n,\gamma_n}(X, Y^{\gamma_n})^{-1}\overline{\mathsf{X}}_{\gamma_n}^{-1}$ converges narrowly to some probability measure Υ on $\mathsf{X}\times\mathbb{R}$. By Skorokhod's theorem, we can define a stochastic process $((\mathsf{x}_n, \mathsf{y}_n) : n \in \mathbb{N})$ on some probability space $(\Omega', \mathscr{F}', \mathbb{P}')$ into $C(\mathbb{R}_+, \mathsf{X}\times\mathbb{R})$, whose distribution for a fixed n coincides with $\bar{\mathbb{P}}^{a_n,\gamma_n}(X, Y^{\gamma_n})^{-1}\overline{\mathsf{X}}_{\gamma_n}^{-1}$, and s.t. for every $\omega \in \Omega'$, $(\mathsf{x}_n(\omega), \mathsf{y}_n(\omega)) \to (\mathsf{z}(\omega), \mathsf{w}(\omega))$, where (z, w) is a r.v. defined on the same space. In particular, the first marginal distribution of $\bar{\mathbb{P}}^{a_n,\gamma_n}(X, Y^{\gamma_n})^{-1}\overline{\mathsf{X}}_{\gamma_n}^{-1}$ coincides with $\bar{\mathbb{P}}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1}$. Thus, the first point is proven.

For every $\gamma \in (0, \gamma_0]$, introduce the mapping

$$\begin{aligned}\Gamma_\gamma : C(\mathbb{R}_+, \mathsf{X}) &\to C(\mathbb{R}_+, \mathbb{R})\\ \mathsf{x} &\mapsto \left(t \mapsto \int_0^t (\gamma^{-1}\boldsymbol{d}(\mathsf{x}(\gamma\lfloor u/\gamma\rfloor)))^{1+\varepsilon/2}\mathbb{1}_{\|\mathsf{x}(\gamma\lfloor u/\gamma\rfloor)\|\le R} du\right)\,.\end{aligned}$$

We denote by $\underline{\mathsf{X}}_\gamma^{-1} : \mathbb{R}^{\mathbb{N}} \to C(\mathbb{R}_+, \mathbb{R})$ the piecewise linear interpolated process, defined in the same way as X_γ only substituting $\mathbb{R}$ with X in the definition. It is straightforward to show that $\underline{\mathsf{X}}_\gamma \circ Y^{\gamma_n} = \Gamma_\gamma \circ \mathsf{X}_\gamma$. For every n, by definition of the couple $(\mathsf{x}_n, \mathsf{y}_n)$, the distribution under $\mathbb{P}'$ of the r.v. $\Gamma_{\gamma_n}(\mathsf{x}_n) - \mathsf{y}_n$ is equal to the distribution of $\Gamma_{\gamma_n}\circ\mathsf{X}_{\gamma_n} - \underline{\mathsf{X}}_{\gamma_n}\circ Y^{\gamma_n}$ under $\bar{\mathbb{P}}^{a_n,\gamma_n}$. Therefore, $\mathbb{P}'$-a.e. and for every n, $\mathsf{y}_n = \Gamma_{\gamma_n}(\mathsf{x}_n)$. This implies that, $\mathbb{P}'$-a.e., $\Gamma_{\gamma_n}(\mathsf{x}_n)$ converges (uniformly on compact set) to w. On that event, this implies that for every $T \ge 0$, $\Gamma_{\gamma_n}(\mathsf{x}_n)(T) \to \mathsf{w}(T)$, which is finite. Hence, $\sup_n \Gamma_{\gamma_n}(\mathsf{x}_n)(T) < \infty$ on that event, which proves the second point.

4.8.3 Proof of Lem. 4.5.5

Define $c_a := \sup_{x\in B(R)\cap\mathcal{D}} \int \|A_0(s,x)\|^{1+\varepsilon/2} d\mu(s)$ and $c_b := \sup_{x\in B(R)} \int \|B(s,x)\|^{1+\varepsilon/2} d\mu(s)$ (these constants being finite by Assumptions 4.3.1 and 4.3.3). By the same derivations as those leading to Eq. (4.30), we obtain

$$\begin{aligned}&\int \|v_n(s,t)\|^{1+\varepsilon/2} d\mu(s) \le c_b\\ &\int \|w_n(s,t)\|^{1+\varepsilon/2} d\mu(s) \le C\left(\frac{\boldsymbol{d}(u_n(t))^{1+\varepsilon/2}}{\gamma^{1+\varepsilon/2}}\mathbb{1}_{\|u_n(t)\|\le R} + c_a + c_b\right).\end{aligned}$$

The proof is concluded by applying Lem. 4.5.4.

4.8.4 Proof of Lem. 4.5.6

The sequence $((v_n, w_n, \|w_n(\cdot,\cdot)\|, \|v_n(\cdot,\cdot)\|))$ converges weakly to $(v, w, \tilde{v}, \tilde{w})$ in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}^2\times\mathbb{R}^2}$ along some subsequence (*n.b.*: compactness and sequential compactness are the same notions in the weak topology of $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}^2\times\mathbb{R}^2}$). We still denote by $((v_n, w_n, \|v_n\|, \|w_n\|))$ this subsequence. By Mazur's theorem, there

exists a function $J : \mathbb{N} \to \mathbb{N}$ and a sequence of sets of weights $\{\alpha_{k,n} : n \in \mathbb{N}, k = n \dots, J(n) \ : \ \alpha_{k,n} \geq 0, \sum_{k=n}^{J(n)} \alpha_{k,n} = 1\}$ such that the sequence of functions

$$(\bar{v}_n, \bar{w}_n, \hat{v}_n, \hat{w}_n) \ : \ (s,t) \mapsto \sum_{k=n}^{J(n)} \alpha_{k,n}(v_k(s,t), w_k(s,t), \|v_k(s,t)\|, \|w_k(s,t)\|)$$

converges strongly to $(v, w, \tilde{v}, \tilde{w})$ in that space, as $n \to \infty$. Taking a further subsequence (which we still denote by $(\bar{v}_n, \bar{w}_n, \hat{v}_n, \hat{w}_n)$) we obtain the $\mu \otimes \lambda_T$-almost everywhere convergence of $(\bar{v}_n, \bar{w}_n, \hat{v}_n, \hat{w}_n)$ to $(v, w, \tilde{v}, \tilde{w})$. Consider a negligible set $\mathscr{N} \in \mathscr{G} \otimes \mathscr{B}([0,T])$ such that for all $(s,t) \notin \mathscr{N}$,

$$(\bar{v}_n(s,t), \bar{w}_n(s,t), \hat{v}_n(s,t), \hat{w}_n(s,t)) \to (v(s,t), w(s,t), \tilde{v}(s,t), \tilde{w}(s,t))$$

and $\tilde{v}(s,t), \tilde{w}(s,t)$ are finite. We shall prove that for every $(s,t) \notin \mathscr{N}$, $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(s,\cdot))$. First consider the case where $\|\mathsf{z}(t)\| > R$. Since $u_n(t) \to \mathsf{z}(t)$, there exists a positive integer n_0 such that for every $n \geq n_0$, $\|u_n(t)\| > R$. Hence, $v_n(s,t)$ and $w_n(s,t)$ are equal to zero for every $n \geq n_0$ and similarly for $\bar{v_n}(s,t)$ and $\bar{w_n}(s,t)$. For every $(s,t) \notin \mathscr{N}$ such that $\|\mathsf{z}(t)\| > R$, $(v+w)(s,t) = 0$ and $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(s,\cdot))$. Then, if $\|\mathsf{z}(t)\| = R$, $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(s,\cdot))$ obviously. Finally, assume that $\|\mathsf{z}(t)\| < R$. In this case the condition $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(s,\,.\,))$ is equivalent to:

$$(\mathsf{z}(t), -(v+w)(s,t)) \in \mathrm{gr}(A(s,\cdot) + B(s,\cdot))\,. \tag{4.31}$$

Besides, there exists a positive integer n_0 such that for every $n \geq n_0$, $\|u_n(t)\| < R$. To show Eq. (4.31), consider $p \in D(s)$, $q_a \in A(s,p)$ and $q_b \in B(s,p)$. Decompose:

$$\langle q_b + \bar{v}_n(s,t) + q_a + \bar{w}_n(s,t), p - \mathsf{z}(t)\rangle = A_n + B_n + C_n\,, \tag{4.32}$$

where

$$\begin{aligned}
A_n &= \sum_{k=n}^{J(n)} \alpha_{k,n}\langle q_a + w_k(s,t), p - J_{\gamma_k}(s, u_k(t) - \gamma_k B(s, u_k(t)))\rangle \\
B_n &= \sum_{k=n}^{J(n)} \alpha_{k,n}\langle q_b + v_k(s,t), p - u_k(t)\rangle \\
C_n &= \sum_{k=n}^{J(n)} \alpha_{k,n}\langle q_a + w_k(s,t), J_{\gamma_k}(s, u_k(t) - \gamma_k B(s, u_k(t))) - u_k(t)\rangle\,.
\end{aligned}$$

The left hand side of (4.32) converges to $\langle q_a + w(s,t), p - \mathsf{z}(t)\rangle + \langle q_b + v(s,t), p - \mathsf{z}(t)\rangle$. The terms A_n and B_n are nonnegative by monotonicity of $A(s,\cdot)$ and $B(s,\cdot)$ for every $n \geq n_0$. Moreover,

$$\begin{aligned}
C_n &= \sum_{k=n}^{J(n)} \alpha_{k,n}\langle q_a + w_k(s,t), J_{\gamma_k}(s, u_k(t) - \gamma_k B(s, u_k(t))) - u_k(t)\rangle \\
&= \sum_{k=n}^{J(n)} \alpha_{k,n}\langle q_a + w_k(s,t), J_{\gamma_k}(s, u_k(t) - \gamma_k B(s, u_k(t))) - (u_k(t) - \gamma_k B(s, u_k(t)))\rangle \\
&\quad + \sum_{k=n}^{J(n)} \alpha_{k,n}\langle q_a + w_k(s,t), -\gamma_k B(s, u_k(t))\rangle \\
&= \sum_{k=n}^{J(n)} \alpha_{k,n}\gamma_k\langle q_a + w_k(s,t), w_k(s,t) + v_k(s,t)\rangle.
\end{aligned}$$

We conclude that $C_n \to 0$ and $\langle q_a + w(s,t) + q_b + v(s,t), p - \mathsf{z}(t)\rangle \geq 0$. As $A(s,\cdot) + B(s,\cdot) \in \mathscr{M}(\mathsf{X})$, this implies that Eq. (4.31) holds.

Chapter 5

A Constant Step Stochastic Douglas-Rachford Algorithm with Application to Non Separable Regularizations

The Douglas Rachford algorithm is a classical algorithm to solve a composite optimization problem. It involves the computation of the proximity operators of the two functions separately and enjoys more numerical stability than the proximal gradient algorithm. In this chapter, we bring the tools from the previous chapter to study a stochastic version of the constant step Douglas Rachford algorithm. In this algorithm, a random realization of both functions is used at each iteration to perform a Douglas Rachford step. Application to structured regularizations and distributed optimization is provided. Theoretical results are supported by the technical report in Appendix A.

5.1 Introduction

Many applications in the fields of machine learning [35] and signal processing [44] require the solution of the programming problem

$$\min_{x \in \mathsf{X}} F(x) + G(x) \tag{5.1}$$

where X is an Euclidean space, $F, G \in \Gamma_0(\mathsf{X})$. In these contexts, F often represents a cost function and G a regularization term. The Douglas-Rachford algorithm is one of the most popular approach towards solving Problem (5.1). Given $\gamma > 0$, the algorithm is written

$$\begin{aligned} y_{n+1} &= \operatorname{prox}_{\gamma F}(x_n) \\ z_{n+1} &= \operatorname{prox}_{\gamma G}(2y_{n+1} - x_n) \\ x_{n+1} &= x_n + z_{n+1} - y_{n+1}. \end{aligned} \tag{5.2}$$

Assuming that a standard qualification condition holds and that the set of solutions $\arg\min F + G$ of (5.1) is not empty, the sequence $(y_n)_n$ converges to an element in $\arg\min F + G$ as $n \to +\infty$ ([78, 56]).

In this chapter, we study the case where F and G are integral functionals of the form

$$F(x) = \mathbb{E}_\xi(f(\xi, x)), \qquad G(x) = \mathbb{E}_\xi(g(\xi, x))$$

where ξ is a random variable (r.v) from some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to a measurable space $(\Xi, \mathcal{G})$, with distribution μ, and where $f, g : \Xi \times \mathsf{X} \to (-\infty, +\infty]$ are normal convex integrand. In this context, the stochastic Douglas Rachford algorithm aims to solve Problem (5.1) by iterating

$$\begin{aligned} y_{n+1} &= \mathrm{prox}_{\gamma f(\xi_{n+1},\cdot)}(x_n) \\ z_{n+1} &= \mathrm{prox}_{\gamma g(\xi_{n+1},\cdot)}(2y_{n+1} - x_n) \\ x_{n+1} &= x_n + z_{n+1} - y_{n+1}\,, \end{aligned} \tag{5.3}$$

where $(\xi_n)_n$ is a sequence of i.i.d copies of the random variable ξ and $\gamma > 0$ is the constant step size. Compared to the "deterministic" Douglas Rachford algorithm (5.2), the stochastic Douglas Rachford algorithm (5.3) is an online method. The constant step size used make it implementable in adaptive signal processing or online machine learning contexts. In this algorithm, the function F (resp. G) is replaced at each iteration n by a random realization $f(\xi_n, \cdot)$ (resp. $g(\xi_n, \cdot)$). It can be implemented in the case where F (resp. G) cannot be computed in its closed form or in the case where the computation of its proximity operator is demanding. Compared to other online optimization algorithm like the stochastic subgradient algorithm, the algorithm (5.3) benefits from the numerical stability of stochastic proximal methods [108, 122].

Stochastic version of the Douglas Rachford algorithm have been considered in [44, 116]. These papers consider the case where G is deterministic, *i.e* is not written as an expectation and F is written as an expectation that reduces to a sum. The algorithms [105, 49] are generalizations of a partially stochastic Douglas Rachford algorithm where G is deterministic. The convergence of these algorithms is obtained under a summability assumption of the noise over the iterations.

In this chapter we provide theoretical basis for the algorithm (5.3) and convergence results based on the technical report [110]. We also provide applications to optimization problems regularized by the overlapping group lasso and we provide an application to a target tracking problem involving distributed optimization (based on [89]).

Chapter organization. The next section 5.2 is devoted to the statement of the main convergence result. In Sec. 5.3, an outline of the proof of the result in Sec. 5.2 is provided. Finally, the algorithm (5.3) is implemented to solve a regularized problem (resp. a distributed optimization problem) in Sec. 5.4 (resp. Sec. 5.5).

From now on, we shall state explicitly the dependence of the iterates of the algorithm in the step size and the starting point. Namely, we shall denote $(x_n^{\nu,\gamma})_n$ the sequence $(x_n)_n$ generated by the stochastic Douglas Rachford algorithm (5.3) with step γ, such that the distribution of $x_0^{\nu,\gamma}$ over X is ν. If $\nu = \delta_a$, where δ_a is the Dirac measure at the point $a \in \mathsf{X}$, we shall prefer the notation $x_n^{a,\gamma}$.

5.2 Main convergence theorem

For every $s \in \Xi$, the domain of $g(s, \cdot)$ is denoted $D(s)$ and $\mathcal{D}$ is their μ-essential intersection (see Sec. 2.3).

Consider the following assumptions.

Assumption 5.2.1. For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K} \cap \mathcal{D}} \int \|\partial_0 g(s, x)\|^{1+\varepsilon}\, \mu(ds) < \infty.$$

Assumption 5.2.2. For μ-a.e $s \in \Xi$, $f(s,\cdot)$ is differentiable. Moreover, there exists $\varepsilon > 0, x_0 \in \mathsf{X}$ such that

$$\int \|\nabla f(s,x_0)\|^{1+\varepsilon}\,\mu(ds) < \infty\,.$$

Assumption 5.2.3. There exists $L > 0$ such that $\nabla f(s,\cdot)$ is μ-a.e, a L-Lipschitz continuous function.

Assumption 5.2.4. $\forall x \in \mathsf{X}$, $\int d(x, D(s))^2\,\mu(ds) \geq C\boldsymbol{d}(x)^2$, where $\boldsymbol{d}(\cdot)$ is the distance function to $\mathcal{D}$.

Assumption 5.2.5. For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon, C, \gamma_0 > 0$ such that for all $\gamma \in (0,\gamma_0]$ and all $x \in \mathcal{K}$,

$$\frac{1}{\gamma^{1+\varepsilon}} \int \|\operatorname{prox}_{\gamma g(s,\cdot)}(x) - \Pi_{\mathrm{cl}(D(s))}(x)\|^{1+\varepsilon}\,\mu(ds) < C\,.$$

Assumption 5.2.6. There exists $x_\star \in \arg\min F + G$ and $\varphi \in \mathfrak{S}^2_{\partial g(\cdot,x_\star)}$ (see Sec. 2.3) such that $\nabla f(\cdot,x_\star) \in \mathcal{L}^2(\Xi,\mathsf{X})$ and $\int \nabla f(s,x_\star)\,\mu(ds) + \int \varphi(s)\,\mu(ds) = 0$.

Assumption 5.2.7. The function $F+G$ satisfies one of the following properties:

(a) $F+G$ is coercive *i.e* $F(x)+G(x) \longrightarrow_{\|x\|\to+\infty} +\infty$

(b) $F+G$ is supercoercive *i.e* $\frac{F(x)+G(x)}{\|x\|} \longrightarrow_{\|x\|\to+\infty} +\infty$.

Assumption 5.2.8. There exists $\gamma_0 > 0$, such that for all $\gamma \in (0,\gamma_0]$ and all $x \in \mathsf{X}$,

$$\begin{aligned}&\int \|\nabla f_\gamma(s,x)\| + \frac{1}{\gamma}\|\operatorname{prox}_{\gamma g(s,\cdot)}(x) - \Pi_{\mathrm{cl}(D(s))}(x)\|\mu(ds)\\ &\leq C(1+|F^\gamma(x)+G^\gamma(x)|)\,.\end{aligned}$$

where $f_\gamma(s,\cdot)$ is the Moreau envelope of $f(s,\cdot)$.

Theorem 5.2.1. Let Assumptions 5.2.1– 5.2.8 hold true. Then, for each probability measure ν over X having a finite second moment, for any $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{P}\left(d(x_k^{\nu,\gamma}, \arg\min(F+G)) > \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,.$$

Moreover, if Assumption 5.2.7–(b) holds true, then

$$\begin{aligned}\limsup_{n\to\infty}\ \mathbb{P}\left(d\left(\bar{x}_n^{\nu,\gamma}, \arg\min(F+G)\right) \geq \varepsilon\right) &\xrightarrow[\gamma\to 0]{} 0, \text{ and}\\ \limsup_{n\to\infty}\ d\left(\mathbb{E}(\bar{x}_n^{\nu,\gamma}), \arg\min(F+G)\right) &\xrightarrow[\gamma\to 0]{} 0\,.\end{aligned}$$

where $\bar{x}_n^{\nu,\gamma} = \frac{1}{n}\sum_{k=1}^n x_k^{\nu,\gamma}$.

Loosely speaking, the theorem states that, with high probability, the iterates $(x_n^{\nu,\gamma})_n$ stay close to the set of solutions $\arg\min F+G$ as $n\to\infty$ and $\gamma\to 0$. This theorem is reminiscent of Corollary 4.4.3. Moreover, Assumptions 5.2.1– 5.2.8 are reminiscent of Assumptions C1– C4 of Chap. 4. Note that Assumption 5.2.1 combined with Assumption 5.2.3 implies that for every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x\in\mathcal{K}} \int \|\nabla f(s,x)\|^{1+\varepsilon}\,\mu(ds) < \infty\,.$$

Since $f(s,\cdot), g(s,\cdot) \in \Gamma_0(\mathsf{X})$, and $f(s,\cdot)$ is differentiable, [103]

$$\partial(F+G)(x) = \nabla F(x) + \partial G(x) = \mathbb{E}_\xi(\nabla f(\xi,x)) + \mathbb{E}_\xi(\partial g(\xi,x)),$$

where the set $\mathbb{E}(\partial g(\xi,x))$ is defined by its selection integral, see Eq. 2.4. Therefore, using Fermat's rule, if $x \in \arg\min F+G$, then there exists $\varphi \in \mathcal{L}^1(\Xi,\mathsf{X})$, such that $\varphi(s) \in \partial g(s,x)$ μ-a.s, and $\int \nabla f(s,x)\,\mu(ds) + \int \varphi(s)\,\mu(ds) = 0$. We refer to $(\nabla f(\cdot,x),\varphi)$ as a *representation* of the solution x. Assumption 5.2.6 ensures the existence of $x_\star \in \arg\min F+G$ with a representation $\nabla f(\cdot,x),\varphi \in \mathcal{L}^2(\Xi,\mathsf{X})$.

5.3 Outline of the convergence proof

This section is devoted to sketching the proof of the convergence of the stochastic Douglas Rachford algorithm. The approach follows the same steps as Chap. 4 and is detailed in the Technical Report [110]. The first step of the proof is to study the dynamical behavior of the iterates $(x_n^{a,\gamma})_n$ where $a \in \mathcal{D}$. Consider the continuous time stochastic process $\mathsf{x}^{a,\gamma}$ obtained by linearly interpolating with time interval γ the iterates $(x_n^{a,\gamma})$:

$$\mathsf{x}^{a,\gamma}(t) = x_n^{a,\gamma} + (t-n\gamma)\frac{x_{n+1}^{a,\gamma} - x_n^{a,\gamma}}{\gamma}, \tag{5.4}$$

for all $t \geq 0$ such that $n\gamma \leq t < (n+1)\gamma$, for all $n \in \mathbb{N}$. Let Assumptions 5.2.1–5.2.5 [1] hold true. Consider the set $C(\mathbb{R}_+,\mathsf{X})$ of continuous functions from $\mathbb{R}_+$ to X equipped with the topology of uniform convergence on the compact intervals. It is shown that the continuous time stochastic process $\mathsf{x}^{a,\gamma}$ converges weakly over $\mathbb{R}_+$ (*i.e* in distribution in $C(\mathbb{R}_+,\mathsf{X})$) as $\gamma \to 0$. Moreover, the limit is proven to be the unique absolutely continuous function x over $\mathbb{R}_+$ satisfying $\mathsf{x}(0) = a$ and for almost every $t \geq 0$, the Differential Inclusion (DI),

$$\dot{\mathsf{x}}(t) \in -(\nabla F + \partial G)(\mathsf{x}(t)), \tag{5.5}$$

(see Sec. 2.2.2). The semiflow associated with the DI is denoted Φ. The weak convergence of $(\mathsf{x}^{a,\gamma})$ to x is not enough to study the long term behavior of the iterates $(x_n^{a,\gamma})_n$. The second step of the proof is to prove a stability result for the Feller Markov chain $(x_n^{a,\gamma})_n$. Denote by P_γ its transition kernel. The deterministic counterpart of this step of the proof is the so-called *Fejér monotonicity* of the sequence (x_n) of the algorithm (5.2). Even if some work has been done [24, 48] to generalize the latter Fejér monotonicity to the stochastic setting, there is no immediate way to adapt it to our framework. As an alternative, we assume Hypotheses 5.2.3-5.2.6, and prove the existence of positive numbers α, C and γ_0, such that for every $\gamma \in (0,\gamma_0]$,

$$\begin{aligned}\mathbb{E}_n\|x_{n+1}^{a,\gamma} - x_\star\|^2 \leq &\|x_n^{a,\gamma} - x_\star\|^2 \\ &- \alpha\gamma(F^\gamma + G^\gamma)(x_n^{a,\gamma}) + \gamma C.\end{aligned} \tag{5.6}$$

In this inequality, $\mathbb{E}_n$ denotes the conditional expectation with respect to the sigma-algebra $\sigma(x_0^\gamma,\ldots,x_n^\gamma)$ and

$$F^\gamma(x) = \int f_\gamma(s,x)\,\mu(ds), \quad G^\gamma(x) = \int g_\gamma(s,x)\,\mu(ds).$$

Since $\gamma \mapsto F^\gamma(x) + G^\gamma(x)$ is decreasing (see Sec. A or Chap. 4, End of the proof of Prop. 4.4.1), the function $F^\gamma + G^\gamma$ in Eq. (5.6) can be replaced by $F^{\gamma_0} + G^{\gamma_0}$. Besides, the coercivity of $F+G$

[1] In the case where the domains are common, *i.e* $s \mapsto D(s)$ is μ-a.s constant, the moment Assumptions 5.2.1 and 5.2.2 are sufficient to state the dynamical behavior result. See Sec. 5.5 for an applicative context where the domains $D(s)$ are distinct.

(Assumption 5.2.7) implies the coercivity of $F^{\gamma_0} + G^{\gamma_0}$ (see Sec. A or Chap. 4, End of the proof of Prop. 4.4.1). Therefore, assuming 5.2.3–5.2.7 and setting $\Psi = F^{\gamma_0} + G^{\gamma_0}$, there exist positive numbers α, C and γ_0, such that for every $\gamma \in (0, \gamma_0]$,

$$\mathbb{E}_n \|x_{n+1}^{a,\gamma} - x_\star\|^2 \leq \|x_n^{a,\gamma} - x_\star\|^2 - \alpha\gamma\Psi(x_n^{a,\gamma}) + \gamma C. \tag{5.7}$$

Equation (5.7) can alternatively be seen as a tightness result. This is the purpose of Chap. 3, Sec. 3.8.3. It implies that the set I_γ of invariant measures of the Markov kernel P_γ is not empty for every $\gamma \in (0, \gamma_0]$, and that the set

$$\text{Inv} = \cup_{\gamma \in (0,\gamma_0]} I_\gamma \tag{5.8}$$

is *tight*.

It remains to characterize the cluster points of Inv as $\gamma \to 0$. To that end, the dynamical behavior result and the stability result are combined. Let Assumptions 5.2.1– 5.2.8 hold true. [2] Then, the set Inv is tight, and, as $\gamma \to 0$, every cluster point of Inv is an invariant measure for the semiflow Φ. The Theorem 5.2.1 is a consequence of this fact.

5.4 Application to structured regularization

In this section is provided an application of the stochastic Douglas Rachford (5.3) algorithm to solve a regularized optimization problem. The code is available at the address https://github.com/adil-salim/Stochastic-DR. Consider problem (5.1), where F is a cost function that is written as an expectation, and G is a regularization term. Towards solving (5.1), many approaches involve the computation of the proximity operator of the regularization term G. In the case where G is a structured regularization term, its proximity operator is often difficult to compute. We shall concentrate on the case where G is an overlapping group regularization. In this case, the computation of the proximity operator of G is known to be a bottleneck [132]. We shall apply the algorithm (5.3) to overcome this difficulty.

Consider $\mathsf{X} = \mathbb{R}^N$, $N \in \mathbb{N}^\star$, and $g \in \mathbb{N}^\star$. Consider g subsets of $\{1, \dots, N\}$, $S_1, \dots, S_g$, possibly overlapping. Set $G(x) = \sum_{j=1}^g \|x_{S_j}\|$, where x_{S_j} denotes the restriction of x to the set of index S_j and $\|\cdot\|$ denotes the Euclidean norm. Set $F(x) = \mathbb{E}_{(\xi,\eta)}(h(\eta\langle x, \xi\rangle))$ where h denotes the hinge loss $h(z) = \max(0, 1-z)$ and (ξ, η) is a r.v defined on some probability space with values in $\mathsf{X} \times \{-1, +1\}$. In this case, the problem (5.1) is also called the SVM classification problem, regularized by the overlapping group lasso. It is assumed that the user is provided with i.i.d copies $((\xi_n, \eta_n))_n$ of the r.v (ξ, η) online.

To solve this problem, we implement a stochastic Douglas Rachford strategy. To that end, the regularization G is rewritten $G(x) = \mathbb{E}_J(g\|x_{S_J}\|)$ where J is an uniform r.v over $\{1, \dots, g\}$. At each iteration n of the stochastic Douglas Rachford algorithm, the user is provided with the realization (ξ_n, η_n) and sample a group J_n uniformly in $\{1, \dots, g\}$. Then, a Douglas Rachford step is done, involving the computation of the proximity operators of the functions $g_n : x \mapsto \|x_{S_{J_n}}\|$ and $f_n : x \mapsto h(\eta_n\langle x, \xi_n\rangle)$.

This strategy is compared with a partially stochastic Douglas Rachford algorithm, deterministic in the regularization G, where the fast subroutine Fog-Lasso [132] is used to compute the proximity operator of the regularization G. At each iteration n, the user is provided with (ξ_n, η_n). Then, a Douglas Rachford step is done, involving the computation of the proximity operators of the functions G and $f_n : x \mapsto h(\eta_n\langle x, \xi_n\rangle)$. Figure 5.1 demonstrates the advantage of treating the regularization term in a stochastic way.

In Fig. 5.1 "Stochastic D-R" denotes the stochastic Douglas Rachford algorithm and "Partially stochastic D-R" denotes the partially stochastic Douglas Rachford where the subroutine FoG-Lasso [132]

[2] Assumptions 5.2.4, 5.2.5 and 5.2.8 are not needed if the domains $D(s)$ are common *i.e* if $s \mapsto D(s)$ is constant.

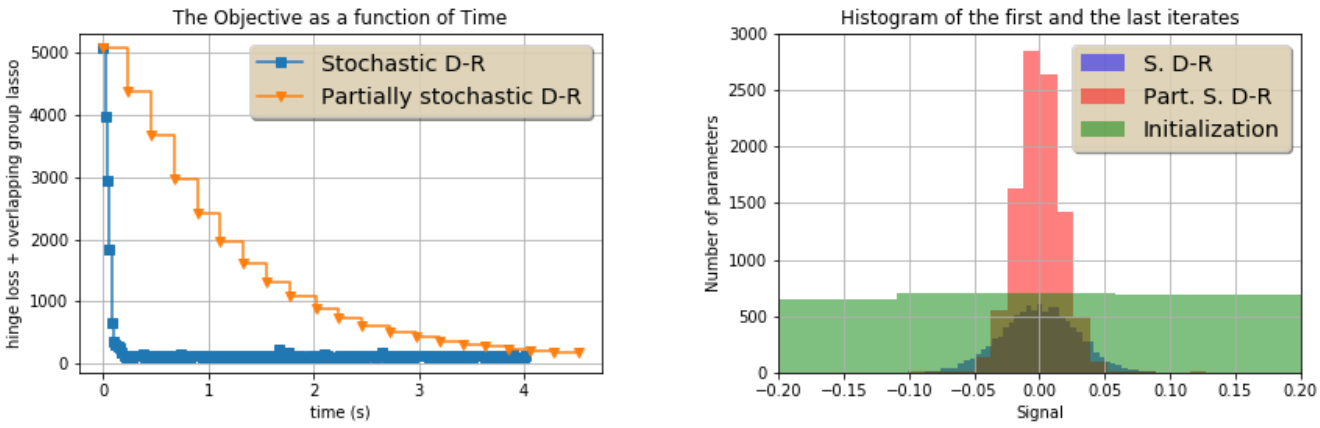

Figure 5.1: Left: The objective function $F + G$ as a function of time in seconds for each algorithm. Right: Histogram of the Initialization and the last iterates of the Stochastic D-R (S. D-R) and the partially stochastic D-R (Part. S. D-R).

is used at each iteration to compute the true proximity operator of the regularization G. Fig. 5.1 also shows the appearance of the first and the last iterates. Even if a best performing procedure [132] is used to compute $\mathrm{prox}_{\gamma G}$, we observe on Fig. 5.1 that Stochastic D-R takes advantage of being a stochastic method. This advantage is known to be twofold ([34]). First, the iteration complexity of Stochastic D-R is moderate because $\mathrm{prox}_{\gamma G}$ is never computed. Then, Stochastic D-R is faster than its partially deterministic counterpart which uses Fog-Lasso [132] as a subroutine, especially in the first iterations of the algorithms. Moreover, Stochastic D-R seems to perform globally better. This is because every proximity operators in Stochastic D-R can be efficiently computed ([12]). Contrary to the proximity operator of G [132], the proximity operator of g_n is easily computable. The proximity operator of f_n is easily computable as well.[3]

5.5 Application to distributed optimization

We now consider an application of the stochastic Douglas Rachford algorithm to a distributed optimization problem. More precisely, a slowly moving underwater target has to be located. To this end, M transmitters and N receivers are spatially distributed. We consider the case of a two-dimensional space, although the sequel can be easily extended to a three-dimensional space. If the receivers process the measurements they receive at the same time and at each time step, we shall say that the receivers operate synchronously. For more flexibility, we shall assume that the receivers only process the measurements at random instants, *i.e* operate asynchronously. Networks of sensors often work under this additional restriction. The processing takes the form of a computation (of a proximity operator w.r.t the measurements they receive) and/or a communication of the estimated position of the target by a receiver to another receiver. More precisely, assume that $N = 1$. It can be shown that the position $(x(t), y(t)) \in \mathbb{R}^2$ of the target at time t can be estimated by solving an optimization problem of the form

$$\min_{(x,y)\in\mathbb{R}^2} \|A(t)(x,y)^T - b(t)\|^2 \tag{5.9}$$

[3]Even if $h(x) = \log(1 + \exp(-x))$ (logistic regression), the proximity operator of f_n is easily computable, see [44].

where $A(t)$ and $b(t)$ encode the positions of the receiver and the transmitters, as well as the measurements of the receiver at time t. If several receivers are used, *i.e* $N > 1$, consider a connected graph $G = (V, E)$ where the set $V = \{1, \ldots, N\}$ of vertices is the set of receivers and E is the set of edges. Receivers are allowed to communicate along the edges of E. In this case, the network of receivers has to solve at each instant t the following problem to estimate the position of the target:

$$\min_{(x,y)\in\mathbb{R}^2} \sum_{i\in V} \|A_i(t)(x, y)^T - b_i(t)\|^2. \tag{5.10}$$

This problem need to be solved asynchronously and can be seen as a consensus problem : at each instant a receiver will update its estimation of the position of the target and share it with a neighbor in G. Indeed, Problem (5.10) is equivalent to

$$\min_{(x_1,y_1),\ldots,(x_N,y_N)\in\mathbb{R}^2} \sum_{i\in V} \|A_i(t)(x_i, y_i)^T - b_i(t)\|^2 + \sum_{\{i,j\}\in E} \iota_S((x_i, y_i), (x_j, y_j)) \tag{5.11}$$

where $S = \{(z, z') \in \mathbb{R}^2 \times \mathbb{R}^2, z = z'\}$. For every $v \in V$ and every $e \in E$ consider the convex functions $g((v, e), \cdot) : (\mathbb{R}^2)^N \to (-\infty, +\infty]$ defined by $g((v, e), ((x_1, y_1), \ldots, (x_N, y_N))) = \iota_S((x_i, y_i), (x_j, y_j))$ where $e = \{i, j\}$ and $f((v, e), \cdot) : (\mathbb{R}^2)^N \to (-\infty, +\infty)$ defined by $f((v, e), ((x_1, y_1), \ldots, (x_N, y_N))) = \|A_v(t)(x_v, y_v)^T - b_v(t)\|^2$. Finally, consider a r.v $\xi = (\xi(1), \xi(2))$ where $\xi(1)$ has a uniform distribution over V and $\xi(2)$ has a uniform distribution over E. Problem (5.11) is equivalent to finding a minimizer of $\mathbb{E}_\xi(f(\xi, \cdot) + g(\xi, \cdot))$ for which we apply stochastic Douglas Rachford algorithm (5.3). At each iteration of this algorithm, a randomly chosen receiver update its estimated value of the position of the target by computing a proximity operator w.r.t. its measurements. Then, two randomly chosen neighbors share their estimated value. Namely, they compute the mean of their values and update their values with the mean. The resulting algorithm is asynchronous and distributed. Moreover, it could be easily made adaptive : if the measurements are corrupted by a zero-mean noise, the proposed algorithm still converges because it is an instance of stochastic Douglas Rachford (5.3).

In our numerical simulation, we considered two transmitters and six receivers, whose positions in 2D Cartesian coordinates are: $t_1 = [0, 0]$, $t_2 = [2000, 2000]$, $r_1 = [-1000, -1000]$, $r_2 = [1500, -1000]$, $r_3 = [-1000, 1000]$, $r_4 = [1500, 1000]$, $r_5 = [1500, 2500]$, and $r_t = [2500, 1500]$, respectively. The receivers form nodes of the connected graph G with edges

$$E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}, \{4, 5\}, \{4, 6\}\}.$$

The initial position of the target is $[500, 500]$. The target is moving according to a spiral. We show the tracking ability of the proposed asynchronous adaptive distributed algorithm. This algorithm is compared to its synchronous analogue (which can also be cast as an instance of stochastic Douglas Rachford (5.3)). Figure 5.2 shows the true track of the target, and the tracks estimated by the two algorithms. Between two sample points of the true track (i.e. between two blue star markers on blue curve), we allowed 50 iterations for both the algorithms, and it is sufficient to track continuously the target with good accuracies. In spite of using only two nodes in estimation at each iteration, the asynchronous algorithm, after certain initial lag, performs almost similar to the synchronous one that involved all six nodes at each iteration.

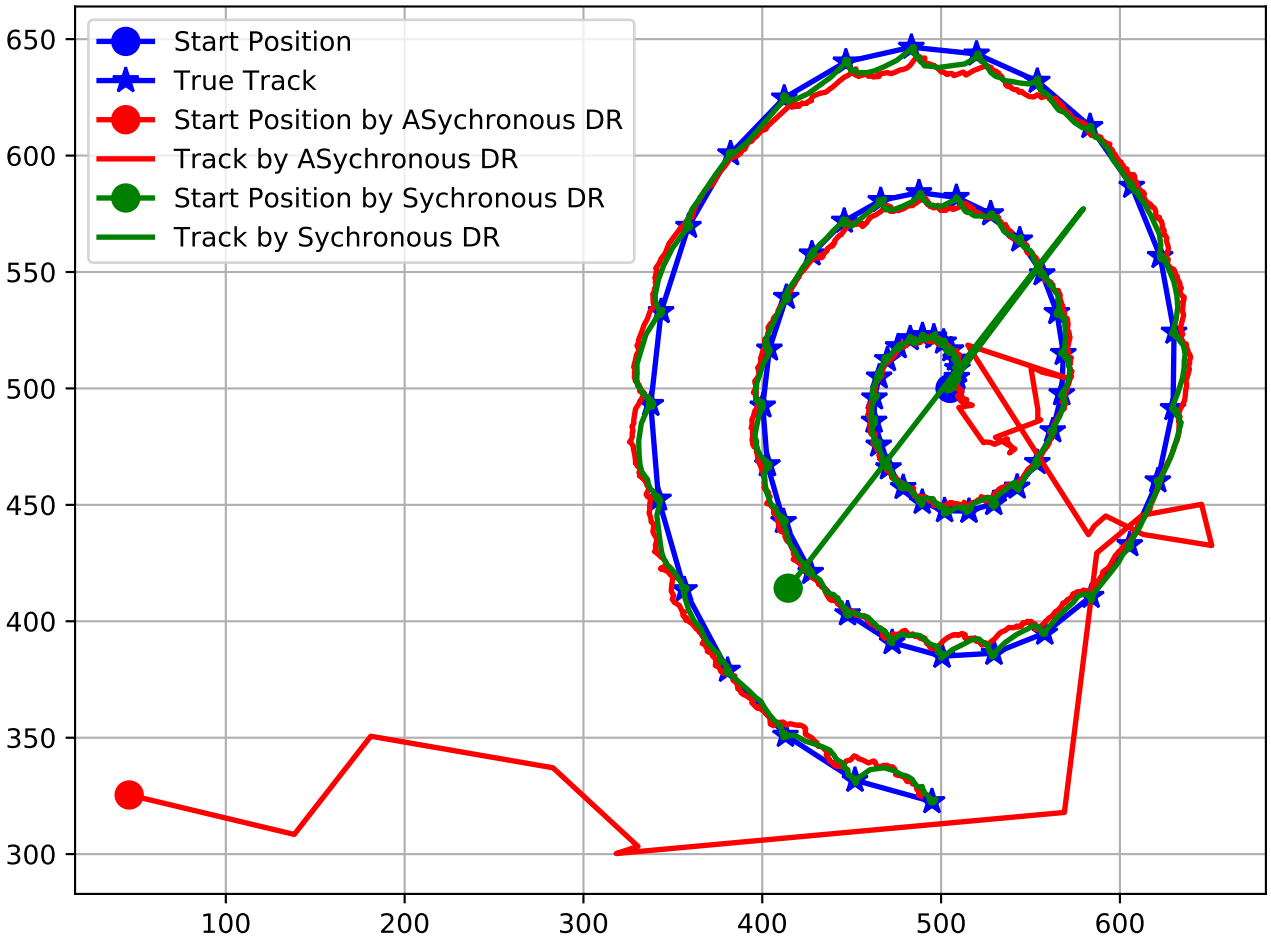

Figure 5.2: Numerical simulation results on tracking slowly moving target.

Part II

Applications using vanishing step sizes

Chapter 6

Introduction to Stochastic Approximations with a decreasing step size involving Maximal Monotone Operators

In the first part of this thesis, we introduced the constant step stochastic approximation framework (1.4) for two kinds of underlying DI. We first studied the DI induced by an usc operator. Then, by considering first the model of the stochastic Forward Backward and then the model of the stochastic Douglas Rachford, we studied DI induced by a monotone operator.

In the second part of this thesis, we are interested in decreasing step sizes stochastic approximation algorithms. More precisely, we are motivated by applications of the stochastic Forward Backward algorithm with vanishing step sizes. The paper [24] performs a theoretical study of the stochastic Forward Backward (FB) algorithm. The authors brought tools from dynamical systems and stochastic approximation to prove the almost sure convergence of the iterates of the stochastic FB algorithm.

In this chapter, we recall the main results and the proof technique of [24]. Although the results of [24] are prior to this thesis, this chapter is the entry point for our applications, which will be studied in the two next chapters.

Chapter organization. In the next section, the stochastic Forward Backward algorithm with decreasing step is presented. Then, the main convergence result of this algorithm is stated in Sec. 6.2. Finally, a sketch of the convergence proof is provided in Sec. 6.3.

6.1 The stochastic Forward-Backward algorithm

Consider two random monotone operators $A, B : (\Xi, \mathscr{G}, \mu) \to \mathscr{M}(\mathsf{X})$ such that for every $s \in \Xi$, $\mathrm{dom}(B(s)) = \mathsf{X}$. Denoting $A(s,x)$ the image of x by the operator $A(s)$, recall that $s \mapsto A(s,x)$ is measurable, and similarly for B. Consider a function $b : \Xi \times \mathsf{X} \to \mathsf{X}$ such that b is $\mathscr{G} \otimes \mathscr{B}(\mathsf{X})$-measurable and for every $x \in \mathsf{X}$, $b(s,x) \in B(s,x)$ for μ-a.e. $s \in \Xi$. A possible choice for $b(s,x)$ is $B_0(s,x)$, the least norm element in $B(s,x)$. Assume moreover that $b(\cdot,x)$ is μ-integrable for all x. Under this hypothesis, $b(\cdot,x) \in \mathfrak{S}^1_{B(\cdot,x)}$ and $B(\cdot,x)$ is μ-integrable for every $x \in \mathsf{X}$, we set $\mathcal{B}(x) := \int B(s,x)\mu(ds)$, where a selection integral is used (see Eq. (2.4)). Note that $\mathrm{dom}\,\mathcal{B} = \mathsf{X}$. Denote $\mathcal{D}$ the essential intersection of

the domains $D(s) = \mathrm{dom}(A(s))$ (see Eq. (2.5)). Assuming that $\mathcal{D} \neq \emptyset$ and that $A(\cdot, x)$ is integrable for every $x \in \mathcal{D}$, we denote the selection integral $\mathcal{A}(x) := \int A(s,x)\mu(ds)$.

Let (ξ_n) be an i.i.d. sequence of random variables from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\Xi, \mathscr{G})$ with the distribution μ. Let x_0 be an X-valued random variable with probability distribution ν, and assume that x_0 and (ξ_n) are independent. Starting from x_0, our purpose is to study the behavior of the iterates

$$x_{n+1} = J_{\gamma_{n+1}}(\xi_{n+1}, x_n - \gamma_{n+1} b(\xi_{n+1}, x_n)), \quad n \in \mathbb{N}\,, \tag{6.1}$$

for a given sequence of positive step size $(\gamma_n) \in \ell^2 \setminus \ell^1$, where we recall the notation $J_\gamma(s,\cdot) := (I + \gamma A(s))^{-1}(\cdot)$ for every $s \in \Xi$.

In the deterministic case where the functions $A(s,\cdot)$ and $B(s,\cdot)$ are replaced with deterministic maximal monotone operators $\mathsf{A}(\cdot)$ and $\mathsf{B}(\cdot)$, with B assumed single-valued and the step size $\gamma_n \equiv \gamma$ assumed to be constant, the algorithm coincides with the well-known Forward-Backward algorithm. If B is cocoercive, and $\gamma > 0$ not too large, the Forward-Backward algorithm converges to an element of $Z(\mathsf{A} + \mathsf{B})$, provided this set is not empty [12, Th. 25.8]. In the stochastic case who is of interest here, the sequence (γ_n) has to converge to zero in order to make the stochastic Forward Backward (6.1) converging to an element of $Z(\mathcal{A} + \mathcal{B})$.

6.2 Almost sure convergence of the iterates

In this section, we present the main convergence result of the stochastic Forward Backward algorithm with decreasing step size (6.1).

For every $x_\star \in Z(\mathcal{A} + \mathcal{B})$, define the set of $2p$-integrable representations of the zero $x_\star$:

$$\mathcal{R}_{2p}(x_\star) = \left\{(\varphi, \psi) \in \mathfrak{S}^{2p}_{A(\cdot,x_\star)} \times \mathfrak{S}^{2p}_{B(\cdot,x_\star)} : \int \varphi d\mu + \int \psi d\mu = 0\right\}. \tag{6.2}$$

Consider the following assumptions.

Assumption 6.2.1. The sequence of positive step sizes satisfies $(\gamma_n) \in \ell^2 \setminus \ell^1$ and $\frac{\gamma_{n+1}}{\gamma_n} \to 0$.

Assumption 6.2.2. The monotone operator $\mathcal{A}$ is maximal.

Assumption 6.2.3. There exists an integer $p \geq 1$ and $x_\star \in Z(\mathcal{A} + \mathcal{B})$ such that $\mathcal{R}_{2p}(x_\star) \neq \emptyset$.

Assumption 6.2.4. For every $x_\star \in Z(\mathcal{A} + \mathcal{B})$, $\mathcal{R}_2(x_\star) \neq \emptyset$.

Assumption 6.2.5. For any compact set $\mathcal{K}$ of X, there exists $\varepsilon \in (0, 1]$ such that

$$\sup_{x \in \mathcal{K} \cap \mathcal{D}} \int \|A_0(s,x)\|^{1+\varepsilon} \mu(ds) < \infty.$$

Moreover, there exists $x_0 \in \mathcal{D}$ such that

$$\int \|A_0(s,x_0)\|^{1+1/\varepsilon} \mu(ds) < \infty.$$

Assumption 6.2.6. There exists $C > 0$ such that for any $x \in \mathsf{X}$,

$$\int d(x, D(s))^2 \mu(ds) \geq C \boldsymbol{d}(x)^2$$

where $\boldsymbol{d}(\cdot)$ is the distance function to $\mathcal{D}$.

Assumption 6.2.7. There exists $C > 0$ such that for any $x \in \mathsf{X}$ and any $\gamma > 0$,

$$\frac{1}{\gamma^4} \int \|J_\gamma(s,x) - \Pi_{\mathrm{cl}(D(s))}(x)\|^4 \mu(ds) \leq C(1 + \|x\|^{2p})$$

where the integer p is specified in Assumption 6.2.3

Assumption 6.2.8. There exists $M : \Xi \to \mathbb{R}_+$ such that M^{2p} is μ-integrable, and for all $x \in \mathsf{X}$, $\|b(s,x)\| \leq M(s)(1 + \|x\|)$. Moreover, there exists a constant $C > 0$ such that $\int \|b(s,x)\|^4 \mu(ds) \leq C(1 + \|x\|^{2p})$.

Theorem 6.2.1 ([24]). Assume that Assumptions 6.2.1–6.2.8 hold. Then, there exists a random variable $X_\star$ such that $\mathbb{P}(X_\star \in Z(\mathcal{A}+\mathcal{B})) = 1$ and such that the sequence of empirical means $(\bar{x}_n)_n$ converges a.s. to $X_\star$. Moreover, if $\mathcal{A}+\mathcal{B}$ is demipositive, then $x_n \to_{n\to+\infty} X_\star$ a.s.

6.3 General Approach

In this section, we present the general approach used to prove the convergence of the stochastic Forward Backward (6.1) with decreasing step size. The approach relies on related the iterates of (6.1) with the DI associated to the monotone operator $\mathcal{A}+\mathcal{B}$ (see Sec. 2.2.2). More precisely, let us endow the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the filtration $(\mathscr{F}_n)$ defined as $\mathscr{F}_n = \sigma(\xi_1, \ldots, \xi_n)$, and we write $\mathbb{E}_n = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$. In particular, $\mathbb{E}_0 = \mathbb{E}$.

The principle of the proof of convergence of the algorithm (6.1) is the following. Given $a \in \mathsf{X}$, consider the Differential Inclusion (DI) associated with the monotone operator $\mathcal{A}+\mathcal{B}$ (see Sec. 2.2.2)

$$\begin{cases} \dot{\mathsf{z}}(t) & \in & -(\mathcal{A}+\mathcal{B})(\mathsf{z}(t)) \\ \mathsf{z}(0) & = & a\,. \end{cases} \tag{6.3}$$

Since $\mathcal{B}$ is maximal ([24, Prop. 3.1]), $\operatorname{dom}\mathcal{B} = \mathsf{X}$ and $\mathcal{A}$ is maximal (6.2.2), $\mathcal{A}+\mathcal{B}$ is maximal ([12, Corollary 24.4]). Denote Φ the semiflow with (6.3). Let us introduce the following function I from $\mathsf{X}^{\mathbb{N}}$ to the space of $C(\mathbb{R}_+, \mathsf{X})$. For $x = (x_n) \in \mathsf{X}^{\mathbb{N}}$, the function $\mathsf{x} = \mathsf{I}(x)$ is the continuous interpolated process obtained from x as

$$\mathsf{x}(t) = x_n + \frac{x_{n+1} - x_n}{\gamma_{n+1}}(t - \tau_n) \tag{6.4}$$

for $t \in [\tau_n, \tau_{n+1})$, where $\tau_n = \sum_{k=1}^n \gamma_k$. Consider the interpolated function $\mathsf{x} = \mathsf{I}((x_n))$ where (x_n) is the sequence satisfying (6.1). The paper [24] proves the two following facts:

- The sequence $(\|x_n - x_\star\|)$ is almost surely convergent for each $x_\star \in Z(\mathcal{A}+\mathcal{B})$,
- The process $\mathsf{x}(t)$ is an almost sure Asymptotic Pseudo Trajectory (APT) of the semiflow Φ see Chap. 1 or [15]. Namely, for each $T > 0$,

$$\sup_{u\in[0,T]} \|\mathsf{x}(t+u) - \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(\mathsf{x}(t)), u)\| \xrightarrow[t\to\infty]{\text{a.s.}} 0. \tag{6.5}$$

Taken together, these two results lead to the a.s. convergence of the empirical means

$$\bar{x}_n = \frac{\sum_{k=1}^n \gamma_k x_k}{\tau_n}$$

to some r.v. $X_\star$ supported by the set $Z(\mathcal{A}+\mathcal{B})$, as is shown by [24, Cor. 3.2]. Moreover, if $\mathcal{A}+\mathcal{B}$ is demipositive (for example if $\mathcal{A}+\mathcal{B} = \partial G, G \in \Gamma_0(\mathsf{X})$), then x_n also converges to $X_\star$.

The following proposition can be found in [22, Prop. 1] or [24, Prop. 6.1]:

Proposition 6.3.1. Let Assumptions 6.2.1 and 6.2.4 hold true. Then the following facts hold true:

1. For each $x_\star \in Z(\mathcal{A}+\mathcal{B})$, the sequence $(\|x_n - x_\star\|)$ converges almost surely.
2. The sequence (x_n) is bounded almost surely and in $\mathcal{L}^2(\Omega, \mathscr{F}, \mathbb{P}; \mathsf{X})$.
3. $\mathbb{E}\left[\sum_{n=1}^{\infty} \gamma_{n+1}^2 \int \|A_{\gamma_{n+1}}(s, x_n - \gamma_{n+1} b(s, x_n))\|^2 \mu(ds)\right] < \infty$.

It remains to establish the almost sure APT. We just provide here the main arguments of this part of the proof, since it can be found in [24].

Let us write

$$x_{n+1} = x_n - \gamma_{n+1} b(\xi_{n+1}, x_n) - \gamma_{n+1} A_{\gamma_{n+1}}(\xi_{n+1}, x_n - \gamma_{n+1} b(\xi_{n+1}, x_n)), \tag{6.6}$$

and let us also define the function

$$h_\gamma(s, x) = -b(s, x) - A_\gamma(s, x - \gamma b(s, x)) \tag{6.7}$$

Using Assumptions 6.2.5 and 6.2.8, $\int \|h_\gamma(s, x)\| \mu(ds) < \infty$ and we define:

$$H_\gamma(x) = \int h_\gamma(s, x) \mu(ds)\,.$$

Note that $x_{n+1} = x_n + \gamma_{n+1} h_{\gamma_{n+1}}(\xi_{n+1}, x_n)$. Defining the $(\mathscr{F}_n)$ martingale

$$M_n = \sum_{k=1}^{n} x_k - \mathbb{E}_{k-1}[x_k]$$

it is clear that $x_{n+1} = x_n + \gamma_{n+1} H_{\gamma_{n+1}}(x_n) + (M_{n+1} - M_n)$. Let us rewrite this equation in a form involving the continuous process $\mathsf{x} = \mathsf{I}((x_n))$. Defining $\mathsf{M} = \mathsf{I}((M_n))$, and writing

$$r(t) = \max\{k \geq 0 \,:\, \tau_k \leq t\}, \quad t \geq 0,$$

we obtain

$$\begin{aligned} \mathsf{x}(\tau_n + t) - \mathsf{x}(\tau_n) = &- \int_0^t H_{\gamma_{r(\tau_n+u)+1}}(x_{r(\tau_n+u)})\, du \\ &+ \mathsf{M}(\tau_n + t) - \mathsf{M}(\tau_n)\,. \end{aligned} \tag{6.8}$$

The first argument of the proof of the almost sure APT is a compactness argument on the sequence of continuous processes $(\mathsf{x}(\tau_n + \cdot))_n$.

Specifically, we show that on a $\mathbb{P}$-probability one set, this sequence is equicontinuous and bounded. By Ascoli's theorem, this sequence admits accumulation points in the topology of the uniform convergence on the compacts of $\mathbb{R}_+$. As a second step, we show that these accumulation points are solutions to the differential inclusion (6.3), which is in fact a reformulation of the almost sure APT property (6.5).

Since

$$\begin{aligned} \mathbb{E}[\|x_{n+1} - \mathbb{E}_n x_{n+1}\|^2] &= \gamma_n^2 \mathbb{E}[\|b(\xi_{n+1}, x_n) - \int b(s, x_n)\mu(ds) \\ &\quad + A_{\gamma_{n+1}}(\xi_{n+1}, x_n - \gamma_{n+1} b(\xi_{n+1}, x_n)) - \int A_{\gamma_{n+1}}(s, x_n - \gamma_{n+1} b(s, x_n))\mu(ds)\|^2] \\ &\leq 4\gamma_{n+1}^2 \mathbb{E}[\int \|b(s, x_n)\|^2 \mu(ds) + \int \|A_{\gamma_{n+1}}(s, x_n - \gamma_{n+1} b(s, x_n))\|^2 \mu(ds)]\,, \end{aligned}$$

we obtain by Prop. 6.3.1 that $\sup_n \mathbb{E}[\|M_n\|^2] < \infty$. Thus, the martingale M_n converges almost surely, which implies that the sequence of stochastic processes $(\mathsf{M}(\tau_n + \cdot) - \mathsf{M}(\tau_n))_n$ converges almost surely to zero, uniformly on $\mathbb{R}_+$.

Using Assumptions 6.2.5, 6.2.8 and the fact that the sequence (x_n) is almost surely bounded by Prop. 6.3.1, it is easily seen that

$$\sup_n \|H_{\gamma_{n+1}}(x_n)\| < \infty \quad \text{a.s.},$$

provided that $D(s) = \mathsf{X}$ (we shall call this context the full domain case) for μ-a.e. $s \in \Xi$. The case where $\mathcal{D} \neq \mathsf{X}$ introduces more technicalities in this part of the proof, that we have chosen to omit. Inspecting (6.8), we thus obtain that the sequence $(\mathsf{x}(\tau_n + \cdot))_n$ is equicontinuous and bounded with probability one.

In order to characterize its cluster points, choose $T > 0$, and consider an elementary event on the probability one set where $(\mathsf{x}(\tau_n + \cdot))_n$ is equicontinuous and bounded. With a small notational abuse, let (n) be a subsequence along which $(\mathsf{x}(\tau_n + \cdot))_n$ converges on $[0, T]$ to some continuous function $\mathsf{z}(t)$. This function then is written as

$$\mathsf{z}(t) - \mathsf{z}(0) = -\lim_{n\to\infty} \int_0^t du \int_\Xi \mu(ds)\, h_{\gamma_{r(\tau_n+u)+1}}(s, x_{r(\tau_n+u)}).$$

Using Assumptions 6.2.5, 6.2.8 and the fact that the sequence (x_n) is bounded by Prop. 6.3.1 it is easy to see that there exists $\varepsilon > 0$ such that

$$\sup_n \int \|h_{\gamma_{n+1}}(s, x_n)\|^{1+\varepsilon} \mu(ds) < \infty.$$

As a consequence, the sequence of functions $(h_{\gamma_{r(\tau_n+u)+1}}(s, x_{r(\tau_n+u)}))_n$ in the parameters (s, u) is bounded in the Banach space $\mathcal{L}^{1+\varepsilon}(d\mu \otimes du)$ where du is the Lebesgue measure on $[0, T]$. Since the unit ball of $\mathcal{L}^{1+\varepsilon}(d\mu \otimes du)$ is weakly compact in this space by the Banach-Alaoglu theorem, since this space is reflexive, we can extract a subsequence (still denoted as (n)) such that $h_{\gamma_{r(\tau_n+u)+1}}(s, x_{r(\tau_n+u)})$ converges weakly in $\mathcal{L}^{1+\varepsilon}(d\mu \otimes du)$, as $n \to \infty$, to a function $Q(s, u)$. The remainder of the proof consists in showing that Q can be written as $Q(s, u) = a(s, u) + \beta(s, u)$ where $a(s, u) \in A(s, \mathsf{z}(u))$ and $\beta(s, u) \in B(s, \mathsf{z}(u))$ for $d\mu \otimes du$-almost all (s, u). Indeed, once this result is established, it becomes clear that $\mathsf{z}(t)$ is an absolutely continuous function whose derivative satisfies almost everywhere the inclusion (6.3). We just provide here the main argument of this part of the proof. Let us focus on the sequence of functions of $(s, u) \in \Xi \times [0, T]$ defined by

$$A_{\gamma_{r(\tau_n+u)+1}}(s, x_{r(\tau_n+u)} - \gamma_{r(\tau_n+u)+1} b(s, x_{r(\tau_n+u)}))$$

and indexed by n. This sequence is bounded in $\mathcal{L}^{1+\varepsilon}(d\mu \otimes du)$ on a probability one set, as a function of (s, u), for the same reasons as those explained above for $(h_{\gamma_{r(\tau_n+u)+1}}(s, x_{r(\tau_n+u)}))_n$. We need to show that any weak limit point $a(s, u)$ of this sequence satisfies $a(s, u) \in A(s, \mathsf{z}(u))$ for $d\mu \otimes du$-almost all (s, u). Using the fact that $\mathsf{x}(\tau_n + \cdot) \to \mathsf{z}(\cdot)$ almost surely, along with the inequality $\langle A_\gamma(s, x) - w, J_\gamma(s, x) - v\rangle \geq 0$, valid for all $x, v \in \mathsf{X}$ and $w \in A(s, v)$, we show that $\langle a(s, u) - w, \mathsf{z}(u) - v\rangle \geq 0$ for $d\mu \otimes du$-almost all (s, u). Since $v \in \mathsf{X}$ and $w \in A(s, v)$ are arbitrary, we get that $a(s, u) \in A(s, \mathsf{z}(u))$ using the maximality of the monotone operator $A(s)$. The APT property is shown.

Chapter 7

A Primal Dual Algorithm for Stochastic Composite Optimization under Stochastic Linear Constraints

In this chapter, we propose a new stochastic primal-dual algorithm for solving a composite optimization problem under linear constraints. We assume that all the functions to be minimized are given as statistical expectations, as well as the matrices and the vector defining the linear constraints. These expectations are unknown but revealed across the time through i.i.d realizations of a random variable. The proposed algorithm can be seen as a stochastic Forward Backward to find a zero of the sum of two monotone operators. The two monotone operators are given as selection integrals and are not subdifferentials in general. We prove that the sequence of empirical means of the iterates converges to a saddle point of the Lagrangian function. The proposed algorithm is tested experimentally to solve a decentralized optimization problem over a real-life graph of computing agents.

7.1 Introduction

Many applications in machine learning, statistics or signal processing require the solution of the following optimization problem. Given three Euclidean spaces X, Z, and V, find a point $(x, z) \in \mathsf{X} \times \mathsf{Z}$ such that

$$\mathsf{F}(x) + \mathsf{G}(x) + \mathsf{P}(z) + \mathsf{Q}(z) \text{ is minimum on } \{(x, z) \in \mathsf{X} \times \mathsf{Z},\ \mathsf{A}x + \mathsf{B}z = \mathsf{c}\}, \tag{7.1}$$

where $\mathsf{F}, \mathsf{G}, \mathsf{P}, \mathsf{Q}$ are convex functions such that F and P have a full domain, A belongs to the set $\mathcal{L}(\mathsf{X}, \mathsf{V})$ of $\mathsf{X} \to \mathsf{V}$ linear operators, $\mathsf{B} \in \mathcal{L}(\mathsf{Z}, \mathsf{V})$, and c is a vector in V. In order to solve Problem (7.1), primal-dual methods typically generate a sequence of primal estimates $(x_n, z_n)_{n\in\mathbb{N}}$ and a sequence of dual estimates $(\lambda_n)_{n\in\mathbb{N}}$ jointly converging to a saddle point of the Lagrangian function. For every saddle point of the Lagrangian function (x, z, λ), (x, z) is a solution of the primal problem (7.1) and λ is a solution of a dual problem of (7.1). Conversely, assume that the following qualification condition

$$\mathsf{c} \in \mathrm{ri}\,(\mathsf{A}\,\mathrm{dom}\,\mathsf{G} + \mathsf{B}\,\mathrm{dom}\,\mathsf{Q})$$

holds true, where ri is the relative interior of a set [12]. Then, there exists a saddle point of the Lagrangian function and for every solution (x, z) of (7.1) and every dual solution λ, (x, z, λ) is a saddle point of the Lagrangian. There is a rich literature on such algorithms which cannot be exhaustively listed [62, 42, 51, 124].

In this chapter, it is assumed that the quantities that enter the minimization problem are likely to be unavailable or difficult to compute numerically. More precisely, it is assumed that the functions $\mathsf{F}, \mathsf{G}, \mathsf{P}$ and Q are defined as expectations of random functions w.r.t a probability measure μ. Namely, these functions take the forms $\mathsf{F}(x) = \mathbb{E}_\xi(f(\xi, x))$, $\mathsf{G}(x) = \mathbb{E}_\xi(g(\xi, x))$, $\mathsf{P}(z) = \mathbb{E}_\xi(p(\xi, z))$, and $\mathsf{Q}(z) = \mathbb{E}_\xi(q(\xi, z))$ where f, g, p, q are normal convex integrands and ξ is some random variable with distribution μ. In addition, it is assumed that the operators A and B are written as $\mathsf{A} = \mathbb{E}_\xi A(\xi)$ and $\mathsf{B} = \mathbb{E}_\xi B(\xi)$, where A and B are measurable functions with values in $\mathcal{L}(\mathsf{X}, \mathsf{V})$ and $\mathcal{L}(\mathsf{Z}, \mathsf{V})$ respectively. Finally c takes the form $\mathsf{c} = \mathbb{E}_\xi c(\xi)$. As an extreme case, *none* of these expectations are available to the observer. What is given to this observer are the functions f, g, p, q, A, B, and c, and a sequence of i.i.d random vectors (ξ_n) with the probability distribution μ. A new stochastic primal dual algorithm based on this data is proposed to solve Problem (7.1).

The convergence proof for this algorithm relies on Th. 6.2.1 of Chap. 6. The algorithm is built around an instantiation of the stochastic Forward-Backward algorithm involving random monotone operators that was introduced in [24] (see also Chap. 6). Existing methods typically allow to handle subproblems of Problem (7.1) in which some quantities used to define (7.1) are assumed to be available or set equal to zero [92, 104, 105, 133, 131, 48, 49, 122]. In particular, the new algorithm generalizes the stochastic gradient algorithm (in the case where only F is non zero), the stochastic proximal point algorithm [95, 122, 22] (only G is non zero), and the stochastic proximal gradient algorithm [2, 3, 27, 49] (only $\mathsf{F} + \mathsf{G}$ is non zero).

With [131], the proposed algorithm is one of the first methods that allow to tackle stochastic constraints online. While [131] is focused on convex and compact constraints, this chapter consider the case of unbounded linear constraints. Handling stochastic constraints online is suitable in various fields of machine learning like Neyman-Pearson classification or online portfolio optimization. For example, in online Markowitz portfolio optimization ([37]) one has to solve the minimization problem

$$\min_{x \in \mathbb{R}^d} \mathbb{E}_\xi(\langle x, \xi \rangle^2) \qquad \text{s.t.} \qquad x \in \Delta \quad \text{and} \quad \langle x, \mathbb{E}_\xi(\xi) \rangle = r, \tag{7.2}$$

where ξ is a random vector in $\mathbb{R}^d$, $r > 0$ and Δ is the simplex of $\mathbb{R}^d$. Authors usually assume $\mathbb{E}_\xi(\xi)$ to be known or estimated [95, 133]. The new primal dual algorithm is also an alternative to efficient methods ([1]) in huge scale convex optimization

$$\min_{x \in \mathsf{X}} \mathsf{F}(x) + \mathsf{G}(x) + \mathsf{Q}(\mathsf{A}x) \tag{7.3}$$

where the functions F, G and Q are intractable and matrix vector operations involving A are also intractable. In many cases, F and G are cost functions and $\mathsf{Q}(\mathsf{A}x)$ a structured regularization term that must be handled by splitting Q and A. Finally, in distributed optimization, in the context where a network of computing agents is required to minimize a cost function, the proposed algorithm allows to design a fully asynchronous and adaptive algorithm in which, at each iteration, only one randomly chosen agent becomes active.

Chapter organization. The chapter is organized as follow. The next section is devoted to rigorously state the main problem and the main algorithm. In Sec. 7.3 the convergence proof of the algorithm is given. Application to distributed optimization is discussed in Sec. 7.4.

7.2 Problem description

Any space of linear operators between two Euclidean space is endowed with the operator norm $\|\cdot\|$ and the resulting Borel field, and can be identified with a space of matrices.

Let $f : \Xi \times \mathsf{X} \to (-\infty, \infty]$ be a normal convex integrand, and assume that $\int |f(s,x)|\,\mu(ds) < \infty$ for all $x \in \mathsf{X}$. Consider the convex function $\mathsf{F}(x)$ defined on X as the Lebesgue integral $\mathsf{F}(x) = \mathbb{E}_\xi f(\xi, x)$. The interchange property holds : $\partial \mathsf{F}(x) = \int \partial f(s,x)\mu(ds)$ (see Ex. 3 of Chap. 2).

Let $g : \Xi \times \mathsf{X} \to (-\infty, \infty]$ be a normal convex integrand, and let $\mathsf{G}(x) = \mathbb{E}_\xi g(\xi, x)$, where the integral is defined as the sum

$$\int_{\{s\,:\,g(s,x)\in[0,\infty)\}} g(s,x)\,\mu(ds) + \int_{\{s\,:\,g(s,x)\in]-\infty,0[\}} g(s,x)\,\mu(ds) + I(x)\,,$$

and

$$I(x) = \begin{cases} +\infty, & \text{if } \mu(\{s : g(s,x) = \infty\}) > 0, \\ 0, & \text{otherwise}\,, \end{cases}$$

and where the convention $(+\infty) + (-\infty) = +\infty$ is used. Assume $\mathsf{G}(x) > -\infty$ for all x [125], and assume that G is proper. We shall assume that the interchange property holds for g (see Ex. 3 of Chap. 2 for sufficient conditions).

To proceed with our problem statement, we introduce two normal convex integrands $p, q : \Xi \times \mathsf{Z} \to (-\infty, \infty]$ and assume that the function p (resp. q) has verbatim the same properties as f (resp. g), after replacing the space X with Z. We also write $\mathsf{P}(x) = \mathbb{E}_\xi p(\xi, x)$ and $\mathsf{Q}(x) = \mathbb{E}_\xi q(\xi, x)$.

Finally, let $A : \Xi \to \mathcal{L}(\mathsf{X}, \mathsf{V})$ and $B : \Xi \to \mathcal{L}(\mathsf{Z}, \mathsf{V})$ be operator-valued random variables, and let $c : \Xi \to \mathsf{V}$ be a vector-valued random variable. Let us assume that $\|A(\cdot)\|$, $\|B(\cdot)\|$, and $\|c(\cdot)\|$ are μ-integrable, and let us introduce the Lebesgue integrals $\mathsf{A} = \mathbb{E}_\xi A(\xi)$, $\mathsf{B} = \mathbb{E}_\xi B(\xi)$, and $\mathsf{c} = \mathbb{E}_\xi c(\xi)$.

Having introduced these functions, our purpose is to find a solution $(x, z) \in \mathsf{X} \times \mathsf{Z}$ of Problem (7.1), where the set of such points is assumed non empty. To solve this problem, the observer is given the functions f, g, p, q, A, B, and c, and a sequence of i.i.d random variables $(\xi_n)_{n\in\mathbb{N}}$ from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\Xi, \mathscr{G})$ with the probability distribution μ.

We shall denote as $\widetilde{\nabla} f(s,x)$ a measurable subgradient of $f(s,\cdot)$ at x. More precisely, $\widetilde{\nabla} f : (\Xi \times \mathsf{X}, \mathscr{G} \otimes \mathscr{B}(\mathsf{X})) \to (\mathsf{X}, \mathscr{B}(\mathsf{X}))$ is a measurable function such that for each $x \in \mathsf{X}$, $\widetilde{\nabla} f(\cdot, x) \in \mathfrak{S}^1_{\partial f(\cdot,x)}$ (recall that this set is non empty). A possible choice for $\widetilde{\nabla} f(s,x)$ is $\partial_0 f(s,x)$ (see [24, §2.3 and §3.1] for the measurability issues). The notation $\widetilde{\nabla} p(s,x)$ will have a similar meaning.

Turning back to Problem (7.1), our purpose will be to find a saddle point of the Lagrangian $((x,z),\lambda) \mapsto \mathsf{F}(x) + \mathsf{G}(x) + \mathsf{P}(z) + \mathsf{Q}(z) + \langle \lambda, \mathsf{A}x + \mathsf{B}z - \mathsf{c}\rangle$. Denoting as $\mathcal{Z} \subset \mathsf{X} \times \mathsf{Z} \times \mathsf{V}$ the set of these saddle points, an element $((x,z),\lambda)$ of $\mathcal{Z}$ is characterized by the inclusions

$$\left\{ \begin{array}{rcl} 0 & \in & \partial \mathsf{F}(x) + \partial \mathsf{G}(x) + \mathsf{A}^T\lambda, \\ 0 & \in & \partial \mathsf{P}(z) + \partial \mathsf{Q}(z) + \mathsf{B}^T\lambda, \\ 0 & = & -\mathsf{A}x - \mathsf{B}z + \mathsf{c}\,. \end{array} \right. \tag{7.4}$$

The algorithm proposed here consists in the following iterations applied to the random vector $(x_n, z_n, \lambda_n) \in \mathsf{X} \times \mathsf{Z} \times \mathsf{V}$. Given a sequence of positive weights $(\gamma_n)_{n\in\mathbb{N}}$, set

The convergence of this algorithm is stated by the following theorem.

Theorem 7.2.1. Consider the Problem (7.1), and let the following assumptions hold true.

1. The step size sequence satisfies $(\gamma_n) \in \ell^2 \setminus \ell^1$, and $\gamma_{n+1}/\gamma_n \to 1$ as $n \to \infty$.

Algorithm 1 The Main Algorithm

$$
\begin{aligned}
x_{n+1} &= \operatorname{prox}_{\gamma_{n+1} g(\xi_{n+1},\cdot)}\left(x_n - \gamma_{n+1}(\widetilde{\nabla} f(\xi_{n+1}, x_n) + A(\xi_{n+1})^T \lambda_n)\right), \\
z_{n+1} &= \operatorname{prox}_{\gamma_{n+1} q(\xi_{n+1},\cdot)}\left(z_n - \gamma_{n+1}(\widetilde{\nabla} p(\xi_{n+1}, z_n) + B(\xi_{n+1})^T \lambda_n)\right), \\
\lambda_{n+1} &= \lambda_n + \gamma_{n+1}\left(A(\xi_{n+1}) x_n + B(\xi_{n+1}) z_n - c(\xi_{n+1})\right).
\end{aligned}
$$

2. There exists an integer $m \geq 2$ that satisfies the following conditions:
   - The functions $A(\cdot)$, $B(\cdot)$ and $c(\cdot)$ are in $\mathcal{L}^{2m}(\mu)$.
   - There exists a point $(x_\star, z_\star, \lambda_\star) \in \mathcal{Z}$, and four functions $\varphi_f \in \mathfrak{S}^{2m}_{\partial f(\cdot,x_\star)}$, $\varphi_g \in \mathfrak{S}^{2m}_{\partial g(\cdot,x_\star)}$, $\varphi_p \in \mathfrak{S}^{2m}_{\partial p(\cdot,z_\star)}$, and $\varphi_q \in \mathfrak{S}^{2m}_{\partial q(\cdot,z_\star)}$, for which
   $$\int \varphi_f d\mu + \int \varphi_g d\mu + \mathsf{A}^T \lambda_\star = 0, \text{ and } \int \varphi_p d\mu + \int \varphi_q d\mu + \mathsf{B}^T \lambda_\star = 0. \tag{7.5}$$
   The last assumption is verified for $m = 1$ and for each point $(x_\star, z_\star, \lambda_\star) \in \mathcal{Z}$.
3. For any compact set $\mathcal{K}$ of $\operatorname{dom} \partial \mathsf{G}$, there exist $\varepsilon \in (0, 1]$ and $x_0 \in \operatorname{dom} \partial \mathsf{G}$ such that
$$\sup_{x \in \mathcal{K}} \int \|\partial_0 g(s, x)\|^{1+\varepsilon} \mu(ds) < +\infty, \text{ and } \int \|\partial_0 g(s, x_0)\|^{1+1/\varepsilon} \mu(ds) < +\infty.$$
4. Writing $D_{\partial g}(s) = \operatorname{dom} \partial g(s, \cdot)$, there exists $C > 0$ such that for all $x \in \mathsf{X}$,
$$\int \operatorname{dist}(x, D_{\partial g}(s))^2 \mu(ds) \geq C \operatorname{dist}(x, \operatorname{dom} \partial \mathsf{G})^2.$$
5. There exists $C > 0$ such that for any $x \in \mathsf{X}$ and any $\gamma > 0$,
$$\int \| \operatorname{prox}_{\gamma g(s,\cdot)}(x) - \Pi_{\mathrm{cl}(D_{\partial g}(s))}(s, x)\|^4 \mu(ds) \leq C\gamma^4 (1 + \|x\|^{2m}),$$
where m is the integer provided by Assumption 2.

Assumptions similar to 3–5 are made on the function q.

6. There exists a measurable $\Xi \to \mathbb{R}_+$ function β such that β^{2m} is μ-integrable, where m is the integer provided by Assumption 2, and such that for all $x \in \mathsf{X}$,
$$\|\widetilde{\nabla} f(s, x)\| \leq \beta(s)(1 + \|x\|).$$
Moreover, there exists a constant $C > 0$ such that $\int \|\widetilde{\nabla} f(s, x)\|^4 \mu(ds) \leq C(1 + \|x\|^{2m})$.
A similar assumption is made on the function p.

Consider the sequence of iterates (x_n, z_n, λ_n) produced by the algorithm 1, and define the averaged estimates
$$\bar{x}_n = \frac{\sum_{k=1}^n \gamma_k x_k}{\sum_{k=1}^n \gamma_k}, \ \bar{z}_n = \frac{\sum_{k=1}^n \gamma_k z_k}{\sum_{k=1}^n \gamma_k}, \text{ and } \bar{\lambda}_n = \frac{\sum_{k=1}^n \gamma_k \lambda_k}{\sum_{k=1}^n \gamma_k}.$$
Then, the sequence $(\bar{x}_n, \bar{z}_n, \bar{\lambda}_n)$ converges almost surely (a.s.) to a random variable (X, Z, Λ) supported by $\mathcal{Z}$.

Let us now discuss our assumptions. Assumption 1 is standard in the decreasing step case. Assumption 2 is a moment assumption that is generally easy to check. Note that this assumption requires the set of saddle points $\mathcal{Z}$ to be non empty. Notice the relation between Equations (7.5) and the first two inclusions in (7.4). Focusing on the first inclusion, there exist $a \in \partial F(x_\star) = \int \partial f(s, x_\star)\mu(ds)$ and $b \in \partial G(x_\star) = \int \partial g(s, z_\star)\mu(ds)$ such that $0 = a + b + \mathsf{A}^T\lambda_\star$. Then, Assumption 2 states that there are two measurable selections φ_f and φ_g of $\partial f(\cdot, x_\star)$ and $\partial g(\cdot, z_\star)$ respectively which are both in $\mathcal{L}^{2m}(\mu)$ and which satisfy $a = \mathbb{E}_\mu \varphi_f$ and $b = \mathbb{E}_\mu \varphi_g$. Not also that the larger is m, and the weaker is Assumption 5.

Assumption 3 is relatively weak and easy to check. This assumption on the functions g and q is much weaker than Assumption 6, which assumes that the growth of $\widetilde{\nabla} f(s, \cdot)$ and $\widetilde{\nabla} p(s, \cdot)$ is not faster than linear. This is due to the fact that g and q enter the algorithm 1 through the proximity operator while the functions f and p are used explicitly in this algorithm. This use of the functions f and p is reminiscent of the well-known Robbins-Monro algorithm (see, *e.g.* [54]), where a linear growth is needed to ensure the algorithm stability. Note that Assumption 6 is satisfied under the more restrictive assumption that $\nabla f(s, \cdot)$ is L-Lipschitz continuous without any bounded gradient assumption.

Assumption 4 is quite weak, and is studied *e.g* in [90], see also Assumption (4.3.6) of Chap. 4. Let us finally discuss Assumption 5. As $\gamma \to 0$, it is known that $\mathrm{prox}_{\gamma g(s,\cdot)}(x)$ converges to $\Pi_g(s, x)$ for every (s, x). Assumption 5 provides a control on the convergence rate. This assumption holds under the sufficient condition that for μ-almost every s and for every $x \in \mathrm{dom}\, \partial g(s, \cdot)$,

$$\|\partial_0 g(s, x)\| \leq \beta(s)(1 + \|x\|^{m/2}) \,,$$

where β is a positive random variable with a finite fourth moment [22].

7.3 Proof of Th. 7.2.1

We now enter the proof of Th. 7.2.1. Let us set $\mathsf{Y} = \mathsf{X} \times \mathsf{Z} \times \mathsf{V}$, and endow this Euclidean space with the standard inner product. By writing $(x, z, \lambda) \in \mathsf{Y}$, it will be understood that $x \in \mathsf{X}$, $z \in \mathsf{Z}$, and $\lambda \in \mathsf{V}$.

For each $s \in \Xi$, define the set-valued operator $M(s)$ on Y as

$$M(s, (x, z, \lambda)) = \begin{bmatrix} \partial g(s, x) \\ \partial q(s, z) \\ c(s) \end{bmatrix} ,$$

where $M(s, (x, y, \lambda))$ is the image of (x, y, λ) by $M(s)$. Fixing $s \in \Xi$, the operator $M(s, (x, z, \lambda))$ coincides with the subdifferential of the normal convex integrand $g(s, x) + q(s, z) + c(s)\lambda$ with respect to (x, z, λ). Thus, the map $s \mapsto M(s)$ is a random monotone operator over Y. Let us also define the operator $M'(s)$ as

$$M'(s, (x, z, \lambda)) = \begin{bmatrix} \partial f(s, x) + A(s)^T\lambda \\ \partial p(s, z) + B(s)^T\lambda \\ -A(s)x - B(s)z \end{bmatrix} .$$

We can write $M'(s) = M'_1(s) + M'_2(s)$, where

$$M'_1(s, (x, y, \lambda)) = \begin{bmatrix} \partial f(s, x) \\ \partial p(s, z) \\ 0 \end{bmatrix} ,$$

and $M_2'(s)$ is the linear skew-symmetric operator that can be written in a block-wise matrix form in Y as

$$M_2'(s) = \begin{bmatrix} 0 & 0 & A(s)^T \\ 0 & 0 & B(s)^T \\ -A(s) & -B(s) & 0 \end{bmatrix}.$$

For each $s \in \Xi$, both these operators belong to $\mathscr{M}(\mathsf{Y})$, and $\operatorname{dom} M_2'(s) = \mathsf{Y}$. Thus, $M(s) \in \mathscr{M}(\mathsf{Y})$ by [12, Cor. 24.4]. Moreover, since both M_1' and M_2' are measurable, M' is also a random monotone operator over Y.

Now, since the interchange property holds for f, g, p, and q, we see that the operators $\mathsf{M}(x) = \int M(s,x)\mu(ds)$ and $\mathsf{M}' = \int M'(s,x)\mu(ds)$ (where the selection integral (2.4) is used) satisfy

$$\mathsf{M}(x,z,\lambda) = \begin{bmatrix} \partial \mathsf{G}(x) \\ \partial \mathsf{Q}(z) \\ \mathsf{c} \end{bmatrix}, \text{ and } \mathsf{M}'(x,z,\lambda) = \begin{bmatrix} \partial \mathsf{F}(x) + \mathsf{A}^T\lambda \\ \partial \mathsf{P}(z) + \mathsf{B}^T\lambda \\ -\mathsf{A}x - \mathsf{B}z \end{bmatrix}.$$

For the same reasons as for the operators $M(s)$ and $M'(s)$, it holds that M, M′, and $\mathsf{M} + \mathsf{M}'$ belong to $\mathscr{M}(\mathsf{Y})$. Moreover, recalling the system of inclusions (7.4), we also obtain that $\mathcal{Z} = Z(\mathsf{M} + \mathsf{M}')$.

Defining the function

$$b(s,(x,z,\lambda)) = \begin{bmatrix} \widetilde{\nabla} f(s,x) + A(s)^T\lambda \\ \widetilde{\nabla} p(s,z) + B(s)^T\lambda \\ -A(s)x - B(s)z \end{bmatrix}$$

(obviously, $b(s,(x,z,\lambda)) \in M'(s,(x,z,\lambda))$ μ-a.e.), let us consider the following version of the Forward-Backward algorithm

$$(x_{n+1}, z_{n+1}, \lambda_{n+1}) = (I + \gamma_{n+1} M(\xi_{n+1}, \cdot))^{-1} \left((x_n, z_n, \lambda_n) - \gamma_{n+1} b(\xi_{n+1}, (x_n, z_n, \lambda_n))\right).$$

One can easily check that this is exactly Algorithm 1. On the other hand, this algorithm is an instance of the random Forward-Backward algorithm studied in [24] (see Chap. 6). By checking the assumptions of Th. 6.2.1 of Chap. 6 one sees that they are verified under the assumptions of Th. 7.2.1. This completes the proof.

Remark 1. The convergence stated by Th. 7.2.1 concerns the averaged sequence $(\bar{x}_n, \bar{z}_n, \bar{\lambda}_n)$. One can ask whether the sequence (x_n, z_n, λ_n) itself converges to $\mathcal{Z}$. This question is answered positively by Th. 6.2.1 of Chap. 6 in the case where the operator $\mathsf{M} + \mathsf{M}'$ is demipositive. Unfortunately, the demipositivity of $\mathsf{M} + \mathsf{M}'$ is not always guaranteed: take $\mathsf{X} = \mathsf{V} = \mathbb{R}$ and $\mathsf{Z} = \{0\}$, and assume that $\mathsf{F} = \mathsf{G} = 0$, $\mathsf{A} = I$, and $\mathsf{c} = 0$. Then, $\mathsf{M} + \mathsf{M}'$ is a linear operator which can be represented by the 2×2 matrix $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. This operator is not demipositive, being a $-\pi/2$ rotation (see, *e.g.*, [96] or Sec. 2.2.2).

Remark 2. In the deterministic setting, with a constant step size, applying the Forward Backward algorithm to the monotone operators M′ and M doesn't lead to a converging algorithm in general because M′ lack the so-called cocoercivity property [12]. This property is not needed if a decreasing step size is used (see [96, 24] or Sec. 6.2).

7.4 Application to distributed optimization

In this section, Algorithm 1 is illustrated in the context of distributed optimization.

Consider an integer $N > 0$ and a connected graph $G = (V, E)$ where $V = \{1, \cdots, N\}$ is the set of nodes and E is the set of edges. At each node $i \in V$ is located the agent i and two local cost functions $L_i, R_i \in \Gamma_0(\mathbb{R})$ (the space $\mathbb{R}$ can be easily replaced by any Euclidean space) only known by the agent i. The aim of the network of agents is to solve

$$\min_{x \in \mathbb{R}} \sum_{i \in V} L_i(x) + \sum_{i \in V} R_i(x) \tag{7.6}$$

distributively. Moreover, it is assumed that L_i is unknown but revealed through i.i.d realizations of a random variable $\theta(i)$ with distribution $\nu(i)$ over a space Θ, *i.e* $L_i(x) = \mathbb{E}(\ell_i(\theta(i), x))$. A similar assumption is made on R_i : $R_i(x) = \mathbb{E}(r_i(\theta(i), x))$.

Consider an arbitrary orientation of the edges of E and denote by $\hat{E}$ the resulting set of oriented edges. Denote by $\mathsf{A} : \mathbb{R}^V \to \mathbb{R}^{\hat{E}}$ the incidence matrix of the graph G defined for every $x \in \mathbb{R}^V$, and every $e = (e(1), e(2)) \in \hat{E}$ by $\mathsf{A}x(e) = x(e(1)) - x(e(2))$. Problem (7.6) is equivalent to Problem (7.1) with $\mathsf{X} = \mathbb{R}^V$, $\mathsf{P}, \mathsf{Q}, \mathsf{B}, \mathsf{c}$ set equal to zero, $\mathsf{F}(x) = \sum_{i \in V} L_i(x(i))$ and $\mathsf{G}(x) = \sum_{i \in V} R_i(x(i))$.

To apply Algorithm 1 to Problem (7.6), we consider $V \times \Theta^V$ as the probability space Ξ. The random matrix A is defined by : for every $i, j \in V, \theta \in \Theta^V$ the column j of $A(i, \theta)$ is N times the column j of A if $i = j$ and is zero else. If μ is the probability distribution over $V \times \Theta^V$ defined by $\mu = U \otimes \nu(1) \otimes \cdots \otimes \nu(N)$ where U is the uniform distribution over V, it is easy to check that $\int A d\mu = \mathsf{A}$. Moreover, $f : ((i, \theta), x) \mapsto N\ell_i((i, \theta(i)), x(i))$ and $g : ((i, \theta, x) \mapsto Nr_i((i, \theta(i)), x(i))$ are normal convex integrands over $(V \times \Theta^V) \times \mathbb{R}^V$ and it is easy to check that $\int f(\cdot, x) d\mu = \mathsf{F}(x)$ and $\int g(\cdot, x) d\mu = \mathsf{G}(x)$. Thanks to the stochastic handling of the constraints, applying Algorithm 1 leads to a distributed and asynchronous algorithm: at each iteration, only one random chosen agent i becomes active and process its local data $\theta(i)$. Its work is decomposed in two parts : first, the agent i does a computation involving its local variable $x(i)$ and then it sends a message to its neighbors in G, which is not instantaneously processed by the neighbors. The message is sent through the dual variables and is stored by the neighbors, waiting for the next time the neighbors will wake up. Another version of the algorithm in which i sends a message to a random subset of its neighborhood can also be casted in our framework. Moreover, Algorithm 1 leads to an adaptive algorithm since L_i and R_i are revealed across time. Applying the method [131] to the same problem also leads to an algorithm with these properties. The two algorithms are compared in Fig. 7.1 in the context of distributed median (resp. mean) estimation over the Facebook graph (see [74], 4039 nodes and 88234 edges).

In these contexts, each agent $i \in V$ is associated with a real value Y_i (which is unknown but revealed across time through i.i.d realizations in the case of mean computation). The network aims to infer the median (resp. mean) value distributively. In our framework, this corresponds to the case where $\ell_i \equiv 0$, $r_i(\theta, x) = |x - Y_i|$ (resp. $r_i(\theta, x) = |x - Y_i|^2 + x\theta$ where ν is a zero mean probability measure).

Fig. 7.1 shows that both methods are converging, and Algorithm 1 performs slightly better. Indeed, if M is the number of edges, the algorithm [131] uses $N + 2M$ optimization variables whereas our method uses $N + M$ variables. However, both methods are asynchronous and require a bounded amount of memory (see *e.g.* [23]). Finally, our method exhibits less fluctuations. The general framework (7.1) allows to treat the cost function r_i through its proximity operator, leading to a more stable algorithm (see [122, 108]).

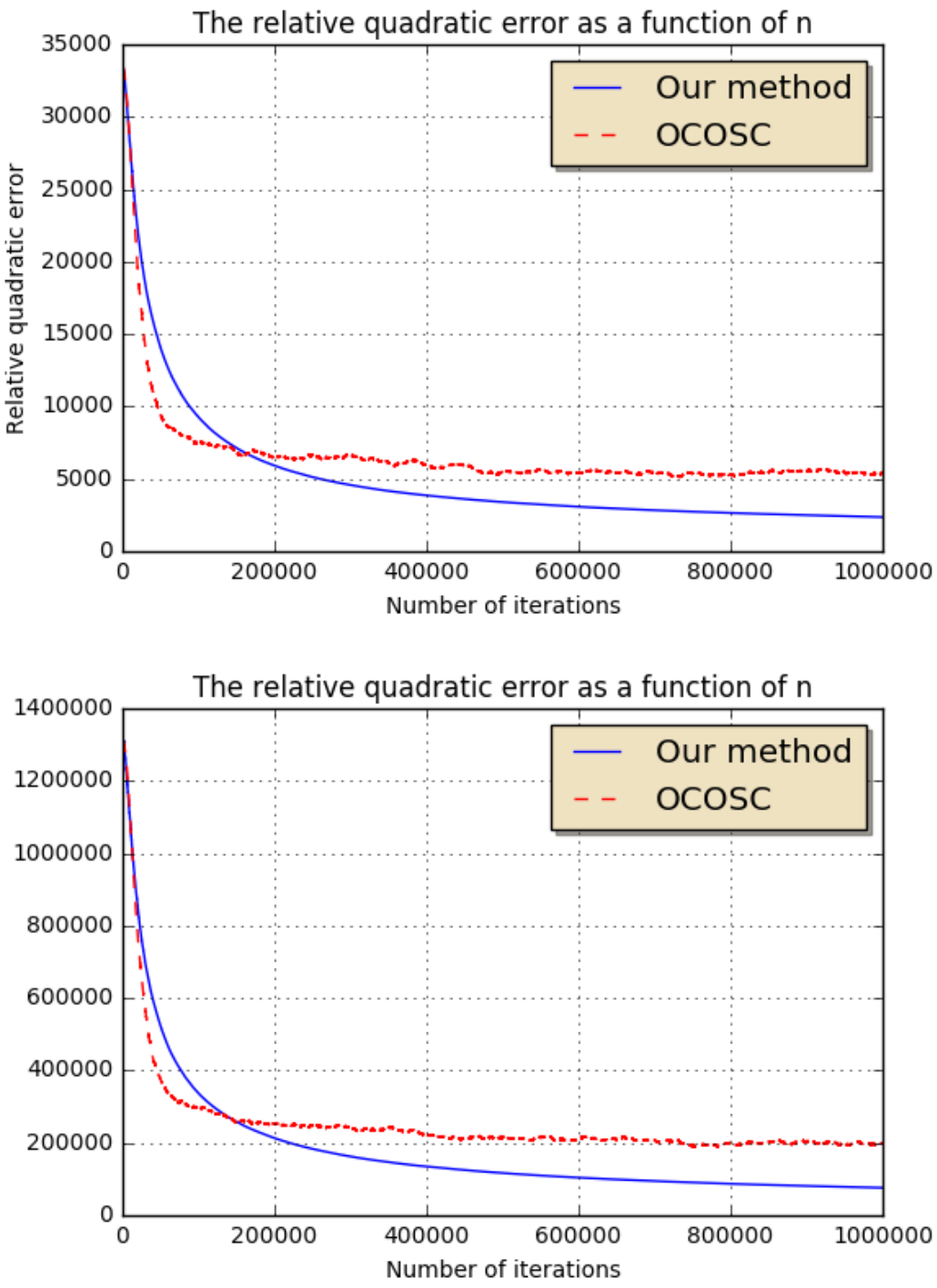

Figure 7.1: First: Distributed median computation. Second: Distributed mean computation. The relative quadratic error at iteration n is defined by $\|x_n - x_\star\|^2/\|x_\star\|^2$ where $x_\star \in \mathbb{R}^V$ such that $x_\star(i) = m$ for every $i \in V$ and m is the solution of Problem (7.6). The algorithm OCOSC is the method introduced in [131].

Chapter 8

Snake: a Stochastic Proximal Gradient Algorithm for Regularized Problems over Large Graphs

In this chapter, we study a regularized optimization problem over a large unstructured graph, where the regularization term is tied to the graph geometry. Typical regularization examples include the Total Variation and the Laplacian regularizations over the graph. When the graph is a simple path without loops, efficient off-the-shelf algorithms can be used. However, when the graph is large and unstructured, such algorithms cannot be used directly. It has been already seen at the end of Chap. 5 that stochastic proximal methods can be successfully applied to regularized problems involving structured regularizations like the overlapping group lasso. This kind of problems are solved by randomizing the proximity operator of the regularization. In this chapter, we propose an algorithm referred to as "Snake" to solve such regularized problems over general graphs. The algorithm consists in properly selecting random simple paths in the graph and performing the proximal gradient algorithm over these simple paths. This algorithm is an instance of a new general stochastic proximal gradient algorithm, whose convergence is proven. Applications to trend filtering and graph inpainting are provided among others and numerical experiments are conducted over large graphs.

8.1 Introduction

Many applications in the fields of machine learning [57, 134], signal and image restoration [41], or trend filtering [120, 127, 93, 69, 75, 119] require the solution of the following optimization problem. On an undirected graph $G = (V, E)$ with no self loops, where $V = \{1, \dots, N\}$ represents a set of N nodes ($N \in \mathbb{N}^*$) and E is the set of edges, find

$$\min_{x \in \mathbb{R}^V} F(x) + R(x, \phi), \tag{8.1}$$

where F is a convex and differentiable function on $\mathbb{R}^V$ representing a data fitting term, and the function $x \mapsto R(x, \phi)$ represents a regularization term of the form

$$R(x, \phi) = \sum_{\{i,j\} \in E} \phi_{\{i,j\}}(x(i), x(j)) \,,$$

where $\phi = (\phi_e)_{e \in E}$ is a family of convex and symmetric $\mathbb{R}^2 \to \mathbb{R}$ functions. The regularization term $R(x, \phi)$ will be called a ϕ-regularization in the sequel. These ϕ-regularizations often promote the sparsity

or the smoothness of the solution. For instance, when $\phi_e(x, x') = w_e|x - x'|$ where $w = (w_e)_{e \in E}$ is a vector of positive weights, the function $R(\cdot, \phi)$ coincides with the weighted Total Variation (TV) norm. This kind of regularization is often used in programming problems over a graph which are intended to recover a piecewise constant signal across adjacent nodes [93, 127, 69, 75, 119, 9, 13, 43]. Another example is the Laplacian regularization $\phi_e(x, x') = (x - x')^2$, or its normalized version obtained by rescaling x and x' by the degrees of each node in e respectively. Laplacian regularization tends to smoothen the solution in accordance with the graph geometry [57, 134].

The proximal gradient algorithm is one of the most popular approaches towards solving Problem (8.1). This algorithm produces the sequence of iterates

$$x_{n+1} = \mathrm{prox}_{\gamma R(\cdot,\phi)}(x_n - \gamma \nabla F(x_n)) \,, \tag{8.2}$$

where $\gamma > 0$ is a fixed step. When $F, G \in \Gamma_0(\mathbb{R}^V)$ and F is smooth, the sequence (x_n) converges to a minimizer of (8.1), assuming this minimizer exists and that γ is enough small.

Implementing the proximal gradient algorithm requires the computation of the proximity operator applied to $R(\cdot, \phi)$ at each iteration. When N is large, this computation is in general affordable only when the graph exhibits a simple structure. For instance, when $R(\cdot, \phi)$ is the TV norm, the so-called *taut-string* algorithm is an efficient algorithm for computing the proximity operator when the graph is one-dimensional (1D) [50] (see Fig. 8.1) or when it is a two-dimensional (2D) regular grid [9]. Similar observations can be made for the Laplacian regularization [45], where, *e.g.*, the discrete cosine transform can be implemented. When the graph is large and unstructured, these algorithms cannot be used, and the computation of the proximity operator is more difficult ([127, 117]).

This problem is addressed in this chapter. Towards obtaining a simple algorithm, we first express the functions $F(\cdot)$ and $R(\cdot, \phi)$ as the expectations of functions defined on the set of random walks in the graph, paving the way for a *randomized* version of the proximal gradient algorithm. Stochastic online algorithms in the spirit of this algorithm are often considered as simple and reliable procedures for solving high dimensional machine learning problems, including in the situations where the randomness is not inherent to these problems [33, 34]. One specificity of the algorithm developed here lies in that it reconciles two requirements: on the one hand, the random versions of $R(\cdot, \phi)$ should be defined on *simple paths*, *i.e.*, on walks without loops (see Fig. 8.1), in a way to make benefit of the power of the existing fast algorithms for computing the proximity operator. Owing to the existence of a procedure for selecting these simple paths, we term our algorithm as the "Snake" algorithm. On the other hand, the expectations of the functions handled by the optimization algorithm coincide with $F(\cdot)$ and $R(\cdot, \phi)$ respectively (up to a multiplicative constant), in such a way that the algorithm does not introduce any bias on the estimates.

There often exists efficient methods to compute the proximity operator of ϕ-regularization over 1D-graphs. The algorithm Snake randomly selects simple paths in a general graph in order to apply the latter 1D efficient methods over a general graph.

Actually, the algorithm Snake will be an instance of a new general stochastic approximation algorithm that we develop in this chapter. In some aspects, this general stochastic approximation algorithm is itself a generalization of the random Forward-Backward algorithm studied in [24].

Before presenting our approach, we provide an overview of the literature dealing with our problem. First consider the case where $R(\cdot, \phi)$ coincides with the TV norm. As said above, fast methods exist when the graph has a simple structure. We refer the reader to [9] for an overview of iterative solvers of Problem (8.1) in these cases. In [71], the author introduces a dynamical programming method to compute the proximity operator on a 1D-graph with a complexity of order $\mathcal{O}(N)$. Still in the 1D case, Condat [50] revisited recently an algorithm that is due to Mammen and Van De Geer [81] referred to

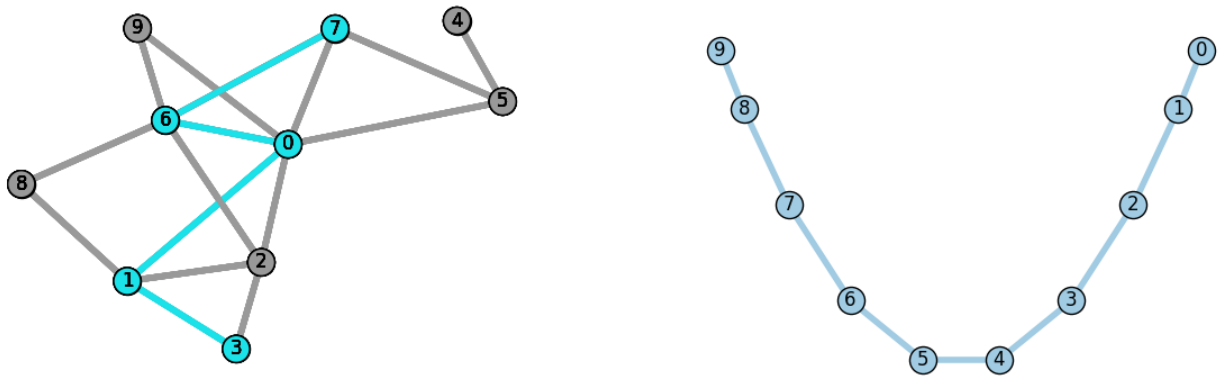

Figure 8.1: Left: General graph on which is colored the simple path 3-1-0-6-7. Right: 1D-graph.

as the taut-string algorithm. The complexity of this algorithm is $O(N^2)$ in the worst-case scenario, and $O(N)$ in the most realistic cases. The taut-string algorithm is linked to a total variation regularized problem in [52]. This algorithm is generalized to 2D-grids, weighted TV norms and ℓ^p TV norms by Barbero and Sra in [9]. To generalize to 2D-grids, the TV regularization can be written as a sum of two terms on which one can apply 1D methods, according to [46] and [70]. Over general graphs, there is no immediate way to generalize the taut string method. The problem of computing the TV-proximity operator over a general graph is addressed in [127].

The authors of [127] suggest to solve the problem using a projected Newton algorithm applied to the dual problem. It is observed that, empirically, this methods performs better than other concurrent approaches. As a matter of fact, this statement holds when the graph has a moderate size. As far as large graphs are concerned, the iteration complexity of the projected Newton method can be a bottleneck. To address this problem, the authors of [13] and [63] propose to solve the problem distributively over the nodes using the Alternating Direction Method of Multipliers (ADMM).

In [119] the authors propose to compute a decomposition of the graph in 1D-graphs and then solve Problem (8.1) by means of the TV-proximity operators over these 1D-graphs. Although the decomposition of the graph is fast in many applications, the algorithm [119] relies on an offline decomposition of the whole graph that needs a global knowledge of the graph topology. The Snake algorithm obtains this decomposition online. In [75], the authors propose a working set strategy to compute the TV-proximity operator. At each iteration, the graph is cut in two well-chosen subgraphs and a reduced problem of (8.1) is deduced from this cut. The reduced problem is then solved efficiently. This method has shown speed-ups when G is an image (*i.e* a two dimensional grid). Although the decomposition of the graph is not done during the preprocessing time, the algorithm [75] still needs a global knowledge of the graph topology during the iterations. On the contrary, the Snake algorithm only needs a local knowledge. Finally, in [93], the authors propose to replace the computation of the TV-proximity operator over the graph G by the computation of the TV-proximity operator over a well chosen 1D-subgraph of G. This produces an approximation of the solution whereas the Snake algorithm is proven to converge to the exact solution.

In the case where $R(\cdot, \phi)$ is the Laplacian regularization, the computation of the proximity operator of R reduces to the resolution of a linear system $(\mathcal{L} + \alpha I)x = b$ where $\mathcal{L}$ is the Laplacian matrix of

the graph G and I the identity matrix. On an 1D-graph, the latter resolution can be done efficiently and relies on an explicit diagonalization of $\mathcal{L}$ ([45]) by means of the discrete cosine transform, which take $\mathcal{O}(N\log(N))$ operations. Over general graphs, the problem of computing the proximity operator of the Laplacian regularization is introduced in [134]. There exist fast algorithms to solve it due to [118]. They are based on recursively preconditioning the conjugate gradient method using graph theoretical results [117]. Nevertheless, the preconditioning phase which may be demanding over very large graphs. Compared to [117], our online method Snake requires no preprocessing step.

8.2 Outline of the approach and chapter organization

The starting point of our approach is a new stochastic optimization algorithm that has its own interest. This algorithm will be presented succinctly here, and more rigorously in Sec. 8.3 below. Given an integer $L > 0$, let $\xi = (\xi^1, \dots, \xi^L)$ be a random vector where the ξ^i are valued in some measurable space. Consider the problem

$$\min_x \sum_{i=1}^{L} \mathbb{E}_\xi \left[f_i(\xi^i, x) + g_i(\xi^i, x) \right] \tag{8.3}$$

where f_i, g_i are normal convex integrands, and $f_i(\xi^i, \cdot)$ are assumed to be differentiable. Given $\gamma > 0$, define the operator $\mathsf{T}_{\gamma,i}(s,x) = \mathrm{prox}_{\gamma g_i(s,\cdot)}(x - \gamma \nabla f_i(s,x))$. Given a sequence (ξ_n) of independent copies of ξ, and a sequence of positive steps $(\gamma_n) \in \ell^2 \setminus \ell^1$, we consider the algorithm

$$x_{n+1} = \mathsf{T}_{\gamma_{n+1}}(\xi_{n+1}, x_n)\,, \tag{8.4}$$

where

$$\mathsf{T}_\gamma((s^1, \dots, s^L), \cdot) = \mathsf{T}_{\gamma,L}(s^L, \cdot) \circ \cdots \circ \mathsf{T}_{\gamma,1}(s^1, \cdot)$$

and where $\circ$ stands for the composition of functions: $f \circ g(x) = f(g(x))$. In other words, an iteration of this algorithm consists in the composition of L random proximal gradient iterations. The case where $L = 1$ was treated in [24] (see also Chap. 6).

Assuming that the set of minimizers of the problem is non empty, Th. 8.3.1 below states that the sequence (x_n) converges almost surely to a (possibly random) point of this set. A sketch of the proof of this theorem can be found in Sec. 8.3.3. It follows the same canvas as the approach of [24] or Chap. 6, with the difference that we are now dealing with possibly different functions (f_i, g_i) and non-independent noises ξ^i for $i \in \{1, \dots, L\}$.

We now want to exploit this stochastic algorithm to develop a simple procedure leading to a solution of Problem (8.1). This will be done in Sec. 8.4 and will lead to the Snake algorithm. The first step is to express the function $R(\cdot, \phi)$ as the expectation of a function with respect to a finite random walk. Given an integer $M > 0$ and a finite walk $s = (v_0, v_1, \dots, v_M)$ of length M on the graph G, where $v_i \in V$ and $\{v_i, v_{i+1}\} \in E$, write

$$R(x, \phi_s) = \sum_{i=1}^{M} \phi_{\{v_{i-1}, v_i\}}(x(v_{i-1}), x(v_i))\,.$$

Now, pick a node at random with a probability proportional to the degree (*i.e.*, the number of neighbors) of this node. Once this node has been chosen, pick another one at random uniformly among the neighbors of the first node. Repeat the process of choosing neighbors M times, and denote as $\xi \in V^{M+1}$ the random walk thus obtained. With this construction, we get that $\frac{1}{|E|} R(x, \phi) = \frac{1}{M} \mathbb{E}_\xi[R(x, \phi_\xi)]$ using some elementary Markov chain formalism (see Prop. 8.4.1 below).

In these conditions, a first attempt of the use of Algorithm (8.4) is to consider Problem (8.1) as an instance of Problem (8.3) with $L = 1$, $f_1(\xi, x) = \frac{1}{|E|}F(x)$, and $g_1(\xi, x) = \frac{1}{M}R(x, \phi_\xi)$. Given an independent sequence (ξ_n) of walks having the same distribution as the random vector ξ and a sequence (γ_n) of steps in $\ell^2 \setminus \ell^1$, Algorithm 8.4 boils down to the stochastic version of the proximal gradient algorithm

$$x_{n+1} = \operatorname{prox}_{\gamma_{n+1}\frac{1}{M}R(\cdot,\phi_{\xi_{n+1}})}(x_n - \gamma_{n+1}\frac{1}{|E|}\nabla F(x_n))\,. \tag{8.5}$$

By Th. 8.3.1 (or by [24]), the iterates x_n converge almost surely to a solution of Problem (8.1).

However, although simpler than the deterministic algorithm (8.2), this algorithm is still difficult to implement for many regularization functions. As said in the introduction, the walk ξ is often required to be a simple path. Obviously, the walk generation mechanism described above does not prevent ξ from having repeated nodes. A first way to circumvent this problem would be to generate ξ as a loop-erased walk on the graph. Unfortunately, the evaluation of the corresponding distribution is notoriously difficult. The generalization of Prop. 8.4.1 to loop-erased walks is far from being immediate.

As an alternative, we identify the walk ξ with the concatenation of at most M simple paths of maximal length that we denote as $\xi^1, \ldots, \xi^M$, these random variables being valued in the space of all walks in G of length at most M:

$$\xi = (\xi^1, \xi^2, \ldots, \xi^M)\,.$$

Here, in the most frequent case where the number of simple paths is strictly less than M, the last ξ^i's are conventionally set to a trivial walk, *i.e.*, a walk with one node and no edge. We also denote as $\ell(\xi^i)$ the length of the simple path ξ^i, *i.e.*, the number of edges in ξ^i. We now choose $L = M$, and for $i = 1, \ldots, L$, we set $f_i(\xi^i, x) = \frac{\ell(\xi^i)}{L|E|}F(x)$ and $g_i(\xi^i, x) = \frac{1}{L}R(x, \phi_{\xi^i})$ if $\ell(\xi^i) > 0$, and $f_i(\xi^i, x) = g_i(\xi^i, x) = 0$ otherwise. With this construction, we show in Sec. 8.4 that $\frac{1}{|E|}(F(x) + R(x, \phi)) = \sum_{i=1}^{L} \mathbb{E}_\xi[f_i(\xi^i, x) + g_i(\xi^i, x)]$ and that the functions f_i and g_i fulfill the general assumptions required for the Algorithm (8.4) to converge to a solution of Problem (8.1). In summary, at each iteration, we pick up a random walk of length L according to the procedure described above, split it into simple paths of maximal length, and then we successively apply the proximal gradient algorithm to these simple paths.

Chapter organization. The next section introduces the generalized stochastic proximal gradient algorithm. Then, in Sec. 8.4, this algorithm is applied to ϕ-regularized problems to obtain the Snake algorithm. The Snake algorithm relies on the computation of the proximity operator of the ϕ-regularization over 1D-graphs. Section 8.5 gives examples of ϕ-regularizations for which the latter computation can be done efficiently (TV regularization and Laplacian regularization). Finally, we simulate the Snake algorithm in several application contexts. First, we study the so called graph trend filtering [127]. Then, we consider the graph inpainting problem [43, 57, 134] and the resolution of Laplacian systems [117]. These contexts are the purpose of Sec. 8.6. Finally, a conclusion and some future research directions are provided in Sec. 8.7.

8.3 A General Stochastic Proximal Gradient Algorithm

8.3.1 Problem and General Algorithm

In this section, we consider the general problem

$$\min_{x\in\mathsf{X}} \sum_{i=1}^{L} \mathbb{E}\left[f_i(\xi^i, x) + g_i(\xi^i, x)\right] \tag{8.6}$$

where L is a positive integer, the $\xi^i : \Omega \to \Xi$ are random variables (r.v.), and the functions $f_i : \Xi\times\mathsf{X} \to \mathbb{R}$ and $g_i : \Xi \times \mathsf{X} \to \mathbb{R}$ satisfy the following assumption:

Assumption 8.3.1. The following holds for all $i \in \{1, \dots, L\}$:

1. The f_i and g_i are normal convex integrands.
2. For every $x \in \mathsf{X}$, $\mathbb{E}[|f_i(\xi^i, x)|] < \infty$ and $\mathbb{E}[|g_i(\xi^i, x)|] < \infty$.
3. The function $f_i(\xi^i, \cdot)$ is a.s. differentiable. We denote as $\nabla f_i(\xi^i, \cdot)$ its gradient w.r.t. the first variable.

Remark 3. In this chapter, we assume that the functions $g_i(\xi^i, \cdot)$ have a.s. a full domain. This assumption can be relaxed with some effort, along the ideas developed in [24].

Let ξ be the random vector $\xi = (\xi^1, \dots, \xi^L)$ with values in Ξ^L and let $(\xi_n : n \in \mathbb{N}^*)$ be a sequence of i.i.d. copies of ξ, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For all $n \in \mathbb{N}^*$, $\xi_n = (\xi_n^1, \dots, \xi_n^L)$. Finally, let (γ_n) be a positive sequence. Our aim is to analyze the convergence of the iterates (x_n) recursively defined by:

$$x_{n+1} = \mathsf{T}_{\gamma_{n+1}}(\xi_{n+1}, x_n)\,, \tag{8.7}$$

as well as the intermediate variables $\bar{x}_{n+1}^i$ $(i = 0, \dots, L)$ defined by $\bar{x}_{n+1}^0 = x_n$, and

$$\bar{x}_{n+1}^i = \mathsf{T}_{\gamma_{n+1},i}(\xi_{n+1}^i, \bar{x}_{n+1}^{i-1})\,, \quad i = 1, \dots, L\,. \tag{8.8}$$

In particular, $x_{n+1} = x_{n+1}^L = T_{\gamma_{n+1},L}(\xi_{n+1}^L, \bar{x}_{n+1}^{L-1})$.

In the special case where the functions g_i, are all constant with respect to their first variable (the algorithm is deterministic) and the functions f_i are equal to zero, the above iterations were studied by Passty in [94]. In the special case where $L = 1$, the algorithm boils down to the stochastic proximal gradient algorithm, whose detailed convergence analysis can be found in [24] (see also Chap. 6, [22], and [126] as an earlier work). In this case, the iterates take the simpler form

$$x_{n+1} = \mathrm{prox}_{\gamma_{n+1}g_1(\xi_{n+1},\cdot)}(x_n - \gamma_{n+1}\nabla f_1(\xi_{n+1}, x_n))\,, \tag{8.9}$$

and converge a.s. to a minimizer of the function $x \mapsto \mathbb{E}_\xi[f_1(x, \xi) + g_1(x,\xi)]$ under the convenient hypotheses.

It is worth noting that the present algorithm (8.7) cannot be written as an instance of (8.9). Indeed, the operator T_γ is a composition of L (random) operators, whereas the stochastic forward backward algorithm (8.9) has a simpler structure. This composition raises technical difficulties that need to be specifically addressed. Among these difficulties is the dependency of the intermediate variables.

8.3.2 Almost sure convergence

We make the following assumptions.

Assumption 8.3.2. The positive sequence (γ_n) satisfies the conditions

$$\sum \gamma_n = +\infty \quad \text{and} \quad \sum \gamma_n^2 < \infty\,,$$

(*i.e.*, $(\gamma_n) \in \ell^2 \setminus \ell^1$). Moreover, $\frac{\gamma_{n+1}}{\gamma_n} \to 1$.

Assumption 8.3.3. The following holds for all $i \in \{1, \dots, L\}$:

1. There exists a measurable map $K_i : \Xi \to \mathbb{R}_+$ s.t. the following holds $\mathbb{P}$-a.e.: for all x, y in X,
$$\|\nabla f_i(\xi^i, x) - \nabla f_i(\xi^i, y)\| \le K_i(\xi^i)\|x - y\|\,.$$
2. For all $\alpha > 0$, $\mathbb{E}[K_i(\xi^i)^\alpha] < \infty$.

We denote by $\mathcal{Z}$ the set of minimizers of Problem (8.6). Thanks to Assumption 8.3.1, the qualification conditions hold, ensuring that a point $x_\star$ belongs to $\mathcal{Z}$ if and only if

$$0 \in \sum_{i=1}^{L} \nabla \mathbb{E}[f_i(\xi^i, x_\star)] + \partial \mathbb{E}[g_i(\xi^i, x_\star)]\,.$$

The (sub)differential and the expectation operators can be interchanged [102], and the above optimality condition also reads

$$0 \in \sum_{i=1}^{L} \mathbb{E}[\nabla f_i(\xi^i, x_\star)] + \mathbb{E}[\partial g_i(\xi^i, x_\star)]\,, \tag{8.10}$$

where $\mathbb{E}[\partial g_i(\xi^i, x_\star)]$ is the selection integral of the random set $\partial g_i(x_\star, \xi^i)$ (see (2.4)). Therefore, the optimality condition (8.10) means that there exist L integrable mappings $\varphi_1, \dots, \varphi_L$ satisfying a.s. $\varphi_i(\xi^i) \in \partial g_i(\xi^i, x_\star)$ and s.t.

$$0 = \sum_{i=1}^{L} \mathbb{E}[\nabla f_i(\xi^i, x_\star)] + \mathbb{E}[\varphi_i(\xi^i)]\,. \tag{8.11}$$

Recalling (6.2), we say that the family $(\nabla f_i(\xi^i, x_\star), \varphi_i(\xi^i))_{i=1,\dots,L}$ is a *representation* of the minimizer $x_\star$. In addition, if for some $\alpha \ge 1$ and every $i = 1, \dots, L$, $\mathbb{E}[\|\nabla f_i(\xi^i, x_\star)\|^\alpha] < \infty$ and $\mathbb{E}[\|\varphi(\xi^i)\|^\alpha] < \infty$, we say that the minimizer $x_\star$ admits a α-integrable representation.

Assumption 8.3.4.
1. The set $\mathcal{Z}$ is not empty.
2. For every $x_\star \in \mathcal{Z}$, there exists $\varepsilon > 0$ s.t. $x_\star$ admits a $(2+\varepsilon)$-integrable representation, which is denoted $(\nabla f_i(\xi^i, x_\star), \varphi_i(\xi^i))_{i=1,\dots,L}$.

We denote by $\partial_0 g_i(\xi^i, x)$ the least norm element in $\partial g_i(\xi^i, x)$.

Assumption 8.3.5. For every compact set $\mathcal{K} \subset$ X, there exists $\eta > 0$ such that for all $i = 1, \dots, L$,

$$\sup_{x \in \mathcal{K}} \mathbb{E}[\|\partial_0 g_i(x, \xi^i)\|^{1+\eta}] < \infty\,.$$

We can now state the main result of this section, which will be proven in Sec. 8.3.3.

Theorem 8.3.1. Let Assumptions 8.3.1–8.3.5 hold true. There exists a r.v. $X_\star$ s.t. $\mathbb{P}(X_\star \in \mathcal{Z}) = 1$ and s.t. (x_n) converges a.s. to $X_\star$ as $n \to \infty$. Moreover, for every $i = 0, \dots, L-1$, $\bar{x}_n^i$ converges a.s. to $X_\star$.

8.3.3 Sketch of the Proof of Th. 8.3.1

We start with some notations. We endow the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with the filtration $(\mathscr{F}_n)$ defined as $\mathscr{F}_n = \sigma(\xi_1, \dots, \xi_n)$, and we write $\mathbb{E}_n = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$. In particular, $\mathbb{E}_0 = \mathbb{E}$. We also define $G_i(x) = \mathbb{E}_\xi[g_i(\xi^i, x)]$ and $F_i(x) = \mathbb{E}_\xi[f_i(\xi^i, x)]$ for every $x \in \mathsf{X}$. We denote by μ_i and μ the probability distributions of ξ^i and ξ respectively. Finally, C and η will refer to positive constants whose values can change from an equation to another. The constant η can be chosen arbitrarily small.

In [24] or Chap. 6, the case $L = 1$ is studied (Algorithm (8.9)). Here we shall reproduce the main steps of the approach of [24] or Chap. 6, only treating in detail the specificities of the case $L > 1$. Note also that the formalism of random monotone operators is not needed here.

Given $a \in \mathsf{X}$, consider the Differential Inclusion (DI) associated with $\sum_{i=1}^L \nabla F_i + \partial G_i$:

$$\begin{cases} \dot{\mathsf{z}}(t) & \in & -\sum_{i=1}^L (\nabla F_i(\mathsf{z}(t)) + \partial G_i(\mathsf{z}(t))) \\ \mathsf{z}(0) & = & a\,. \end{cases} \tag{8.12}$$

and denote Φ the associated semiflow (see 2.2.2). Consider the interpolated function $\mathsf{x} = \mathsf{I}((x_n))$ (see (6.4)) where (x_n) is given by (8.7). We shall prove the two following facts:

- The sequence $(\|x_n - x_\star\|)$ is almost surely convergent for each $x_\star \in \mathcal{Z}$ (Prop. 8.3.3);
- The process $\mathsf{x}(t)$ is an almost sure Asymptotic Pseudo Trajectory (APT) of the semi-flow Φ, see (6.5). Namely, for each $T > 0$,

$$\sup_{u \in [0,T]} \|\mathsf{x}(t+u) - \Phi(\mathsf{x}(t), u)\| \xrightarrow[t\to\infty]{\text{a.s.}} 0, \tag{8.13}$$

Taken together, these two results lead to the a.s. convergence of (x_n) to some r.v. $X^\star$ supported by the set $\mathcal{Z}$, as is shown by [24, Cor. 3.2]. The convergence of the $(\bar{x}_n^i)_n$ stated by Th. 8.3.1 will be shown in the course of the proof.

Denoting g^γ the Moreau envelope of the convex function g, the mapping $\mathsf{T}_{\gamma,i}$ can be rewritten as

$$\mathsf{T}_{\gamma,i}(s,x) = x - \gamma \nabla f_i(s,x) - \gamma \nabla g_i^\gamma(s, x - \gamma \nabla f_i(s,x)) \tag{8.14}$$

The following lemma is proven in Appendix 8.8.1.

Lemma 8.3.2. For $i = 1, \dots, L$, let

$$\bar{x}^i = (T_{\gamma,i}(s^i, \cdot) \circ \cdots \circ T_{\gamma,1}(s^1, \cdot))(x).$$

Then, with Assumption 8.3.3, there exists a measurable map $\kappa : \Xi^L \to \mathbb{R}_+$ s.t. $\mathbb{E}_\xi[\kappa(\xi)^\alpha] < \infty$ for all $\alpha \geq 1$ and s.t. for all $\bar{s} = (s^1, \dots, s^L) \in \Xi^L$,

$$\begin{aligned} &\|\nabla f_i(s^i, \bar{x}^{i-1})\| \leq \kappa(\bar{s}) \sum_{k=1}^i \|\nabla f_k(s^k, x)\| + \|\nabla g_k^\gamma(s^k, x)\| \\ &\|\nabla g_i^\gamma(s^i, \bar{x}^{i-1} - \gamma \nabla f_i(s^i, \bar{x}^{i-1}))\| \\ &\qquad \leq \kappa(\bar{s}) \sum_{k=1}^i \|\nabla f_k(s^k, x)\| + \|\nabla g_k^\gamma(s^k, x)\|. \end{aligned}$$

Recall that we are studying the iterations $\bar{x}^i_{n+1} = \mathsf{T}_{\gamma_{n+1},i}(\xi^i_{n+1}, \bar{x}^{i-1}_{n+1})$, for $i = 1, \dots, L$, $n \in \mathbb{N}^*$, with $\bar{x}^0_{n+1} = x_n$ and $x_{n+1} = \bar{x}^L_{n+1}$. In this section and in Appendix 8.8, we shall write for conciseness, for any $x_\star \in \mathcal{Z}$,

$$\begin{aligned}
\nabla g_i^\gamma &= \nabla g_i^{\gamma_{n+1}}(\xi^i_{n+1}, \bar{x}^{i-1}_{n+1} - \gamma_{n+1}\nabla f_i(\xi^i_{n+1}, \bar{x}^{i-1}_{n+1})),\\
\mathrm{prox}_{\gamma g_i} &= \mathrm{prox}_{\gamma g_i(\xi^i_{n+1},\cdot)}(\bar{x}^{i-1}_{n+1} - \gamma_{n+1}\nabla f_i(\xi^i_{n+1}, \bar{x}^{i-1}_{n+1})),\\
\nabla f_i &= \nabla f_i(\xi^i_{n+1}, \bar{x}^{i-1}_{n+1}),\\
\nabla f_i^\star &= \nabla f_i(\xi^i_{n+1}, x_\star) \text{ where } x_\star \in \mathcal{Z},\\
\varphi_i &= \varphi_i(\xi^i_{n+1}), \text{ (see Assumption 8.3.4) and}\\
\gamma &= \gamma_{n+1}.
\end{aligned}$$

The following proposition is analogous to Prop. 1 of Chap. 6:

Proposition 8.3.3. Let Assumptions 8.3.2–8.3.4 hold true. Then the following facts hold true:

1. For each $x_\star \in \mathcal{Z}$, the sequence $(\|x_n - x_\star\|)$ converges almost surely.
2. $\mathbb{E}\left[\sum_{i=1}^L \sum_{n=1}^\infty \gamma^2(\|\nabla g_i^\gamma\|^2 + \|\nabla f_i\|^2)\right] < \infty$.
3. For each i, $\bar{x}^i_{n+1} - x_n \to 0$ almost surely.

This proposition is shown in Appendix 8.8.2.

The proof of the APT property follows the same lines as in Sec. 6.3, using the following definition of the function h:

$$h_\gamma((s^1, \dots, s^L), x) = -\sum_{i=1}^L \left[\nabla f_i(s^i, \bar{x}^{i-1}) + \nabla g_i^\gamma(s^i, \bar{x}^{i-1} - \gamma \nabla f_i(s^i, \bar{x}^{i-1}))\right],$$

where we recall the notation $\bar{x}^i = (T_{\gamma,i}(s^i, \cdot) \circ \cdots \circ T_{\gamma,1}(s^1, \cdot))(x)$. Note that

$$x_{n+1} = x_n + \gamma_{n+1} h_{\gamma_{n+1}}(\xi_{n+1}, x_n).$$

Also note that since every functions f_i and g_i are assumed to take finite values, it is sufficient to prove the APT property in the so-called full domain case, see Sec. 6.3.

8.4 The Snake Algorithm

8.4.1 Notations

Let $\ell \geq 1$ be an integer. We refer to a walk of length ℓ over the graph G as a sequence $s = (v_0, v_1, \dots, v_\ell)$ in $V^{\ell+1}$ such that for every $i = 1, \dots, \ell$, the pair $\{v_{i-1}, v_i\}$ is an edge of the graph. A walk of length zero is a single vertex.

We shall often identify s with the graph $\mathcal{G}(s)$ whose vertices and edges are respectively given by the sets $\mathcal{V}(s) = \{v_0, \dots, v_\ell\}$ and $\mathcal{E}(s) = \{\{v_0, v_1\}, \dots, \{v_{\ell-1}, v_\ell\}\}$.

Let $L \geq 1$. We denote by Ξ the set of all walks over G with length $\leq L$. This is a finite set. Let $\mathscr{G}$ be the set of all subsets of Ξ. We consider the measurable space $(\Xi, \mathscr{G})$.

Let $s = (v_0, v_1, \dots, v_\ell) \in \Xi$ with $0 < \ell \leq L$. We abusively denote by ϕ_s the family of functions $(\phi_{\{v_{i-1},v_i\}})_{i=1,\dots,\ell}$. We refer to the ϕ_s-regularization of x as the ϕ_s-regularization on the graph s of the restriction of x to s that is

$$R(x, \phi_s) = \sum_{i=1}^{\ell} \phi_{\{v_{i-1},v_i\}}(x(v_{i-1}), x(v_i)) \,.$$

Besides, $R(x, \phi_s)$ is defined to be 0 if s is a single vertex (that is $\ell = 0$).

We say that a walk is a *simple path* if there is no repeated node *i.e*, all elements in s are different or if s is a single vertex. Throughout this chapter, we assume that when s is a simple path, the computation of $\mathrm{prox}_{R(\cdot,\phi_s)}$ can be done easily.

8.4.2 Writing the Regularization Function as an Expectation

One key idea of this chapter is to write the function $R(x, \phi)$ as an expectation in order to use a stochastic approximation algorithm, as described in Sec. 8.3.

Denote by $\deg(v)$ the degree of the node $v \in V$, *i.e.*, the number of neighbors of v in G. Let π be the probability measure on V defined as

$$\pi(v) = \frac{\deg(v)}{2|E|}, \quad v \in V \,.$$

Define the probability transition kernel P on V^2 as $P(v, w) = \mathbb{1}_{\{v,w\}\in E}/\deg(v)$ if $\deg(v) > 0$, and $P(v, w) = \mathbb{1}_{v=w}$ otherwise, where $\mathbb{1}$ is the indicator function.

We refer to a Markov chain (indexed by $\mathbb{N}$) over V with initial distribution π and transition kernel P as an infinite random walk over G. Let $(v_k)_{k\in\mathbb{N}}$ be an infinite random walk over G defined on the canonical probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\Omega = V^{\mathbb{N}}$. The first node v_0 of this walk is randomly chosen in V according to the distribution π. The other nodes are drawn recursively according to the conditional probability $\mathbb{P}(v_{k+1} = w \,|\, v_k) = P(v_k, w)$. In other words, conditionally to v_k, the node v_{k+1} is drawn uniformly from the neighborhood of v_k. Setting an integer $L \geq 1$, we define the random variable ξ from $(v_k)_{k\in\mathbb{N}}$ as $\xi = (v_0, v_1, \dots, v_L)$.

Proposition 8.4.1. For every $x \in \mathbb{R}^V$,

$$\frac{1}{|E|} R(x, \phi) = \frac{1}{L} \mathbb{E}_\xi[R(x, \phi_\xi)] \,. \tag{8.15}$$

Proof. It is straightforward to show that π is an invariant measure of the Markov chain $(v_k)_{k\in\mathbb{N}}$. Moreover, $\mathbb{P}(v_k = w, v_{k-1} = v) = \pi(v)P(v, w) = \mathbb{1}_{\{v,w\}\in E}/(2|E|)$, leading to the identity

$$\mathbb{E}\left[\phi_{\{v_{k-1},v_k\}}(x(v_{k-1}), x(v_k))\right] = \frac{1}{|E|} R(x, \phi) \,,$$

which completes the proof by summing and using the symmetry of $\phi_e, \forall e \in E$. □

This proposition shows that Problem (8.1) is written equivalently

$$\min_{x\in\mathbb{R}^V} \frac{1}{|E|} F(x) + \mathbb{E}[\frac{1}{L} R(x, \phi_\xi)]. \tag{8.16}$$

Hence, applying the stochastic proximal gradient algorithm to solve (8.16) leads to a new algorithm to solve (8.1), which was mentioned in Sec. 8.2, Eq. (8.5):

$$x_{n+1} = \mathrm{prox}_{\gamma_{n+1}\frac{1}{L}R(\cdot,\phi_{\xi_{n+1}})}(x_n - \gamma_{n+1}\frac{1}{|E|}\nabla F(x_n))\,. \tag{8.17}$$

Although the iteration complexity is reduced in (8.17) compared to (8.2), the computation of the proximity operator of the ϕ-regularization over the random subgraph ξ_{n+1} in the algorithm (8.17) can be difficult to implement. This is due to the possible presence of loops in the random walk ξ. As an alternative, we split ξ into several simple paths. We will then replace the proximity operator over ξ by the series of the proximity operators over the simple paths induced by ξ, which are efficiently computable.

8.4.3 Splitting ξ into Simple Paths

Let $(v_k)_{k\in\mathbb{N}}$ be an infinite random walk on $(\Omega, \mathcal{F}, \mathbb{P})$. We recursively define a sequence of stopping time $(\tau_i)_{i\in\mathbb{N}}$ as $\tau_0 = 1$ and for all $i \geq 0$,

$$\tau_{i+1} = \min\{k \geq \tau_i : v_k \in \{v_{\tau_i - 1}, \ldots, v_{k-1}\}\}$$

if the above set is nonempty, and $\tau_{i+1} = +\infty$ otherwise. We now define the stopping times t_i for all $i \in \mathbb{N}$ as $t_i = \min(\tau_i, L+1)$. Finally, for all $i \in \mathbb{N}^*$ we can consider the random variable ξ^i on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\Xi, \mathscr{G})$ defined by

$$\xi^i = (v_{t_{i-1}-1}, v_{t_{i-1}}, \ldots, v_{t_i - 1}).$$

We denote by N the smallest integer n such that $t_n = L+1$. We denote by $\ell(\xi^i)$ the length of the simple path ξ^i.

Example 6. Given a graph with vertices $V = \{a, b, c, \ldots, z\}$ and a given edge set that is not useful to describe here, consider $\omega \in \Omega$ and the walk $\xi(\omega) = (c, a, e, g, a, f, a, b, h)$ with length $L = 8$. Then, $t_0(\omega) = 1$, $t_1(\omega) = 4$, $t_2(\omega) = 6$, $t_3(\omega) = t_4(\omega) = \ldots = 9$, and $\xi(\omega)$ can be decomposed into $N(\omega) = 3$ simple paths and we have $\xi^1(\omega) = (c, a, e, g)$, $\xi^2(\omega) = (g, a, f)$, $\xi^3(\omega) = (f, a, b, h)$ and $\xi^4(\omega) = \ldots = \xi^8(\omega) = (h)$. Their respective lengths are $\ell(\xi^1(\omega)) = 3$, $\ell(\xi^2(\omega)) = 2$, $\ell(\xi^3(\omega)) = 3$ and $\ell(\xi^i(\omega)) = 0$ for all $i = 4, \ldots, 8$. We identify $\xi(\omega)$ with $(\xi^1(\omega), \ldots, \xi^8(\omega))$.

It is worth noting that, by construction, ξ^i is a simple path. Moreover, the following statements hold:

- We have $1 \leq N \leq L$ a.s.
- These three events are equivalent for all i: $\{\xi^i$ is a single vertex$\}$, $\{\ell(\xi^i) = 0\}$ and $\{i \geq N+1\}$
- The last element of ξ^N is a.s. v_L
- $\sum_{i=1}^{L} \ell(\xi^i) = L$ a.s.

In the sequel, we identify the random vector $(\xi^1, \ldots, \xi^L)$ with the random variable $\xi = (v_0, \ldots, v_L)$. As a result, ξ is seen as a r.v with values in Ξ^L.

Our notations are summarized in Table 8.1. For every $i = 1, \ldots, L$, define the functions f_i, g_i on

Table 8.1: Useful Notations

$G = (V, E)$	Graph with no self-loop
s	walk on G
(v_i)	infinite random walk
$\xi = (\xi^1, \dots, \xi^L)$	random walk of length L
ξ^i	random simple path
$\ell(\xi^i)$	length of ξ^i
$R(x, \phi)$	$\phi-$regularization of x on G
$R(x, \phi_s)$	$\phi-$regularization of x along the walk s

$\mathbb{R}^V \times \Xi$ in such a way that

$$f_i(\xi^i, x) = \frac{\ell(\xi^i)}{L|E|} F(x) \tag{8.18}$$

$$g_i(\xi^i, x) = \frac{1}{L} R(x, \phi_{\xi^i}) \,. \tag{8.19}$$

Note that when $i > N(\omega)$ then $f_i(\xi^i(\omega), x) = g_i(\xi^i(\omega), x) = 0$.

Proposition 8.4.2. For every $x \in \mathbb{R}^V$, we have

$$\frac{1}{|E|}(F(x) + R(x, \phi)) = \sum_{i=1}^{L} \mathbb{E}\left[f_i(\xi^i, x) + g_i(\xi^i, x)\right] \,. \tag{8.20}$$

Proof. For every $\omega \in \Omega$ and every $x \in \mathbb{R}^V$,

$$\frac{1}{L} R(x, \phi_{\xi(\omega)}) = \frac{1}{L} \sum_{i=1}^{N(\omega)} R(x, \phi_{\xi^i(\omega)}) = \sum_{i=1}^{L} g_i(\xi^i(\omega), x) \,.$$

Integrating, and using Prop. 8.4.1, it follows that $\sum_{i=1}^{L} \mathbb{E}[g_i(\xi^i, x)] = \frac{1}{|E|} R(x, \phi)$. Moreover, we have $\sum_{i=1}^{L} f_i(\xi^i(\omega), x) = \frac{1}{|E|} F(x)$. This completes the proof. □

8.4.4 Main Algorithm

Prop. 8.4.2 suggests that minimizers of Problem (8.1) can be found by minimizing the right-hand side of (8.20). This can be achieved by means of the stochastic approximation algorithm provided in Sec. 8.3. The corresponding iterations (8.7) read as $x_{n+1} = \mathsf{T}_{\gamma_{n+1}}(\xi_{n+1}, x_n)$ where (ξ_n) are i.i.d copies of ξ. For every $i = 1, \dots, L-1$, the intermediate variable $\bar{x}_{n+1}^i$ given by Eq. (8.8) satisfies

$$\bar{x}_{n+1}^i = \mathrm{prox}_{\gamma_n g_i(\xi_{n+1}^i, \cdot)}(\bar{x}_n^{i-1} - \gamma_n \nabla f_i(\xi_{n+1}^i, \bar{x}_n^{i-1})) \,.$$

Theorem 8.4.3. Let Assumption 8.3.2 hold true. Assume that the convex function F is differentiable and that ∇F is Lipschitz continuous. Assume that Problem (8.1) admits a minimizer. Then, there exists a r.v. $X_\star$ s.t. $X_\star(\omega)$ is a minimizer of (8.1) for all ω $\mathbb{P}$-a.e., and s.t. the sequence (x_n) defined above converges a.s. to $X_\star$ as $n \to \infty$. Moreover, for every $i = 0, \dots, L-1$, $\bar{x}_n^i$ converges a.s. to $X_\star$.

Table 8.2: Proposed Snake algorithm.

procedure SNAKE(x_0, L)
  $z \leftarrow x_0$
  $e \leftarrow$RND_ORIENTED_EDGE
  $n \leftarrow 0$
  $\ell \leftarrow L$
  while stopping criterion is not met **do**
    $c, e \leftarrow$ SIMPLE_PATH(e, ℓ)
    $z \leftarrow$ PROX1D($z - \gamma_n \frac{\text{LENGTH}(c)}{L|E|} \nabla F(z), c, \frac{1}{L}\gamma_n$)
    $\ell \leftarrow \ell -$ LENGTH(c)
    if $\ell = 0$ **then**
      $e \leftarrow$ RND_ORIENTED_EDGE
      $\ell \leftarrow L$
      $n \leftarrow n + 1$ ▷ x_n is z at this step
    end if
  end while
  return z
end procedure

Table 8.3: SIMPLE_PATH procedure.

procedure SIMPLE_PATH(e, ℓ)
  $c \leftarrow e$
  $w \leftarrow$ UNIFORM_NEIB($e[-1]$)
  while $w \notin c$ and LENGTH(c)$< \ell$ **do**
    $c \leftarrow [c, w]$
    $w \leftarrow$ UNIFORM_NEIB(w)
  end while
  return $c, [c[-1], w]$
end procedure

Proof. It is sufficient to verify that the mappings f_i, g_i defined by (8.18) and (8.19) respectively fulfill Assumptions 8.3.1–8.3.5 of Th. 8.3.1. Then, Th. 8.3.1 gives the conclusion. Assumptions 8.3.1 and 8.3.3 are trivially satisfied. It remains to show, for every minimizer $x_\star$, the existence of a $(2+\varepsilon)$-representation, for some $\varepsilon > 0$. Any such $x_\star$ satisfies Eq. (8.11) where $(\nabla f_i(\xi^i, x_\star), \varphi_i(\xi^i))_{i=1,\dots,L}$ is a representation of the minimizer $x_\star$. By definition of f_i and g_i, it is straightforward to show that there exists a deterministic constant $C_\star$ depending only on $x_\star$ and the graph G, such that $\|\nabla f_i(\xi^i, x_\star)\| < C_\star$ and $\|\varphi_i(\xi^i)\| < C_\star$. This proves Assumption 8.3.4. Assumption 8.3.5 can be easily checked by the same arguments. □

Consider the general ϕ-regularized problem (8.1), and assume that an efficient procedure to compute the proximity operator of the ϕ-regularization over an 1D-graph is available. The sequence (x_n) is generated by the algorithm SNAKE (applied with the latter 1D efficient procedure) and is summarized in Table 8.2. Recall the definition of the probability π on V and the transition kernel P on V^2. The procedure presented in this table calls the following subroutines.

- If c is a finite walk, $c[-1]$ is the last element of c and LENGTH(c) is its length as a walk that is

$|c| - 1$.

- The procedure RND_ORIENTED_EDGE returns a tuple of two nodes randomly chosen (v, w) where $v \sim \pi$ and $w \sim P(v, \cdot)$.
- For every $x \in \mathbb{R}^V$, every simple path s and every $\alpha > 0$, PROX1D(x, s, α) is any procedure that returns the quantity $\mathrm{prox}_{\alpha R(\cdot, \phi_s)}(x)$.
- The procedure UNIFORM_NEIB(v) returns a random vertex drawn uniformly amongst the neighbors of the vertex v that is with distribution $P(v, \cdot)$.
- The procedure SIMPLE_PATH(e, ℓ), described in Table 8.3, generates the first steps of a random walk on G with transition kernel P initialized at the vertex $e[-1]$, and prefaced by the first node in e. It represents the ξ^i's of the previous section. The random walk is stopped when one node is repeated, or until the maximum number of samples $\ell + 1$ is reached. The procedure produces two outputs, the walk and the oriented edge $c, (c[-1], w)$. In the case where the procedure stopped due to a repeated node, c represents the simple path obtained by stopping the walk before the first repetition occurs, while w is the vertex which has been repeated (referred to as the pivot node). In the case where no vertex is repeated, it means that the procedure stopped because the maximum length was achieved. In that case, c represents the last simple path generated, and the algorithm doesn't use the pivot node w.

Remark 4. Although Snake converges for every value of the hyperparameter L, a natural question is the influence of L on the behavior of the algorithm. In the case where $R(\,\cdot\,, \phi)$ is the TV regularization, [50] notes that, empirically, the taut-string algorithm used to compute the proximity operator has a complexity of order $O(L)$. The same holds for the Laplacian regularization. Hence, parameter L controls the complexity of every iteration. On the other hand, in the reformulation of Problem (8.1) into the stochastic form (8.15), the random variable $|E|R(x, \phi_\xi)/L$ is an unbiased estimate of $R(x, \phi)$. By the ergodic theorem, the larger L, the more accurate is the approximation. Hence, there is a trade-off between complexity of an iteration and precision of the algorithm. This trade-off is standard in the machine learning literature. It often appears while sampling mini-batches in order to apply the stochastic gradient algorithm to minimize a finite sum (see [33, 34]). The choice of L is somehow similar to the problem of the choice of the length of the mini-batches in this context.

Providing a theoretical rule that would optimally select the value of L is a difficult task that is beyond the scope of this chapter. Nevertheless, in Sec. 8.6, we provide a detailed analysis of the influence of L on the numerical performance of the algorithm.

8.5 Proximity operator over 1D-graphs

We now provide some special cases of ϕ-regularizations, for which the computation of the proximity operator over 1D-graphs is easily tractable. Specifically, we address the case of the total variation regularization and the Laplacian regularization which are particular cases of ϕ-regularizations.

8.5.1 Total Variation norm

In the case where $\phi_{\{i,j\}}(x, x') = w_{\{i,j\}}|x - x'|$, $R(x, \phi)$ reduces to the weighted TV regularization

$$R(x, \phi) = \sum_{\{i,j\} \in E} w_{\{i,j\}} |x(i) - x(j)|$$

and in the case where $\phi_{\{i,j\}}(x,x') = |x-x'|$, $R(x,\phi)$ reduces to the its unweighted version

$$R(x,\phi) = \sum_{\{i,j\}\in E} |x(i)-x(j)|\,.$$

As mentioned above, there exists a fast method, the taut string algorithm, to compute the proximity operator of these $\phi-$regularizations over a 1D-graph ([9, 50]).

8.5.2 Laplacian regularization

In the case where $\phi_{\{i,j\}}(x,x') = w_{\{i,j\}}(x-x')^2$, $R(x,\phi)$ reduces to the Laplacian regularization that is

$$R(x,\phi) = \sum_{\{i,j\}\in E} w_{\{i,j\}}(x(i)-x(j))^2.$$

Its unweighted version is

$$R(x,\phi) = \sum_{\{i,j\}\in E} (x(i)-x(j))^2.$$

In the case where $\phi_{\{i,j\}}(x,x') = w_{\{i,j\}}(x/\sqrt{\deg(i)} - x'/\sqrt{\deg(j)})^2$,

$$R(x,\phi) = \sum_{\{i,j\}\in E} w_{\{i,j\}} \left(\frac{x(i)}{\sqrt{\deg(i)}} - \frac{x'(i)}{\sqrt{\deg(j)}} \right)^2$$

is the normalized Laplacian regularization.

We now explain one method to compute the proximity operator of the unweighted Laplacian regularization over an 1D-graph. The computation of the proximity operator of the normalized Laplacian regularization can be done similarly. The computation of the proximity operator of the weighted Laplacian regularization over an 1D-graph is as fast as the computation of the proximity operator of the unweighted Laplacian regularization over an 1D-graph, using for example Thomas' algorithm.

The proximity operator of a point $y \in \mathbb{R}^{\ell+1}$ is obtained as a solution to a quadratic programming problem of the form:

$$\min_{x\in\mathbb{R}^{\ell+1}} \frac{1}{2}\|x-y\|^2 + \lambda \sum_{k=1}^{\ell} (x(k-1)-x(k))^2\,,$$

where $\lambda > 0$ is a scaling parameter. Writing the first order conditions, the solution x satisfies

$$(I + 2\lambda\mathcal{L})x = y \tag{8.21}$$

where $\mathcal{L}$ is the Laplacian matrix of the 1D-graph with $\ell+1$ nodes and I is the identity matrix in $\mathbb{R}^{\ell+1}$. Using [45], $\mathcal{L}$ can be diagonalized explicitly. In particular, $I + 2\lambda\mathcal{L}$ has eigenvalues

$$1 + 4\lambda\left(1 - \cos\left(\frac{\pi k}{\ell+1}\right)\right),$$

and eigenvectors $e_k \in \mathbb{R}^{\ell+1}$

$$e_k(j) = \frac{1}{2(\ell+1)} \cos\left(\pi\frac{kj}{\ell+1} - \pi\frac{k}{2(\ell+1)}\right),$$

for $0 \le k < n$. Hence, $x = C^*\Lambda^{-1}Cy$, where Λ gathers the eigenvalues of $I+2\lambda\mathcal{L}$ and the operators C and C^* are the discrete cosine transform operator and the inverse discrete cosine transform respectively. Therefore, x can be found in $O(\ell\log(\ell))$ operations.

8.6 Examples

We now give some practical instances of Problem (8.1) by particularizing F and the ϕ-regularization in (8.1). The ϕ-regularizations considered in this section will be among the ϕ-regularizations mentioned in Sec. 8.5. We also provide some simulations to compare our method to existing algorithms. The code is available at the address https://github.com/adil-salim/Snake.

8.6.1 Trend Filtering on Graphs

Consider a vector $y \in \mathbb{R}^V$. The Graph Trend Filtering (GTF) estimate on V with parameter k set to one (see [127] for the definition of the parameter) is defined in [127] by

$$\hat{y} = \arg\min_{x \in \mathbb{R}^V} \frac{1}{2}\|x - y\|^2 + \lambda \sum_{\{i,j\} \in E} |x(i) - x(j)|, \tag{8.22}$$

where $\lambda > 0$ is a scaling parameter. In the GTF context, the vector y represents a sample of noisy data over the graph G and the GTF estimate represents a denoised version of y. When G is an 1D or a 2D-graph, the GTF boils down to a well known context [120, 41]. When G is a general graph, the GTF estimate is studied in [127] and [69]. The estimate $\hat{y}$ is obtained as the solution of a TV-regularized risk minimization with $F(x) = \frac{1}{2}\|x - y\|^2$ where y is fixed. We address the problem of computing the GTF estimate on two real life graphs from [74] and one sampled graph. The first one is the Facebook graph which is a network of 4039 nodes and 88234 edges extracted from the Facebook social network. The second one is the Orkut graph with 3072441 nodes and 117185083 edges. Orkut was also an on-line social network. The third graph is sampled according to a Stochastic Block Model (SBM). Namely we generate a graph of 4000 nodes with four well-separated clusters of 1000 nodes (also called "communities") as depicted in Fig. 8.2. Then we draw independently N^2 Bernoulli r.v. $E(i,j)$, encoding the edges of the graph (an edge between nodes i and j is present iff $E(i,j) = 1$), such that $\mathbb{P}[E(i,j) = 1] = P(c_i, c_j)$ where c_i denotes the community of the node i and where

$$\begin{cases} P(c,c') = & 0.1 \text{ if } c = c' \\ P(c,c') = & 0.005 \text{ otherwise.} \end{cases}$$

This model is called the stochastic block model for the matrix P [68]. It amounts to a blockwise Erdös-Rényi model with parameters depending only on the blocks. It leads to 81117 edges.

We assume that every node is provided with an unknown value in $\mathbb{R}$ (the set of all these values being referred to as the *signal* in the sequel). In our example, the value $y(i)$ at node i is generated as $y(i) = l(c_i) + \sigma\epsilon_i$ where l is a mapping from the communities to a set of levels (in Fig. 8.2, $l(i)$ is an integer in $[0, 255]$), and ϵ denotes a standard Gaussian white noise with $\sigma > 0$ as its standard deviation. In Fig. 8.2, we represent an example of the signal y (left figure) along with the "initial" values $l(c_i)$ represented in grayscale at every node.

Over the two real life graphs, the vector y is sampled according to a standard Gaussian distribution of dimension $|V|$. The parameter λ is set such that $\mathbb{E}[\frac{1}{2}\|x - y\|^2] = \mathbb{E}[\lambda \sum_{\{i,j\} \in E} |x(i) - x(j)|]$ if x, y are two independent r.v with standardized Gaussian distribution. The initial guess x_0 is set equal to y. The step size γ_n is set equal to $|V|/(10n)$ for the two real life graphs and $|V|/(5n)$ for the SBM realization graph. We ran the Snake algorithm for different values of L, except over the Orkut graph where $L = |V|$. The dual problem of (8.22) is quadratic with a box constraint. The Snake algorithm is compared to the well-known projected gradient (PG) algorithm for the dual problem. To solve the dual

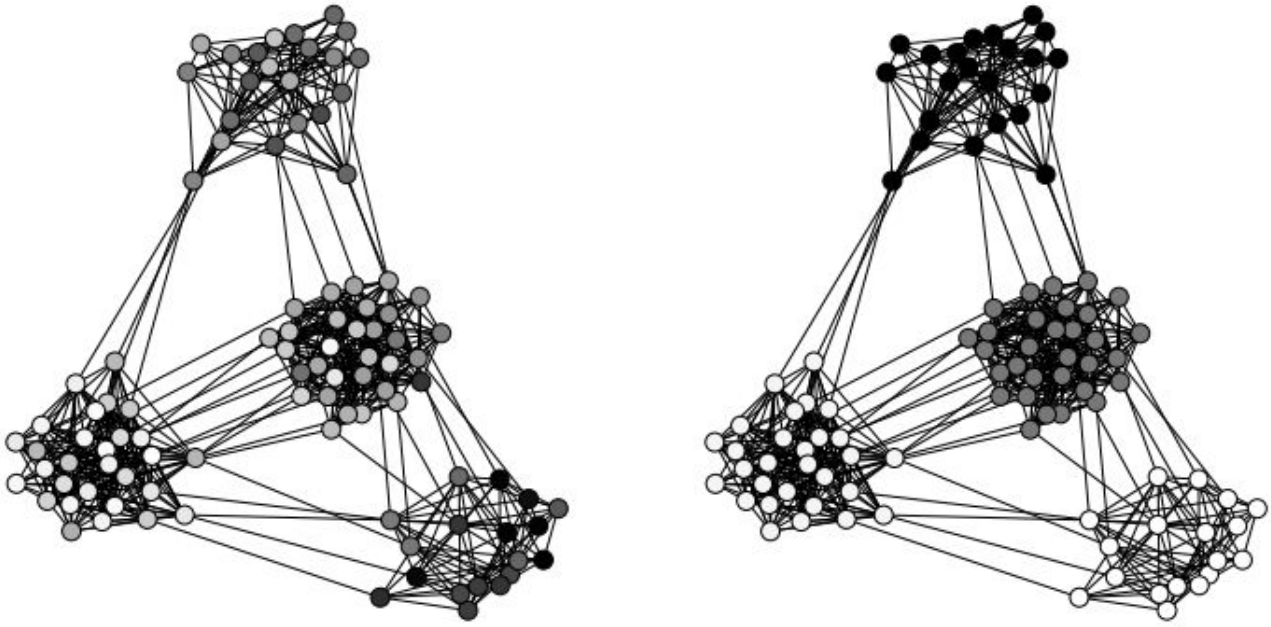

Figure 8.2: The signal is the grayscale of the node. The graph is sampled according to a SBM. Left: Noised signal over the nodes. Right: Sought signal.

problem of (8.22), we use L-BFGS-B [39] as suggested in [127]. Note that, while running on the Orkut graph, the algorithm L-BFGS-B leads to a memory error from the solver [39] in SciPy (using one thread of a 2800 MHz CPU and 256GB RAM).

Fig. 8.3 shows the objective function as a function of time for each algorithm.

In the case of the TV regularization, we observe that Snake takes advantage of being an online method, which is known to be twofold ([33, 34]). First, the iteration complexity is controlled even over large general graphs: the complexity of the computation of the proximity operator is empirically linear [50]. On the contrary, the projected gradient algorithm involves a matrix-vector product with complexity $O(|E|)$. Hence, *e.g* the projected gradient algorithm has an iteration complexity of at least $O(|E|)$. The iteration complexity of Snake can be set to be moderate in order to frequently get iterates while running the algorithm. Then, Snake is faster than L-BFGS-B and the projected gradient algorithms for the dual problem in the first iterations of the algorithms.

Moreover, for the TV regularization, Snake seems to perform globally better than L-BFGS-B and the projected gradient. This is because Snake is a proximal method where the proximity operator is efficiently computed ([12]).

The parameter L seems to have a minor influence on the performance of the algorithm since, in Fig. 8.3 the curves corresponding to different values of L are closely superposed.

Over the three graphs, the value $L = O(|V|)$ is a good value, if not the best value to use the Snake algorithm. One can show that, while sampling the first steps of the infinite random walk over G from the node, say v, the expected time of return to the random node v is $|V|$. Hence, the value $L = |V|$ allows Snake to significantly explore the graph during one iteration.

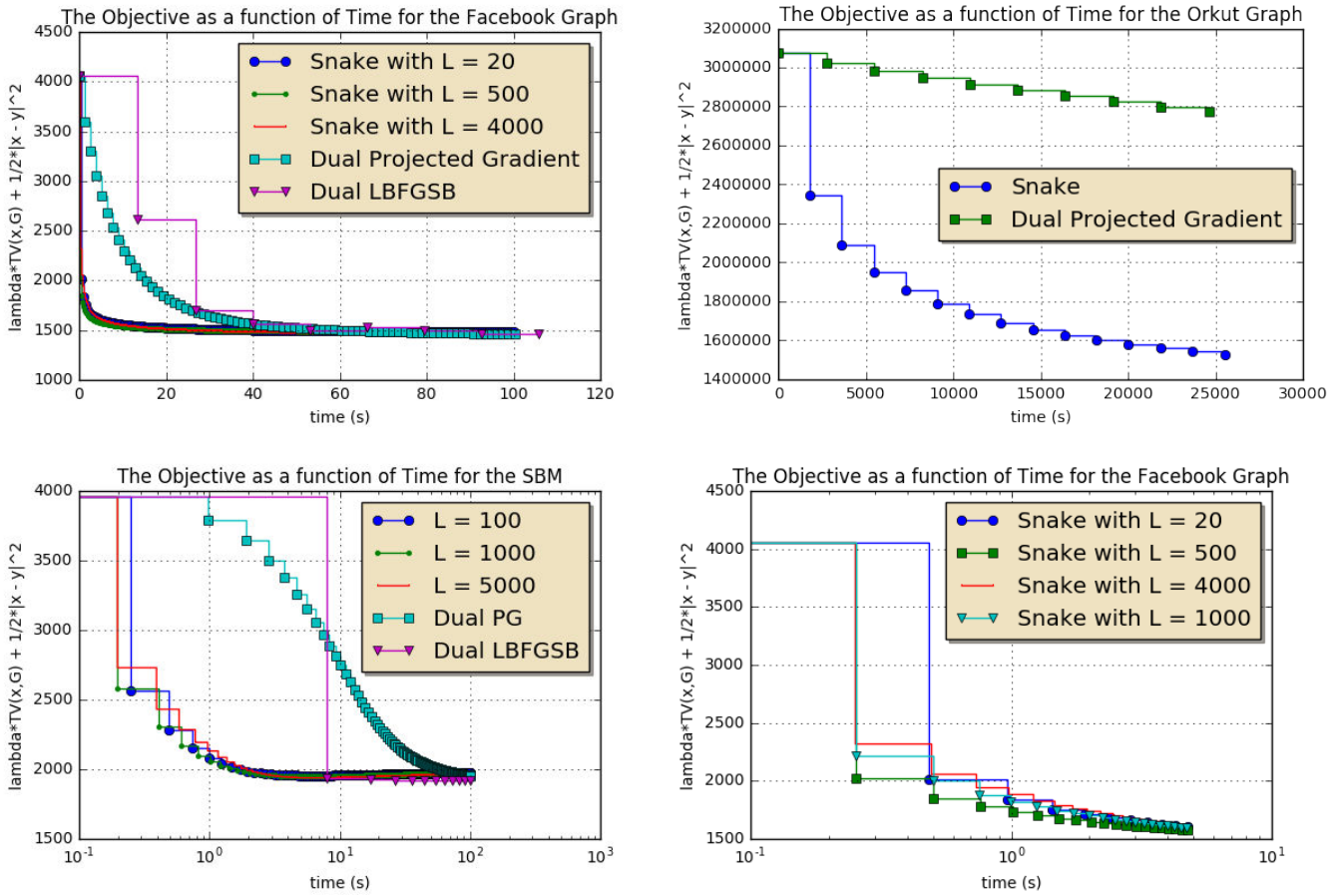

Figure 8.3: The algorithm Snake for the TV regularization is applied to different graphs, with different values of the parameter L.

8.6.2 Graph Inpainting

The problem of graph inpainting has been studied in [43, 57, 134] and can be expressed as follows. Consider a vector $y \in \mathbb{R}^V$, a subset $O \subset V$. Let $\bar{O}$ be its complementary in V. The harmonic energy minimization problem is defined in [134] by

$$\begin{aligned} \min_{x \in \mathbb{R}^V} \quad & \sum_{\{i,j\} \in E} (x(i) - x(j))^2 \\ \text{subject to} \quad & x(i) = y(i), \forall i \in O. \end{aligned}$$

This problem is interpreted as follows. The signal $y \in \mathbb{R}^V$ is partially observed over the nodes and the aim is to recover y over the non observed nodes. The subset $O \subset V$ is the set of the observed nodes and $\bar{O}$ the set of unobserved nodes. An example is shown in Fig. 8.4.

Denote by $G_{\bar{O}} = (\bar{O}, E_{\bar{O}})$ the subgraph of G induced by $\bar{O}$. Namely, $\bar{O}$ is the set of vertices, and the set $E_{\bar{O}}$ is formed by the edges $\{i,j\} \in E$ s.t. $i \in \bar{O}$ and $j \in \bar{O}$. The harmonic energy minimization is equivalent to the following Laplacian regularized problem over the graph $G_{\bar{O}}$:

$$\min_{x \in \mathbb{R}^{\bar{O}}} \quad F(x) + \sum_{\substack{\{i,j\} \in E_{\bar{O}} \\ i<j}} (x(i) - x(j))^2 \tag{8.23}$$

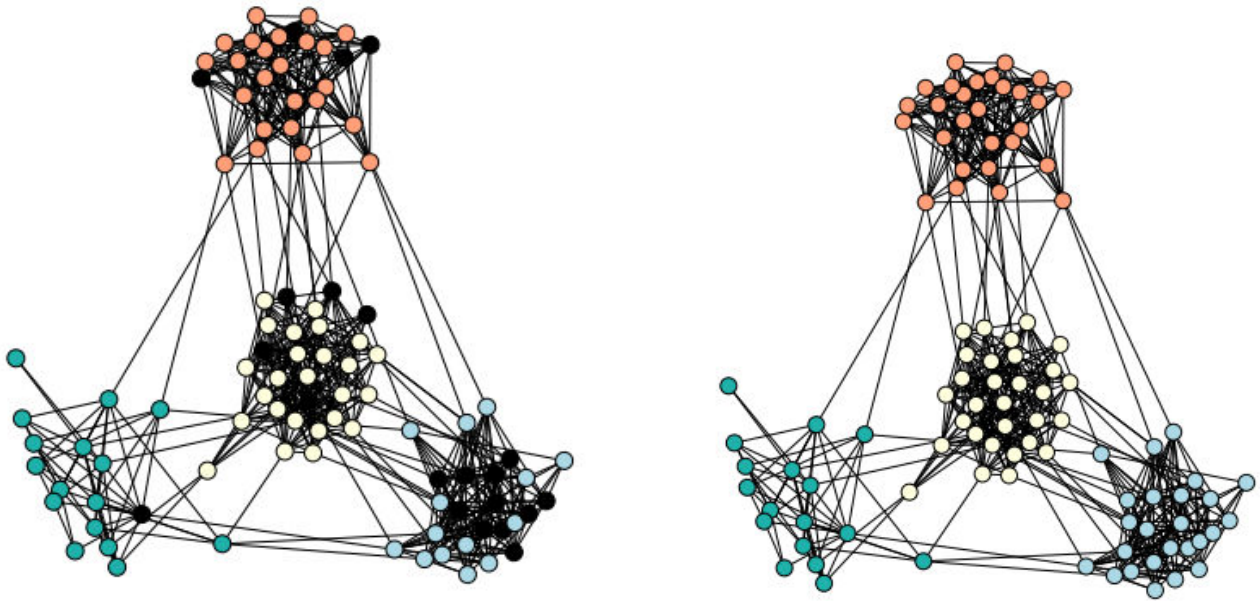

Figure 8.4: Left: Partially observed data (unobserved nodes are black, data is the color of nodes). Right: Fully observed data (the color is observed for all nodes).

where

$$F(x) = \sum_{\substack{i\in\bar{O}, j\in O \\ \{i,j\}\in E}} (x(i) - y(j))^2 \,.$$

The signal y is sampled according to a standardized Gaussian distribution of dimension $|V|$. We compared the Snake algorithm to existing algorithm over the Orkut graph. The set V is divided in two parts of equal size to define O and $\bar{O}$. The initial guess is set equal to zero over the set of unobserved nodes $\bar{O}$, and to the restriction of y to O over the set of observed nodes O. We compare our algorithm with the conjugate gradient.

Fig. 8.5 represents the objective function $\sum_{\{i,j\}\in E}(x(i) - x(j))^2$ as a function of time. Over the Facebook graph, the parameter L is set equal to $|V|/10$. The step size γ_n are set equal to $|V|/(10n)$. Over the Orkut graph, L is set equal to $|V|/50$. The step size are set equal to $|V|/(5\sqrt{n})$ on the range displayed in Fig. 8.5. Even if the sequence $(|V|/(5\sqrt{n}))_{n\in\mathbb{N}}$ does not satisfies the Assumption 8.3.2, it is a standard trick in stochastic approximation to take a slowly decreasing step size in the first iterations of the algorithm ([88]). It allows the iterates to be quickly close to the set of solutions without converging to the set of solutions. Then, one can continue the iterations using a sequence of step size satisfying Assumption 8.3.2 to make the algorithm converging. There is a trade-off between speed and precision while choosing the step-size. Snake turns out to be faster in the first iterations. Moreover, as an online method, it allows the user to control the iteration complexity of the algorithm. Since a discrete cosine transform is used, the complexity of the computation of the proximity operator is $O(L\log(L))$. In contrast, the iteration complexity of the conjugate gradient algorithm can be a bottleneck (at least $O(|E|)$) as far as very large graphs are concerned.

Besides, Snake for the Laplacian regularization does not perform globally better than the conjugate gradient. This is because the conjugate gradient is designed to fully take advantage of the quadratic

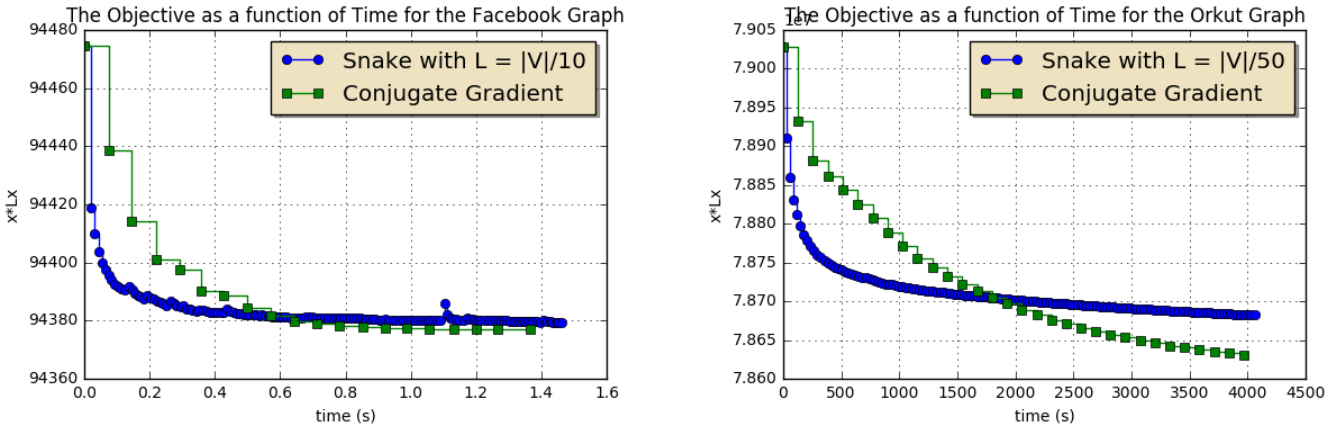

Figure 8.5: Left: Snake applied to the Laplacian regularization over the Facebook Graph. Right: Snake applied to the Laplacian regularization over the Orkut Graph.

structure. On the contrary, Snake is not specific to quadratic problems.

8.6.3 Online Laplacian solver

Let $\mathcal{L}$ the Laplacian of a graph $G = (V, E)$. The resolution of the equation $\mathcal{L}x = b$, where b is a zero mean vector, has numerous applications ([123, 117]). This equation can be solved by minimizing the Laplacian regularized problem

$$\min_{x \in \mathbb{R}^V} -b^* x + \frac{1}{2} x^* \mathcal{L} x.$$

In our experiment, the vector b is randomly chosen using a standardized Gaussian distribution of dimension $|V|$. We compare our algorithm with the conjugate gradient over the Orkut graph.

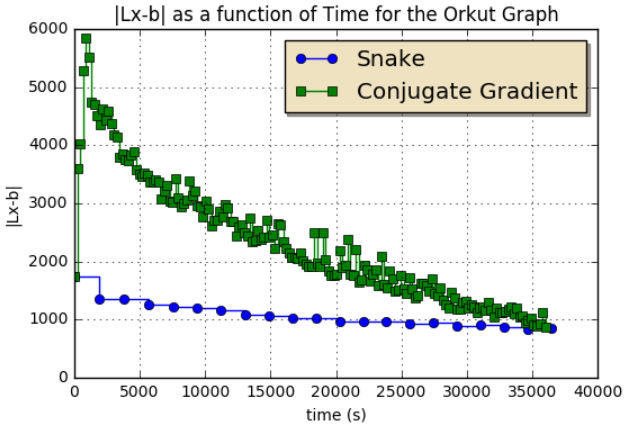

Figure 8.6: Snake applied to the resolution of a Laplacian system over the Orkut graph

Fig. 8.6 represents the quantity $\|\mathcal{L}x_n - b\|$ as a function of time, where x_n is the iterate provided either by Snake or by the conjugate gradient method. The parameter L is set equal to $|V|$. The step

size γ_n are set equal to $|V|/(2n)$. Snake appears to be more stable than the conjugate gradient method, and has a better performance at start up.

8.7 Conclusion

A fast regularized optimization algorithm over large unstructured graphs was introduced in this chapter. This algorithm is a variant of the proximal gradient algorithm that operates on randomly chosen simple paths. It belongs to the family of stochastic approximation algorithms with a decreasing step size. One future research direction consists in a fine convergence analysis of this algorithm, hopefully leading to a provably optimal choice of the total walk length L. Another research direction concerns the constant step analogue of the described algorithm, whose transient behavior could be interesting in many applicative contexts in the fields of statistics and learning.

8.8 Proofs for Sec. 8.3.3

8.8.1 Proof of Lem. 8.3.2

We start by writing $\|\nabla f_i(s^i, \bar{x}^{i-1})\| \leq \|\nabla f_i(s^i, \bar{x}^{i-2})\| + K_i(s^i)\|\bar{x}^{i-1} - \bar{x}^{i-2}\|$, where $K_i(s^i)$ is provided by Assumption 8.3.3. Using the identity $\bar{x}^{i-1} = \mathsf{T}_{\gamma,i-1}(\bar{x}^{i-2})$, where $\mathsf{T}_{\gamma,i}$ is given by (8.14), and recalling that $\nabla g_i^\gamma(s^i, \cdot)$ is γ^{-1}-Lipschitz, we get

$$\|\nabla f_i(s^i, \bar{x}^{i-1})\| \leq \|\nabla f_i(s^i, \bar{x}^{i-2})\| + \gamma K_i(s^i)(2\|\nabla f_{i-1}(s^{i-1}, \bar{x}^{i-2})\| + \|\nabla g_{i-1}^\gamma(s^{i-1}, \bar{x}^{i-2})\|).$$

Similarly,

$$\begin{aligned}\|\nabla g_i^\gamma(s^i, \bar{x}^{i-1} - \gamma\nabla f_i(s^i, \bar{x}^{i-1}))\| \leq \\ \|\nabla f_i(s^i, \bar{x}^{i-1})\| + 2\|\nabla f_{i-1}(s^{i-1}, \bar{x}^{i-2})\| + \|\nabla g_i^\gamma(s^i, \bar{x}^{i-2})\| + \|\nabla g_{i-1}^\gamma(s^{i-1}, \bar{x}^{i-2})\|.\end{aligned}$$

Iterating down to $\bar{x}^0 = x$, we get the result since for every i, since all the moments of $K_i(\xi^i)$ are finite.

8.8.2 Proof of Prop. 8.3.3

Let $x_\star$ be an arbitrary element of $\mathcal{Z}$. Let $i \in \{1, \ldots, L\}$. We start by writing

$$\begin{aligned}\|\bar{x}_{n+1}^i - x_\star\|^2 &= \|\bar{x}_{n+1}^i - \bar{x}_{n+1}^{i-1}\|^2 + \|\bar{x}_{n+1}^{i-1} - x_\star\|^2 + 2\langle \bar{x}_{n+1}^i - \bar{x}_{n+1}^{i-1}, \bar{x}_{n+1}^{i-1} - x_\star\rangle \\ &= \|\bar{x}_{n+1}^{i-1} - x_\star\|^2 + \gamma^2\|\nabla f_i + \nabla g_i^\gamma\|^2 - 2\gamma\langle \nabla f_i^\star + \varphi_i, \bar{x}_{n+1}^{i-1} - x_\star\rangle \\ &\quad - 2\gamma\langle \nabla f_i - \nabla f_i^\star, \bar{x}_{n+1}^{i-1} - x_\star\rangle - 2\gamma\langle \nabla g_i^\gamma - \varphi_i, \bar{x}_{n+1}^{i-1} - x_\star\rangle \\ &= \|\bar{x}_{n+1}^{i-1} - x_\star\|^2 + A_1 + A_2 + A_3 + A_4.\end{aligned}$$

Most of the proof consists in bounding the A_i's. We shall repeatedly use Young's inequality $|\langle a, b\rangle| \leq \eta\|a\|^2 + C\|b\|^2$, where $\eta > 0$ is a constant chosen as small as desired, and $C > 0$ is fixed accordingly. Starting with A_1, we have

$$A_1 \leq \gamma^2(1+\eta)\|\nabla g_i^\gamma\|^2 + C\gamma^2\|\nabla f_i\|^2.$$

We have $A_3 \leq 0$ by the convexity of f_L. We can write

$$A_4 = -2\gamma\langle \nabla g_i^\gamma - \varphi_i, \mathrm{prox}_{\gamma g_i} - x_\star\rangle - 2\gamma\langle \nabla g_i^\gamma - \varphi_i, \bar{x}_{n+1}^{i-1} - \gamma\nabla f_i - \mathrm{prox}_{\gamma g_i}\rangle - 2\gamma\langle \nabla g_i^\gamma - \varphi_i, \gamma\nabla f_i\rangle$$

By monotonicity of ∂g_i, the first term at the right hand side is ≤ 0. Since $\bar{x}_{n+1}^{i-1} - \gamma \nabla f_i - \mathrm{prox}_{\gamma g_i} = \gamma \nabla g_i^\gamma$. Thus,

$$\begin{aligned} A_4 &\leq -2\gamma^2 \|\nabla g_i^\gamma\|^2 + 2\gamma^2 \langle \varphi_i, \nabla g_i^\gamma + \nabla f_i \rangle - 2\gamma^2 \langle \nabla g_i^\gamma, \nabla f_i \rangle \\ &\leq -(2-\eta)\gamma^2 \|\nabla g_i^\gamma\|^2 + C\gamma^2 \|\nabla f_i\|^2 + C\gamma^2 \|\varphi_i\|^2 \end{aligned}$$

As regards A_2, we have

$$A_4 = -2\gamma \langle \nabla f_i^\star + \varphi_i, x_n - x_\star \rangle - 2\gamma \langle \nabla f_i^\star + \varphi_i, \bar{x}_{n+1}^{i-1} - x_n \rangle.$$

Gathering these inequalities, we get

$$\begin{aligned} \|\bar{x}_{n+1}^i - x_\star\|^2 \leq \|\bar{x}_{n+1}^{i-1} - x_\star\|^2 - (1-\eta)\gamma^2 \|\nabla g_i^\gamma\|^2 + C\gamma^2 \|\nabla f_i\|^2 + C\gamma^2 \|\varphi_i\|^2 \\ - 2\gamma \langle \nabla f_i^\star + \varphi_i, x_n - x_\star \rangle - 2\gamma \langle \nabla f_i^\star + \varphi_i, \bar{x}_{n+1}^{i-1} - x_n \rangle. \end{aligned}$$

Iterating over i, we get

$$\begin{aligned} \|\bar{x}_{n+1}^i - x_\star\|^2 \leq \|x_n - x_\star\|^2 - (1-\eta)\gamma^2 \sum_{k=1}^i \|\nabla g_k^\gamma\|^2 + C\gamma^2 \sum_{k=1}^i \|\nabla f_k\|^2 + C\gamma^2 \sum_{k=1}^i \|\varphi_k\|^2 \\ - 2\gamma \sum_{k=1}^i \langle \nabla f_k^\star + \varphi_k, x_n - x_\star \rangle - 2\gamma \sum_{k=1}^i \langle \nabla f_k^\star + \varphi_k, \bar{x}_{n+1}^{k-1} - x_n \rangle. \end{aligned}$$

The summand in the last term can be written as

$$\begin{aligned} -2\gamma \langle \nabla f_k^\star + \varphi_k, \bar{x}_{n+1}^{k-1} - x_n \rangle &= -2\gamma \sum_{\ell=1}^{k-1} \langle \nabla f_k^\star + \varphi_k, \bar{x}_{n+1}^\ell - \bar{x}_{n+1}^{\ell-1} \rangle \\ &= -2\gamma^2 \sum_{\ell=1}^{k-1} \langle \nabla f_k^\star + \varphi_k, \nabla f_\ell + \nabla g_\ell^\gamma \rangle \\ &\leq \gamma^2 C \|\nabla f_k^\star\|^2 + \gamma^2 C \|\varphi_k\|^2 + \gamma^2 C \sum_{\ell=1}^{k-1} \|\nabla f_\ell\|^2 + \gamma^2 \eta \sum_{\ell=1}^{k-1} \|\nabla g_\ell^\gamma\|^2. \end{aligned}$$

where we used $|\langle a, b \rangle| \leq \eta \|a\|^2 + C\|b\|^2$ as above. Therefore, for all $i = 1, \ldots, L$,

$$\begin{aligned} \|\bar{x}_{n+1}^i - x_\star\|^2 \leq \|x_n - x_\star\|^2 - (1-\eta)\gamma^2 \sum_{k=1}^i \|\nabla g_k^\gamma\|^2 + C\gamma^2 \sum_{k=1}^i \|\nabla f_k\|^2 \\ + C\gamma^2 \sum_{k=1}^i \|\nabla f_k^\star\|^2 + C\gamma^2 \sum_{k=1}^i \|\varphi_k\|^2 - 2\gamma \langle \sum_{k=1}^i \nabla f_k^\star + \varphi_k, x_n - x_\star \rangle. \end{aligned} \tag{8.24}$$

Consider the case $i = L$. Using Assumption 8.3.4,

$$\begin{aligned} \mathbb{E}_n \left[\|\bar{x}_{n+1}^L - x_\star\|^2 \right] \leq \|x_n - x_\star\|^2 - (1-\eta)\gamma^2 \mathbb{E}_n \left[\sum_{k=1}^L \|\nabla g_k^\gamma\|^2 \right] \\ + C\gamma^2 \sum_{k=1}^L \mathbb{E}_n [\|\nabla f_k\|^2] - 2\gamma \mathbb{E}_n \left[\langle \sum_{k=1}^L \nabla f_k^\star + \varphi_k, x_n - x_\star \rangle \right] + C\gamma^2. \end{aligned}$$

The last term at the right hand side is zero since

$$\mathbb{E}_n\left[\langle\sum_{k=1}^{L}\nabla f_k^\star+\varphi_k, x_n-x_\star\rangle\right]=\langle\mathbb{E}\left[\sum_{k=1}^{L}\nabla f_k^\star+\varphi_k\right], x_n-x_\star\rangle=0$$

by definition of $\nabla f_k^\star$ and φ_k. Besides, using Assumption 8.3.3, for all k we have

$$\mathbb{E}_n[\|\nabla f_k\|^2]\leq C\mathbb{E}_n[\|\nabla f_k^\star\|^2]+C\mathbb{E}_n[K_k^2(\xi_{n+1}^k)\|\bar{x}_{n+1}^{k-1}-x_\star\|^2].$$

Then,

$$\begin{aligned}\mathbb{E}_n[\|x_{n+1}-x_\star\|^2]\leq\|x_n-x_\star\|^2-(1-\eta)\gamma^2\mathbb{E}_n\left[\sum_{k=1}^{L}\|\nabla g_k^\gamma\|^2\right]\\+C\gamma^2\sum_{k=1}^{L}\mathbb{E}_n\left[K_k^2(\xi_{n+1}^k)\|\bar{x}_{n+1}^{k-1}-x_\star\|^2\right]+C\gamma^2.\end{aligned}\tag{8.25}$$

We shall prove by induction that for all r.v P_k which is a monomial expression of the r.v

$$K_k^2(\xi_{n+1}^k),\dots,K_L^2(\xi_{n+1}^L),$$

there exists $C>0$ such that

$$\mathbb{E}_n\left[P_k\|\bar{x}_{n+1}^{k-1}-x_\star\|^2\right]\leq C(1+\|x_n-x_\star\|^2),\tag{8.26}$$

for all $k=1,\dots,L$. Note that such a r.v P_k is independent of $\mathscr{F}_n$, non-negative and for all $\alpha>0$, $\mathbb{E}[P_k^\alpha]<\infty$ by Assumption 8.3.3. Using Assumption 8.3.3, the induction hypothesis 8.26 is satisfied if $k=1$. Assume that it holds true until the step $k-1$ for some $k\leq L$. Using 8.24 and Assumption 8.3.3,

$$\begin{aligned}\mathbb{E}_n\left[P_k\|\bar{x}_{n+1}^{k-1}-x_\star\|^2\right]\leq C\|x_n-x_\star\|^2\\+C\gamma^2\mathbb{E}_n\left[P_k\sum_{\ell=1}^{k-1}\|\nabla f_\ell\|^2\right]\\+C\gamma^2\mathbb{E}_n\left[P_k\sum_{\ell=1}^{k-1}\|\varphi_\ell\|^2+\|\nabla f_\ell^\star\|^2\right]\\-2\gamma\mathbb{E}_nP_k\langle\sum_{\ell=1}^{k-1}\nabla f_\ell^\star+\varphi_\ell, x_n-x_\star\rangle.\end{aligned}\tag{8.27}$$

The last term at the right hand side can be bounded as

$$\begin{aligned}&-2\gamma\mathbb{E}_nP_k\langle\sum_{\ell=1}^{k-1}\nabla f_\ell^\star+\varphi_\ell, x_n-x_\star\rangle\\\leq&C\|x_n-x_\star\|^2+C\mathbb{E}_n\left[P_k\sum_{\ell=1}^{k-1}\|\nabla f_\ell^\star\|^2+\|\varphi_\ell\|^2\right]\\\leq&C\|x_n-x_\star\|^2+C\end{aligned}\tag{8.28}$$

using Hölder inequality and Assumption 8.3.4. For all $\ell=1,\dots,k-1$,

$$\begin{aligned}\mathbb{E}_n[P_k\|\nabla f_\ell\|^2]\leq C\mathbb{E}_n[P_k\|\nabla f_\ell^\star\|^2]+C\mathbb{E}_n\left[P_kK_\ell^2(\xi_{n+1}^\ell)\|\bar{x}_{n+1}^{\ell-1}-x_\star\|^2\right]\\\leq C(1+\|x_n-x_\star\|^2)\end{aligned}\tag{8.29}$$

where we used Hölder inequality and Assumption 8.3.4 for the first term at the right hand side and the induction hypothesis (8.26) at the step ℓ with the r.v $P_\ell := P_k K_\ell^2(\xi_{n+1}^\ell)$ for the second term.

Plugging (8.28) and (8.29) into (8.27) and using again Hölder inequality and Assumption 8.3.4 we find that (8.26) holds true at the step k. Hence (8.26) holds true for all $k = 1, \dots, L$. Finally, plugging (8.26) into (8.25) with $P_k = K_k^2(\xi_{n+1}^k)$ for all $k = 1, \dots, L$ we get

$$\mathbb{E}_n[\|x_{n+1} - x_\star\|^2] \leq (1 + C\gamma^2)\|x_n - x_\star\|^2 - (1 - \eta)\gamma^2 \mathbb{E}_n \left[\sum_{k=1}^{L} \|\nabla g_k^\gamma\|^2 \right] + C\gamma^2.$$

By the Robbins-Siegmund lemma [100], used along with $(\gamma_n) \in \ell^2$, we get that $(\|x_n - x_\star\|)$ converges almost surely, showing the first point.

By taking the expectations at both sides of this inequality, we also obtain that $(\mathbb{E}\|x_n - x_\star\|^2)$ converges, $\sup_n \mathbb{E}\|x_n - x_\star\|^2 < \infty$, and $\mathbb{E} \sum_n \gamma_{n+1}^2 \sum_{i=1}^L \|\nabla g_i^\gamma\|^2 < \infty$. As $\sup_n \mathbb{E}\|x_n - x_\star\|^2 < \infty$, we have by Assumption 8.3.3 that $\sup_n \mathbb{E}\|\nabla f_1\|^2 < \infty$. Using Lem. 8.3.2 and iterating, we easily get that $\mathbb{E} \sum_n \gamma_{n+1}^2 \sum_{i=1}^L \|\nabla f_i\|^2 < \infty$ for all i.

Since $\|\bar{x}_{n+1}^1 - x_n\| \leq \gamma \|\nabla f_1\| + \gamma \|\nabla g_1^\gamma\|$, we get that $\sum_n \mathbb{E}\|\bar{x}_{n+1}^1 - x_n\|^2 < \infty$. By Borel-Cantelli's lemma, we get that $\bar{x}_{n+1}^1 - x_n \to 0$ almost surely. The almost sure convergence of $\bar{x}_{n+1}^i - x_n$ to zero is shown similarly, and the proof of Prop. 8.3.3 is concluded.

Chapter 9

Conclusion and Prospects

In this thesis, we first generalized the Ordinary Differential Equation method to Differential Inclusions for constant step size stochastic approximation. Two kinds of DI are considered : DI involving an upper semicontinous operator and DI built upon a maximal monotone operator with possibly empty values. For each DI, several discretization schemes are considered. These schemes include the explicit implicit Euler scheme (Forward Backward algorithm) and a Douglas Rachford method. As randomness is involved in every stage of the discretization, we brought tools from probability theory to study the resulting algorithms. First, the dynamical behavior of the iterates is studied using weak convergence of stochastic processes techniques. This result is not enough to study the long-run behavior of the methods. Studying the sequence of iterates as a Markov chain, we provided a stability criterion that allowed to state the asymptotic behavior of the algorithm. We finally showed that this criterion is satisfied in many use cases, including the stochastic proximal gradient algorithm. In the second part of this thesis, we designed and applied generalizations of the stochastic proximal gradient algorithm to solve two kinds of problems. We first considered the saddle point problem of finding primal dual optimal points of a stochastic optimization problem. The optimization problem is linearly constrained by matrices that are also written as expectations. Then, we proposed an algorithm to address regularized optimization problems over large and general graphs. The regularization term is tied to the graph geometry and our proximal method allows to handle it stochastically.

In this thesis, we chose to tackle general problems using general compactness techniques that give asymptotic results. Non asymptotic bounds could be obtain for particular subproblems. Such bounds has already be obtained from a dynamical system point of view for several algorithms, including Langevin algorithm or FISTA. The underlying (stochastic) differential equation can often be cast in the framework of Hamiltonian dynamics. Another direction of research is the adaptation of the tools used in this thesis to study optimization in measure spaces. Such problems arise in machine learning and can be at the core of sampling methods. Finally, many algorithms in the fields of control or reinforcement learning can be seen as stochastic approximation algorithm, with a more general assumption on the noise (*i.e* a case where the sequence of random variables is not i.i.d). The algorithms studied in this thesis could be analyzed under these more general assumptions.

Appendix A

Technical Report : Stochastic Douglas Rachford

Notations

A.1 Statement of the Problem

Consider ξ a random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into an arbitrary measurable space $(\Xi, \mathcal{G})$ with distribution μ. Let $f, g : \Xi \times \mathsf{X} \to (-\infty, +\infty]$ two normal convex integrands and assume that $f(\xi, x)$ is integrable. We define F and G by

$$F(x) = \mathbb{E}_\xi(f(\xi, x))$$
$$G(x) = \mathbb{E}_\xi(g(\xi, x)) \ .$$

Denote by $(\xi_n)_n$ an i.i.d sequence of copies of ξ. In the sequel, we use the notation $f_n := f(\xi_n, \cdot)$ and $g_n := g(\xi_n, \cdot)$. The adaptive Douglas-Rachford algorithm is given by

$$u_{n+1} = \mathrm{prox}_{\gamma, f_{n+1}}(x_n)$$
$$z_{n+1} = \mathrm{prox}_{\gamma, g_{n+1}}(2u_{n+1} - x_n)$$
$$x_{n+1} = x_n + z_{n+1} - u_{n+1} \, .$$

We denote by $D(s)$ the domain of $g(s, \cdot)$, and by $\mathcal{D}$ the set defined by the relation $x \in \mathcal{D} \iff x \in D(\xi)$ a.s. We denote by $\boldsymbol{d}(x) = d(x, \mathcal{D})$. We also denote $F^\gamma(x) = \int f_\gamma(s, x)\mu(ds)$ and $G^\gamma(x) = \int g_\gamma(s, x)\mu(ds)$. We assume that $f(\xi, \cdot)$ has a.s a full domain (equal to X) and is continuously differentiable. Under these assumptions, $Z(\partial(G+F)) = Z(\partial G + \nabla F) = Z(\mathbb{E}(\partial g(\xi, \cdot)) + \mathbb{E}(\nabla f(\xi, \cdot)))$, see Chap. 2.

A.1.1 Useful facts

We first observe that the process (x_n) described by Eq. (4.4) is a homogeneous Markov chain with transition kernel denoted by P_γ. The kernel P_γ and the initial measure ν determine completely the probability distribution of the process (x_n), seen as a $(\Omega, \mathscr{F}) \to (\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes \mathbb{N}})$ random variable. We shall denote this probability distribution on $(\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes \mathbb{N}})$ as $\mathbb{P}^{\nu,\gamma}$. We denote by $\mathbb{E}^{\nu,\gamma}$ the corresponding expectation. When $\nu = \delta_a$ for some $a \in \mathsf{X}$, we shall prefer the notations $\mathbb{P}^{a,\gamma}$ and $\mathbb{E}^{a,\gamma}$ to $\mathbb{P}^{\delta_a,\gamma}$ and

$\mathbb{E}^{\delta_a,\gamma}$. From nom on, (x_n) will denote the canonical process on the canonical space $(\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes\mathbb{N}})$. We denote by $\mathscr{F}_n$ the sub-σ-field of $\mathscr{F}$ generated by the family $\{x_0, \{\xi_k^\gamma : 1 \le k \le n\}\}$, and we write $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$ for $n \in \mathbb{N}$. In the remainder of the paper, C will always denote a positive constant that does not depend on the time n nor on γ. This constant may change from a line of calculation to another. In all our derivations, γ will lie in the interval $(0, \gamma_0]$ where γ_0 is a fixed constant which is chosen as small as needed. Then, we observe that the Markov kernels P_γ are Feller, *i.e.*, they take the set $C_b(\mathsf{X})$ of the real, continuous, and bounded functions on X to $C_b(\mathsf{X})$. Indeed, for each $f \in C_b(\mathsf{X})$, Eq. (4.5) shows that $P_\gamma(\cdot, f) \in C_b(\mathsf{X})$ by the continuity of $\mathrm{prox}_{\gamma g(s,\cdot)}$ and $\mathrm{prox}_{\gamma f(s,\cdot)}$, and by dominated convergence. For each $\gamma > 0$, we denote as

$$\mathcal{I}(P_\gamma) := \{\pi \in \mathcal{M}(\mathsf{X}) \,:\, \pi = \pi P_\gamma\}$$

the set of invariant probability measures of P_γ. Define the family of kernels $\mathcal{P} := \{P_\gamma\}_{\gamma\in(0,\gamma_0]}$, and let

$$\mathcal{I}(\mathcal{P}) := \bigcup_{\gamma\in(0,\gamma_0]} \mathcal{I}(P_\gamma)$$

be the set of distributions π such that $\pi = \pi P_\gamma$ for at least one P_γ with $\gamma \in (0, \gamma_0]$.

Finally, we shall often refer to the Differential Inclusion (DI)

$$\begin{cases} \dot{\mathsf{x}}(t) & \in & -(\partial F + \partial G)(\mathsf{x}(t)) \\ \mathsf{x}(0) & = & x_0. \end{cases} \tag{A.1}$$

and to the associated semiflow Φ.

A.2 Theorem

H1 There exists $x_\star \in Z(\partial G + \nabla F)$ admitting a $\mathcal{L}^2(\mu)$ representation (φ, ψ) *i.e* $\exists \varphi, \psi \in \mathcal{L}^2(\mu)$, such that $\varphi(s) \in \partial g(s, x_\star)$ μ-a.s, $\psi(s) = \nabla f(s, x_\star)$ μ-a.s and $\mathbb{E}(\varphi(\xi) + \psi(\xi)) = 0$.

H2 There exists $L > 0$ s.t. $\nabla f(s, \cdot)$ is a.s L-Lipschitz continuous.

H3 The function $F + G$ satisfies one of the following properties:

(a) $F + G$ is coercive.

(b) $F + G$ is supercoercive.

H4 For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x\in\mathcal{K}\cap\mathcal{D}} \int \|\partial_0 g(s, x)\|^{1+\varepsilon}\, \mu(ds) < \infty,$$

H5 For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x\in\mathcal{K}} \int \|\nabla f(s, x)\|^{1+\varepsilon}\, \mu(ds) < \infty\,.$$

H6 For all $\gamma \in (0, \gamma_0]$ and all $x \in \mathsf{X}$,

$$\int \left(\|\nabla f_\gamma(s, x)\| + \frac{1}{\gamma}\|\,\mathrm{prox}_{\gamma g(s,\cdot)}(x) - \Pi_{\mathrm{cl}(D(s))}(x)\|\right) \mu(ds) \le C(1 + |F^\gamma(x) + G^\gamma(x)|)\,.$$

H7 $\forall x \in \mathsf{X},\ \int d(x, D(s))^2 \, \mu(ds) \geq C \boldsymbol{d}(x)^2$.

H8 For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{\gamma \in (0,\gamma_0], x \in \mathcal{K}} \frac{1}{\gamma^{1+\varepsilon}} \int \| \operatorname{prox}_{\gamma g(s,\cdot)}(x) - \Pi_{\mathrm{cl}(D(s))}(x) \|^{1+\varepsilon} \, \mu(ds) < \infty \, .$$

Theorem A.2.1. Let Hypotheses H1–H8 hold true. Then, for each probability measure ν having a finite second moment, for any $\varepsilon > 0$,

$$\limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma} \left(d(x_k, \arg\min(F+G)) > \varepsilon \right) \xrightarrow[\gamma \to 0]{} 0 \, .$$

Moreover, if Hypothesis H3–(b) is satisfied, then

$$\limsup_{n \to \infty} \ \mathbb{P}^{\nu,\gamma} \left(d\left(\bar{x}_n, \arg\min(F+G)\right) \geq \varepsilon \right) \xrightarrow[\gamma \to 0]{} 0, \ \text{and}$$
$$\limsup_{n \to \infty} \ d\left(\mathbb{E}^{\nu,\gamma}(\bar{x}_n), \arg\min(F+G)\right) \xrightarrow[\gamma \to 0]{} 0 \, .$$

where $\bar{x}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$.

A.3 Proof of Th. A.2.1

In this section, we study the iterations given by the adaptive Douglas Rachford algorithm. Let $\gamma_0 > 0$, $a \in \mathsf{X}$ and $(\xi_n)_{n \in \mathbb{N}}$ be an i.i.d sequence of random variables from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\Xi, \mathcal{G})$ with distribution μ. The adaptive Douglas Rachford algorithm with step size $\gamma > 0$ writes $x_0 = a$ and for all $n \in \mathbb{N}$,

$$x_{n+1} = x_n - \gamma \nabla f_\gamma(\xi_{n+1}, x_n) - \gamma \nabla g_\gamma(\xi_{n+1}, x_n - 2\gamma \nabla f_\gamma(\xi_{n+1}, x_n)). \tag{A.2}$$

Define

$$h_\gamma(s, x) := -\nabla f_\gamma(s, x) - \nabla g_\gamma(s, x - 2\gamma \nabla f_\gamma(s, x)).$$

The algorithm (A.2) can be rewritten as

$$x_{n+1} = x_n + \gamma h_\gamma(\xi_{n+1}, x_n). \tag{A.3}$$

In Sec. A.3.1, we show that the linearly interpolated process constructed from the sequence (x_n) with time frame γ converges narrowly as $\gamma \to 0$ to the DI solution in the topology of uniform convergence on compact sets. The main result of this section is Th. A.3.1, which has its own interest. To prove this theorem, we establish the tightness of the linearly interpolated process (Lem. A.3.2), then we show that the limit points coincide with the DI solution (Lem. A.3.3–A.3.5). In Sec. A.3.2, we start by establishing the inequality (A.15), which implies the tightness of the set of invariant measures $\mathcal{I}(\mathcal{P})$ in Lem A.3.7. Then, we show that the cluster points of $\mathcal{I}(\mathcal{P})$ are invariant measures for the flow induced by the DI (Lem A.3.9). In the different domains case, this lemma requires that the invariant measures of P_γ put most of their weights in a thickening of the domain $\mathcal{D}$ of order γ. This fact is established by Lem. A.3.8.

A.3.1 Dynamical behavior

For every $\gamma > 0$, we introduce the linearly interpolated process

$$\mathsf{X}_\gamma : (\mathsf{X}^{\mathbb{N}}, \mathscr{B}(\mathsf{X})^{\otimes\mathbb{N}}) \to (C(\mathbb{R}_+, \mathsf{X}), \mathscr{B}(C(\mathbb{R}_+, \mathsf{X})))$$

defined for every $x = (x_n : n \in \mathbb{N})$ in $\mathsf{X}^{\mathbb{N}}$ as

$$\mathsf{X}_\gamma(x) \, : t \mapsto x_{\lfloor \frac{t}{\gamma} \rfloor} + (t/\gamma - \lfloor t/\gamma \rfloor)(x_{\lfloor \frac{t}{\gamma} \rfloor + 1} - x_{\lfloor \frac{t}{\gamma} \rfloor}) \, .$$

This map will be referred to as the linearly interpolated process. When $x = (x_n)$ is the process with the probability measure $\mathbb{P}^{\nu,\gamma}$ defined above, the distribution of the r.v. X_γ is $\mathbb{P}^{\nu,\gamma}\mathsf{X}_\gamma^{-1}$. The set $C(\mathbb{R}_+, \mathsf{X})$ of continuous functions from $\mathbb{R}_+$ to X is equipped with the topology of uniform convergence on the compact intervals, who is known to be compatible with the distance d defined as

$$\mathsf{d}(\mathsf{x}, \mathsf{y}) := \sum_{n \in \mathbb{N}^*} 2^{-n} \left(1 \wedge \sup_{t \in [0,n]} \|\mathsf{x}(t) - \mathsf{y}(t)\| \right) .$$

If S is a subset of X and $\varepsilon > 0$, we denote by $S_\varepsilon := \{a \in \mathsf{X} : d(a, S) < \varepsilon\}$ the ε-neighborhood of S. The aim of the beginning section is to establish the following result:

Theorem A.3.1. Let Assumptions H4, H5, H7 and H8 hold true. Then, for every $\eta > 0$, for every compact set $\mathcal{K} \subset \mathsf{X}$ s.t. $\mathcal{K} \cap \mathcal{D} \neq \emptyset$,

$$\forall M \geq 0, \sup_{a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}} \mathbb{P}^{a,\gamma} \left(\mathsf{d}(\mathsf{X}_\gamma, \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), \cdot)) > \eta \right) \xrightarrow[\gamma \to 0]{} 0. \tag{A.4}$$

Choose a compact set $\mathcal{K} \subset \mathsf{X}$ s.t. $\mathcal{K} \cap \mathrm{cl}(\mathcal{D}) \neq \emptyset$. Choose $R > 0$ s.t. $\mathcal{K}$ is contained in the ball of radius R. For every $x = (x_n : n \in \mathbb{N})$ in $\mathsf{X}^{\mathbb{N}}$, define $\tau_R(x) := \inf\{n \in \mathbb{N} : x_n > R\}$ and introduce the measurable mapping $B_R : \mathsf{X}^{\mathbb{N}} \to \mathsf{X}^{\mathbb{N}}$, given by

$$B_R(x) : n \mapsto x_n \mathbb{1}_{n < \tau_R(x)} + x_{\tau_R(x)} \mathbb{1}_{n \geq \tau_R(x)} \, .$$

Consider the image measure $\bar{\mathbb{P}}^{a,\gamma} := \mathbb{P}^{a,\gamma} B_R^{-1}$, which corresponds to the law of the *truncated* process $B_R(x)$ and denote by $\bar{\mathbb{E}}^{a,\gamma}$ the corresponding mathematical expectation. The crux of the proof consists in showing that for every $\eta > 0$ and every $M > 0$,

$$\sup_{a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}} \bar{\mathbb{P}}^{a,\gamma} \left(\mathsf{d}(\mathsf{X}_\gamma, \Phi(\Pi_{\mathrm{cl}(\mathcal{D})}(a), \cdot)) > \eta \right) \xrightarrow[\gamma \to 0]{} 0. \tag{A.5}$$

Eq. (A.5) is the counterpart of Th. 3.5.1. Once it has been proven, the conclusion follows verbatim from the End of the proof of Th. 3.5.1, Chap. 3. Our aim is thus to establish Eq. (A.5). The proof follows the same steps as the proof of Th. 3.5.1 up to some confined changes. Here, the steps of the proof which do not need any modification are recalled rather briefly (we refer the reader to Chap. 3 for the details). On the other hand, the parts which require an adaptation are explicitly stated as lemmas, whose detailed proofs are provided.

Define $h_{\gamma,R}(s,x) := h_\gamma(s,x)\mathbb{1}_{\|x\| \leq R}$. First, we recall the following decomposition, established in Chap. 3:

$$\mathsf{X}_\gamma = \Pi_0 + \mathsf{G}_{\gamma,R} \circ \mathsf{X}_\gamma + \mathsf{X}_\gamma \circ M_{\gamma,R} \, ,$$

$\bar{\mathbb{P}}^{a,\gamma}$ almost surely, where $\Pi_0 : \mathsf{X}^{\mathbb{N}} \to C(\mathbb{R}_+, \mathsf{X})$, $\mathsf{G}_{\gamma,R} : C(\mathbb{R}_+, \mathsf{X}) \to C(\mathbb{R}_+, \mathsf{X})$ and $M_{\gamma,R} : \mathsf{X}^{\mathbb{N}} \to \mathsf{X}^{\mathbb{N}}$ are the mappings respectively defined by

$$\begin{aligned}
&\Pi_0(x) : t \mapsto x_0 \\
&M_{\gamma,R}(x) : n \mapsto (x_n - x_0) - \gamma \sum_{k=0}^{n-1} \int h_{\gamma,R}(s, x_k)\mu(ds) \\
&\mathsf{G}_{\gamma,R}(\mathsf{x}) : t \mapsto \int_0^t \int h_{\gamma,R}(s, \mathsf{x}(\gamma \lfloor u/\gamma \rfloor))\, \mu(ds)du \ ,
\end{aligned}$$

for every $x = (x_n : n \in \mathbb{N})$ and every $\mathsf{x} \in C(\mathbb{R}_+, \mathsf{X})$.

Lemma A.3.2. For all $\gamma \in (0, \gamma_0]$ and all $x \in \mathsf{X}^{\mathbb{N}}$, define $Z^{\gamma}_{n+1}(x) := \gamma^{-1}(x_{n+1} - x_n)$. There exists $\varepsilon > 0$ such that:

$$\sup_{n\in\mathbb{N}, a\in\mathcal{K}\cap\mathcal{D}_{\gamma M}, \gamma\in(0,\gamma_0]} \bar{\mathbb{E}}^{a,\gamma}\left(\left(\|Z^{\gamma}_n\| + \frac{\boldsymbol{d}(x_n)}{\gamma}\mathbb{1}_{\|x_n\|\le R}\right)^{1+\varepsilon}\right) < +\infty$$

Proof. Let ε be the smallest of the three constants (also named ε) in Assumptions H4, H5 and H8 respectively where $\mathcal{K}$ is the ball of center 0 and radius R. For every a, γ, the following holds for $\bar{\mathbb{P}}^{a,\gamma}$-almost all $x = (x_n : n \in \mathbb{N})$:

$$\begin{aligned}
\boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_{n+1}\|\le R} &= \boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_{n+1}\|\le R}(\mathbb{1}_{\|x_n\|\le R} + \mathbb{1}_{\|x_n\|>R}) = \boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_{n+1}\|\le R}\mathbb{1}_{\|x_n\|\le R} \\
&\le \boldsymbol{d}(x_{n+1})\mathbb{1}_{\|x_n\|\le R} \\
&= \|x_{n+1} - \Pi_{\mathcal{D}}(x_{n+1})\|\mathbb{1}_{\|x_n\|\le R} \\
&\le \|x_{n+1} - \Pi_{\mathcal{D}}(x_n)\|\mathbb{1}_{\|x_n\|\le R}\,.
\end{aligned}$$

Using the notation $\mathbb{E}^{a,\gamma}_n = \bar{\mathbb{E}}^{a,\gamma}(\,.\,|x_0, \ldots, x_n)$, we thus obtain:

$$\mathbb{E}^{a,\gamma}_n(\boldsymbol{d}(x_{n+1})^{1+\varepsilon}\mathbb{1}_{\|x_{n+1}\|\le R}) \le \int \|x_n + \gamma h_\gamma(s, x_n) - \Pi_{\mathcal{D}}(x_n)\|^{1+\varepsilon}\mathbb{1}_{\|x_n\|\le R}\, d\mu(s)\,.$$

By the convexity of $\|\cdot\|^{1+\varepsilon}$, for all $\alpha \in (0, 1)$,

$$\|x + y\|^{1+\varepsilon} = \frac{1}{\alpha^{1+\varepsilon}}\left\|\alpha x + (1-\alpha)\frac{\alpha}{1-\alpha}y\right\|^{1+\varepsilon} \le \alpha^{-\varepsilon}\|x\|^{1+\varepsilon} + (1-\alpha)^{-\varepsilon}\|y\|^{1+\varepsilon}\,.$$

Therefore, by setting $\delta_\gamma(s, x) := \|x + \gamma h_\gamma(s, x) - \Pi_{D(s)}(x)\|$,

$$\begin{aligned}
\mathbb{E}^{a,\gamma}_n(\boldsymbol{d}(x_{n+1})^{1+\varepsilon}\mathbb{1}_{\|x_{n+1}\|\le R}) \le \alpha^{-\varepsilon}&\int \delta_\gamma(s, x_n)^{1+\varepsilon}\mathbb{1}_{\|x_n\|\le R}\, d\mu(s) \\
&+ (1-\alpha)^{-\varepsilon}\int \|\Pi_{D(s)}(x_n) - \Pi_{\mathcal{D}}(x_n)\|^{1+\varepsilon}\mathbb{1}_{\|x_n\|\le R}\, d\mu(s)\,.
\end{aligned}$$

Note that for every $s \in \Xi$, $x \in \mathsf{X}$,

$$\delta_\gamma(s, x) = \|\operatorname{prox}_{\gamma g(s,\cdot)}(x - 2\gamma\nabla f_\gamma(s, x)) + \gamma\nabla f_\gamma(s, x) - \Pi_{D(s)}(x) + \operatorname{prox}_{\gamma g(s,\cdot)}(x) - \operatorname{prox}_{\gamma g(s,\cdot)}(x)\|$$

Hence,

$$\delta_\gamma(s, x) \le 3\gamma\|\nabla f_\gamma(s, x)\| + \|\operatorname{prox}_{\gamma g(\cdot,s)}(x) - \Pi_{D(s)}(x)\|$$

And, by Assumptions H4 and H5, there exists a deterministic constant $C > 0$ s.t.

$$\sup_n \int \delta_\gamma(s, x_n)^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \, d\mu(s) \le C\gamma^{1+\varepsilon} .$$

Moreover, since $\Pi_{\mathrm{cl}(D(s))}$ is a firmly non expansive operator [12, Chap. 4], it holds that for all $u \in \mathrm{cl}(\mathcal{D})$, and for μ-almost all s,

$$\|\Pi_{\mathrm{cl}(D(s))}(x_n) - u\|^2 \le \|x_n - u\|^2 - \|\Pi_{\mathrm{cl}(D(s))}(x_n) - x_n\|^2.$$

Taking $u = \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)$, we obtain that

$$\|\Pi_{\mathrm{cl}(D(s))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2 \le \boldsymbol{d}(x_n)^2 - d(x_n, D(s))^2. \tag{A.6}$$

Making use of Assumption H7, and assuming without loss of generality that $\varepsilon \le 1$, we obtain

$$\begin{aligned}\int \|\Pi_{\mathrm{cl}(D(s))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^{1+\varepsilon} \, d\mu(s) &\le \Big(\int \|\Pi_{\mathrm{cl}(D(s))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2 \, d\mu(s) \Big)^{(1+\varepsilon)/2} \\ &\le \alpha' \boldsymbol{d}(x_n)^{1+\varepsilon} ,\end{aligned}$$

for some $\alpha' \in [0, 1)$. Choosing α close enough to zero, we obtain that there exists $\rho \in [0, 1)$ such that

$$\mathbb{E}_n^{a,\gamma} \left(\frac{\boldsymbol{d}(x_{n+1})^{1+\varepsilon}}{\gamma^{1+\varepsilon}} \mathbb{1}_{\|x_{n+1}\| \le R} \right) \le \rho \frac{\boldsymbol{d}(x_n)^{1+\varepsilon}}{\gamma^{1+\varepsilon}} \mathbb{1}_{\|x_n\| \le R} + C.$$

Taking the expectation at both sides, iterating, and using the fact that $\boldsymbol{d}(x_0) = \boldsymbol{d}(a) < M\gamma$, we obtain that

$$\sup_{n \in \mathbb{N}, a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}, \gamma \in (0, \gamma_0]} \bar{\mathbb{E}}^{a,\gamma} \left(\left(\frac{\boldsymbol{d}(x_n)}{\gamma} \right)^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \right) < +\infty . \tag{A.7}$$

Since $\nabla g_\gamma(s, \cdot)$ is γ^{-1}-Lipschitz continuous, $\|h_\gamma(s, x)\| \le \|\nabla g_\gamma(s, x)\| + 3\|\nabla f_\gamma(s, x)\|$. Moreover, choosing measurably $\tilde{x} \in \mathcal{D}$ in such a way that $\|x - \tilde{x}\| \le 2\boldsymbol{d}(x)$, we obtain $\|\nabla g_\gamma(s, x)\| \le \|\partial g_0(\tilde{x}, s)\| + 2\frac{\boldsymbol{d}(x)}{\gamma}$. Therefore, there exists R' depending only on R and $\mathcal{D}$ s.t.

$$\|\nabla g_\gamma(s, x)\| \mathbb{1}_{\|x\| \le R} \le \|\partial_0 g(s, \tilde{x})\| \mathbb{1}_{\|\tilde{x}\| \le R'} + 2\frac{\boldsymbol{d}(x)}{\gamma} \mathbb{1}_{\|x\| \le R} .$$

In the following, C is a positive constant that can change from a line to another. Choosing $\varepsilon > 0$ enough small and using Assumption H4, H5 and Eq. (A.7), we have

$$\begin{aligned}\bar{\mathbb{E}}_n^{a,\gamma}(\|Z_{n+1}^\gamma\|^{1+\varepsilon}) &= \int \|h_\gamma(s, x_n)\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \, d\mu(s) \\ &\le \int \left(\|\nabla g_\gamma(s, x_n)\| + 3\|\nabla f_\gamma(s, x_n)\| \right)^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \, d\mu(s) \\ &\le C \int \|\nabla g_\gamma(s, x_n)\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} + \|\nabla f_\gamma(s, x_n)\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \, d\mu(s) \\ &\le C \int \|\partial_0 g(s, \tilde{x}_n)\|^{1+\varepsilon} \mathbb{1}_{\|\tilde{x}_n\| \le R'} \, d\mu(s) + C \int \|\nabla f_\gamma(s, x_n)\|^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \, d\mu(s) \\ &\quad + C \frac{\boldsymbol{d}(x_n)}{\gamma}^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} \\ &\le C + C \frac{\boldsymbol{d}(x_n)}{\gamma}^{1+\varepsilon} \mathbb{1}_{\|x_n\| \le R} .\end{aligned} \tag{A.8}$$

Taking expectations, the bound (3.21) is established.

□

Using Lem. 3.6.2, the uniform integrability condition (3.21) implies[1] that $\{\bar{\mathbb{P}}^{a,\gamma}\mathsf{X}_\gamma^{-1} : a \in \mathcal{K}\cap\mathcal{D}_{\gamma M}, \gamma \in (0,\gamma_0]\}$ is tight, and for any $T>0$,

$$\sup_{a\in\mathcal{K}\cap\mathcal{D}_{\gamma M}} \bar{\mathbb{P}}^{a,\gamma}(\|\mathsf{X}_\gamma \circ M_{\gamma,R}\|_{\infty,T} > \varepsilon) \xrightarrow{\gamma\to 0} 0\,, \tag{A.9}$$

where the notation $\|\mathsf{x}\|_{\infty,T}$ stands for the uniform norm of x on $[0,T]$.

Lemma A.3.3. For an arbitrary sequence (a_n,γ_n) such that $a_n \in \mathcal{K}\cap\mathcal{D}_{\gamma_n M}$ and $\gamma_n \to 0$, there exists a subsequence (still denoted as (a_n,γ_n)) such that $(a_n,\gamma_n)\to(a^*,0)$ for some $a^*\in\mathcal{K}\cap\mathrm{cl}(\mathcal{D})$, and there exist r.v. z and $(\mathsf{x}_n : n\in\mathbb{N})$ defined on some probability space $(\Omega',\mathscr{F}',\mathbb{P}')$ into $C(\mathbb{R}_+,\mathsf{X})$ s.t. x_n has the distribution $\bar{\mathbb{P}}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1}$ and $\mathsf{x}_n(\omega)\to\mathsf{z}(\omega)$ for all $\omega\in\Omega'$. Moreover, defining

$$u_n(t) := \mathsf{x}_n(\gamma_n\lfloor t/\gamma_n\rfloor)\ ,$$

the sequence (a_n,γ_n) and (x_n) can be chosen in such a way that the following holds $\mathbb{P}'$-a.e.

$$\sup_n \int_0^T \left(\frac{\boldsymbol{d}(u_n(t))}{\gamma_n}\mathbb{1}_{\|u_n(t)\|\le R}\right)^{1+\frac{\varepsilon}{2}} dt < +\infty \quad (\forall T>0)\,, \tag{A.10}$$

where $\varepsilon>0$ is the constant introduced in Lem. A.3.2.

Proof. The first point can be obtained by straightforward application of Prokhorov and Skorokhod's theorems. However, to verify the second point, we need to construct the sequences more carefully. Choose $\varepsilon>0$ as in Lem. A.3.2. We define the process $Y^\gamma : \mathsf{X}^{\mathbb{N}}\to\mathbb{R}^{\mathbb{N}}$ s.t. for every $n\in\mathbb{N}$,

$$Y_n^\gamma(x) := \sum_{k=0}^{n-1} \frac{\boldsymbol{d}(x_k)^{1+\varepsilon/2}}{\gamma^{\varepsilon/2}}\mathbb{1}_{\|x_k\|\le R}\,,$$

and we denote by $(X,Y^\gamma) : \mathsf{X}^{\mathbb{N}}\to(\mathsf{X}\times\mathbb{R})^{\mathbb{N}}$ the process given by $(X,Y^\gamma)_n(x) := (x_n, Y_n^\gamma(x))$. We define for every n, $\tilde{Z}_{n+1}^\gamma := \gamma^{-1}((X,Y^\gamma)_{n+1} - (X,Y^\gamma)_n)$. By Lem. A.3.2, it is easily seen that

$$\sup_{n\in\mathbb{N},a\in\mathcal{K}\cap\mathcal{D}_{\gamma M},\gamma\in(0,\gamma_0]} \bar{\mathbb{E}}^{a,\gamma}\left(\|\tilde{Z}_n^\gamma\|\mathbb{1}_{\|\tilde{Z}_n^\gamma\|>A}\right) \xrightarrow{A\to+\infty} 0\,.$$

We now apply Lem. 3.6.2, only replacing X by $\mathsf{X}\times\mathbb{R}$ and $\bar{\mathbb{P}}^{a,\gamma}$ by $\bar{\mathbb{P}}^{a,\gamma}(X,Y^\gamma)^{-1}$. By this lemma, the family $\{\bar{\mathbb{P}}^{a,\gamma}(X,Y^\gamma)^{-1}\overline{\mathsf{X}}_\gamma^{-1} : a\in\mathcal{K}\cap\mathcal{D}_{\gamma M},\gamma\in(0,\gamma_0]\}$ is tight, where $\overline{\mathsf{X}}_\gamma^{-1} : (\mathsf{X}\times\mathbb{R})^{\mathbb{N}}\to C(\mathbb{R}_+,\mathsf{X}\times\mathbb{R})$ is the piecewise linear interpolated process, defined in the same way as X_γ only substituting $\mathsf{X}\times\mathbb{R}$ with X in the definition. By Prokhorov's theorem, one can choose the subsequence (a_n,γ_n) s.t. $\bar{\mathbb{P}}^{a_n,\gamma_n}(X,Y^{\gamma_n})^{-1}\overline{\mathsf{X}}_{\gamma_n}^{-1}$ converges narrowly to some probability measure Υ on $\mathsf{X}\times\mathbb{R}$. By Skorokhod's theorem, we can define a stochastic process $((\mathsf{x}_n,\mathsf{y}_n) : n\in\mathbb{N})$ on some probability space $(\Omega',\mathscr{F}',\mathbb{P}')$ into $C(\mathbb{R}_+,\mathsf{X}\times\mathbb{R})$, whose distribution for a fixed n coincides with $\bar{\mathbb{P}}^{a_n,\gamma_n}(X,Y^{\gamma_n})^{-1}\overline{\mathsf{X}}_{\gamma_n}^{-1}$, and s.t. for every $\omega\in\Omega'$, $(\mathsf{x}_n(\omega),\mathsf{y}_n(\omega))\to(\mathsf{z}(\omega),\mathsf{w}(\omega))$, where (z,w) is a r.v. defined on the same space. In particular, the first marginal distribution of $\bar{\mathbb{P}}^{a_n,\gamma_n}(X,Y^{\gamma_n})^{-1}\overline{\mathsf{X}}_{\gamma_n}^{-1}$ coincides with $\bar{\mathbb{P}}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1}$. Thus, the first point is proven.

For every $\gamma\in(0,\gamma_0]$, introduce the mapping

$$\begin{aligned}\Gamma_\gamma : C(\mathbb{R}_+,\mathsf{X}) &\to C(\mathbb{R}_+,\mathbb{R})\\ \mathsf{x} &\mapsto \left(t\mapsto\int_0^t(\gamma^{-1}\boldsymbol{d}(\mathsf{x}(\gamma\lfloor u/\gamma\rfloor)))^{1+\varepsilon/2}\mathbb{1}_{\|\mathsf{x}(\gamma\lfloor u/\gamma\rfloor)\|\le R}du\right)\ .\end{aligned}$$

[1] Lem. 3.6.2 was actually shown with condition $[a\in\mathcal{K}]$ instead of $[a\in\mathcal{K}\cap\mathcal{D}_{\gamma M}]$, but the proof can be easily adapted to the latter case.

We denote by $\underline{\mathsf{X}}_\gamma^{-1} : \mathbb{R}^{\mathbb{N}} \to C(\mathbb{R}_+, \mathbb{R})$ the piecewise linear interpolated process, defined in the same way as X_γ only substituting X with $\mathbb{R}$ in the definition. It is straightforward to show that $\underline{\mathsf{X}}_\gamma \circ Y^{\gamma_n} = \Gamma_\gamma \circ \mathsf{X}_\gamma$. For every n, by definition of the couple $(\mathsf{x}_n, \mathsf{y}_n)$, the distribution under $\mathbb{P}'$ of the r.v. $\Gamma_{\gamma_n}(\mathsf{x}_n) - \mathsf{y}_n$ is equal to the distribution of $\Gamma_{\gamma_n} \circ \mathsf{X}_{\gamma_n} - \underline{\mathsf{X}}_{\gamma_n} \circ Y^{\gamma_n}$ under $\bar{\mathbb{P}}^{a_n, \gamma_n}$. Therefore, $\mathbb{P}'$-a.e. and for every n, $\mathsf{y}_n = \Gamma_{\gamma_n}(\mathsf{x}_n)$. This implies that, $\mathbb{P}'$-a.e., $\Gamma_{\gamma_n}(\mathsf{x}_n)$ converges (uniformly on compact set) to w. On that event, this implies that for every $T \geq 0$, $\Gamma_{\gamma_n}(\mathsf{x}_n)(T) \to \mathsf{w}(T)$, which is finite. Hence, $\sup_n \Gamma_{\gamma_n}(\mathsf{x}_n)(T) < \infty$ on that event, which proves the second point. □

Define

$$v_n(s,t) := -\nabla f_{\gamma_n}(s, u_n(t)) \mathbb{1}_{\|u_n(t)\| \leq R} \,.$$

and

$$w_n(s,t) := -\nabla g_{\gamma_n}(s, u_n(t) - 2\gamma \nabla f_{\gamma_n}(s, u_n(t))) \mathbb{1}_{\|u_n(t)\| \leq R} \,.$$

Thanks to the convergence (A.9), the following holds $\mathbb{P}'$-a.e.:

$$\mathsf{z}(t) = \mathsf{z}(0) + \lim_{n \to \infty} \int_0^t \int_\Xi v_n(s,u) + w_n(s,u) \, \mu(ds) \, du \qquad (\forall t \geq 0) \,. \tag{A.11}$$

We now select an $\omega \in \Omega'$ s.t. the events (A.10) and (A.11) are all realized, and omit the dependence in ω in the sequel. Otherwise stated, u_n, v_n and w_n are handled from now on as deterministic functions, and no longer as random variables. The aim of the next lemmas is to analyze the integrand $v_n(s,u) + w_n(s,u)$. Consider some $T > 0$ and let λ_T represent the Lebesgue measure on the interval $[0,T]$. To simplify notations, we set $\mathcal{L}^{1+\varepsilon}_{\mathsf{X}} := \mathcal{L}^{1+\varepsilon}(\Xi \times [0,T], \mathscr{G} \otimes \mathscr{B}([0,T]), \mu \otimes \lambda_T; \mathsf{X})$.

Lemma A.3.4. The sequences $(v_n : n \in \mathbb{N}), (w_n : n \in \mathbb{N})$ form bounded subsets of $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}}$.

Proof. By the same derivations as those leading to Eq. (A.8), there exists $C > 0$ such that

$$\int \left(\|v_n(s,t)\|^{1+\varepsilon/2} + \|w_n(s,t)\|^{1+\varepsilon/2} \right) d\mu(s) \leq C + C \frac{\boldsymbol{d}(u_n(t))^{1+\varepsilon/2}}{\gamma^{1+\varepsilon/2}} \mathbb{1}_{\|u_n(t)\| \leq R} \,.$$

The proof is concluded by applying Lem. A.3.3.

□

The sequence of mappings $((s,t) \mapsto (v_n(s,t), w_n(s,t)))$ is bounded in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}^2}$ and therefore admits a weak cluster point in that space. We denote by (v, w) such a cluster point, where $v : \Xi \times [0,T] \to \mathsf{X}$ and $w : \Xi \times [0,T] \to \mathsf{X}$. Let $H_R(s,x) := \nabla f(s,x) + \partial g(s,x))$ if $\|x\| < R$, $\{0\}$ if $\|a\| > R$, and $H_R(s,x) := \mathsf{X}$ otherwise. Denote the corresponding selection integral as $\mathsf{H}_R(a) = \int H_R(s,a) \, \mu(ds)$.

Lemma A.3.5. For every (s,t) $\mu \otimes \lambda_T$-a.e., $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(s, \,.\,))$.

Proof. To simplify notations, we now omit the dependence in (s,t) in the sequel and write $u_n := u_n(t)$, $v_n := v_n(s,t)$, $w_n := w_n(s,t)$, $h_\gamma := h_\gamma(s, \cdot)$, $\partial g := \partial g(s, \cdot)$, $\nabla f := \nabla f(s, \cdot)$, $\gamma := \gamma_n$, $\mathrm{prox}_{\gamma f} := \mathrm{prox}_{\gamma f(s,\cdot)}$, $\nabla f_\gamma := \nabla f_\gamma(s, \cdot)$, $\mathsf{z} := \mathsf{z}(t)$. Moreover, we write $\widetilde{\mathrm{prox}}_{\gamma g}(x) := \mathrm{prox}_{\gamma g(s,\cdot)}(x - 2\gamma \nabla f_\gamma(s,x))$ and $\widetilde{\nabla} g_\gamma := \nabla g_\gamma(s, x - 2\gamma \nabla f_\gamma(s,x))$, for all $x \in \mathsf{X}$.

There exists a subsequence of (γ_n) that is decreasing and such that $d_n := \sup_{t \in [0,T]} \|u_n(t) - \mathsf{z}(t)\|$ is decreasing to zero. We still denote by (γ_n) such a subsequence. The sequence

$$((v_n, w_n, \|v_n\|, \|w_n\|))_n$$

converges weakly to $(v, w, \tilde{v}, \tilde{w})$ in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}^2\times\mathbb{R}^2}$ along some subsequence (*n.b.*: compactness and sequential compactness are the same notions in the weak topology of $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}\times\mathbb{R}}$). We still denote this subsequence by $((v_n, w_n, \|v_n\|, \|w_n\|))_n$. By Mazur's theorem, there exists a function $J : \mathbb{N} \to \mathbb{N}$ and a sequence of sets of weights $\{\alpha_{k,n} : n \in \mathbb{N}, k = n \ldots, J(n) \ : \ \alpha_{k,n} \geq 0, \sum_{k=n}^{J(n)} \alpha_{k,n} = 1\}$ such that the sequence of functions

$$(\bar{v}_n, \bar{w}_n, \tilde{v}_n, \tilde{w}_n) \ : \ (s,t) \mapsto \sum_{k=n}^{J(n)} \alpha_{k,n}(v_k(s,t), w_k(s,t), \|v_k(s,t)\|, \|w_k(s,t)\|)$$

converges strongly to $(v, w, \tilde{v}, \tilde{w})$ in that space, as $n \to \infty$. Taking a further subsequence (which we still denote by $(\bar{v}_n, \bar{w}_n, \tilde{v}_n, \tilde{w}_n)$) we obtain the $\mu \otimes \lambda_T$-almost everywhere convergence of $(\bar{v}_n, \bar{w}_n, \tilde{v}_n, \tilde{w}_n)$ to $(\bar{v}, \bar{w}, \tilde{v}, \tilde{w})$. Consider a negligible set $\mathscr{N} \in \mathscr{G}\otimes\mathscr{B}([0,T])$ such that for all $(s,t) \notin \mathscr{N}$, $(\bar{v}_n, \bar{w}_n, \tilde{v}_n, \tilde{w}_n) \to (v, w, \tilde{v}, \tilde{w})$ and $\tilde{v}, \tilde{w}$ are finite.

If $\|z(t)\| = R$, obviously $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(\cdot, s))$. If $\|z(t)\| > R$, then, $(v+w)(s,t) = 0$ since $\|u_n(t)\| > R$ for n enough large. We just need to consider the case where $\|z(t)\| < R$. Besides, the condition $(\mathsf{z}(t), (v+w)(s,t)) \in \mathrm{gr}(H_R(\cdot, s))$ is equivalent to:

$$(\mathsf{z}(t), -(v+w)(s,t)) \in \mathrm{gr}(\partial(f(\cdot, s) + g(\cdot, s))) = \mathrm{gr}(\nabla f(\cdot, s) + \partial g(\cdot, s)) \,. \tag{A.12}$$

To show Eq. (A.12), consider an arbitrary $(p, q) \in \mathrm{gr}(\nabla f(\cdot, s) + \partial g(\cdot, s))$. There exists $(q_f, q_g) \in \mathsf{X}^2$ such that $q = q_f + q_g$, $(p, q_f) \in \mathrm{gr}(\nabla f(\cdot, s))$ and $(p, q_g) \in \mathrm{gr}(\partial g(\cdot, s))$.

Recall that $-h_\gamma(x) = \nabla f_\gamma(x) + \nabla g_\gamma(x - 2\gamma\nabla f_\gamma(x))$. We start by decomposing $\langle x - p, -h_\gamma(x) - q\rangle$ for any $x \in \mathsf{X}$. On the one hand,

$$\langle x + \gamma h_\gamma(x) - p, -h_\gamma(x) - q\rangle = -\gamma\langle h_\gamma(x), q\rangle - \gamma\|h_\gamma(x)\|^2 + \langle x - p, -h_\gamma(x) - q\rangle$$

On the other hand,

$$\begin{aligned}
&\langle x - \gamma\nabla f_\gamma(x) - \gamma\widetilde{\nabla} g_\gamma(x) - p, \nabla f_\gamma(x) + \widetilde{\nabla} g_\gamma(x) - (q_f + q_g)\rangle \\
=&\langle \mathrm{prox}_{\gamma f}(x) - \gamma\widetilde{\nabla} g_\gamma(x) - p, \nabla f_\gamma(x) - q_f\rangle + \langle \widetilde{\mathrm{prox}}_{\gamma g}(x) + \gamma\nabla f_\gamma(x) - p, \widetilde{\nabla} g_\gamma(x) - q_g\rangle \\
=&\langle \mathrm{prox}_{\gamma f}(x) - p, \nabla f_\gamma(x) - q_f\rangle + \langle \widetilde{\mathrm{prox}}_{\gamma g}(x) - p, \widetilde{\nabla} g_\gamma(x) - q_g\rangle \\
&- \gamma\langle\widetilde{\nabla} g_\gamma(x), \nabla f_\gamma(x) - q_f\rangle + \gamma\langle\nabla f_\gamma(x), \widetilde{\nabla} g_\gamma(x) - q_g\rangle \\
=&\langle \mathrm{prox}_{\gamma f}(x) - p, \nabla f_\gamma(x) - q_f\rangle + \langle \widetilde{\mathrm{prox}}_{\gamma g}(x) - p, \widetilde{\nabla} g_\gamma(x) - q_g\rangle \\
&+ \gamma\langle\widetilde{\nabla} g_\gamma(x), q_f\rangle - \gamma\langle\nabla f_\gamma(x), q_g\rangle.
\end{aligned}$$

Using the monotonicity of ∇f and ∂g, we finally have

$$\begin{aligned}
0 \leq&\langle x - p, -h_\gamma(x) - q\rangle \\
&- \gamma\langle\tilde{\nabla} g_\gamma(x), q_f\rangle + \gamma\langle\nabla f_\gamma(x), q_g\rangle - \gamma\langle h_\gamma(x), q\rangle.
\end{aligned} \tag{A.13}$$

As $\|\mathsf{z}\| < R$, it holds that $\|u_n\| < R$ for every n large enough. Thus, $-v_n = \nabla f_{\gamma_n}(u_n)$ and $-w_n =$

$\nabla g_{\gamma_n}(u_n - 2\gamma_n \nabla f_{\gamma_n}(u_n))$. Using (A.13) with u_n instead of x and γ_n instead of γ, we have

$$\begin{aligned}
0 \leq & \sum_{k=n}^{J(n)} \alpha_{k,n} \left(\langle z - p, -(v_k + w_k) - q \rangle + \langle u_k - z, -(v_k + w_k) - q \rangle \right) \\
& + \sum_{k=n}^{J(n)} \alpha_{k,n} \gamma_k \left(\langle w_k, q_f \rangle - \langle v_k, q_g \rangle - \langle h_{\gamma_k}(u_k), q \rangle \right) \\
\leq & \langle z - p, -(\bar{v}_n + \bar{w}_n) - q \rangle + \sum_{k=n}^{J(n)} \alpha_{k,n} d_k \left(\|v_k\| + \|w_k\| + \|q\| \right) \\
& + \sum_{k=n}^{J(n)} \alpha_{k,n} \gamma_k \left(\|w_k\| \|q_f\| + \|v_k\| \|q_g\| + \|v_k\| \|q\| + \|w_k\| \|q\| \right) \\
\leq & \langle z - p, -(\bar{v}_n + \bar{w}_n) - q \rangle + d_n \left(\tilde{v}_n + \tilde{w}_n + \|q\| \right) \\
& + \gamma_n \left(\tilde{w}_n \|q_f\| + \tilde{v}_n \|q_g\| + \tilde{v}_n \|q\| + \tilde{w}_n \|q\| \right). \qquad \text{(A.14)}
\end{aligned}$$

Letting $n \to +\infty$, since (v_n) and (w_n) are a.e bounded sequences in X, we conclude that $\langle \mathsf{z} - p, -(v + w) - q \rangle \geq 0$ a.e. As $\nabla f + \partial g \in \mathscr{M}$, this implies that $(\mathsf{z}, (v + w)) \in \nabla f + \partial g$ a.e.

□

By Lem. A.3.5 and Fubini's theorem, there is a λ_T-negligible set s.t. for every t outside this set, $v(\cdot, t)$ is an integrable selection of $H_R(\cdot, \mathsf{z}(t))$. Moreover, as v is a weak cluster point of v_n in $\mathcal{L}^{1+\varepsilon/2}_{\mathsf{X}}$, it holds that

$$\mathsf{z}(t) = \mathsf{z}(0) + \int_0^t \int_\Xi v(s,u) + w(s,u)\, \mu(ds)\, du\,, \qquad (\forall t \in [0,T])\,.$$

By the above equality, z is a solution to the DI $\dot{\mathsf{x}} \in \mathsf{H}_R(\mathsf{x})$ with initial condition $\mathsf{z}(0) = a^*$. Denoting by $\Phi_R(a^*)$ the set of such solutions, this reads $\mathsf{z} \in \Phi_R(a^*)$. As $a^* \in \mathcal{K} \cap \mathrm{cl}(\mathcal{D})$, one has $\mathsf{z} \in \Phi_R(\mathcal{K} \cap \mathrm{cl}(\mathcal{D}))$ where we use the notation $\Phi_R(S) := \cup_{a \in S} \Phi_R(a)$ for every set $S \subset \mathsf{X}$. Extending the notation $\mathsf{d}(x, \mathsf{S}) := \inf_{\mathsf{y} \in \mathsf{S}} \mathsf{d}(\mathsf{x}, \mathsf{y})$, we obtain that $\mathsf{d}(\mathsf{x}_n, \Phi_R(\mathcal{K} \cap \mathrm{cl}(\mathcal{D}))) \to 0$. Thus, for every $\eta > 0$, we have shown that $\bar{\mathbb{P}}^{a_n, \gamma_n}(\mathsf{d}(X_{\gamma_n}, \Phi_R(\mathcal{K} \cap \mathrm{cl}(\mathcal{D}))) > \eta) \to 0$ as $n \to \infty$. We have thus proven the following result:

$$\forall \eta > 0, \ \lim_{\gamma \to 0} \sup_{a \in \mathcal{K} \cap \mathcal{D}_{\gamma M}} \bar{\mathbb{P}}^{a,\gamma}(\mathsf{d}(X_\gamma, \Phi_R(\mathcal{K} \cap \mathrm{cl}(\mathcal{D}))) > \eta) = 0\,.$$

Let $T > 0$ and $R > \sup\{\|\Phi(a,t)\| : t \in [0,T], a \in \mathcal{K} \cap \mathrm{cl}(\mathcal{D})\}$ (the latter quantity being finite, see *e.g.* [36]). Consider any solution x to the DI $\dot{\mathsf{x}} \in \mathsf{H}_R(\mathsf{x})$ with initial condition $a \in \mathcal{K} \cap \mathrm{cl}(\mathcal{D})$. Consider the set $F = \{t \in [0,T], \mathsf{x}(t) = \Phi(a,t)\}$. Then, $0 \in F$. Let $\bar{t} = \sup F$ and assume that $\bar{t} < T$. Since F is closed, $\bar{t} \in F$ and we have $\|\mathsf{x}(\bar{t})\| < R$, hence there exists $\varepsilon > 0$ such that $\|\mathsf{x}(t)\| < R$ for all $t \in [\bar{t}, \bar{t} + \varepsilon]$. Then, x and $\Phi(a, \cdot)$ are solutions to the DI $\dot{\mathsf{x}} \in \mathsf{H}(\mathsf{x})$ over $[\bar{t}, \bar{t} + \varepsilon]$ and $\mathsf{x}(\bar{t}) = \Phi(a, \bar{t})$, therefore $\mathsf{x}(t) = \Phi(a,t)$ for all $t \in [\bar{t}, \bar{t} + \varepsilon]$. Hence, $\bar{t} + \varepsilon \in F$. Finally, $\bar{t} = T$ and $F = [0,T]$. By the same arguments as in [28, Section 4 - End of the proof], Th. A.3.1 follows.

A.3.2 Stability of the Markov chain

Theorem A.3.6. Assume hypotheses H1 and H2. Let $x_\star \in Z(\partial G + \nabla F)$ that admits a $\mathcal{L}^2$ representation. Then, there exists $\alpha, C > 0$ such that

$$\mathbb{E}^{\gamma,a}_n \|x_{n+1} - x_\star\|^2 \leq \|x_n - x_\star\|^2 - \alpha\gamma(F^\gamma(x_n) + G^\gamma(x_n)) + \gamma^2 C. \qquad \text{(A.15)}$$

for γ enough close to 0.

Proof. To simplify notations, we now omit the dependence in (s,t) in the sequel and write $u_n := u_n(t)$, $v_n := v_n(s,t)$, $w_n := w_n(s,t)$, $h_\gamma := h_\gamma(s,\cdot)$, $\partial g := \partial g(s,\cdot)$, $\nabla f := \nabla f(s,\cdot)$, $\gamma := \gamma_n$, $\mathrm{prox}_{\gamma f} := \mathrm{prox}_{\gamma f(s,\cdot)}$, $\nabla f_\gamma := \nabla f_\gamma(s,\cdot)$, $\mathsf{z} := \mathsf{z}(t)$. Moreover, we write $\widetilde{\mathrm{prox}}_{\gamma g}(x) := \mathrm{prox}_{\gamma g(s,\cdot)}(x - 2\gamma\nabla f_\gamma(s,x))$ and $\widetilde{\nabla} g_\gamma := \nabla g_\gamma(s, x - 2\gamma\nabla f_\gamma(s,x))$, for all $x \in \mathsf{X}$.

By assumption, there exists a $\mathcal{L}^2$ representation (φ,ψ) of $x_\star$. We write

$$\begin{aligned}\langle \nabla f_\gamma(x), x - x_\star\rangle &= \langle \nabla f_\gamma(x) - \psi, x - x_\star\rangle + \langle \psi, x - x_\star\rangle \\ &= \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle + \langle \nabla f_\gamma(x) - \psi, \gamma\nabla f_\gamma(x)\rangle \\ &\quad + \langle \psi, x - x_\star\rangle \\ &= \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle - \gamma\langle \psi, \nabla f_\gamma(x)\rangle \\ &\quad + \langle \varphi, x - x_\star\rangle + \gamma\|\nabla f_\gamma(x)\|^2.\end{aligned}$$

We also write

$$\begin{aligned}\langle \widetilde{\nabla} g_\gamma(x), x - x_\star\rangle &= \langle \widetilde{\nabla} g_\gamma(x) - \varphi, x - x_\star\rangle + \langle \varphi, x - x_\star\rangle \\ &= \langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle + \langle \widetilde{\nabla} g_\gamma(x) - \varphi, x - \widetilde{\mathrm{prox}}_{\gamma g}(x) - 2\gamma\nabla f_\gamma(x)\rangle \\ &\quad + \langle \varphi, x - x_\star\rangle + \langle \widetilde{\nabla} g_\gamma(x) - \varphi, 2\gamma\nabla f_\gamma(x)\rangle \\ &= \langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle - \gamma\langle \varphi, \widetilde{\nabla} g_\gamma(x)\rangle \\ &\quad + 2\gamma\langle \widetilde{\nabla} g_\gamma(x), \nabla f_\gamma(x)\rangle + \langle \varphi, x - x_\star\rangle + \gamma\|\widetilde{\nabla} g_\gamma(x)\|^2 - 2\gamma\langle \varphi, \nabla f_\gamma(x)\rangle.\end{aligned}$$

Hence,

$$\begin{aligned}&\langle \nabla f_\gamma(x) + \widetilde{\nabla} g_\gamma(x), x - x_*\rangle \\ =&\langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle + \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle \\ &+ \gamma\|\widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\|^2 - \gamma\left\{\langle \varphi + \psi, \nabla f_\gamma(x)\rangle + \langle \varphi, \widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\rangle\right\} \\ &+ \langle \varphi + \psi, x - x_\star\rangle \end{aligned} \tag{A.16}$$

By expanding

$$\|x_{n+1} - x_\star\|^2 = \|x_n - x_\star\|^2 + 2\langle x_{n+1} - x_n, x_n - x_\star\rangle + \|x_{n+1} - x_n\|^2,$$

we obtain

$$\begin{aligned}\|x_{n+1} - x_\star\|^2 = \|x_n - x_\star\|^2 - 2\gamma\langle \widetilde{\nabla} g_\gamma(x_n), x_n - x_\star\rangle - 2\gamma\langle \nabla f_\gamma(x_n), x_n - x_\star\rangle \\ + \gamma^2\|\widetilde{\nabla} g_\gamma(x_n) + \nabla f_\gamma(x_n)\|^2.\end{aligned} \tag{A.17}$$

Using (A.16), we obtain

$$\begin{aligned}\|x_{n+1} - x_\star\|^2 = \|x_n - x_\star\|^2 \\ - 2\gamma\left\{\langle \widetilde{\nabla} g_\gamma(x_n) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x_n) - x_\star\rangle + \langle \nabla f_\gamma(x_n) - \psi, \mathrm{prox}_{\gamma f}(x_n) - x_\star\rangle\right\} \\ - \gamma^2\|\widetilde{\nabla} g_\gamma(x_n) + \nabla f_\gamma(x_n)\|^2 + 2\gamma^2\left\{\langle \varphi + \psi, \nabla f_\gamma(x_n)\rangle + \langle \varphi, \widetilde{\nabla} g_\gamma(x_n) + \nabla f_\gamma(x_n)\rangle\right\} \\ - 2\gamma\langle \varphi + \psi, x_n - x_\star\rangle\end{aligned}$$

where we used $\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle$. Then, since $2\langle a, b\rangle \leq \|a\|^2/2 + 2\|b\|^2$,

$$\begin{aligned}\|x_{n+1} - x_\star\|^2 \leq \|x_n - x_\star\|^2 \\ - 2\gamma \left\{\langle \widetilde{\nabla} g_\gamma(x_n) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x_n) - x_\star\rangle + \langle \nabla f_\gamma(x_n) - \psi, \mathrm{prox}_{\gamma f}(x_n) - x_\star\rangle\right\} \\ - \gamma^2/2 \|\widetilde{\nabla} g_\gamma(x_n) + \nabla f_\gamma(x_n)\|^2 + \gamma^2/2 \|\nabla f_\gamma(x_n)\|^2 \\ + 2\gamma^2\|\varphi\|^2 + 2\gamma^2\|\varphi + \psi\|^2 - 2\gamma\langle \varphi + \psi, x_n - x_\star\rangle.\end{aligned} \tag{A.18}$$

Observe that the term between the braces at the right hand side of the last inequality is nonnegative thanks to the monotonicity of $\nabla f(\cdot, s)$ and $\partial g(\cdot, s)$.

Let $x \in \mathrm{X}$. By the convexity of g_γ and f_γ, we have

$$g_\gamma(x - 2\gamma\nabla f_\gamma(x)) - g_\gamma(x_\star) \leq \langle \widetilde{\nabla} g_\gamma(x), x - 2\gamma\nabla f_\gamma(x) - x_\star\rangle \tag{A.19}$$

and

$$f_\gamma(x) - f_\gamma(x_\star) \leq \langle \nabla f_\gamma(x), x - x_\star\rangle. \tag{A.20}$$

Using the $1/\gamma$-Lipschitz continuity of ∇g_γ we have

$$g_\gamma(x) - g_\gamma(x - 2\gamma\nabla f_\gamma(x)) \leq \langle \widetilde{\nabla} g_\gamma(x), 2\gamma\nabla f_\gamma(x)\rangle + 2\gamma\|\nabla f_\gamma(x)\|^2. \tag{A.21}$$

Summing the inequalities (A.19), (A.20) and (A.21) we obtain

$$f_\gamma(x) - f_\gamma(x_*) + g_\gamma(x) - g_\gamma(x_*) \leq \langle \nabla f_\gamma(x) + \widetilde{\nabla} g_\gamma(x), x - x_*\rangle + 2\gamma\|\nabla f_\gamma(x)\|^2 \tag{A.22}$$

Using (A.16),

$$\begin{aligned}& f_\gamma(x) - f_\gamma(x_*) + g_\gamma(x) - g_\gamma(x_*) \\ \leq\ & 2\gamma\|\nabla f_\gamma(x)\|^2 \\ & + \langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle + \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle \\ & + \gamma\|\widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\|^2 - \gamma\left\{\langle \varphi + \psi, \nabla f_\gamma(x)\rangle + \langle \varphi, \widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\rangle\right\} \\ & + \langle \varphi + \psi, x - x_\star\rangle \\ \leq\ & \frac{3}{2}\gamma\|\nabla f_\gamma(x)\|^2 \\ & + \langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle + \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle \\ & + \frac{3}{2}\gamma\|\widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\|^2 + \langle \varphi + \psi, x - x_\star\rangle + \frac{\gamma}{2}\|\varphi\|^2 + \frac{\gamma}{2}\|\varphi + \psi\|^2 \\ \leq\ & -\frac{3}{2}\gamma\|\nabla f_\gamma(x)\|^2 + 3\left\{\gamma\|\nabla f_\gamma(x)\|^2 - \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle\right\} \\ & + 6\left\{\langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle + \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle\right\} \\ & + \frac{3}{2}\gamma\|\widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\|^2 + \langle \varphi + \psi, x - x_\star\rangle + \frac{\gamma}{2}\|\varphi\|^2 + \frac{\gamma}{2}\|\varphi + \psi\|^2,\end{aligned}$$

since $\langle \widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle$ and $\langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle$ are nonnegative. Using $\psi = \nabla f(x_\star)$, $\nabla f_\gamma(x) = \nabla f(\mathrm{prox}_{\gamma f}(x))$ and Assumption H2, there exists by Baillon-Haddad theorem $c > 0$ such that $c\|\nabla f_\gamma(x) - \psi\|^2 \leq \langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle$. Then,

$$-\langle \nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle \leq -c\|\nabla f_\gamma(x) - \psi\|^2 \leq -c/2\|\nabla f_\gamma(x)\|^2 + c\|\psi\|^2.$$

If $\gamma < c/2$ we finally have

$$\begin{aligned}
&f_\gamma(x) - f_\gamma(x_*) + g_\gamma(x) - g_\gamma(x_*) \\
\leq& -\frac{3}{2}\gamma\|\nabla f_\gamma(x)\|^2 + 3c\|\psi\|^2 \\
&+ 6\left\{\langle\widetilde{\nabla} g_\gamma(x) - \varphi, \widetilde{\mathrm{prox}}_{\gamma g}(x) - x_\star\rangle + \langle\nabla f_\gamma(x) - \psi, \mathrm{prox}_{\gamma f}(x) - x_\star\rangle\right\} \\
&+ \frac{3}{2}\gamma\|\widetilde{\nabla} g_\gamma(x) + \nabla f_\gamma(x)\|^2 + \langle\varphi + \psi, x - x_\star\rangle + \frac{\gamma}{2}\|\varphi\|^2 + \frac{\gamma}{2}\|\varphi + \psi\|^2.
\end{aligned}$$

Using A.18, there exists $\alpha, C, C' > 0$ such that

$$\begin{aligned}
\|x_{n+1} - x_\star\|^2 \leq& \|x_n - x_\star\|^2 \\
&- \alpha\gamma\left\{f_\gamma(x_n) + g_\gamma(x_n) - f_\gamma(x_\star) - g_\gamma(x_\star)\right\} \\
&+ C\gamma^2\left\{\|\varphi\|^2 + \|\psi\|^2 + \|\varphi + \psi\|^2\right\} + C'\gamma\langle\varphi + \psi, x_n - x_\star\rangle.
\end{aligned}$$

Taking the conditional expectation, the last inner product vanishes, and we get the result. □

Lemma A.3.7. If Eq (A.15) hold, then the set of measures $\mathcal{I}(\mathcal{P})$ is tight.

Proof. The following inequalities hold $F^{\gamma_0}(x) + G^{\gamma_0}(x) \leq F^{\gamma}(x) + G^{\gamma}(x) \leq F^{\gamma'}(x) + G^{\gamma'}(x) \leq F(x) + G(x)$ for all $0 \leq \gamma' \leq \gamma \leq \gamma_0$. Moreover H3 $\iff F^{\gamma_0} + G^{\gamma_0}$coercive $\iff F^{\gamma_0} + G^{\gamma_0}$ coercive (see End of the proof of Prop. 4.4.1, Chap. 4). Hence, condition (PH) in Chap. 3 is satisfied. The conclusion follows from Prop. 3.8.7. □

Lemma A.3.8. Let Assumptions H7, H6, and H3 hold true. Then, for all $\varepsilon > 0$, there exists $M > 0$ such that

$$\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi((\mathcal{D}_{M\gamma})^c) \leq \varepsilon.$$

Proof. We start by writing

$$\boldsymbol{d}(x_{n+1}) \leq \|x_{n+1} - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\| \leq \|x_{n+1} - \Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n)\| + \|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|.$$

On the one hand, we have by Assumption H6 and the nonexpansiveness of the resolvent that

$$\begin{aligned}
\mathbb{E}_n^{a,\gamma}\|x_{n+1} - \Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n)\| &\leq \mathbb{E}_n^{a,\gamma}\|\mathrm{prox}_{\gamma g(\cdot,\xi_{n+1})}(x_n) - \Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n)\| + \gamma\mathbb{E}_n^{a,\gamma}\|\nabla f_\gamma(\xi_{n+1}, x_n)\| \\
&\leq C\gamma(1 + F^\gamma(x_n) + G^\gamma(x_n))\,,
\end{aligned}$$

On the other hand, since

$$\|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2 \leq \boldsymbol{d}(x_n)^2 - d(x_n, D(\xi_{n+1}))^2 \quad \text{(see (A.6))},$$

we can make use of Assumption H7 to obtain

$$\mathbb{E}_n^{a,\gamma}\|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\| \leq (\mathbb{E}_n^{a,\gamma}\|\Pi_{\mathrm{cl}(D(\xi_{n+1}))}(x_n) - \Pi_{\mathrm{cl}(\mathcal{D})}(x_n)\|^2)^{1/2} \leq \rho\boldsymbol{d}(x_n)\,,$$

where $\rho \in [0, 1)$. We therefore obtain that $\mathbb{E}_n^{a,\gamma}\boldsymbol{d}(x_{n+1}) \leq \rho\boldsymbol{d}(x_n) + C\gamma(1 + F^\gamma(x_n) + G^\gamma(x_n))$. By iterating, we end up with the inequality

$$\mathbb{E}^{a,\gamma}(\boldsymbol{d}(x_{n+1})) \leq \rho^{n+1}\boldsymbol{d}(a) + C\gamma\sum_{k=0}^{n}\rho^{n-k}(1 + \mathbb{E}^{a,\gamma}(F^\gamma(x_k) + G^\gamma(x_k))). \tag{A.23}$$

From Th. A.3.6 we have

$$\mathbb{E}_n^{\gamma,a}\|x_{n+1}-x_\star\|^2 \le \|x_n-x_\star\|^2 - \alpha\gamma(F^\gamma(x_n)+G^\gamma(x_n)) + \gamma^2 C. \tag{A.24}$$

for all $\gamma \in (0,\gamma_0]$. Using Prop. 3.8.7, it implies

$$\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi(F^{\gamma_0}+G^{\gamma_0}) < \infty.$$

Recall that $q \in \Gamma_0(\mathsf{X})$ is coercive if and only if $\liminf_{\|x\|\to+\infty} q(x)/\|x\| > 0$ ([12, Prop. 11.11 and 11.12]). Noting that $\boldsymbol{d}(x) \le \|x\| + \|\Pi_{\mathrm{cl}(\mathcal{D})}(0)\|$, we obtain that $\sup_{\gamma\in(0,\gamma_0]}\sup_{\pi\in\mathcal{I}(P_\gamma)}\pi(\boldsymbol{d}) < \infty$ using Assumption H3. Moreover, with a small adaptation of the proof of Prop. 3.8.7 to the case of Lem. A.3.6, we can show the slightly stronger result that $\sup_{\gamma\in(0,\gamma_0]}\sup_{\pi\in\mathcal{I}(P_\gamma)}\pi(F^\gamma+G^\gamma) < \infty$. Let $\gamma \in (0,\gamma_0]$ and $\pi \in \mathcal{I}(P_\gamma)$. We can integrate w.r.t π in A.23 to obtain

$$\pi(\boldsymbol{d}) \le \rho^{n+1}\pi(\boldsymbol{d}) + C\gamma\sum_{k=0}^{n}\rho^{n-k}(1+\pi(F^\gamma+G^\gamma)).$$

Using Markov's inequality, we have for every $n \in \mathbb{N}$,

$$\pi((\mathcal{D}_{M\gamma})^c) \le \frac{\pi(\boldsymbol{d})}{M\gamma} \le \frac{\rho^{n+1}}{M\gamma}\pi(\boldsymbol{d}) + \frac{C}{M}\sum_{k=0}^{n}\rho^{n-k}(1+\pi(F^\gamma+G^\gamma)) \le \frac{\rho^{n+1}C}{M\gamma} + \frac{C}{M}.$$

By making $n \to \infty$, we obtain that $\pi((\mathcal{D}_{M\gamma})^c) \le C/M$, and the proof is concluded by taking M as large as required. □

Lemma A.3.9. Let the assumptions of the statement of Th. A.3.1 hold true. Assume that for all $\varepsilon > 0$, there exists $M > 0$ such that

$$\sup_{\gamma\in(0,\gamma_0]} \sup_{\pi\in\mathcal{I}(P_\gamma)} \pi((\mathcal{D}_{M\gamma})^c) \le \varepsilon. \tag{A.25}$$

Then, as $\gamma \to 0$, any cluster point of $\mathcal{I}(\mathcal{P})$ is an element of $\mathcal{I}(\Phi)$.

Proof. The proof is verbatim the same that the proof of Lem. 4.6.2. □

A.3.3 End of the proof

Assume H1-H6. By Lem. A.3.7, $\bigcup_{\gamma\in(0,\gamma_0]}\mathcal{I}(P_\gamma)$ is tight and by Lem. A.3.8 and Lem. A.3.9 any cluster point of $\mathcal{I}(\mathcal{P})$ is an element of $\mathcal{I}(\Phi)$ as $\gamma \to 0$. The rest of the proof follows word-for-word from Sec. 3.8.4.

Bibliography

[1] A. Alacaoglu, Q. Tran-Dinh, O. Fercoq, and V. Cevher. Smooth primal-dual coordinate descent algorithms for nonsmooth convex optimization. In *NIPS*, pages 5854–5863, 2017.

[2] Y. F. Atchade, G. Fort, and E. Moulines. On stochastic proximal gradient algorithms. *ArXiv e-prints, 1402.2365*, February 2014.

[3] Y. F Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.*, 18(1):310–342, 2017.

[4] H. Attouch. Familles d'opérateurs maximaux monotones et mesurabilité. *Ann. Mat. Pura Appl.*, 120(1):35–111, 1979.

[5] H. Attouch, J. Bolte, and B. F Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Math. Program.*, 137(1-2):91–129, 2013.

[6] J.-P. Aubin and A. Cellina. *Differential inclusions*, volume 264 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1984. Set-valued maps and viability theory.

[7] J.-P. Aubin, H. Frankowska, and A. Lasota. Poincaré's recurrence theorem for set-valued dynamical systems. *Ann. Polon. Math.*, 54(1):85–91, 1991.

[8] U. Ayesta, M. Erausquin, M. Jonckheere, and I. M. Verloop. Scheduling in a random environment: stability and asymptotic optimality. *IEEE/ACM Trans. Netw.*, 21(1):258–271, 2013.

[9] A. Barbero and S. Sra. Modular proximal optimization for multidimensional total-variation regularization. *arXiv preprint arXiv:1411.0589*, 2014.

[10] H. H. Bauschke and J. M. Borwein. Legendre functions and the method of random bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.

[11] H. H. Bauschke, J. M. Borwein, and W. Li. Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization. *Math. Program.*, 86(1):135–160, 1999.

[12] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.

[13] W. Ben-Ameur, P. Bianchi, and J. Jakubowicz. Robust distributed consensus using total variation. *IEEE Trans. Automat. Contr.*, 61(6):1550–1564, 2016.

[14] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin, 1999.

[15] M. Benaïm and M. W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *J. Dynam. Differential Equations*, 8(1):141–176, 1996.

[16] M. Benaïm and M. W. Hirsch. Stochastic approximation algorithms with constant step size whose average is cooperative. *Ann. Appl. Probab.*, 9(1):216–241, 1999.

[17] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1):328–348 (electronic), 2005.

[18] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. II. Applications. *Math. Oper. Res.*, 31(4):673–695, 2006.

[19] M. Benaïm and S. J. Schreiber. Ergodic properties of weak asymptotic pseudotrajectories for semiflows. *J. Dynam. Differential Equations*, 12(3):579–598, 2000.

[20] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

[21] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Math. Program.*, 129(2, Ser. B):163–195, 2011.

[22] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM J. Optim.*, 26(4):2235–2260, 2016.

[23] P. Bianchi, G. Fort, and W. Hachem. Performance of a distributed stochastic approximation algorithm. *IEEE Trans. Inf. Theory*, 59(11):7405–7418, 2013.

[24] P. Bianchi and W. Hachem. Dynamical behavior of a stochastic Forward-Backward algorithm using random monotone operators. *J. Optim. Theory Appl.*, 171(1):90–120, 2016.

[25] P. Bianchi, W. Hachem, and A. Salim. Building stochastic optimization algorithms with random monotone operators. In *EUCCO*, 2016.

[26] P. Bianchi, W. Hachem, and A. Salim. Convergence d'un algorithme du gradient proximal stochastique à pas constant et généralisation aux opérateurs monotones aléatoires. In *GRETSI*, 2017.

[27] P. Bianchi, W. Hachem, and A. Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *J. Convex Anal.*, 2019.

[28] P. Bianchi, W. Hachem, and A. Salim. Constant step stochastic approximations involving differential inclusions: Stability, long-run convergence and applications. *Stochastics*, 2019.

[29] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[30] J. Bolte, T.-P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[31] V. S. Borkar. *Stochastic approximation. A dynamical systems viewpoint*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008.

[32] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3. Springer, New York, second edition, 2006. Theory and examples.

[33] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT'2010*, pages 177–186, 2010.

[34] L. Bottou, F. E Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[36] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.

[37] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable markowitz portfolios. *Proc. Natl. Acad. Sci. USA*, 106(30):12267–12272, 2009.

[38] R. E. Bruck, Jr. Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *J. Funct. Anal.*, 18:15–26, 1975.

[39] R. H. Byrd, P. Lu, J. Nocedal, and C. Y. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, 1995.

[40] C. Castaing and M. Valadier. *Convex analysis and measurable multifunctions*. Lecture Notes in Mathematics, Vol. 580. Springer-Verlag, Berlin-New York, 1977.

[41] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9:263–340, 2010.

[42] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[43] S. Chen, A. Sandryhaila, G. Lederman, Z. Wang, J. MF Moura, P. Rizzo, J. Bielak, J. H Garrett, and J. Kovačević. Signal inpainting on graphs via total variation minimization. In *ICASSP*, pages 8267–8271, 2014.

[44] G. Chierchia, A. Cherni, E. Chouzenoux, and J.-C. Pesquet. Approche de Douglas-Rachford aléatoire par blocs appliquée à la régression logistique parcimonieuse. In *GRETSI*, pages 1–4, 2017.

[45] F. RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[46] P. L. Combettes. Iterative construction of the resolvent of a sum of maximal monotone operators. *J. Convex Anal.*, 16(4):727–748, 2009.

[47] P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.*, 25(2):1221–1248, 2015.

[48] P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure Appl. Funct. Anal.*, 1(1):13–37, 2016.

[49] P. L. Combettes and J.-C. Pesquet. Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration. In *EUSIPCO*, pages 1813–1817, 2016.

[50] L. Condat. A direct algorithm for 1D total variation denoising. *IEEE Signal Process. Lett.*, 20(11):1054–1057, 2013.

[51] L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

[52] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Stat.*, pages 1–48, 2001.

[53] J. de Vries. *Elements of topological dynamics*, volume 257 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1993.

[54] B. Delyon. Stochastic Approximation with Decreasing Gain: Convergence and Asymptotic Theory. *Unpublished Lecture Notes, http://perso.univ-rennes1.fr/bernard.delyon/as_cours.ps*, 2000.

[55] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

[56] J. Eckstein and D. P. Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.*, 55(1):293–318, 1992.

[57] A. El Alaoui, X. Cheng, A. Ramdas, M. J Wainwright, and M. I Jordan. Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. In *COLT*, pages 879–906, 2016.

[58] M. Faure and G. Roth. Stochastic approximations of set-valued dynamical systems: convergence with positive probability to an attractor. *Math. Oper. Res.*, 35(3):624–640, 2010.

[59] M. Faure and G. Roth. Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems. *Stoch. Dyn.*, 13(1):1250011, 23, 2013.

[60] J.-C. Fort and G. Pagès. Asymptotic behavior of a Markovian stochastic algorithm with constant step. *SIAM J. Control Optim.*, 37(5):1456–1482 (electronic), 1999.

[61] N. Gast and B. Gaujal. Markov chains with discontinuous drifts have differential inclusion limits. *Perform. Eval.*, 69(12):623 – 642, 2012.

[62] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975.

[63] D. Hallac, J. Leskovec, and S. Boyd. Network lasso: Clustering and optimization in large graphs. In *SIGKDD*, pages 387–396, 2015.

[64] R.Z. Ha'sminskii. The average principle for parabolic and elliptic differential equations and Markov processes with small diffusions. *Theor. Probab. Appl.*, 8:1–21, 1963.

[65] F. Hiai and H. Umegaki. Integrals, conditional expectations, and martingales of multivalued functions. *J. Multivar. Anal.*, 7(1):149 – 182, 1977.

[66] J.-B. Hiriart-Urruty. *About properties of the mean value functional and of the continuous infimal convolution in stochastic convex analysis*, pages 763–789. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

[67] J.-B. Hiriart-Urruty. *Contributions à la programmation mathématique: cas déterministe et stochastique*. Université de Clermont-Ferrand II, Clermont-Ferrand, 1977. Thèse présentée à l'Université de Clermont-Ferrand II pour obtenir le grade de Docteur ès Sciences Mathématiques, Série E, No. 247.

[68] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Soc. Networks*, 5(2):109–137, 1983.

[69] J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. In *COLT*, pages 1115–1146, 2016.

[70] S. Jegelka, F. Bach, and S. Sra. Reflection methods for user-friendly submodular optimization. In *NIPS*, pages 1313–1321, 2013.

[71] N. A Johnson. A dynamic programming algorithm for the fused lasso and ℓ^0-segmentation. *J. Comput. Graph. Stat.*, 22(2):246–260, 2013.

[72] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *ECML-PKDD*, pages 795–811. Springer, 2016.

[73] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[74] Jure L. and Andrej K. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[75] L. Landrieu and G. Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions. In *AISTATS*, pages 1384–1393, 2016.

[76] P.-J. Laurent. *Approximation et optimisation*. Hermann, Paris, 1972. Collection Enseignement des Sciences, No. 13.

[77] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[78] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[79] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Contr.*, 22(4):551–575, 1977.

[80] S. Majewski, B. Miasojedow, and E. Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

[81] E. Mammen and S. Van de Geer. Locally adaptive regression splines. *Ann. Stat.*, 25(1):387–413, 1997.

[82] B. Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4(R3):154–158, 1970.

[83] P. Mertikopoulos, H. Zenati, B. Lecouat, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[84] W. Miller and E. Akin. Invariant measures for set-valued dynamical systems. *Trans. Amer. Math. Soc.*, 351(3):1203–1225, 1999.

[85] G. J Minty et al. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

[86] I. Molchanov. *Theory of random sets*. Probability and its Applications (New York). Springer-Verlag London, Ltd., London, 2005.

[87] J. Moreau. Fonctions convexes duales et points proximaux dans un espace Hilbertien. *CR Acad. Sci. Paris Ser. A Math.*, 255:2897–2899, 1965.

[88] E. Moulines and F. R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS*, pages 451–459, 2011.

[89] R. Mourya, P. Bianchi, A. Salim, and C. Richard. An adaptive distributed asynchronous algorithm with application to target localization. In *CAMSAP 2017*, pages 1–5. IEEE, 2017.

[90] I. Necoara, P. Richtarik, and A. Patrascu. Randomized projection methods for convex feasibility problems: conditioning and convergence rates. *arXiv preprint arXiv:1801.04873*, 2018.

[91] J. Neveu. *Bases mathématiques du calcul des probabilités*. Masson et Cie, Éditeurs, Paris, 1964.

[92] H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *ICML*, pages 80–88, 2013.

[93] O. H. M. Padilla, J. Sharpnack, and J. G Scott. The dfs fused lasso: Linear-time denoising over general graphs. *J. Mach. Learn. Res.*, 18(1):6410–6445, 2017.

[94] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 72(2):383–390, 1979.

[95] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization. *J. Mach. Learn. Res*, 2017.

[96] J. Peypouquet and S. Sorin. Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time. *J. Convex Anal.*, 17(3-4):1113–1163, 2010.

[97] R. R. Phelps. Lectures on maximal monotone operators. *Extracta Math.*, 12(3):193–230, 1997.

[98] S. J Reddi, S. Sra, B. Poczos, and A. J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *NIPS*, pages 1145–1153, 2016.

[99] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22(3):400–407, 1951.

[100] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.

[101] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[102] R. T. Rockafellar and R. J.-B. Wets. On the interchange of subdifferentiation and conditional expectations for convex functionals. *Stochastics*, 7(3):173–182, 1982.

[103] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[104] L. Rosasco, S. Villa, and B. C. Vũ. Convergence of stochastic proximal gradient algorithm. *arXiv preprint arXiv:1403.5074*, 2014.

[105] L. Rosasco, S. Villa, and B. C. Vũ. Stochastic inertial primal-dual algorithms. *arXiv preprint arXiv:1507.00852*, 2015.

[106] L. Rosasco, S. Villa, and B. C. Vũ. A stochastic inertial forward–backward splitting algorithm for multivariate monotone inclusions. *Optimization*, 65(6):1293–1314, 2016.

[107] G. Roth and W. H Sandholm. Stochastic approximations with constant step size and differential inclusions. *SIAM J. Control Optim.*, 51(1):525–555, 2013.

[108] E. K Ryu and S. Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent.

[109] A. Salim, P. Bianchi, and W. Hachem. Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs. In *CAp*, 2017.

[110] A. Salim, P. Bianchi, and W. Hachem. A stochastic Douglas-Rachford algorithm with constant step size. Technical report, see https://adil-salim.github.io/Research, 2017.

[111] A. Salim, P. Bianchi, and W. Hachem. A constant step stochastic douglas-rachford algorithm with application to non separable regularizations. In *ICASSP*, pages 2886–2890, 2018.

[112] A. Salim, P. Bianchi, and W. Hachem. A splitting algorithm for minimization under stochastic linear constraints. In *ISMP*, 2018.

[113] A. Salim, P. Bianchi, and W. Hachem. Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs. *IEEE Trans. Automat. Contr.*, 2019.

[114] A. Salim, P. Bianchi, W. Hachem, and J. Jakubowicz. A stochastic proximal point algorithm for total variation regularization over large scale graphs. In *CDC*, pages 4490–4495, 2016.

[115] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

[116] Z. Shi and R. Liu. Online and stochastic Douglas-Rachford splitting method for large scale machine learning. *arXiv preprint arXiv:1308.4757*, 2013.

[117] D. A Spielman. Algorithms, graph theory, and linear equations in laplacian matrices. In *Proc. ICM*, volume 4, pages 2698–2722, 2010.

[118] D. A Spielman and S.-H. Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.*, 35(3):835–885, 2014.

[119] W. Tansey and J. G Scott. A fast and flexible algorithm for the graph-fused lasso. *arXiv preprint arXiv:1505.06475*, 2015.

[120] R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.*, 42(1):285–323, 2014.

[121] P. Toulis, E. M Airoldi, et al. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Stat.*, 45(4):1694–1727, 2017.

[122] P. Toulis, T. Horel, and E. M Airoldi. Stable robbins-monro approximations through stochastic proximal updates. *arXiv preprint arXiv:1510.00967*, 2015.

[123] N. K Vishnoi. Laplacian solvers and their algorithmic applications. *Theor. Comput. Sci.*, 8(1-2):1–141, 2012.

[124] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Appl. Math.*, 38(3):667–681, 2013.

[125] D. W. Walkup and R. J.-B. Wets. Stochastic programs with recourse. II: On the continuity of the objective. *SIAM J. Appl. Math.*, 17:98–103, 1969.

[126] M. Wang and D. P. Bertsekas. Incremental constraint projection methods for variational inequalities. *Math. Program.*, 150(2, Ser. A):321–363, 2015.

[127] Y.-X. Wang, J. Sharpnack, A. Smola, and R. J Tibshirani. Trend filtering on graphs. *J. Mach. Learn. Res.*, 17(105):1–41, 2016.

[128] V. G Yaji and S. Bhatnagar. Stochastic recursive inclusions with non-additive iterate-dependent markov noise. *Stochastics*, 90(3):330–363, 2018.

[129] J. Yoon and S. J. Hwang. Combined group and exclusive sparsity for deep neural networks. In *ICML*, pages 3958–3966, 2017.

[130] K. Yosida. Functional analysis, berlin, 1965. *Google Scholar*, page 126, 1967.

[131] H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *NIPS*, pages 1427–1437, 2017.

[132] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. In *NIPS*, pages 352–360, 2011.

[133] A. Yurtsever, B. C. Vũ, and V. Cevher. Stochastic three-composite convex minimization. In *NIPS*, pages 4329–4337, 2016.

[134] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.

université PARIS-SACLAY
DÉPARTEMENT
Sciences et Technologies de l'Information et de la Communication

Titre : Opérateurs monotones aléatoires et application à l'optimisation stochastique

Mots clés : Approximation stochastique, optimisation, apprentissage automatique, traitement du signal

Résumé : Cette thèse porte essentiellement sur l'étude d'algorithmes d'optimisation. Les problèmes de programmation intervenant en apprentissage automatique ou en traitement du signal sont dans beaucoup de cas composites, c'est-à-dire qu'ils sont contraints ou régularisés par des termes non lisses. Les méthodes proximales sont une classe d'algorithmes très efficaces pour résoudre de tels problèmes. Cependant, dans les applications modernes de sciences des données, les fonctions à minimiser se représentent souvent comme une espérance mathématique, difficile ou impossible à évaluer. C'est le cas dans les problèmes d'apprentissage en ligne, dans les problèmes mettant en jeu un grand nombre de données ou dans les problèmes de calcul distribué. Pour résoudre ceux-ci, nous étudions dans cette thèse des méthodes proximales stochastiques, qui adaptent les algorithmes proximaux aux cas de fonctions écrites comme une espérance. Les méthodes proximales stochastiques sont d'abord étudiées à pas constant, en utilisant des techniques d'approximation stochastique. Plus précisément, la méthode de l'Equation Differentielle Ordinaire est adaptée au cas d'inclusions differentielles. Afin d'établir le comportement asymptotique des algorithmes, la stabilité des suites d'itérés (vues comme des chaines de Markov) est étudiée. Ensuite, des généralisations de l'algorithme du gradient proximal stochastique à pas décroissant sont mises au point pour resoudre des problèmes composites. Toutes les grandeurs qui permettent de décrire les problèmes à résoudre s'écrivent comme une espérance. Cela inclut un algorithme primal dual pour des problèmes régularisés et linéairement contraints ainsi qu'un algorithme d'optimisation sur les grands graphes.

Title : Random monotone operators and application to stochastic optimization

Keywords : Stochastic approximation, optimization, machine learning, signal processing

Abstract : This thesis mainly studies optimization algorithms. Programming problems arising in signal processing and machine learning are composite in many cases, *i.e* they exhibit constraints and non smooth regularization terms. Proximal methods are known to be efficient to solve such problems. However, in modern applications of data sciences, functions to be minimized are often represented as statistical expectations, whose evaluation is intractable. This cover the case of online learning, big data problems and distributed computation problems. To solve this problems, we study in this thesis proximal stochastic methods, that generalize proximal algorithms to the case of cost functions written as expectations. Stochastic proximal methods are firstly studied with a constant step size, using stochastic approximation techniques. More precisely, the Ordinary Differential Equation method is adapted to the case of differential inclusions. In order to study the asymptotic behavior of the algorithms, the stability of the sequences of iterates (seen as Markov chains) is studied. Then, generalizations of the stochastic proximal gradient algorithm with decreasing step sizes are designed to solve composite problems. Every quantities used to define the optimization problem are written as expectations. This include a primal dual algorithm to solve regularized and linearly constrained problems and an optimization over large graphs algorithm.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France