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

Pascal Bianchi[(1)], Walid Hachem[(2)], and Adil Salim[(1)]
[(1)] LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
[(2)] CNRS / LIGM (UMR 8049), Université Paris-Est Marne-la-Vallée
pascal.bianchi,adil.salim@telecom-paristech.fr,walid.hachem@u-pem.fr

December 14, 2017

Abstract

We consider a Markov chain (x_n) whose kernel is indexed by a scaling parameter $\gamma > 0$, refered 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 Differential Inclusion (DI) involving an upper semicontinuous 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 as $\gamma \to 0$, to the Birkhoff center of the DI. The ergodic behavior of the iterates is also provided. Application examples are investigated. We apply our findings to 1) the problem of nonconvex proximal stochastic optimization and 2) a fluid model of parallel queues.

Keywords: Differential inclusions; Dynamical systems; Stochastic approximation with constant step; Non-convex optimization; Queueing systems.
34A60,47N10, 54C60,62L20.

1 Introduction

In this paper, we consider a Markov chain $(x_n, n \in \mathbb{N})$ with values in $E = \mathbb{R}^N$, where $N \geq 1$ is an integer. 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 paper is to analyze the long term behavior of the Markov chain in the regime where γ is small. The map

$$g_\gamma(x) := \int \frac{y - x}{\gamma} P_\gamma(x, dy)\,, \tag{1}$$

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

$$x_{n+1} = x_n + \gamma\, g_\gamma(x_n) + \gamma\, U_{n+1}\,, \tag{2}$$

where U_{n+1} is a 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 paper is given by iterative models of the form:

$$x_{n+1} = x_n + \gamma\, h_\gamma(\xi_{n+1}, x_n)\,, \tag{3}$$

where $(\xi_n, n \in \mathbb{N}^*)$ is a sequence of independent and identically distributed (iid) random variables defined on a probability space Ξ with probability law μ, and $\{h_\gamma\}_{\gamma\in(0,\gamma_0)}$ is a family of maps on $\Xi \times \mathbb{R}^N \to \mathbb{R}^N$. In this case, the drift g_γ has the form:

$$g_\gamma(x) = \int h_\gamma(s, x)\, \mu(ds)\,. \tag{4}$$

Our results are as follows.

1. **Dynamical behavior.** Assume that the drift g_γ has the form (4). Assume that for μ-almost all s and for every sequence $((\gamma_k, z_k) \in (0, \gamma_0) \times \mathbb{R}^N, k \in \mathbb{N})$ converging to $(0, z)$,

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

   where $H(s, z)$ is a subset of $\mathbb{R}^N$ (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) 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)

   $$\dot{\mathsf{x}}(t) \in \int H(s, \mathsf{x}(t))\mu(ds)\,, \tag{5}$$

   where for every $x \in \mathbb{R}^N$, $\int H(s, x)\mu(ds)$ is the *selection integral* of $H(\,.\,x)$, which is defined as the closure of the set of integrals of the form $\int \varphi d\mu$ where φ is any integrable function such that $\varphi(s) \in H(s, x)$ for μ-almost all s.

2. **Tightness.** As the iterates are not *a priori* supposed to be in a compact subset of $\mathbb{R}^N$, 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 (5). Here, since the flow generated by the DI is in general set-valued, the notion of invariant measure is borrowed from [17].

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 $a \in E$ and a subset $B \subset E$, we prove that for all $\varepsilon > 0$,

   $$\lim_{\gamma\to 0} \limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathrm{Prob}\left(d(x_k, \mathrm{BC}) > \varepsilon\right) = 0\,, \tag{6}$$

   where BC is the Birkhoff center of the flow induced by the DI (5), and Prob 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

   $$\lim_{\gamma\to 0} \limsup_{n\to\infty} \mathrm{Prob}\left(d(\overline{x}_n, \mathrm{co}(L_{\mathrm{av}})) > \varepsilon\right) = 0\,, \tag{7}$$

where $co(L_{av})$ is the convex hull of the limit set of the averaged flow associated with (5) (see Section 4.4).

4. **Applications.** We investigate several application scenarios. 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 [22]. 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. 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 [3, 19]. Whereas the narrow convergence of the interpolated process was studied in [19], 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. As a final example, we explain how our results can be used in the context of monotone operator theory, in order to analyze a stochastic version of the celebrated proximal point algorithm. The algorithm consists in replacing the usual monotone operator by an iid sequence of random monotone operators. The algorithm has been studied in [11, 12] in the context of decreasing step size. Our analysis provide the tools to characterize its behavior in a constant step regime.

Paper organization. In Section 2, we introduce the application examples. In Section 3, we briefly discuss the literature. Section 4 is devoted to the mathematical background and to the notations. The main results are given in Section 5. The tightness of the interpolated process as well as its narrow convergence towards the solution set of the DI (Th. 5.1) are proven in Section 6. Turning to the Markov chain characterization, Prop. 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 Section 7. A general result describing the asymptotic behavior of a functional of the iterates with a prescribed growth is provided by Th. 5.3, and proven in Section 8. Finally, in Section 9, we show how the results pertaining to the ergodic convergence and to the convergence of the iterates (Th. 5.4 and 5.5 respectively) can be deduced from Th. 5.3. Finally, Section 10 is devoted to the application examples. We prove that our hypotheses are satisfied.

2 Examples

Example 2.1. *Non-convex stochastic optimization.* Consider the problem

$$\text{minimize } \mathbb{E}_\xi(\ell(\xi, x)) + r(x) \text{ w.r.t } x \in \mathbb{R}^N , \tag{8}$$

where $\ell(\xi, \, . \,)$ is a (possibly non-convex) differentiable function on $\mathbb{R}^N \to \mathbb{R}$ indexed by a random variable (r.v.) ξ, $\mathbb{E}_\xi$ represents the expectation w.r.t. ξ, and $r : \mathbb{R}^N \to \mathbb{R}$ is a convex function. The problem typically arises in deep neural networks [30, 28]. 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 (8) is the proximal stochastic gradient algorithm, which reads

$$x_{n+1} = \text{prox}_{\gamma r}(x_n - \gamma \nabla \ell(\xi_{n+1}, x_n)) , \tag{9}$$

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 $\mathbb{R}^N \to \mathbb{R}^N$ defined by

$$\text{prox}_{\gamma r} : x \mapsto \arg\min_{y \in \mathbb{R}^N} \left(\gamma \, r(y) + \frac{\|y - x\|^2}{2} \right) .$$

The drift g_γ has the form (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

$$\partial r(x) := \left\{ u \in \mathbb{R}^N \,:\, \forall y \in \mathbb{R}^N,\, r(y) \geq r(x) + \langle u, y - x \rangle \right\}$$

at every point $x \in \mathbb{R}^N$ such that $r(x) < +\infty$, and $\partial r(x) = \emptyset$ elsewhere. We provide a sufficient condition under which the iterates (9) satisfy the Pakes-Has'minskii criterion, which in turn, allows to characterize the long-run behavior of the iterates.

Example 2.2. *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 [3, 19]. 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:

$$x_{n+1}^k = x_n^k + \gamma\, A_{n+1}^k - \gamma\, B_{n+1}^k \mathbb{1}_{\{x_n^k > 0,\, x_n^{k-1} = \cdots = x_n^1 = 0\}}. \tag{10}$$

The Markov chain $x_n = (x_n^1, \ldots, x_n^N)$ admits the representation (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

$$g_\gamma^k(x) = \lambda_k - \eta_k \mathbb{1}_{\{x^k > 0,\, x^{k-1} = \cdots = x^1 = 0\}}, \tag{11}$$

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 $\mathbb{R}_+^N$

$$\mathsf{H}(x) := \begin{cases} \boldsymbol{u}_1 & \text{if } x^{(1)} > 0 \\ \mathrm{co}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) & \text{if } x^1 = \cdots = x^{k-1} = 0 \text{ and } x^k > 0, \end{cases} \tag{12}$$

where co is the convex hull. Clearly, $g_\gamma(x) \in \mathsf{H}(x)$ for every $x \in \gamma\mathbb{N}^N$. In [19, § 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 [19]. More importantly, if the following stability condition

$$\sum_{k=1}^N \frac{\lambda_k}{\eta_k} < 1 \tag{13}$$

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 (6). The ergodic convergence in the sense of (7) can be also established with a small extra effort.

Example 2.3. *Random monotone operators.* As a second application, we consider the problem of finding a zero of a maximal monotone operator $\mathsf{A} : \mathbb{R}^N \to 2^{\mathbb{R}^N}$:

$$\text{Find } x \text{ s.t. } 0 \in A(x)\,. \tag{14}$$

We recall that a set-valued map $\mathsf{A} : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ is said monotone if for every x, y in $\mathbb{R}^N$, and every $u \in \mathsf{A}(x)$, $v \in \mathsf{A}(y)$, $\langle u - v, x - y\rangle \geq 0$. The domain and the graph of A are the respective subsets of $\mathbb{R}^N$ and $\mathbb{R}^N \times \mathbb{R}^N$ defined as $\text{dom}(\mathsf{A}) := \{x \in \mathbb{R}^N : \mathsf{A}(x) \neq \emptyset\}$, and $\text{gr}(\mathsf{A}) := \{(x, y) \in \mathbb{R}^N \times \mathbb{R}^N : y \in \mathsf{A}(x)\}$. We denote by $\text{zer}(\mathsf{A}) := \{x \in \mathbb{R}^N : 0 \in \mathsf{A}(x)\}$ the set of zeroes of A. The operator A is proper if $\text{dom}(\mathsf{A}) \neq \emptyset$. A proper monotone operator A is said maximal if its graph $\text{gr}(\mathsf{A})$ is a maximal element in the inclusion ordering. Denote by I the identity operator, and by A^{-1} the inverse of the operator A, defined by the fact that $(x, y) \in \text{gr}(\mathsf{A}^{-1}) \Leftrightarrow (y, x) \in \text{gr}(\mathsf{A})$. It is well know that A is maximal monotone if and only if, for all $\gamma > 0$, the *resolvent* $:= (I + \gamma\mathsf{A})^{-1}$ is a contraction defined on the whole space (in particular, it is single valued).

Problem (14) arises in several applications such as convex optimization, variational inequalities, or game theory. The celebrated *proximal point algorithm* [26] generates the sequence $(u_n, n \in \mathbb{N})$ defined recursively as $u_{n+1} = (I + \gamma\mathsf{A})^{-1}(u_n)$. The latter sequence converges to a zero of the operator A, whenever such a zero exists. Recent works (see [12] and references therein) have been devoted to the special case where the operator A is defined as the following selection integral

$$\mathsf{A}(x) = \int A(s, x)\mu(ds)\,,$$

where μ is a probability on Ξ and where $\{A(s, \cdot), s \in \Xi\}$ is a family of maximal monotone operators. In this context, a natural algorithm for solving (14) is

$$x_{n+1} = (I + \gamma A(\xi_{n+1}, \,.\,))^{-1}(x_n) \tag{15}$$

where $(\xi_n, n \in \mathbb{N}^*)$ is an iid sequence of r.v. whose law coincides with μ. The asymptotic behavior of (15) is analyzed in [11] under the assumption that the step size γ is decreasing with n. On the other hand, the results of the present paper apply to the case where γ is a constant which does not depend on n. Here, the drift g_γ has the form (4) where the map $-h_\gamma(s, x) = \gamma^{-1}(x - (I + \gamma A(s, \cdot))^{-1}(x))$ is the so-called *Yosida regularization* of the operator $A(s, \cdot)$ at x. As $\gamma \to 0$, it is well known that for every $x \in \text{dom}(A(s, \cdot))$, $-h_\gamma(s, x)$ converges to the element of least norm in $A(s, \cdot)$ [4]. Thanks to our results, it can be shown that under some hypotheses, the interpolated process converges narrowly to the unique solution to the DI

$$\dot{\mathsf{x}}(t) \in -\int A(s, \mathsf{x}(t))\mu(ds)\,, \tag{16}$$

and, under the Pakes-Has'minskii condition, that the iterates x_n converge in the long run to the zeroes of A.

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 [10, 5, 21, 14, 6] among others. In particular, narrow convergence of the interpolated process to the solution of an Ordinary Differential Equation (ODE) is established. The authors of [18] 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 [7], where decreasing steps were considered. A similar setting is considered in [16]. A Markov noise was considered in the recent manuscript [29]. We also mention [17], 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 [11] and [12].

Differential inclusions arise in many applications, which include game theory (see [7, 8], [27] and the references therein), convex optimization [12], queuing theory or wireless communications, where stochastic approximation algorithms with non continuous drifts are frequently used, and can be modelled by differential inclusions [19].

Differential inclusions with a constant step were studied in [27]. The paper [27] extends previous results of [9] to the case of a DI. The key result established in [27] 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. 5.2 of the present paper 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 [27] does not cover certain applications, such as the ones provided in Section 2, for instance. In addition, [27] 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 [27].

4 Background

4.1 General Notations

The notation $C(E, F)$ is used to denote the set of continuous functions from the topological space 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 integer part of x.

Let (E, d) be 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 $\overline{S}$, 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.

Let $E = \mathbb{R}^N$ for some integer $N \geq 1$. We endow the space $C(\mathbb{R}_+, E)$ with the topology of uniform convergence on compact sets. The space $C(\mathbb{R}_+, E)$ is metrizable by the distance d defined for every $\mathsf{x}, \mathsf{y} \in C(\mathbb{R}_+, E)$ by

$$d(\mathsf{x}, \mathsf{y}) := \sum_{n \in \mathbb{N}} 2^{-n} \left(1 \wedge \sup_{t \in [0,n]} \|\mathsf{x}(t) - \mathsf{y}(t)\| \right), \tag{17}$$

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

4.2 Random Probability Measures

Let E denote a metric space and let $\mathscr{B}(E)$ be its Borel σ-field. We denote by $\mathcal{M}(E)$ the set of probability measures on $(E, \mathscr{B}(E))$. The support $\text{supp}(\nu)$ of a measure $\nu \in \mathcal{M}(E)$ is the smallest closed set G such that $\nu(G) = 1$. We endow $\mathcal{M}(E)$ with the topology of narrow convergence: a sequence $(\nu_n, n \in \mathbb{N})$ on $\mathcal{M}(E)$ converges to a measure $\nu \in \mathcal{M}(E)$ (denoted $\nu_n \Rightarrow \nu$) if for every $f \in C_b(E)$, $\nu_n(f) \to \nu(f)$, where $\nu(f)$ is a shorthand for $\int f(x)\nu(dx)$. If E is a Polish space, $\mathcal{M}(E)$ is metrizable by the Lévy-Prokhorov distance, and is a Polish space as well. A subset $\mathcal{G}$ of $\mathcal{M}(E)$ is said tight if for every $\varepsilon > 0$, there exists a compact subset K of E such that for all $\nu \in \mathcal{G}$, $\nu(K) > 1 - \varepsilon$. By Prokhorov's theorem, $\mathcal{G}$ is tight if and only if it is relatively compact in $\mathcal{M}(E)$.

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

4.3 Set-Valued Mappings and Differential Inclusions

A set-valued mapping $\mathsf{H} : E \rightrightarrows F$ is a function on E into the set 2^F of subsets of F. The graph of H is $\text{gr}(\mathsf{H}) := \{(a, b) \in E \times F : y \in \mathsf{H}(a)\}$. The domain of H is $\text{dom}(\mathsf{H}) := \{a \in E : \mathsf{H}(a) \neq \emptyset\}$. The mapping H is said proper if $\text{dom}(\mathsf{H})$ is non-empty. We say that H is single-valued if $\mathsf{H}(a)$ is a singleton for every $a \in E$ (in which case we handle H simply as a function $\mathsf{H} : E \to F$).

Let $\mathsf{H} : E \rightrightarrows E$ be a set-valued map on $E = \mathbb{R}^N$, where N is a positive integer. Consider the differential inclusion:

$$\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))\,. \tag{18}$$

We say that an absolutely continuous mapping $\mathsf{x} : \mathbb{R}_+ \to E$ is a solution to the differential inclusion with initial condition $a \in E$ if $\mathsf{x}(0) = a$ and if (18) holds for almost every $t \in \mathbb{R}_+$. We denote by

$$\Phi_{\mathsf{H}} : E \rightrightarrows C(\mathbb{R}_+, E)$$

the set-valued mapping such that for every $a \in E$, $\Phi_{\mathsf{H}}(a)$ is set of solutions to (18) with initial condition a. We refer to Φ_{H} as the evolution system induced by H. For every subset $A \subset E$, we define $\Phi_{\mathsf{H}}(A) = \bigcup_{a \in A} \Phi_{\mathsf{H}}(a)$.

A mapping $\mathsf{H} : E \rightrightarrows E$ is said *upper semi continuous* (usc) at a point $a_0 \in E$ if for every open set U containing $\mathsf{H}(a_0)$, there exists $\eta > 0$, such that for every $a \in E$, $\|a - a_0\| < \eta$ implies $\mathsf{H}(a) \subset U$. It is said usc if it is usc at every point [1, Chap. 1.4]. In the particular case where H is usc with nonempty compact convex values and satisfies the condition

$$\exists c > 0,\ \forall a \in E,\ \sup\{\|b\| : b \in \mathsf{H}(a)\} \le c(1 + \|a\|)\,, \tag{19}$$

then, $\text{dom}(\Phi_{\mathsf{H}}) = E$, see *e.g.* [1], and moreover, $\Phi_{\mathsf{H}}(E)$ is closed in the metric space $(C(\mathbb{R}_+, E), d)$.

4.4 Invariant Measures of Set-Valued Evolution Systems

Let (E, d) be a metric space. We define the shift operator $\Theta : C(\mathbb{R}_+, E) \to C(\mathbb{R}_+, C(\mathbb{R}_+, E))$ s.t. for every $\mathsf{x} \in C(\mathbb{R}_+, E)$, $\Theta(\mathsf{x}) : t \mapsto \mathsf{x}(t + \cdot)$.

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

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

Definition 4.1. A probability measure $\pi \in \mathcal{M}(E)$ is said invariant for Φ if there exists $\upsilon \in \mathcal{M}(C(\mathbb{R}_+, E))$ s.t.

(i) $\text{supp}(\upsilon) \subset \overline{\Phi(E)}$;

(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 [17] (see also [23]). Note that $\overline{\Phi(E)}$ can be replaced by $\Phi(E)$ whenever the latter set is closed (sufficient conditions for this have been provided above).

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

$$L_{\mathsf{x}} := \bigcap_{t \ge 0} \overline{\mathsf{x}([t, +\infty)}\,.$$

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 $\Phi : E \rightrightarrows C(\mathbb{R}_+, E)$. The limit set $L_{\Phi(a)}$ of a point $a \in E$ for Φ is

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

and $L_\Phi := \bigcup_{a \in E} L_{\Phi(a)}$. 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

$$\mathrm{BC}_\Phi := \overline{\{a \in E \,:\, a \in L_{\Phi(a)}\}} \,.$$

The following result, established in [17] (see also [2]), is a consequence of the celebrated recurrence theorem of Poincaré.

Proposition 4.1. Let $\Phi : E \rightrightarrows C(\mathbb{R}_+, E)$. Assume that $\Phi(E)$ is closed. Let $\pi \in \mathcal{M}(E)$ be an invariant measure for Φ. Then, $\pi(\mathrm{BC}_\Phi) = 1$.

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

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

We define the mapping $\mathrm{av} : C(\mathbb{R}_+, E) \to C(\mathbb{R}_+, E)$ 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 $\mathrm{av}(\Phi) : E \rightrightarrows C(\mathbb{R}_+, E)$ by $\mathrm{av}(\Phi)(a) = \{\mathrm{av}(\mathsf{x}) \,:\, \mathsf{x} \in \Phi(a)\}$ for each $a \in E$.

4.5 The Selection Integral

Let $(\Xi, \mathscr{G}, \mu)$ denote an arbitrary probability space. For $1 \le p < \infty$, we denote by $\mathcal{L}^p(\Xi, \mathscr{G}, \mu; E)$ the Banach space of the measurable functions $\varphi : \Xi \to E$ such that $\int \|\varphi\|^p d\mu < \infty$. For any set-valued mapping $G : \Xi \rightrightarrows E$, we define the set

$$\mathfrak{S}^p_G := \{\varphi \in \mathcal{L}^p(\Xi, \mathscr{G}, \mu; E) \,:\, \varphi(\xi) \in G(\xi) \ \mu - \text{a.e.}\} \,.$$

Any element of $\mathfrak{S}^1_G$ is referred to as an *integrable selection*. If $\mathfrak{S}^1_G \neq \emptyset$, the mapping G is said to be integrable. The *selection integral* [24] of G is the set

$$\int G d\mu := \overline{\left\{\int_\Xi \varphi d\mu \,:\, \varphi \in \mathfrak{S}^1_G\right\}} \,.$$

5 Main Results

5.1 Dynamical Behavior

From now on to the end of this paper, we set $E := \mathbb{R}^N$ where N is a positive integer. Choose $\gamma_0 > 0$. For every $\gamma \in (0, \gamma_0)$, we introduce a probability transition kernel P_γ on $E \times \mathscr{B}(E) \to [0, 1]$. Let $(\Xi, \mathscr{G}, \mu)$ be an arbitrary probability space.

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

i) For every $x \in E$,

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

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

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

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

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

v) For every $T > 0$ and every compact set $K \subset E$,

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

vi) For every compact set $K \subset E$, there exists $\epsilon_K > 0$ such that

$$\sup_{x \in K} \sup_{0<\gamma<\gamma_0} \int \left\| \frac{y - x}{\gamma} \right\|^{1+\epsilon_K} P_\gamma(x, dy) < \infty , \tag{20}$$

$$\sup_{x \in K} \sup_{0<\gamma<\gamma_0} \int \|h_\gamma(s, x)\|^{1+\epsilon_K} \mu(ds) < \infty . \tag{21}$$

Assumption i) implies that the drift has the form (1). As mentioned in the introduction, this is for instance useful in the case of iterative Markov models such as (3). Assumption v) requires implicitly that the set of solutions $\Phi_{\mathsf{H}}(a)$ is non-empty for any value of a. It holds true if, *e.g.*, the linear growth condition (19) on H is satisfied.

On the canonical space $\Omega := E^{\mathbb{N}}$ equipped with the σ-algebra $\mathscr{F} := \mathscr{B}(E)^{\otimes \mathbb{N}}$, we denote by $X : \Omega \to E^{\mathbb{N}}$ 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 $\nu \in \mathcal{M}(E)$ 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 $a \in E$, we shall prefer the notation $\mathbb{P}^{a,\gamma}$ to $\mathbb{P}^{\delta_a,\gamma}$.

The set $C(\mathbb{R}_+, E)$ is equipped with the topology of uniform convergence on the compact intervals, who is known to be compatible with the distance d defined by (17). For every $\gamma > 0$, we introduce the measurable map on $(\Omega, \mathscr{F}) \to (C(\mathbb{R}_+, E), \mathscr{B}(C(\mathbb{R}_+, E)))$, 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 $(C(\mathbb{R}_+, E), \mathscr{B}(C(\mathbb{R}_+, E)))$, the distribution of the r.v. X_γ is $\mathbb{P}^{\nu,\gamma}\mathsf{X}_\gamma^{-1}$.

Theorem 5.1. Suppose that Assumption (RM) is satisfied. Then, for every compact set $K \subset E$, the family $\{\mathbb{P}^{a,\gamma}\mathsf{X}_\gamma^{-1} : a \in K, 0 < \gamma < \gamma_0\}$ is tight. Moreover, for every $\varepsilon > 0$,

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

5.2 Convergence Analysis

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

$$\mathcal{I}(P_\gamma) := \{\pi \in \mathcal{M}(E) : \pi = \pi P_\gamma\}$$

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 $\nu \in \mathcal{M}(E)$ 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

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

and $\mathscr{I}(\mathcal{P}) = \bigcup_{\gamma \in (0,\gamma_0)} \mathscr{I}(P_\gamma)$. We say that a measure $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(E))$ 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 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 $\mathcal{I}(\Phi_{\mathsf{H}})$;

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

In order to explore the consequences of this proposition, we introduce two supplementary assumptions. The first is the so-called Pakes-Has'minskii tightness criterion, who reads as follows [18]:

Assumption (PH)**.** There exists measurable mappings $V : E \to [0, +\infty)$, $\psi : E \to [0, +\infty)$ 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 $\gamma \in (0, \gamma_0)$,

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

We recall that a transition kernel P on $E \times \mathscr{B}(E) \to [0, 1]$ is said *Feller* if the mapping $Pf : x \mapsto \int f(y)P(x, dy)$ is continuous for any $f \in C_b(E)$. If P is Feller, then the set of invariant measures of P is a closed subset of $\mathcal{M}(E)$. 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 5.3. Let Assumptions (RM), (PH) and (FL) be satisfied. Let ψ and V be the functions specified in (PH). Let $\nu \in \mathcal{M}(E)$ s.t. $\nu(V) < \infty$. Let $\mathcal{U} := \bigcup_{\pi \in \mathcal{I}(\Phi)} \operatorname{supp}(\pi)$. Then, for all $\varepsilon > 0$,

$$\limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}(d(X_k, \mathcal{U}) > \varepsilon) \xrightarrow[\gamma \to 0]{} 0\,. \tag{22}$$

Let $N' \in \mathbb{N}^*$ and $f \in C(E, \mathbb{R}^{N'})$. Assume that there exists $M \ge 0$ and $\varphi : \mathbb{R}^{N'} \to \mathbb{R}_+$ such that $\lim_{\|a\| \to \infty} \varphi(a)/\|a\| = \infty$ and

$$\forall a \in E,\;\; \varphi(f(a)) \le M(1 + \psi(a))\,. \tag{23}$$

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 $\varepsilon > 0$,

$$\limsup_{n \to \infty}\; d\left(\mathbb{E}^{\nu,\gamma}(F_n)\,, \mathcal{S}_f\right) \xrightarrow[\gamma \to 0]{} 0\,, \tag{24}$$

$$\limsup_{n \to \infty}\; \mathbb{P}^{\nu,\gamma}\left(d\left(F_n\,, \mathcal{S}_f\right) \ge \varepsilon\right) \xrightarrow[\gamma \to 0]{} 0\,. \tag{25}$$

Theorem 5.4. Let Assumptions (RM), (PH) and (FL) be satisfied. Assume that $\Phi_{\mathsf{H}}(E)$ is closed. Let ψ and V be the functions specified in (PH). Let $\nu \in \mathcal{M}(E)$ 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 $\varepsilon > 0$,

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

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

where $\operatorname{co}(S)$ is the convex hull of the set S.

Theorem 5.5. Let Assumptions (RM), (PH) and (FL) be satisfied. Assume that $\Phi_{\mathsf{H}}(E)$ is closed. Let ψ and V be the functions specified in (PH). Let $\nu \in \mathcal{M}(E)$ s.t. $\nu(V) < \infty$. Then, for all $\varepsilon > 0$,

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

6 Proof of Theorem 5.1

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

Lemma 6.1. Let $\{F_\xi : \xi \in \Xi\}$ be a family of mappings on $E \rightrightarrows E$. Let $T > 0$ and for all $n \in \mathbb{N}$, let $u_n : [0,T] \to E$, $v_n : \Xi \times [0,T] \to E$ 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 $u : [0,T] \to E$.

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

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 $\mathsf{x} \in C(\mathbb{R}_+, E)$.

Lemma 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 $K \subset E$ 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:

$$\sup_{n\in\mathbb{N}^*, a\in K, \gamma\in(0,\gamma_0)} \bar{\mathbb{E}}^{a,\gamma}(\|Z_n^\gamma\| \mathbb{1}_{\|Z_n^\gamma\| > A}) \xrightarrow{A\to+\infty} 0\,. \tag{26}$$

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

$$\sup_{a\in K} \bar{\mathbb{P}}^{a,\gamma}\left(\max_{0\leq n\leq \lfloor \frac{T}{\gamma}\rfloor} \gamma \left\|\sum_{k=0}^{n}\left(Z_{k+1}^\gamma - \bar{\mathbb{E}}^{a,\gamma}(Z_{k+1}^\gamma|\mathscr{F}_k)\right)\right\| > \varepsilon\right) \xrightarrow{\gamma\to 0} 0\,. \tag{27}$$

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 $R > 0$,

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

Noting that $n - m + 1 \leq \frac{\delta}{\gamma}$ and using Markov inequality, we obtain

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

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 [13, Theorem 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

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

and we define $\eta_{n+1}^{a,\gamma,>}$ in a similar way, by replacing $\le$ with $>$ in the right hand side of the above equation. Clearly, for all $a\in K$, $M_{n+1}^{a,\gamma} = \eta_{n+1}^{a,\gamma,\le} + \eta_{n+1}^{a,\gamma,>}$. Thus,

$$\gamma\left\|M_{n+1}^{a,\gamma}\right\| \le \|S_{n+1}^{a,\gamma,\le}\| + \|S_{n+1}^{a,\gamma,>}\|$$

where $S_{n+1}^{a,\gamma,\le} := \gamma\sum_{k=0}^n \eta_{k+1}^{a,\gamma,\le}$ and $S_{n+1}^{a,\gamma,>}$ is defined similarly. Under $\bar{\mathbb{P}}^{a,\gamma}$, the random processes $S^{a,\gamma,\le}$ 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 $S_n^{a,\gamma,\le}$ that

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

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

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

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 $h_{\gamma,R}(s,a) := h_\gamma(s,a)\mathbb{1}_{\|a\|\le R}$. Let $H_R(s,x) := H(s,x)$ if $\|x\| < R$, $\{0\}$ if $\|x\| > R$, and E otherwise. Denote the corresponding selection integral as $\mathsf{H}_R(a) = \int H_R(s,a)\,\mu(ds)$. 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

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

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

Lemma 6.3. Suppose that Assumption (RM) is satisfied. Then, for every compact set $K\subset E$, 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 $\Delta_{\gamma,R} : \Omega\to E^{\mathbb{N}}$ s.t. for all $x\in\Omega$, $\Delta_{\gamma,R}(x)(0) := 0$ and

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

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

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

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

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

from which we obtain the decomposition

$$\mathsf{X}_\gamma(x) = x_0 + \mathsf{G}_{\gamma,R} \circ \mathsf{X}_\gamma(x) + \mathsf{X}_\gamma \circ \Delta_{\gamma,R}(x)\,. \tag{28}$$

The uniform integrability condition (26) is satisfied when letting $\bar{\mathbb{P}}^{a,\gamma} := \mathbb{P}^{a,\gamma} B_R^{-1}$. Indeed,

$$\bar{\mathbb{E}}^{a,\gamma}(\|\gamma^{-1}(X_{n+1} - X_n)\|^{1+\epsilon_K}) = \mathbb{E}^{a,\gamma}\left(\left\|\gamma^{-1}(X_{n+1} - X_n)\right\|^{1+\epsilon_K} \mathbb{1}_{\tau_R(X) > n}\right)\,,$$

and the condition (26) follows from hypothesis (20). On the other hand, as the event $\{\tau_R(X) > n\}$ is $\mathscr{F}_n$-measurable,

$$\begin{aligned}
\bar{\mathbb{E}}^{a,\gamma}(\gamma^{-1}(X_{n+1} - X_n)|\mathscr{F}_n) &= \mathbb{E}^{a,\gamma}(\gamma^{-1}(X_{n+1} - X_n)|\mathscr{F}_n)\mathbb{1}_{\tau_R(X)>n} \\
&= \int h_\gamma(s, X_n)\mu(ds)\mathbb{1}_{\tau_R(X)>n} \\
&= \int h_{\gamma,R}(s, X_n)\mu(ds)\,.
\end{aligned}$$

Thus, Lemma 6.2 implies that for all $\varepsilon > 0$ and $T > 0$,

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

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

$$\sup_{a\in K} \bar{\mathbb{P}}^{a,\gamma}(\|\mathsf{X}_\gamma \circ \Delta_{\gamma,R}\|_{\infty,T} > \varepsilon) \xrightarrow{\gamma\to 0} 0\,, \tag{29}$$

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

As a second consequence of Lemma 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 $\upsilon \in \mathcal{M}(C(\mathbb{R}_+, E))$ 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 $C(\mathbb{R}_+, E)$ 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 (28) and (29), 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 $C(\mathbb{R}_+, E)$ 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

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

By Eq. (21), the sequence $(v_n, n \in \mathbb{N})$ forms a bounded subset of $\mathcal{L}^{1+\epsilon_{K'}} := \mathcal{L}^{1+\epsilon_{K'}}(\Xi \times [0,T], \mathscr{G} \otimes \mathscr{B}([0,T]), \mu \otimes \lambda_T; E)$. By the Banach-Alaoglu theorem, the sequence converges weakly to some mapping $v \in \mathcal{L}^{1+\epsilon_{K'}}$ along some subsequence. This has two consequences. First,

$$\mathsf{z}(t) = \mathsf{z}(0) + \int_0^t \int_\Xi v(s,u)\, \mu(ds)\, du\,, \quad (\forall t \in [0,T])\,. \tag{30}$$

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 Lemma 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 $x \in E$, the same holds for H_R. Equation (30) 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 Theorem 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 $\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n}, \Phi_{\mathsf{H}}(K)) > \varepsilon) \to 0$. 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 $\mathsf{y} \in \Phi_{\mathsf{H}}(K)$ 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 Lemma 6.3, each sequence (a_n,γ_n) chosen as above admits a subsequence along which, for all $\varepsilon > 0$,

$$\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \Phi_{\mathsf{H}}(K)) > \varepsilon) \to 0\,. \tag{31}$$

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 $R_0(T) + d_T(\mathsf{X}_\gamma \circ B_R, \Phi_{\mathsf{H}}(K)) \geq R$. Therefore,

$$\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \mathsf{X}_{\gamma_n}) > \varepsilon) \leq \mathbb{P}(d_T(\mathsf{X}_{\gamma_n} \circ B_R, \Phi_{\mathsf{H}}(K)) \geq R - R_0(T))\,. \tag{32}$$

By (31), the right hand side converges to zero. Using (31) again along with the triangular inequality, it follows that $\mathbb{P}^{a_n,\gamma_n}(d_T(\mathsf{X}_{\gamma_n}, \Phi_{\mathsf{H}}(K)) > \varepsilon) \to 0$, which proves the second statement of the theorem.

We prove the first statement (tightness). By [13, Theorem 7.3], 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 $\limsup_n \mathbb{P}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1}(\{x : w_x^T(\delta) > \varepsilon\}) < \eta$.

First consider the case where $\gamma_n \to 0$. Fixing $T > 0$, letting $R > R_0(T)$ and using (32), 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. The tightness of $(\mathbb{P}^{a_n,\gamma_n}\mathsf{X}_{\gamma_n}^{-1} p_0^{-1})$ follows. In addition, for every $\mathsf{x},\mathsf{y} \in C(\mathbb{R}_+, E)$, it holds by the triangle inequality that $w_{\mathsf{x}}^T(\delta) \leq w_{\mathsf{y}}^T(\delta) + 2 d_T(\mathsf{x},\mathsf{y})$ for every $\delta > 0$. Thus,

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

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 $\gamma > \eta$,

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

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

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

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

$$P_\gamma^k(a,\cdot) = \int P_\gamma(a,dy)P_\gamma^{k-1}(y,\cdot) \tag{33}$$

and $P_\gamma^0(a,\cdot) = \delta_a$. Using (20), 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.

7 Proof of Proposition 5.2

To establish Prop. 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. 4.1.

Given $\varepsilon > 0$, there exists a compact set $K \subset E$ such that $\inf_n \pi_n(K) > 1 - \varepsilon/2$. By Th. 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 $C(\mathbb{R}_+, E)$ 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 $\mathbb{P}^{\pi_n,\gamma_n}(\cdot) = \int_E \mathbb{P}^{a,\gamma_n}(\cdot)\,\pi_n(da)$. Thus,

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

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. 4.1–(iii) is satisfied by π and υ. To establish Prop. 5.2–i), we need to verify the conditions (i) and (ii) of Definition 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. 4.1–(i), we prove that for every $\eta > 0$, $\upsilon_n((\Phi_{\mathsf{H}}(E))_\eta) \to 1$ as $n \to \infty$; the result will follow from the convergence of (υ_n). Fix $\varepsilon > 0$, and let $K \subset E$ be a compact set such that $\inf_n \pi_n(K) > 1 - \varepsilon$. We have

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

By Th. 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

$$\int f(\mathsf{x})\,\upsilon(d\mathsf{x}) = \int f \circ \Theta_t(\mathsf{x})\,\upsilon(d\mathsf{x}) \tag{34}$$

for all $f \in C_b(C(\mathbb{R}_+, E))$ 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, $\Theta_{\eta_n}(\mathsf{x}(\gamma_n\lfloor t/\gamma_n \rfloor + \cdot))$ and $\Theta_t(\mathsf{x})$ are equal in law under $\upsilon_n(d\mathsf{x})$. Thus,

$$\int f \circ \Theta_t(\mathsf{x})\, \upsilon_n(d\mathsf{x}) = \int f \circ \Theta_{\eta_n}(\mathsf{x}(\gamma_n \lfloor t/\gamma_n \rfloor + \cdot))\, \upsilon_n(d\mathsf{x})$$
$$= \int f \circ \Theta_{\eta_n}(\mathsf{x})\, \upsilon_n(d\mathsf{x}). \tag{35}$$

Using Skorokhod's representation theorem, there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and random variables $(\bar{\mathsf{x}}_n, n \in \mathbb{N})$ and $\bar{\mathsf{x}}$ over this probability space, with values in $C(\mathbb{R}_+, E)$, 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, $d(\Theta_{\eta_n}(\bar{\mathsf{x}}_n), \bar{\mathsf{x}}) \longrightarrow_{n\to+\infty} 0$. Hence,

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

Recalling Eq. (35), we have shown that $\int f \circ \Theta_t(\mathsf{x})\, \upsilon_n(d\mathsf{x}) \xrightarrow[n\to\infty]{} \int f \circ \Theta_t(\mathsf{x})\, \upsilon(d\mathsf{x})$. Since $\int f(\mathsf{x})\, \upsilon_n(d\mathsf{x}) \xrightarrow[n\to\infty]{} \int f(\mathsf{x})\, \upsilon(d\mathsf{x})$, the identity (34) holds true. Prop. 5.2–i) is proven.

We now prove Prop. 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 $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(E))$. Since the space $\mathcal{M}(E)$ is separable, Skorokhod's representation theorem shows that there exists a probability space $(\Omega', \mathscr{F}', \mathbb{P}')$, a sequence of $\Omega' \to \mathcal{M}(E)$ random variables (Λ_n) with distributions $\mathfrak{m}_n$, and a $\Omega' \to \mathcal{M}(E)$ 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 $(E^{\mathbb{N}}, \mathscr{F})$ a probability measure $\mathbb{P}^{\Lambda_n(\omega),\gamma_n}$ as we did in Sec. 5.1. By the same argument as in the proof of Prop. 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. 4.1 with $\Lambda(\omega) = \upsilon p_0^{-1}$. Prop. 5.2 is proven.

8 Proof of Theorem 5.3

8.1 Technical lemmas

Lemma 8.1. Given a family $\{K_j, j \in \mathbb{N}\}$ of compact sets of E, the set

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

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

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

Lemma 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. □

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

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

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

Lemma 8.3. Let $\mathcal{L}$ be a family on $\mathcal{M}(\mathcal{M}(E))$. 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 $K_j \subset E$ s.t. for all $\mathfrak{m} \in \mathcal{L}$, $e(\mathfrak{m})(K_j) > 1 - 2^{-2j}$. Define U as the set of measures $\nu \in \mathcal{M}(E)$ s.t. for all $j \geq k$, $\nu(K_j) \geq 1 - 2^{-j}$. By Lemma 8.1, U is compact. For all $\mathfrak{m} \in \mathcal{L}$, the union bound implies that

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

By Lemma 8.2, $\mathfrak{m}\{\nu : \nu(K_j) \geq 1 - 2^{-j}\} \geq 1 - 2^{-j}$. Therefore, $\mathfrak{m}(E \backslash U) \leq \sum_{j=k}^{\infty} 2^{-j} = 2^{-k+1} \leq \varepsilon$. This proves that $\mathcal{L}$ is tight. □

Lemma 8.4. Let $(\mathfrak{m}_n, n \in \mathbb{N})$ be a sequence on $\mathcal{M}(\mathcal{M}(E))$, and consider $\bar{\mathfrak{m}} \in \mathcal{M}(\mathcal{M}(E))$. If $\mathfrak{m}_n \Rightarrow \bar{\mathfrak{m}}$, then $e(\mathfrak{m}_n) \Rightarrow e(\bar{\mathfrak{m}})$.

Proof. For any $f \in C_b(E)$, $\mathcal{T}_f \in C_b(\mathcal{M}(E))$. 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 $\mathcal{M}(\mathcal{M}(E))$ converges narrowly to $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(E))$, 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 8.5. Let $f \in C(E, \mathbb{R}^{N'})$ where $N' \geq 1$ is an integer. Define by $\mathcal{T}_f : \mathcal{M}(E) \to \mathbb{R}$ 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 $\mathcal{M}(\mathcal{M}(E))$ and let $\mathfrak{m} \in \mathcal{M}(\mathcal{M}(E))$. Assume that $\mathfrak{m}_n \Rightarrow \mathfrak{m}$ and

$$\lim_{K \to \infty} \sup_n e(\mathfrak{m}_n)(\|f\| \mathbb{1}_{\|f\| > K}) = 0\,. \tag{36}$$

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. (36), $e(\mathfrak{m})(\|f\|) < \infty$. This implies that for all ν $\mathfrak{m}$-a.e., $\nu(\|f\|) < \infty$. Choose $h \in C_b(\mathbb{R}^{N'})$ 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 (36), $\mathfrak{m}_n \mathcal{T}_f^{-1}(h) = \int h(\nu(f)) d\mathfrak{m}_n(\nu)$. Choose $\varepsilon > 0$. By Eq. (36), 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 $f_K \in C(E, \mathbb{R}^{N'})$ by $f_K(x) = f(x)(1 \wedge K / \|f(x)\|)$. For all $K > K_0$, and for all $n \in \mathbb{N}$,

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

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 $K > K_0$, $\limsup_n |\mathfrak{m}_n \mathcal{T}_f^{-1}(h) - \mathfrak{m} \mathcal{T}_{f_K}^{-1}(h)| \leq L\varepsilon$. 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, $\limsup_n |\mathfrak{m}_n \mathcal{T}_f^{-1}(h) - \mathfrak{m} \mathcal{T}_f^{-1}(h)| \leq L\varepsilon$. 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. □

8.2 Narrow Cluster Points of the Empirical Measures

Let $P : E \times \mathscr{B}(E) \to [0,1]$ be a probability transition kernel. For $\nu \in \mathcal{M}(E)$, 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 $\Lambda_n : \Omega \to \mathcal{M}(E)$ as

$$\Lambda_n(x) := \frac{1}{n+1} \sum_{k=0}^{n} \delta_{x_k} \tag{37}$$

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 $\mathcal{M}(\mathcal{M}(E))$ formed by the measures whose support is included in $\mathcal{I}(P)$.

Proposition 8.6. Let $P : E \times \mathscr{B}(E) \to [0,1]$ be a Feller probability transition kernel. Let $\nu \in \mathcal{M}(E)$.

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 $f \in C_b(E)$, $\mathbb{E}\Lambda_n(Pf) - \mathbb{E}\Lambda_n(f) \to 0$. 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 $f \in C_b(E)$ 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 $\left(\sum_{k=0}^{n-1} k^{-1}(Pf(X_k) - f(X_{k+1}))\right)_n$ 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,

$$\Lambda_n(Pf) - \Lambda_n(f) \to 0 \text{ a.s.} \tag{38}$$

Now consider a subsequence (Λ_{φ_n}) which converges in distribution to some r.v. Λ as n tends to infinity. For a fixed $f \in C_b(E)$, the mapping $\nu \mapsto (\nu(f), \nu(Pf))$ on $\mathcal{M}(\mathbb{R}) \to \mathbb{R}^2$ is continuous. From the mapping theorem, $\Lambda_{\varphi_n}(f) - \Lambda_{\varphi_n}(Pf)$ converges in distribution to $\Lambda(f) - \Lambda(Pf)$. By (38), it follows that $\Lambda(f) - \Lambda(Pf) = 0$ on some event $\mathcal{E}_f \in \mathscr{F}$ of probability one. Denote by $C_\kappa(E) \subset C_b(E)$ the set of continuous real-valued functions having a compact support, and let $C_\kappa(E)$ be equipped with the uniform norm $\|\cdot\|_\infty$. Introduce a dense denumerable subset S of $C_\kappa(E)$. On the probability-one event $\mathcal{E} = \cap_{f \in S}\mathcal{E}_f$, it holds that for all $f \in S$, $\Lambda(f) = \Lambda P(f)$. Now consider $g \in C_\kappa(E)$ and let $\varepsilon > 0$. Choose $f \in S$ such that $\|f - g\|_\infty \le \varepsilon$. Then, almost everywhere on $\mathcal{E}$, $|\Lambda(g) - \Lambda P(g)| \le |\Lambda(f) - \Lambda(g)| + |\Lambda P(f) - \Lambda P(g)| \le 2\varepsilon$. Thus, $\Lambda(g) - \Lambda P(g) = 0$ for every $g \in C_\kappa(E)$. Hence, almost everywhere on $\mathcal{E}$, one has $\Lambda = \Lambda P$. □

8.3 Tightness of the Empirical Measures

Proposition 8.7. Let $\mathcal{P}$ be a family of transition kernels on E. Let $V : E \to [0,+\infty)$, $\psi : E \to [0,+\infty)$ 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 $\nu \in \mathcal{M}(E)$ 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,

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

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

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

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

$$\begin{aligned}\frac{1}{n+1}\sum_{k=0}^{n} P^k(\psi\wedge M) &\le \left(\frac{1}{n+1}\sum_{k=0}^{n} P^k\psi\right)\wedge M\\ &\le \left(\frac{V}{\alpha(P)(n+1)} + c\right)\wedge M\,. \end{aligned}\tag{39}$$

Set $\pi \in \mathcal{I}(\mathcal{P})$, and consider $P \in \mathcal{P}$ such that $\pi = \pi P$. Inequality (39) implies that for every n,

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

By Lebesgue's dominated convergence theorem, $\pi(\psi \wedge M) \le c$. Letting $M \to \infty$ yields $\pi(\psi) \le c$. The tightness of $\mathcal{I}(\mathcal{P})$ follows from the convergence of $\psi(x)$ to ∞ as $\|x\| \to \infty$. Setting $M = +\infty$ in (39), and integrating w.r.t. ν, we obtain

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

which proves the second point. □

Proposition 8.8. We posit the assumptions of Prop. 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. 8.7. By Lemma 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. 8.7 and Lemma 8.3. □

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. 8.7 and Eq. (23), 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

$$\lim_{K\to\infty}\sup_n \mathbb{E}^{\nu,\gamma}\Lambda_n(\|f\|\mathbb{1}_{\|f\|>K}) = 0\,. \tag{40}$$

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 (24). 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.

$$\forall n,\ d\left(\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(f), \mathcal{S}_f\right) > \delta\,. \tag{41}$$

By Prop. 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. 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 (40), $\pi_j(\|f\|) < \infty$ 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. (41), for all $j \in \mathbb{N}$, $d(\pi_j(f), \mathcal{S}_f) \geq \delta$. By Prop. 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

$$\sup_{\pi\in\mathcal{I}(\mathcal{P})} \pi(\|f\|\mathbb{1}_{\|f\|>K}) < \infty\,. \tag{42}$$

Also by Prop. 8.7, the sequence (π_j) is tight. Thus $\pi_j \Rightarrow \pi$ along some subsequence, for some measure π which, by Prop. 5.2, is invariant for Φ_{H}. The uniform integrability condition (42) 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 (24) holds true (and in particular, $\mathcal{S}_f$ must be non-empty).

The proof of the statement (22) follows the same line, by replacing f with the function $\mathbb{1}_{\overline{\mathcal{U}_\epsilon}}$. We briefly explain how the proof adapts, without repeating all the arguments. In this case, $\mathcal{S}_{\mathbb{1}_{\mathcal{U}_\epsilon^c}}$ is the singleton $\{0\}$, and Equation (41) reads $\mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(\mathcal{U}_\epsilon^c) > \delta$. By the Portmanteau theorem, $\limsup_n \mathbb{E}^{\nu,\gamma_j}\Lambda_{\varphi_n^j}(\mathcal{U}_\epsilon^c) \leq \pi_j(\mathcal{U}_\epsilon^c)$ where the $\limsup$ is taken along some subsequence. The contradiction follows from the fact that $\limsup \pi_j(\mathcal{U}_\epsilon^c) \leq \pi(\overline{\mathcal{U}_\epsilon^c}) = 0$ (where the $\limsup$ is again taken along the relevant subsequence).

We prove the statement (25). 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.

$$\forall n,\ \mathbb{P}^{\nu,\gamma_j}\left(d\left(\Lambda_{\varphi_n^j}(f), \mathcal{S}_f\right) \geq \varepsilon\right) > \delta\,. \tag{43}$$

By Prop. 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. 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. (40), Lemma 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. 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. 5.2, belongs to $\mathscr{I}(\Phi_{\mathsf{H}})$. For every j, $e(\mathfrak{m}_j) \in \mathcal{I}(P_{\gamma_j})$. By the uniform integrability condition (42), one can apply Lemma 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 $\mathfrak{m} \in \mathscr{I}(\Phi_{\mathsf{H}})$, it holds that $\mathfrak{m}\mathcal{T}_f^{-1}((\mathcal{S}_f)_\varepsilon^c) = 0$, hence a contradiction.

9 Proofs of Theorems 5.4 and 5.5

9.1 Proof of Theorem 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. 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 $\upsilon \in \mathcal{M}(C(\mathbb{R}_+, E))$ s.t. $\mathrm{supp}(\upsilon) \subset \Phi(E)$ and $\upsilon p_0^{-1} = \pi$. We remark that for all $t > 0$,

$$q = \upsilon(p_0) = \upsilon(p_t) = \upsilon(p_t \circ \mathrm{av})\,, \tag{44}$$

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 $S \in \mathscr{B}(C(\mathbb{R}_+, E))$. 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 $(C(\mathbb{R}_+, E), \mathscr{B}(C(\mathbb{R}_+, E)), \upsilon)$. By Prokhorov's theorem, there exists a sequence $t_n \to \infty$ and a measurable function $g : C(\mathbb{R}_+, E) \to E$ such that $p_{t_n} \circ \mathrm{av}$ converges in distribution to g as $n \to \infty$. By uniform integrability, $\upsilon(p_{t_n} \circ \mathrm{av}) \to \upsilon(g)$. Equation (44) 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 $\mathsf{x} \in \Phi(E)$ 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 $x \in \Phi(E)$. As $\mathrm{supp}(\upsilon) \subset \Phi(E)$, $\mathbb{1}_{\overline{L_\varepsilon}^c \cap K}(p_{t_n} \circ \mathrm{av})$ 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.

9.2 Proof of Theorem 5.5

Recall the definition $\mathcal{U} := \bigcup_{\pi \in \mathcal{I}(\Phi)} \mathrm{supp}(\pi)$. By Th. 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 (37). By Theorem 4.1, $\mathrm{supp}(\pi) \subset \mathrm{BC}_\Phi$ for each $\pi \in \mathcal{I}(\Phi)$. Thus, $\mathcal{U}_\varepsilon \subset (\mathrm{BC}_\Phi)_\varepsilon$. Hence, $\limsup_n \mathbb{E}^{\nu,\gamma} \Lambda_n(((\mathrm{BC}_\Phi)_\varepsilon)^c) \to 0$ as $\gamma \to 0$. This completes the proof.

10 Applications

In this section, we return to the Examples 2.1 and 2.2 of Section 2.

10.1 Non-Convex Optimization

Consider the algorithm (9) to solve problem (8) where $\ell : \Xi \times E \to \mathbb{R}$, $r : E \to \mathbb{R}$ 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 $\ell(\xi, \,.\,)$ is continuously differentiable for every $\xi \in \Xi$, that $\ell(\,.\,, x)$ is μ-integrable for every $x \in E$ and that r is a convex and lower semicontinuous function. We assume that for every compact subset K of E, there exists $\epsilon_K > 0$ s.t.

$$\sup_{x \in K} \int \|\nabla \ell(s,x)\|^{1+\epsilon_K} \mu(ds) < \infty \,. \tag{45}$$

Define $L(x) := \mathbb{E}_\xi(\ell(\xi, x))$. Under Condition (45), 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. Letting $H(s,x) := -\nabla \ell(s,x) - \partial r(x)$, it holds that $H(\,.\,,x)$ is proper, μ-integrable and usc [25], 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 [15, Theorem 3.17, Remark 3.14], for every $a \in E$, the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(x(t))$ admits a unique solution on $[0, +\infty)$ s.t. $\mathsf{x}(0) = a$.

Now consider the iterates x_n given by (9). They satisfy (3) where $h_\gamma(s,x) := \gamma^{-1}(\mathrm{prox}_{\gamma r}(x - \gamma \nabla \ell(s,x)) - x)$. We verify that the map h_γ satisfies Assumption (RM). Let us first recall some known facts about proximity operators. Using [4, Prop. 12.29], the mapping $x \mapsto \gamma^{-1}(x - \mathrm{prox}_{\gamma r}(x))$ coincides with the gradient ∇r_γ of the Moreau enveloppe $r_\gamma : x \mapsto \min_y r(y) + \|y - x\|^2$. By [4, Prop. 23.2], $\nabla r_\gamma(x) \in \partial r(\mathrm{prox}_{\gamma r}(x))$, for every $x \in E$. Therefore,

$$\begin{aligned} h_\gamma(s,x) &= -\nabla r_\gamma(x - \gamma \nabla \ell(s,x)) - \nabla \ell(s,x) && (46)\\ &\in -\partial r(\mathrm{prox}_{\gamma r}(x - \gamma \nabla \ell(s,x))) - \nabla \ell(s,x) \\ &\in -\partial r(x - \gamma h_\gamma(s,x)) - \nabla \ell(s,x)\,. && (47) \end{aligned}$$

In order to show that Assumption (RM)-ii)) is satisfied, we need some estimate on $\|h_\gamma(s,x)\|$. Using Eq. (46) and the fact that ∇r_γ is γ^{-1}-Lipschitz continuous (see [4, Prop. 12.29]), we obtain that

$$\begin{aligned} \|h_\gamma(s,x)\| &\le \|\nabla r_\gamma(x)\| + 2\|\nabla \ell(s,x)\| \\ &\le \|\partial^0 r(x)\| + 2\|\nabla \ell(s,x)\|\,, && (48) \end{aligned}$$

where $\partial^0 r(x)$ the least norm element in $\partial r(x)$ for every $x \in E$, and where the last inequality is due to [4, Prop. 23.43]. As $\partial^0 r$ is locally bounded and ∂r is usc, it follows from Eq. (47) that Assumption (RM)-ii)) is satisfied. The estimate (48) also yields Assumption (RM)-vi)). As a conclusion, Assumption (RM) is satisfied. In particular, the statement of Th. 5.1 holds.

To show that Assumption (PH) is satisfied, we first recall the Proximal Polyak-Lojasiewicz (PPL) condition introduced in [20]. Assume that L is differentiable with a C-Lipschitz continuous gradient. We say that L and r satisfy the (PPL) condition with constant $\beta > 0$ if for every $x \in E$,

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

where

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

The (PPL) 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 [20] for practical cases where the (PPL) condition is satisfied. In our stochastic setting, we introduce the Stochastic PPL condition (SPPL). We say that ℓ and r satisfy the (SPPL) condition if there exists $\beta > 0$ such that for every $x \in E$,

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

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

Proposition 10.1. Assume that the (SPPL) condition is satisfied, that $\gamma \le \frac{1}{C}$ and that

$$\beta(L(x)+r(x)) - W(x) - \frac{1}{4}\|\nabla L(x)\|^2 \longrightarrow_{\|x\|\to+\infty} +\infty.$$

Then (PH) is satisfied.

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

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

Since ∇L is $1/\gamma$-Lipschitz continuous,

$$\begin{aligned}(L+r)(x+\gamma h_\gamma(s,x)) &= L(x+\gamma h_\gamma(s,x)) + r(x) + r(x+\gamma h_\gamma(s,x)) - r(x)\\ &\le (L+r)(x) + \langle \nabla L(x), \gamma h_\gamma(s,x)\rangle + \frac{1}{2\gamma}\|\gamma h_\gamma(s,x)\|^2\\ &\quad + r(x+\gamma h_\gamma(s,x)) - r(x)\\ &\le (L+r)(x) + \langle \nabla \ell(s,x), \gamma h_\gamma(s,x)\rangle + \frac{1}{2\gamma}\|\gamma h_\gamma(s,x)\|^2\\ &\quad + \langle \nabla L(x) - \nabla \ell(s,x), \gamma h_\gamma(s,x)\rangle + r(x+\gamma h_\gamma(s,x)) - r(x)\\ &\le (L+r)(x) - \frac{\gamma}{2} D_{\ell(s,\cdot),r}(x,1/\gamma)\\ &\quad + \gamma\langle \nabla \ell(s,x) - \nabla L(x), \nabla \ell(s,x) + \nabla r_\gamma(x - \gamma\nabla\ell(s,x))\rangle \end{aligned} \tag{49}$$

Recall that for every $x, y \in E$,

$$\begin{aligned}\langle \nabla r_\gamma(x) - \nabla r_\gamma(y), x-y\rangle &= \langle \nabla r_\gamma(x) - \nabla r_\gamma(y), \operatorname{prox}_{\gamma r}(x) - \operatorname{prox}_{\gamma r}(y)\rangle\\ &\quad + \langle \nabla r_\gamma(x) - \nabla r_\gamma(y), \gamma\nabla r_\gamma(x) - \gamma\nabla r_\gamma(y)\rangle\\ &\ge \gamma\|\nabla r_\gamma(x) - \nabla r_\gamma(y)\|^2,\end{aligned}$$

using the monotonicity of ∂r. Hence,

$$\langle \nabla r_\gamma(x-\gamma\nabla\ell(s,x)) - \nabla r_\gamma(x), \gamma\nabla\ell(s,x)\rangle \le -\gamma\|\nabla r_\gamma(x) - \nabla r_\gamma(x-\gamma\nabla\ell(s,x))\|^2.$$

Therefore,

$$\begin{aligned}&\gamma\langle \nabla\ell(s,x) - \nabla L(x), \nabla r_\gamma(x-\gamma\nabla\ell(s,x)) - \nabla r_\gamma(x)\rangle\\ \le& -\gamma\|\nabla r_\gamma(x) - \nabla r_\gamma(x-\gamma\nabla\ell(s,x))\|^2\\ &+ \gamma\|\nabla r_\gamma(x) - \nabla r_\gamma(x-\gamma\nabla\ell(s,x))\|^2 + \frac{\gamma}{4}\|\nabla L(x)\|^2\\ \le& \frac{\gamma}{4}\|\nabla L(x)\|^2,\end{aligned}$$

where we used $\langle x,y\rangle \le \|x\|^2 + \frac{1}{4}\|y\|^2$.

Plugging into (49),

$$\begin{aligned}(L+r)(x+\gamma h_\gamma(s,x)) &\le (L+r)(x) - \frac{\gamma}{2}D_{\ell(s,\cdot),r}(x,1/\gamma)\\ &\quad + \gamma\langle \nabla\ell(s,x) - \nabla L(x), \nabla\ell(s,x)\rangle\\ &\quad + \gamma\langle \nabla\ell(s,x) - \nabla L(x), \nabla r_\gamma(x-\gamma\nabla\ell(s,x)) - \nabla r_\gamma(x)\rangle\\ &\quad + \gamma\langle \nabla\ell(s,x) - \nabla L(x), \nabla r_\gamma(x)\rangle\\ &\le (L+r)(x) - \frac{\gamma}{2}D_{\ell(s,\cdot),r}(x,1/\gamma)\\ &\quad + \gamma\langle \nabla\ell(s,x) - \nabla L(x), \nabla\ell(s,x)\rangle\\ &\quad + \frac{\gamma}{4}\|\nabla L(x)\|^2\\ &\quad + \gamma\langle \nabla\ell(s,x) - \nabla L(x), \nabla r_\gamma(x)\rangle\end{aligned}$$

Integrating with respect to μ, we obtain

$$\int (L+r)(x+\gamma h_\gamma(s,x))\mu(ds) \le (L+r)(x) - \gamma\beta\left((L+r)(x) - \min(L+r)\right) + \gamma W(x) + \frac{\gamma}{4}\|\nabla L(x)\|^2.$$

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

$$\psi = \beta V - W - \frac{1}{4}\|\nabla L\|^2.$$

□

Note that the assumptions of Proposition 10.1 are satisfied if the (SPPL) condition is satisfied, $L(x) + r(x) \to_{\|x\|\to+\infty} +\infty$ and the function $x \mapsto \int \|\nabla \ell(s,x)\|^2 \mu(ds)$ is bounded.

The condition (FL) is naturally satisfied. Identifying the invariant measures of the DI, we finally obtain a long-run convergence result for the algorithm (9). Let $\nu \in \mathcal{M}(E)$ s.t. $\nu(L+r) < \infty$. Let $\mathcal{Z} = \{x \in E,\ \text{s.t } 0 \in \nabla L(x) + \partial r(x)\}$. For all $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{P}^{\nu,\gamma}(d(X_k, \mathcal{Z}) > \varepsilon) \xrightarrow[\gamma\to 0]{} 0\,. \tag{50}$$

10.2 Fluid Limit of a System of Parallel Queues

We now apply the results of this paper to the dynamical system described in Example 2.2 above. For a given $\gamma > 0$, the transition kernel P_γ of the Markov chain (x_n) whose entries are given by Eq. (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 paper readability. The limit behavior of the interpolated process (see Theorem 5.1) is described by the following proposition, which has an analogue in [19]:

Proposition 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 (12).

Proof. To prove this proposition, we mainly need to check that Assumption (RM) is verified. We recall that the Markov chain (x_n) given by Eq. (10) admits the representation (2), where the function $g_\gamma = (g_\gamma^1, \ldots, g_\gamma^N)$ is given by (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 (19), Assumption (RM)–v)) is verified. Finally, the finiteness assumption (20) with $\epsilon_K = 2$ follows from the existence of second moments for the A_n^k, and (21) is immediate. The rest of the proof follows word for word the proof of Theorem 5.1. □

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

Proposition 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 (13) is satisfied, then for all $\varepsilon > 0$,

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

To prove this proposition, we essentially show that the assumptions of Theorem 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 Theorem 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, Proposition 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 (13) is satisfied. Let

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

Given $1 \le k, \ell \le N$, define $f(x) = x^k x^\ell$ on $\gamma\mathbb{N}^2$. Using Eq. (10), the iid property of the process $((A_n^1, \ldots, A_n^N, B_n^1, \ldots, B_n^N), n \in \mathbb{N})$ 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 (13) is satisfied, the function ψ is coercive, and one can straightforwardly check that the statements of Proposition 8.7–i)) and Proposition 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 $P \in \mathcal{P}$, $\{\mathbb{E}^{\nu_\gamma, P_\gamma} \Lambda_n\,, \gamma \in (0,\gamma_0), n \in \mathbb{N}\}$ is tight, since $\sup_{\gamma \in (0,\gamma_0), n \in \mathbb{N}} \mathbb{E}^{\nu_\gamma, P_\gamma} \Lambda_n(\psi) < \infty$. We can now follow the proof of Theorem 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 (13) is satisfied, all the trajectories of the flow Φ_{H} converge to zero, as shown in [19, § 3.2]. □

References

[1] 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.

[2] 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.

[3] U. Ayesta, M. Erausquin, M. Jonckheere, and I. M. Verloop. Scheduling in a random environment: stability and asymptotic optimality. *IEEE/ACM Transactions on Networking (TON)*, 21(1):258–271, 2013.

[4] 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.

[5] 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.

[6] 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.

[7] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1):328–348 (electronic), 2005.

[8] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. II. Applications. *Math. Oper. Res.*, 31(4):673–695, 2006.

[9] M. Benaïm and S. J. Schreiber. Ergodic properties of weak asymptotic pseudotrajectories for semiflows. *J. Dynam. Differential Equations*, 12(3):579–598, 2000.

[10] 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.

[11] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM J. Optim.*, 26(4):2235–2260, 2016.

[12] 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.

[13] 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.

[14] V. S. Borkar. *Stochastic approximation. A dynamical systems viewpoint.* Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008.

[15] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.

[16] 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.

[17] M. Faure and G. Roth. Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems. *Stoch. Dyn.*, 13(1):1250011, 23, 2013.

[18] 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.

[19] N. Gast and B. Gaujal. Markov chains with discontinuous drifts have differential inclusion limits. *Performance Evaluation*, 69(12):623 – 642, 2012.

[20] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[21] 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.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] W. Miller and E. Akin. Invariant measures for set-valued dynamical systems. *Trans. Amer. Math. Soc.*, 351(3):1203–1225, 1999.

[24] I. Molchanov. *Theory of random sets*. Probability and its Applications (New York). Springer-Verlag London, Ltd., London, 2005.

[25] R. R. Phelps. Lectures on maximal monotone operators. *Extracta Math.*, 12(3):193–230, 1997.

[26] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[27] G. Roth and W. H. Sandholm. Stochastic approximations with constant step size and differential inclusions. *SIAM J. Control Optim.*, 51(1):525–555, 2013.

[28] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

[29] V. Yaji and S. Bhatnagar. Stochastic recursive inclusions with non-additive iterate-dependent Markov noise. *arXiv preprint arXiv:1607.04735*, 2016.

[30] J. Yoon and S. J. Hwang. Combined group and exclusive sparsity for deep neural networks. In *International Conference on Machine Learning*, pages 3958–3966, 2017.