Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Stochastic Recursive Inclusions in Two Timescales with Nonadditive Iterate-Dependent Markov Noise

Vinayaka G. Yaji, Shalabh Bhatnagar

informs
http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 45, No. 4, November 2020, pp. 1405–1444
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Stochastic Recursive Inclusions in Two Timescales with Nonadditive Iterate-Dependent Markov Noise

Vinayaka G. Yaji,[a] Shalabh Bhatnagar[a]

[a] Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India

Contact: vgyaji@gmail.com, http://orcid.org/0000-0002-8438-6283 (VGY); shalabh@iisc.ac.in, http://orcid.org/0000-0001-7644-3914 (SB)

Received: October 30, 2016
Revised: October 25, 2017; December 14, 2018; June 17, 2019
Accepted: July 20, 2019
Published Online in Articles in Advance: July 21, 2020

Subject Classifications: 62L20, 26E25, 28B20, 93E20

https://doi.org/10.1287/moor.2019.1037

Abstract. In this paper, we study the asymptotic behavior of a stochastic approximation scheme on two timescales with set-valued drift functions and in the presence of nonadditive iterate-dependent Markov noise. We show that the recursion on each timescale tracks the flow of a differential inclusion obtained by averaging the set-valued drift function in the recursion with respect to a set of measures accounting for both averaging with respect to the stationary distributions of the Markov noise terms and the interdependence between the two recursions on different timescales. The framework studied in this paper builds on a recent work by Ramaswamy and Bhatnagar, by allowing for the presence of nonadditive iterate-dependent Markov noise. As an application, we consider the problem of computing the optimum in a constrained convex optimization problem, where the objective function and the constraints are averaged with respect to the stationary distribution of an underlying Markov chain. Further, the proposed scheme neither requires the differentiability of the objective function nor the knowledge of the averaging measure.

Funding: This work was supported under the ICPS Programme by the Department of Science and Technology, Government of India, as well as the Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bangalore.

Keywords: stochastic approximation • set-valued maps • iterate-dependent Markov noise • differential inclusions • constrained convex optimization

1. Introduction

Stochastic approximation, introduced by Robbins and Monro [37] as a tool for statistical computation, is today a main paradigm for online algorithms for system identification, adaptive control, and optimization. A comprehensive and detailed account of the field can be found in texts by Kushner and Yin [25] and Borkar [11]. Standard stochastic approximation scheme is given by

$$X_{n+1} - X_n - a(n)M_{n+1} = a(n)h(X_n), \tag{1}$$

where $\{X_n\}_{n\geq 0}$ (iterate sequence) and $\{M_n\}_{n\geq 0}$ (noise sequence, usually assumed to be unbiased with restrictions on its variance) are sequences of $\mathbb{R}^d$-valued random variables, and $h:\mathbb{R}^d \to \mathbb{R}^d$ is a Lipschitz continuous function. Further, $\{a(n)\}_{n\geq 0}$ is a sequence of positive real numbers denoting the step-size sequence. The convergence analysis of Recursion (1), where one wishes to show that the iterates X_n converge to $x^* \in \mathbb{R}^d$, belonging to the zero set of the function $h(\cdot)$, is usually accomplished using *the martingale method* (see Duflo [17]) or the *ordinary differential equation*, or *o.d.e.*, *method* (introduced by Derevitskii and Fradkov [16] and later generalized by Benaïm et al. [4] to include set-valued dynamical systems). The martingale method involves showing that the norm distance between the iterate and the desired solution goes to zero using the generalized martingale convergence theorem, whereas the o.d.e. method involves the identification of a dynamical system whose solution trajectories the iterates are shown to track. The method of choice usually depends on the ease of verification of various assumptions involved, in a particular application setting. For example, in stochastic approximation schemes used to solve stochastic variational inequality problems (SVIP), the martingale method has been found to be extremely useful (see Jiang and Xu [20], Koshal et al. [23], Iusem et al. [19]), whereas the o.d.e. method is applied in the analysis of learning algorithms and gradient-based schemes (Sutton et al. [40], Borkar and Meyn [13], Bhatnagar and Prashanth [7]). Different noise models also have been, most notably the martingale noise model and the Markov noise model. The study of stochastic approximations with Markov noise terms was pioneered by Metivier and Priouret [28], where the assumptions on the law of the Markov noise terms guarantees the existence of a solution to Poisson equation, which is then used to convert the Markov noise case to stochastic approximation with Martingale noise terms. An extension of the above method when the drift function is discontinuous can be found in Fort et al. [18]. An alternate set of Markov noise conditions that does not require the existence of solution to the Poisson equation was studied by Borkar [10] (which we extend to our framework presented later).

Under the o.d.e. method of analysis, different step sizes for different components of the iterate in Recursion (1) introduces different timescales, if the step sizes satisfy a certain *rate* condition. Such schemes are known as multitimescale stochastic approximation schemes and are well studied in the presence of martingale noise with set-valued or single-valued dynamical systems (Ramaswamy and Bhatnagar [36], Perkins and Leslie [34], Borkar [9]) and Markov noise with single-valued dynamical system (Karmakar and Bhatnagar [21]). In this paper, we present a two-timescale stochastic approximation scheme with set-valued maps and nonadditive iterate-dependent Markov noise and study its asymptotic behavior. The framework studied in this paper, to the best of our knowledge, is the most general two-timescale stochastic approximation framework studied to date.

1.1. Overview of Two-Timescale Stochastic Approximation Schemes

The standard two-timescale stochastic approximation scheme is given by

$$Y_{n+1} - Y_n - b(n)M^{(2)}_{n+1} = b(n)h_2(X_n, Y_n), \tag{2a}$$

$$X_{n+1} - X_n - a(n)M^{(1)}_{n+1} = a(n)h_1(X_n, Y_n), \tag{2b}$$

where $n \geq 0$ denotes the iteration index, $\{X_n\}_{n\geq 0}$ is a sequence of $\mathbb{R}^{d_1}$-valued random variables, $\{Y_n\}_{n\geq 0}$ is a sequence of $\mathbb{R}^{d_2}$-valued random variables, for any $i \in \{1,2\}$, $h_i : \mathbb{R}^{d_1+d_2} \to \mathbb{R}^{d_i}$ is a Lipschitz continuous function, and for every $i \in \{1,2\}$, $\{M^{(i)}_n\}_{n\geq 1}$ is a $\mathbb{R}^{d_i}$-valued square-integrable martingale difference sequence. The step-size sequences $\{a(n)\}_{n\geq 0}$ and $\{b(n)\}_{n\geq 0}$ are sequences of positive real numbers chosen such that they satisfy $\lim_{n\to\infty} \frac{b(n)}{a(n)} = 0$ in addition to the Monte Carlo step-size conditions. The condition $\lim_{n\to\infty} \frac{b(n)}{a(n)} = 0$ ensures that after a large number of iterations, the time step of Recursion (2a) is much smaller than that of (2b). Thus, the Recursion (2a) appears to be static with respect to the Recursion (2b). Borkar [9], using the dynamical systems approach studied by Benaim [3], showed the above intuition to hold. More precisely, the faster timescale Recursion (2b), was shown to track the ordinary differential equation (o.d.e.) given by

$$\frac{dx}{dt} = h_1(x, y_0), \tag{3}$$

for some $y_0 \in \mathbb{R}^{d_2}$, and assuming that for every $y \in \mathbb{R}^{d_2}$, o.d.e. (3) admits a unique globally asymptotically stable equilibrium point, say $\lambda(y)$, the slower timescale Recursion (2a) was shown to track the o.d.e. given by,

$$\frac{dy}{dt} = h_2(\lambda(y), y). \tag{4}$$

Further, the map $y \to \lambda(y)$ was assumed to be Lipschitz continuous.

An important application of the above stochastic approximation scheme is in the computation of a saddle point of a function. Given a function $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, $(x^*, y^*) \in \mathbb{R}^{d_1+d_2}$ ($x^* \in \mathbb{R}^{d_1}$ and $y^* \in \mathbb{R}^{d_2}$, respectively) is a saddle point of the function $f(\cdot,\cdot)$ if

$$\inf_{x\in\mathbb{R}^{d_1}} \sup_{y\in\mathbb{R}^{d_2}} f(x,y) = \sup_{y\in\mathbb{R}^{d_2}} \inf_{x\in\mathbb{R}^{d_1}} f(x,y) = f(x^*, y^*). \tag{5}$$

From Bertsekas [6, proposition 5.5.6], we know that the function $f(\cdot,\cdot)$ admits a saddle point if for every $(x,y) \in \mathbb{R}^d$,

(1) $-f(x,\cdot)$ and $f(\cdot,y)$ are convex functions, and

(2) the sub level sets of functions $x \to \sup_{y\in\mathbb{R}^{d_2}} f(x,y)$ and $y \to -\inf_{x\in\mathbb{R}^{d_1}} f(x,y)$ are compact sets.

Over the years, significant effort has been devoted to developing algorithms to compute such points (see Nedić and Ozdaglar [32] and Benzi et al. [5] and the references therein). Most of the solutions proposed in the literature require the computation of partial derivatives of the function $f(\cdot,\cdot)$. However, in practice, the closed-form expressions of the partial derivatives are often not known or are expensive to compute, and in such cases, one often estimates the partial derivatives using values of the objective function (see Spall [39] for one such estimation method). The two-timescale stochastic approximation scheme can be used to compute a saddle point with noisy partial derivative values by setting $h_1(x,y) := -\nabla_x f(x,y)$ and $h_2(x,y) := \nabla_y f(x,y)$, where ∇_x and ∇_y denote the partial derivative operators with respect to x and y, respectively. In this setting, the sequences

$\{M_n^{(1)}\}_{n\geq 1}$ and $\{M_n^{(2)}\}_{n\geq 1}$ could contain the partial derivative estimation errors, and the map $\lambda(\cdot)$ denotes correspondence between $y \in \mathbb{R}^{d_2}$ and the minimum of the function $f(\cdot, y)$. The vector field associated with o.d.e. (4) is now given by $\nabla_y f(x,y)|_{x=\lambda(y)}$, which can be shown to be the same as $\nabla_y f(\lambda(y), y)$ using the *envelope theorem* from mathematical economics (see Milgrom and Segal [30]). Thus, the slower timescale maximizes the function $y \to \inf_{x\in\mathbb{R}^{d_1}} f(x,y) = f(\lambda(y), y)$, thereby at the limit, the iterates of Recursion (2) converge to a saddle point of the function $f(\cdot,\cdot)$.

In some cases, the function whose saddle point needs to be computed is itself averaged with respect to a certain probability measure. For example, consider the function $f: \mathbb{R}^{d_1+d_2} \times \mathcal{S} \to \mathbb{R}$, where $\mathcal{S}$ is a compact metric space, and for some probability measure μ on $\mathcal{S}$, one wishes to compute the saddle point of the function $f_\mu : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$, where for every $(x,y) \in \mathbb{R}^{d_1+d_2}$, $f_\mu(x,y) := \int_{\mathcal{S}} f(x,y,s)\mu(ds)$. If one has access to i.i.d. samples with probability measure μ, then the saddle point problem above can be solved using Recursion (2). But if access to such samples is not available and one uses Markov chain Monte Carlo methods to sample from the measure μ, then Recursion (2) has a nonadditive iterate-dependent Markov noise component. The recursion in this case takes the following form:

$$Y_{n+1} - Y_n - b(n)M_{n+1}^{(2)} = b(n)h_2\left(X_n, Y_n, S_n^{(2)}\right), \tag{6a}$$

$$X_{n+1} - X_n - a(n)M_{n+1}^{(1)} = a(n)h_1\left(X_n, Y_n, S_n^{(1)}\right), \tag{6b}$$

where $\{S_n^{(1)}\}_{n\geq 1}$ and $\{S_n^{(2)}\}_{n\geq 1}$ denote the Markov noise terms taking values in an appropriate state space. The Recursion (6) was studied by Karmakar and Bhatnagar [21] under assumptions similar to those of Borkar [9] that include the Lipschitz continuity of the maps h_1, h_2, and λ.

Often, the maps h_1 and h_2 in Recursion (2) are not Lipschitz continuous, and the map λ is not even single valued (i.e., the o.d.e. (3) has a globally asymptotically stable equilibrium set). This motivates one to study the two-timescale recursion with set-valued drift functions. The recursion then takes the following form:

$$Y_{n+1} - Y_n - b(n)M_{n+1}^{(2)} \in b(n)H_2(X_n, Y_n), \tag{7a}$$

$$X_{n+1} - X_n - a(n)M_{n+1}^{(1)} \in a(n)H_1(X_n, Y_n), \tag{7b}$$

where H_1 and H_2 are set-valued maps and other quantities have similar interpretation to those in Recursion (2). The above recursion was studied by Ramaswamy and Bhatnagar [36], and the map λ was allowed to be set valued and upper semicontinuous.

1.2. Contributions of This Paper and Comparisons with the State of the Art

In this paper, we study the asymptotic behavior of the recursion given by

$$Y_{n+1} - Y_n - b(n)M_{n+1}^{(2)} \in b(n)H_2\left(X_n, Y_n, S_n^{(2)}\right), \tag{8a}$$

$$X_{n+1} - X_n - a(n)M_{n+1}^{(1)} \in a(n)H_1\left(X_n, Y_n, S_n^{(1)}\right), \tag{8b}$$

where H_1 and H_2 are set-valued maps and $\{S_n^{(1)}\}_{n\geq 0}$ and $\{S_n^{(2)}\}_{n\geq 0}$ are the Markov noise terms taking values in compact metric spaces $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$, respectively. We show that the fast timescale Recursion (8b) tracks the flow of the differential inclusion (DI) given by

$$\frac{dx}{dt} \in \cup_{\mu\in D^{(1)}(x,y)} \int_{\mathcal{S}^{(1)}} H_1\left(x, y_0, s^{(1)}\right)\mu\left(ds^{(1)}\right), \tag{9}$$

for some $y_0 \in \mathbb{R}^{d_2}$, where $D^{(1)}(x,y)$ denotes the set of stationary distributions of the Markov noise terms $\{S_n^{(1)}\}_{n\geq 0}$ for every $(x,y) \in \mathbb{R}^d$, and the integral above denotes the integral of a set-valued map with respect to measure μ. (This is explained in Section 2.2, see Definition 6.) Further, we assume that for every $y \in \mathbb{R}^{d_2}$, the above DI admits a unique globally attracting set $\lambda(y)$. The map $y \to \lambda(y)$ is also assumed to be upper semicontinuous (see Definition 1 in Section 2.1). The slower timescale Recursion (8a) is shown to track the flow of the DI given by

$$\frac{dy}{dt} \in \cup_{\mu\in D(y)} \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H_2\left(x, y, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right), \tag{10}$$

where $y \to D(y)$ denotes a set-valued map taking values in the space of probability measures on $\mathcal{S}^2$, and the map $D(\cdot)$ is defined such that it captures both the equilibration of the fast timescale iterates to $\lambda(\cdot)$ and the averaging due to the Markov noise terms $\{S_n^{(2)}\}_{n\geq 0}$.

In comparison with the two-timescale framework studied by Karmakar and Bhatnagar [21], our work allows for the drift functions (i.e., H_1 and H_2) to be set valued, and the map $\lambda(\cdot)$ is allowed to be set valued and upper semicontinuous, which is much weaker than the requirement imposed in Karmakar and Bhatnagar [21] of being single valued and Lipschitz continuous. The generalization to the set-valued case allows one to analyze Recursion (6) when the drift functions h_1 and h_2 are single valued and are just measurable, because the graph of such a map can be embedded in the graph of a upper semicontinuous set-valued map as in Borkar [11, chapter 5.3(iv)]. We refer the reader to Borkar [11, chapter 5.3] for several other scenarios in which the study of stochastic approximation scheme with set-valued maps becomes essential.

Our work further generalizes the two-timescale framework studied in Ramaswamy and Bhatnagar [36] by allowing for the presence of Markov noise terms. The analysis in this paper does not extend in a straightforward manner from that by Ramaswamy and Bhatnagar [36] and requires results from set-valued map approximations, parametrization, integration, and the use of probability measure-valued functions. However, the method of analysis adopted in this paper can be appropriately adapted to obtain the same convergence guarantees as in Ramaswamy and Bhatnagar [36] when the Markov noise terms are absent.

1.3. Overview of the Analysis and Organization of the Paper

It is known that continuous, convex, and compact set-valued maps taking values in a finite dimensional space admit a continuous single-valued parametrization. The properties of the set-valued drift function ensure that the drift functions H_1 and H_2 are convex and compact set-valued maps and are upper semicontinuous. However, such maps do not admit a continuous parametrization. We can work around this problem by enlarging the graph of the drift function because the graph of drift function can be embedded in the graph of a continuous, convex, and compact set-valued map that admits a continuous single-valued parametrization. Thus, a sequence of continuous single-valued maps can be obtained that approximate the set-valued drift function from above. This enables us to write the Inclusion (8) in the form of Recursion (6) with an additional parameter. The results needed to accomplish the above are stated in Section 2.1.

Before proceeding further, one needs to identify the mean fields that the Recursion (8) is expected to track. To this end, we need some results from the theory of integration of set-valued maps that are reviewed in Section 2.2. Further, the measurablility and integrability properties of the drift functions of the recursion are investigated, and the characterization of the integral of a continuous set-valued map in terms of its parametrization is established.

In Section 2.3, we compile some definitions and results from the theory of differential inclusions that are needed later to characterize the asymptotic behavior of Recursion (8). Further, in Section 2.4, we state the assumptions and the main result of the analysis of single timescale stochastic recursive inclusions with nonadditive iterate-dependent Markov noise from Yaji and Bhatnagar [41], and in Section 2.5, we define and compile some results needed from theory of the space of probability measure-valued functions.

In Section 3, we state and motivate the assumptions under which the Recursion (8) is analyzed. Using the results from integration of set-valued maps reviewed in Section 2.2, the mean fields are defined and the main convergence result is stated. The mean fields defined in Section 3 possess some properties that ensure the existence of solutions (of their associated differential inclusions). These properties are established in Section 4. This section also shows that appropriate modifications of the continuous set-valued maps that approximate the drift functions (obtained in Section 2.1) approximate the mean fields, which play an important role in the analysis later.

The analysis of Recursion (8) consists of two parts. In Section 5.1, the Recursion (8) is analyzed along the faster timescale. The Recursion (8) when viewed along the faster timescale appears to be a single timescale stochastic recursive inclusion with nonadditive iterate-dependent Markov noise. In Section 5.1, we show that Recursion (8) viewed along the faster timescale satisfies all the assumptions associated with the single timescale recursion presented in Section 2.4. Applying the main result of single timescale analysis, we conclude that the faster timescale iterates converge to $\lambda(y)$ for some $y \in \mathbb{R}^{d_2}$. In Section 5.2, the slower timescale recursion is analyzed. It is shown that the linearly interpolated sample path of the slower timescale iterates (defined in Section 5.2.1) tracks an appropriate DI. Continuous functions tracking the flow of a dynamical system are known as *asymptotic pseudotrajectories* (see Benaïm et al. [4] for definition and related results). The asymptotic pseudotrajectory argument in this paper presented in Section 5.2.2 comprises the following steps:

(1) The first step is to get rid of the additive noise terms, $\{M_n^{(2)}\}_{n\geq 1}$. This involves defining an o.d.e. with an appropriate piecewise constant vector field and showing that the limit points of the shifted linearly interpolated trajectory of the slower timescale iterates coincide with the limit points of the solutions of this o.d.e. in the space of continuous functions on $[0,\infty)$ taking values in $\mathbb{R}^{d_2}$. Further, a simple argument gives us that the set of limit points of the shifted linearly interpolated trajectories of the slower timescale iterates is nonempty.

(2) The second step is to show that the limit point obtained in the first step is in fact a solution of DI (10). This is accomplished using probability measure-valued functions reviewed in Section 2.5. This method also has been used in analyzing stochastic approximation schemes as in Recursion (6) by Karmakar and Bhatnagar [21] and by Borkar [10]. But the analysis in these references made explicit use of the Lipschitz property of the underlying drift functions. We observe that continuity is sufficient to carry out this analysis. This is also where our analysis significantly differs from that of Ramaswamy and Bhatnagar [36]. That the equilibration of the faster timescale is also accomplished using probability measures simplifies the proof compared with that of Ramaswamy and Bhatnagar [36].

In Section 5.2.3, the limit sets of the slower timescale iterates are characterized in terms of the dynamics of DI (10). In addition to the above, using the convergence of the faster timescale iterates to $\lambda(\cdot)$ obtained in Section 5.1, we obtain the main convergence result of this paper (i.e., Theorem 4ii).

In Section 7, as applications, we provide methods for computing a solution to constrained and unconstrained optimization problems. In the first application, we present an algorithm to compute an ϵ-optimal solution of a constrained convex optimization, where the optimization problem is obtained by averaging the quantities involved with respect to the stationary distribution of an underlying Markov chain. Such problems arise in optimal control where the controller must find an optimum parameter where the changes in state of the underlying system can be modeled by a Markov chain. The cost function and system constraints are dependent on the state of the system, and the controller seeks to find the optimum of the long-run average of the cost function while satisfying the long-run average constraints. In such applications, the stationary distribution of the system states is not known, but one has access to a sample path of system state changes. We propose a two-timescale scheme, which performs primal ascent along the faster timescale and dual descent along the slower timescale with the knowledge of the current state at a given iteration. Using the theory presented in this paper, it is shown that the limit set of the iterates of the proposed two-timescale scheme are contained in the set of Lagrangian saddle points of the underlying averaged constrained convex optimization problem. Further, the algorithm does not assume the differentiability of the objective function and requires only a noisy estimate of the subgradient. In the second application, we present an algorithm to compute an ϵ-optimal solution of an unconstrained optimization problem, using newtons method, where on the faster timescale, the approximation of the newton update is computed without the need for explicitly inverting the hessian matrix of the objective function. Further, we allow for a general nondiminishing noise model for the error in hessian and gradient estimation.

In Section 8, we conclude by providing a few directions for future research and outline certain extensions where we believe the analysis remains the same.

2. Background

In this section, we will briefly review some results needed from the theory of set-valued maps and differential inclusions, present a brief outline of the analysis of the single timescale version of stochastic recursive inclusions with nonadditive iterate-dependent Markov noise, and define the space of probability measure-valued functions, with a metrizable topology, which are needed later in the analysis of the two-timescale recursion.

Throughout this paper, $\mathcal{S}$ denotes a compact metric space and the metric on $\mathcal{S}$ is denoted by $d_{\mathcal{S}}$. Further, let $1 \leq d_1 \in \mathbb{Z}$, $1 \leq d_2 \in \mathbb{Z}$, $d := d_1 + d_2$, and (x,y) denote a generic element in $\mathbb{R}^d$, where $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$.

2.1. Upper Semicontinuous Set-Valued Maps and Their Approximation

First, we will recall the notions of upper semicontinuity, lower semicontinuity, and continuity of set-valued maps. These notions are taken from Aubin and Cellina [2, chapter 1.1].

Definition 1. A set-valued map $F : \mathbb{R}^d \times \mathcal{S} \rightarrow \{\text{subsets of } \mathbb{R}^k\}$ is

- *Upper semicontinuous* (u.s.c.) if, for every $(x_0, y_0, s_0) \in \mathbb{R}^d \times \mathcal{S}$, for every $\epsilon > 0$, there exists $\delta > 0$ (depending on (x_0, y_0, s_0) and ϵ) such that,

$$\| (x,y) - (x_0,y_0) \| < \delta,\ d_{\mathcal{S}}(s,s_0) < \delta \Rightarrow F(x,y,s) \subseteq F(x_0,y_0,s_0) + \epsilon U,$$

where U denotes the closed unit ball in $\mathbb{R}^k$.

- *Lower semicontinuous* (l.s.c) if, for every $(x_0, y_0, s_0) \in \mathbb{R}^d \times \mathcal{S}$, for every $z_0 \in F(x_0, y_0, s_0)$, for every sequence $\{(x_n, y_n, s_n)\}_{n\geq 1}$ converging to (x_0, y_0, s_0), there exists a sequence $\{z_n \in F(x_n, y_n, s_n)\}$ converging to z_0.
- *Continuous* if it is both u.s.c. and l.s.c.

For set-valued maps taking compact set values, we have the above mentioned notion of u.s.c. to be equivalent to the standard notion of u.s.c. (see Aubin and Cellina [2, p. 45]). In this paper, we will encounter set-valued maps that are compact set valued, and hence we have chosen to state the above as the definition of upper semicontinuity.

Set-valued maps studied later satisfy certain properties under which we will be able to approximate them with a family of continuous single-valued maps with an additional parameter. These properties are natural extensions of the properties imposed on maps studied by Benaïm et al. [4] and Ramaswamy and Bhatnagar [36] to the case of stochastic recursive inclusions with Markov noise, and we choose to call such maps *stochastic approximation maps* (SAM). The definition of SAM is stated below.

Definition 2 (SAM)**.** A set-valued map $F : \mathbb{R}^d \times \mathcal{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ is a stochastic approximation map if,

(a) for every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $F(x, y, s)$ is a convex and compact subset of $\mathbb{R}^k$,

(b) for every $(x_0, y_0, s_0) \in \mathbb{R}^d \times \mathcal{S}$, for every $\mathbb{R}^d \times \mathcal{S}$ sequence, say $\{(x_n, y_n, s_n)\}_{n\geq 1}$ converging to (x_0, y_0, s_0) and a sequence $\{z_n \in F(x_n, y_n, s_n)\}_{n\geq 0}$ converging to $z \in \mathbb{R}^k$, we have that $z \in F(x_0, y_0, s_0)$, and

(c) there exists $K > 0$ such that for every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $\sup_{z\in F(x,y,s)} \|z\| \leq K(1 + \|(x, y)\|)$.

For SAM appearing in two-timescale stochastic recursive inclusions, the condition (c), which is stated above, is replaced by an equivalent condition:

(c)′ there exists $K > 0$ such that for every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $\sup_{z\in F(x,y,s)} \|z\| \leq K(1 + \|x\| + \|y\|)$.

The condition (b) in the definition of SAM tells us that the graph of the set-valued map F, $\mathcal{G}(F)$, defined as

$$\mathcal{G}(F) := \{(x, y, s, z) : z \in F(x, y, s),\ (x, y, s) \in \mathbb{R}^d \times \mathcal{S}\} \subseteq \mathbb{R}^d \times \mathcal{S} \times \mathbb{R}^k,$$

is closed and hence the said condition is known as the *closed graph property*. The condition (c) (or $(c)'$) is known as the *pointwise boundedness condition*, and it makes sure that the "size" of the sets linearly grows with the distance from the origin. This is the only condition where we differ from the conditions imposed by Benaïm et al. [4] and Ramaswamy and Bhatnagar [36]. It is easy to show that when the Markov noise component is absent, condition (c) (or $(c)'$) imposed in this paper is the same as the one in Benaïm et al. [4] (Ramaswamy and Bhatnagar [36]).

As a consequence of the properties possessed by a SAM, F, one can show that the map F is u.s.c. This claim follows from arguments similar to those in Aubin and Cellina [2, chapter 1.1, corollary 1] and is stated as a lemma below.

Lemma 1 (u.s.c.)**.** *A set-valued map F, which is a SAM, is u.s.c.*

The graph of a convex and compact u.s.c. set-valued map can be embedded in the graph of a sequence of decreasing continuous set-valued maps. The following statement is made precise in the following lemma.

Lemma 2 (Continuous embedding)**.** *For any set-valued map F that is a SAM, there exists a sequence of set-valued maps $\{F^{(l)}\}_{l\geq 1}$ such that for every $l \geq 1$, $F^{(l)} : \mathbb{R}^d \times \mathcal{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ is continuous and satisfies the following.*

i. *For every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $F^{(l)}(x, y, s)$ is a convex and compact subset of $\mathbb{R}^k$.*

ii. *For every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $F(x, y, s) \subseteq F^{(l+1)}(x, y, s) \subseteq F^{(l)}(x, y, s)$.*

iii. *There exists $K^{(l)} > 0$ such that for every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $\sup_{z\in F^{(l)}(x,y,s)} \|z\| \leq K^{(l)}(1 + \|(x, y)\|)$. (If the set-valued map F satisfies condition (c)′ instead of (c) in the definition of SAM, we have $\sup_{z\in F^{(l)}(x,y,s)} \|z\| \leq K^{(l)}(1 + \|x\| + \|y\|)$).*

Furthermore, for every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, $\cap_{l\geq 1} F^{(l)}(x, y, s) = F(x, y, s)$.

The statement of the above lemma can be found in Aubin and Cellina [2, p. 39] and the proof is similar to the proof by Aubin and Cellina [2, chapter 1.13, theorem 1] (a brief outline can be found in Yaji and Bhatnagar [41, appendix A]). The following are some useful observations from the proof of Lemma 2.

1. $\sup_{l\geq 1} K^{(l)}$ is finite and let $\tilde{K} := \sup_{l\geq 1} K^{(l)}$, and
2. for every $(x, y, s) \in \mathbb{R}^d \times \mathcal{S}$, for every $\epsilon > 0$, there exists L (depending on ϵ and (x, y, s)), such that for every $l \geq L$, $F^{(l)}(x, y, s) \subseteq F(x, y, s) + \epsilon U$ where U denotes the closed unit ball in $\mathbb{R}^k$.

Continuous set-valued maps admit a parametrization by which we mean that a continuous single-valued map can be obtained that represents the set-valued map in the sense made precise in the lemma below, which follows from Aubin and Cellina [2, chapter 1.7, theorem 2].

Lemma 3 (Parametrization)**.** *Let F be a SAM, and for every $l \geq 1$, let the set-valued map $F^{(l)}$ be as in Lemma 2. Then for every $l \geq 1$, there exists a continuous single-valued map $f^{(l)} : \mathbb{R}^d \times \mathcal{S} \times U \to \mathbb{R}^k$, where U denotes the closed unit ball in $\mathbb{R}^k$, such that*

i. *for every $(x,y,s) \in \mathbb{R}^d \times \mathcal{S}$, $F^{(l)}(x,y,s) = f^{(l)}(x,y,s,U)$ where $f^{(l)}(x,y,s,U) = \{f^{(l)}(x,y,s,u) : u \in U\}$, and*

ii. *for $K^{(l)}$ as in Lemma 2iii, for every $(x,y,s,u) \in \mathbb{R}^d \times \mathcal{S} \times U$, we have that $\|f^{(l)}(x,y,s,u)\| \leq K^{(l)}(1 + \|(x,y)\|)$ (If the set-valued map F satisfies condition (c)′ instead of (c) in the definition of SAM, we have $\|f^{(l)}(x,y,s,u)\| \leq K^{(l)}(1 + \|x\| + \|y\|)$.)*

Throughout this paper, we will use U to denote the closed unit ball in $\mathbb{R}^k$, where the dimension k will be made clear by the context.

Remark 1. To summarize, as a consequence of Lemma 2, we have that corresponding to a set-valued SAM (which is upper semicontinuous), there exists a sequence of continuous set-valued maps, denoted by $F^{(l)}$, within which the SAM may be embedded and the sequence of continuous set-valued maps $\{F^{(l)}\}_{l\geq 1}$ approximate F as explained in Lemma 2. Further, as a consequence of Lemma 3, we have that corresponding to every member of the sequence, there exists an appropriately defined continuous single-valued map denoted by $f^{(l)}$.

Combining Lemmas 2 and 3 we obtain the approximation theorem stated below.

Theorem 1 (Approximation)**.** *For any SAM F, there exists a sequence of continuous functions $\{f^{(l)}\}_{l\geq 1}$ such that for every $l \geq 1$, $f^{(l)} : \mathbb{R}^d \times \mathcal{S} \times U \to \mathbb{R}^k$ is such that,*

i. *for every $(x,y,s) \in \mathbb{R}^d \times \mathcal{S}$, $F(x,y,s) \subseteq f^{(l+1)}(x,y,s,U) \subseteq f^{(l)}(x,y,s,U)$ and $f^{(l)}(x,y,s,U)$ is a convex and compact subset of $\mathbb{R}^k$, and*

ii. *for $K^{(l)}$ as in Lemma 2iii, for every $(x,y,s,u) \in \mathbb{R}^d \times \mathcal{S} \times U$, we have that $\|f^{(l)}(x,y,s,u)\| \leq K^{(l)}(1 + \|(x,y)\|)$ (If the set-valued map F satisfies condition (c)′ instead of (c) in the definition of SAM, we have $\|f^{(l)}(x,y,s,u)\| \leq K^{(l)}(1 + \|x\| + \|y\|)$.)*

Furthermore, for every $(x,y,s) \in \mathbb{R}^d \times \mathcal{S}$, $F(x,y,s) = \cap_{l\geq 1} f^{(l)}(x,y,s,U)$.

2.2. Measurable Set-Valued Maps and Integration

Here, we will review concepts of measurability and integration of set-valued maps. These concepts will be needed to define the limiting differential inclusion, which the recursion studied in this paper is expected to track.

Let $(\mathcal{W}, \mathscr{F}_{\mathcal{W}})$ denote a measurable space and $F : \mathcal{W} \to \{\text{subsets of } \mathbb{R}^k\}$ be a set-valued map such that, for every $w \in \mathcal{W}$, $F(w)$ is a nonempty closed subset of $\mathbb{R}^k$. Throughout this subsection, F refers to the set-valued map as defined above.

Definition 3 (Measurable set-valued map)**.** A set-valued map F is measurable if for every $C \subseteq \mathbb{R}^k$, closed,

$$F^{-1}(C) := \{w \in \mathcal{W} : F(w) \cap C \neq \emptyset\} \in \mathscr{F}_{\mathcal{W}}.$$

We refer the reader to Li et al. [27, theorem 1.2.3] for other notions of measurability and their relation to the definition above.

Definition 4 (Measurable selection)**.** A function $f : \mathcal{W} \to \mathbb{R}^k$ is a measurable selection of a set-valued map F if f is measurable and for every $w \in \mathcal{W}$, $f(w) \in F(w)$.

For a set-valued map F, let $\mathscr{S}(F)$ denote the set of all measurable selections. The next lemma summarizes some standard results about measurable set-valued maps and their measurable selections.

Lemma 4. *For any measurable set-valued map F,*

i. *$\mathscr{S}(F) \neq \emptyset$.*

ii. *(Castaing representation) there exists $\{f_n\}_{n\geq 1} \subseteq \mathscr{S}(F)$ such that, for every $w \in \mathcal{W}$, $F(w) = cl(\{f_n(w)\}_{n\geq 1})$, where $cl(\cdot)$ denotes the closure of a set.*

We refer the reader to Li et al. [27, theorems 1.2.6 and 1.2.7] for the proofs of Lemma 4i and ii, respectively.

Definition 5 (μ-integrable set-valued map)**.** Let μ be a probability measure on $(\mathcal{W}, \mathscr{F}_{\mathcal{W}})$. A measurable set-valued map F is said to be μ-integrable if there exists $f \in \mathscr{S}(F)$, which is μ-integrable.

Definition 6 (Aumann's integral). Let μ be a probability measure on $(\mathscr{W}, \mathscr{F}_{\mathscr{W}})$. The integral of a μ-integrable set-valued map F is defined as,

$$\int_{\mathscr{W}} F(w)\mu(dw) := \left\{ \int_{\mathscr{W}} f(w)\mu(dw) : \ f \in \mathscr{S}(F), \ f \ is \ \mu - integrable \right\}.$$

The next lemma states a useful result on the properties of the integral of a set-valued map, which is convex and compact set valued.

Lemma 5. *Let μ be a probability measure on $(\mathscr{W}, \mathscr{F}_{\mathscr{W}})$ and F a μ-integrable set-valued map such that for every $w \in \mathscr{W}$, $F(w)$ is convex and compact. Then, $\int_{\mathscr{W}} F(w)\mu(dw)$ is a convex and closed subset of $\mathbb{R}^k$.*

For a proof of the above lemma, we refer the reader to Li et al. [27, theorem 2.2.2].

Now, we will briefly investigate the measurability properties of a SAM. First, we will define slices of a SAM, F, for each $(x,y) \in \mathbb{R}^d$ and for each $y \in \mathbb{R}^{d_2}$. As shown in Lemma 2, when F is a SAM there exists $\{F^{(l)}\}_{l\geq 1}$, a sequence of continuous set-valued maps that approximate F, and for every $l \geq 1$, the set-valued map $F^{(l)}$ can be parameterized with single-valued maps $f^{(l)}$ as in Lemma 3. We will define similar slices of $F^{(l)}$ and $f^{(l)}$ as well.

Definition 7. Let $F : \mathbb{R}^d \times \mathscr{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ be a SAM. Let $\{F^{(l)}\}_{l\geq 1}$ and $\{f^{(l)}\}_{l\geq 1}$ be as in Lemmas 2 and 3, respectively.

i. For every $(x,y) \in \mathbb{R}^d$, define $F_{(x,y)} : \mathscr{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ such that for every $s \in \mathscr{S}$, $F_{(x,y)}(s) := F(x,y,s)$.

ii. For every $l \geq 1$, for every $(x,y) \in \mathbb{R}^d$, define $F^{(l)}_{(x,y)} : \mathscr{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ such that for every $s \in \mathscr{S}$, $F^{(l)}_{(x,y)}(s) := F^{(l)}(x,y,s)$.

iii. For every $l \geq 1$, for every $(x,y) \in \mathbb{R}^d$, define $f^{(l)}_{(x,y)} : \mathscr{S} \times U \longrightarrow \mathbb{R}^k$ such that for every $(s,u) \in \mathscr{S} \times U$, $f^{(l)}_{(x,y)}(s,u) := f^{(l)}(x,y,s,u)$.

iv. For every $y \in \mathbb{R}^{d_2}$, define $F_y : \mathbb{R}^{d_1} \times \mathscr{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ such that for every $(x,s) \in \mathbb{R}^{d_1} \times \mathscr{S}$, $F_y(x,s) := F(x,y,s)$.

v. For every $l \geq 1$, for every $y \in \mathbb{R}^{d_2}$, define $F^{(l)}_y : \mathbb{R}^{d_1} \times \mathscr{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ such that for every $(x,s) \in \mathbb{R}^{d_1} \times \mathscr{S}$, $F^{(l)}_y(x,s) := F^{(l)}(x,y,s)$.

vi. For every $l \geq 1$, for every $y \in \mathbb{R}^{d_2}$, define $f^{(l)}_y : \mathbb{R}^{d_1} \times \mathscr{S} \times U \longrightarrow \mathbb{R}^k$ such that for every $(x,s,u) \in \mathbb{R}^{d_1} \times \mathscr{S} \times U$, $f^{(l)}_y(x,s,u) := f^{(l)}(x,y,s,u)$.

The next two lemmas summarize properties that the slices inherit from the maps F, $F^{(l)}$ and $f^{(l)}$. Let $\mathscr{B}(\mathscr{S})$ denote the Borel sigma algebra associated with the metric space $(\mathscr{S}, d_{\mathscr{S}})$.

Lemma 6. *Let $F : \mathbb{R}^d \times \mathscr{S} \longrightarrow \{\text{subsets of } \mathbb{R}^k\}$ be a SAM. Let $\{F^{(l)}\}_{l\geq 1}$ and $\{f^{(l)}\}_{l\geq 1}$ be as in Lemmas 2 and 3, respectively. For every $(x,y) \in \mathbb{R}^d$, let $F_{(x,y)}$, $F^{(l)}_{(x,y)}$ and $f^{(l)}_{(x,y)}$ denote the slices as in Definition 7. Then, for every $(x,y) \in \mathbb{R}^d$,*

i. $F_{(x,y)}$ is a measurable set-valued map and for every $s \in \mathscr{S}$, $F_{(x,y)}(s)$ is a convex and compact subset of $\mathbb{R}^k$. Further, there exists $C_{(x,y)} = K(1 + \|(x,y)\|) > 0$ such that for every $s \in \mathscr{S}$, $\sup_{z \in F_{(x,y)}(s)} \|z\| \leq C_{(x,y)}$. (If F satisfies condition (c)′ instead of condition (c) in the definition of SAM, we have $C_{(x,y)} = K(1 + \|x\| + \|y\|)$).

ii. for every $l \geq 1$, $F^{(l)}_{(x,y)}$ is a measurable set-valued map and for every $s \in \mathscr{S}$, $F^{(l)}_{(x,y)}(s)$ is a convex and compact subset of $\mathbb{R}^k$. Further, there exists $C^{(l)}_{(x,y)} = K^{(l)}(1 + \|(x,y)\|) > 0$ such that for every $s \in \mathscr{S}$, $\sup_{z \in F^{(l)}_{(x,y)}(s)} \|z\| \leq C^{(l)}_{(x,y)}$. (If F satisfies condition (c)′ instead of condition (c) in the definition of SAM, we have $C^{(l)}_{(x,y)} = K^{(l)}(1 + \|x\| + \|y\|)$).

iii. for any probability measure μ on $(\mathscr{S}, \mathscr{B}(S))$, every measurable selection of $F_{(x,y)}$ is μ-integrable and hence $F_{(x,y)}$ is μ-integrable.

iv. for every $l \geq 1$, for any probability measure μ on $(\mathscr{S}, \mathscr{B}(S))$, every measurable selection of $F^{(l)}_{(x,y)}$ is μ-integrable and hence $F^{(l)}_{(x,y)}$ is μ-integrable.

v. for every $l \geq 1$, $f^{(l)}_{(x,y)}$ is continuous and for every $s \in \mathscr{S}$, $f^{(l)}_{(x,y)}(s,U) = F^{(l)}_{(x,y)}(s)$ and $\sup_{u \in U} \|f^{(l)}_{(x,y)}(s,u)\| \leq C^{(l)}_{(x,y)}$ where $C^{(l)}_{(x,y)}$ is as in part ii of this lemma.

The proof of the above lemma is similar to that of Yaji and Bhatnagar [41, lemma 4.1], and we will provide a brief outline here. Fix $(x,y) \in \mathbb{R}^d$. In order to show that $F_{(x,y)}$ is measurable, one needs to establish that $F^{-1}_{(x,y)}(C) \in \mathscr{B}(\mathscr{S})$ for any $C \subseteq \mathbb{R}^k$ closed. Using the closed graph property of F, one can show that $F^{-1}_{(x,y)}(C)$ is closed subset of $\mathscr{S}$ and hence is in $\mathscr{B}(\mathscr{S})$. The bound $C_{(x,y)}$ and the claim that $F_{(x,y)}(s)$ is convex and compact for every $s \in \mathscr{S}$ follows from conditions (c) (or (c)′) and (a) in the definition of SAM, respectively. Because all measurable selections of $F_{(x,y)}$ are bounded, they are μ-integrable for any probability measure μ on $(\mathscr{S}, \mathscr{B}(\mathscr{S}))$.

The arguments are exactly same for the claims associated with the slices of approximating maps $F^{(l)}$, for every $l \geq 1$. Finally, part v of the above lemma follows from the properties of maps $f^{(l)}$ stated in Lemma 3.

Let μ be a probability measure on $(\mathbb{R}^{d_1} \times \mathcal{S}, \mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S}))$, where $\mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S})$ denotes the Borel sigma algebra on metric space $\mathbb{R}^{d_1} \times \mathcal{S}$ with metric $\max\{\|x - x'\|, d_{\mathcal{S}}(s, s')\}$ for every $(x, s),\ (x', s') \in \mathbb{R}^{d_1} \times \mathcal{S}$ (in fact $\mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S})$ is the same as the product sigma algebra $\mathscr{B}(\mathbb{R}^{d_1}) \otimes \mathcal{S}$). The *support* of the measure μ denoted by $\text{supp}(\mu)$ is defined as a closed subset of $\mathbb{R}^{d_1} \times \mathcal{S}$ such that

1. $\mu(\text{supp}(\mu)) = 1$, and
2. for any other closed set $A \subseteq \mathbb{R}^{d_1} \times \mathcal{S}$ such that $\mu(A) = 1$, we have $\text{supp}(\mu) \subseteq A$.

For any probability measure μ on $\mathbb{R}^{d_1} \times \mathcal{S}$, such a set always exists and is unique (see Parthasarathy [33, chapter 2, theorem 2.1]).

Lemma 7. *Let $F : \mathbb{R}^d \times \mathcal{S} \to \{\text{subsets of } \mathbb{R}^k\}$ be a SAM satisfying condition (c)′ instead of condition (c) in the definition of SAM. Let $\{F^{(l)}\}_{l \geq 1}$ and $\{f^{(l)}\}_{l \geq 1}$ be as in Lemmas 2 and 3, respectively. For every $(x, y) \in \mathbb{R}^d$, let F_y, $F_y^{(l)}$ and $f_y^{(l)}$ denote the slices as in Definition 7. Then, for every $y \in \mathbb{R}^{d_2}$,*

i. *F_y is a measurable set-valued map, and for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $F_y(x, s)$ is a convex and compact subset of $\mathbb{R}^k$. Further, for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $\sup_{z \in F_y(x,s)} \|z\| \leq K_y(1 + \|x\|)$, where $K_y := \max\{K, K\|y\|\}$ and K is as in condition (c)′ in the definition of SAM.*

ii. *for every $l \geq 1$, $F_y^{(l)}$ is a measurable set-valued map, and for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $F_y^{(l)}(x, s)$ is a convex and compact subset of $\mathbb{R}^k$. Further, for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $\sup_{z \in F_y(x,s)} \|z\| \leq K_y^{(l)}(1 + \|x\|)$, where $K_y^{(l)} := \max\{K^{(l)}, K^{(l)}\|y\|\}$ and $K^{(l)}$ is as in Lemma 2iii.*

iii. *for every probability measure μ on $(\mathbb{R}^{d_1} \times \mathcal{S}, \mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S}))$ such that $\text{supp}(\mu)$ is a compact subset of $\mathbb{R}^{d_1} \times \mathcal{S}$, every measurable selection of F_y is μ-integrable and hence F_y is μ-integrable.*

iv. *for every $l \geq 1$, for every probability measure μ on $(\mathbb{R}^{d_1} \times \mathcal{S}, \mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S}))$ such that $\text{supp}(\mu)$ is a compact subset of $\mathbb{R}^{d_1} \times \mathcal{S}$, every measurable selection of $F_y^{(l)}$ is μ-integrable and hence $F_y^{(l)}$ is μ-integrable.*

v. *for every $l \geq 1$, $f_y^{(l)}$ is continuous and for every $(x,s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $f_y^{(l)}(x,s,U) = F_y^{(l)}(x,s)$ and $\sup_{u \in U} \|f_y^{(l)}(x,s,u)\| \leq K_y^{(l)}(1 + \|x\|)$, where $K_y^{(l)}$ is as in part ii of this lemma.*

The proof of parts i, ii, and v of the above lemma are similar to the corresponding parts in Lemma 6. We will provide a proof of part iii, and the proof of part iv is exactly the same.

Proof. Fix $y \in \mathbb{R}^{d_2}$.

iii. Consider $f \in \mathscr{S}(F_y)$. By part i of this lemma, we have that $\|f(x,s)\| \leq K_y(1 + \|x\|)$. Because $\text{supp}(\mu)$ is a compact subset of $\mathbb{R}^{d_1} \times \mathcal{S}$, there exists $M > 0$ such that for every $x \in \mathbb{R}^{d_1}$ for which there exists $s \in \mathcal{S}$ satisfying $(x, s) \in \text{supp}(\mu)$, we have $\|x\| \leq M$. Hence, $\|\int_{\mathbb{R}^{d_1} \times \mathcal{S}} f(x,s)\mu(dx,ds)\| = \|\int_{\text{supp}(\mu)} f(x,s)\mu(dx,ds)\| \leq \int_{\text{supp}(\mu)} \|f(x,s)\|\mu(dx,ds) \leq \int_{\text{supp}(\mu)} K_y(1+\|x\|)\mu(dx,ds) \leq K_y(1+M)$. Therefore, every measurable selection $f \in \mathscr{S}(F_y)$ is μ-integrable and hence F_y is μ-integrable. □

By Lemma 6iv and v, we know that $F_{(x,y)}^{(l)}$ is a μ-integrable set-valued map for any probability measure μ on $(\mathcal{S}, \mathscr{B}(\mathcal{S}))$, and $f_{(x,y)}^{(l)}$ is a continuous parametrization of $F_{(x,y)}^{(l)}$ for every $l \geq 1$ and for every $(x, y) \in \mathbb{R}^d$. Similarly, by Lemma 7iv and v, we know that $F_y^{(l)}$ is μ-integrable for any probability measure μ on $(\mathbb{R}^{d_1} \times \mathcal{S}, \mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S}))$ with compact support, and $f_y^{(l)}$ is a continuous parametrization of $F_y^{(l)}$ for every $l \geq 1$ and for every $y \in \mathbb{R}^{d_2}$. A natural question to ask is about the relation between integral of map $F_{(x,y)}^{(l)}$ (or $F_y^{(l)}$) and the integral of its parametrization $f_{(x,y)}^{(l)}$ (or $f_y^{(l)}$). The next lemma answers this question. Before stating the lemma, we introduce the following notation, which will be used throughout this paper.

Let $\mathcal{P}(\cdots)$ denote the space of probability measures on a Polish space "$\cdots$" with the Prohorov topology (also known as the "topology of convergence in distribution"; see Borkar [12, chapter 2] for details). For any probability measure $\nu \in \mathcal{P}(\mathcal{S} \times U)$, let $\nu_{\mathcal{S}} \in \mathcal{P}(\mathcal{S})$ denote the image of measure ν under the projection $\mathcal{S} \times U \to \mathcal{S}$ (i.e., for any $A \in \mathscr{B}(\mathcal{S})$, $\nu_{\mathcal{S}}(A) = \int_{A \times U} \mu(ds, du)$). Similarly, for any probability measure $\nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S} \times U)$, let $\nu_{\mathbb{R}^{d_1} \times \mathcal{S}}$, $\nu_{\mathcal{S}}$ and $\nu_{\mathbb{R}^{d_1}}$ belonging to $\mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S})$, $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(\mathbb{R}^{d_1})$, respectively, denote the image of measure ν under the projections $\mathbb{R}^{d_1} \times \mathcal{S} \times U \to \mathbb{R}^{d_1} \times \mathcal{S}$, $\mathbb{R}^{d_1} \times \mathcal{S} \times U \to \mathcal{S}$, and $\mathbb{R}^{d_1} \times \mathcal{S} \times U \to \mathbb{R}^{d_1}$, respectively.

Lemma 8. *Let $F : \mathbb{R}^d \times \mathcal{S} \to \{\text{subsets of } \mathbb{R}^k\}$ be a SAM. Let $\{F^{(l)}\}_{l \geq 1}$ and $\{f^{(l)}\}_{l \geq 1}$ be as in Lemmas 2 and 3, respectively. For every $l \geq 1$, for every $(x, y) \in \mathbb{R}^d$, let $F_{(x,y)}^{(l)}$, $f_{(x,y)}^{(l)}$, and for every $y \in \mathbb{R}^{d_2}$, let $F_y^{(l)}$, $f_y^{(l)}$ denote the slices as in Definition 7.*

i. *For every $l \geq 1$, for every $(x, y) \in \mathbb{R}^d$, for any probability measure $\mu \in \mathcal{P}(\mathcal{S})$,*

$$\int_{\mathcal{S}} F_{(x,y)}^{(l)}(s)\mu(ds) = \left\{ \int_{\mathcal{S} \times U} f_{(x,y)}^{(l)}(s,u)\nu(ds,du) : \nu \in \mathcal{P}(\mathcal{S} \times U),\ \nu_{\mathcal{S}} = \mu \right\}.$$

ii. *Suppose F satisfies condition (c)′ instead of condition (c) in the definition of SAM. Then, for every $l \geq 1$, for every $y \in \mathbb{R}^{d_2}$, for any probability measure $\mu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S})$ with compact support,*

$$\int_{\mathbb{R}^{d_1} \times \mathcal{S}} F_y^{(l)}(dx, ds)\mu(dx, ds) = \left\{ \int_{\mathbb{R}^{d_1} \times \mathcal{S} \times U} f_y^{(l)}(x, s, u)\nu(dx, ds, du) : \nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S} \times U),\ \nu_{\mathbb{R}^{d_1} \times \mathcal{S}} = \mu \right\}.$$

Remark 2. For any $\nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S} \times U)$ with $\nu_{\mathbb{R}^{d_1} \times \mathcal{S}} = \mu$, the support of the measure ν is contained in $\mathrm{supp}(\mu) \times U$ because by Borkar [12, chapter 3, corollary 3.1.2], there exists a μ a.s. unique measurable map $q : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathcal{P}(U)$ such that $\nu(dx, ds, du) = q(x, s, du)\mu(dx, ds)$ and $1 = \nu(\mathbb{R}^{d_1} \times \mathcal{S} \times U) = \int_{\mathbb{R}^{d_1} \times \mathcal{S} \times U} \nu(dx, ds, du) = \int_{\mathbb{R}^{d_1} \times\ \mathcal{S}} [\int_U q(x, s, du)]$ $\mu(dx, ds) = \int_{\mathrm{supp}(\mu)} [\int_U q(x, s, du)]\mu(dx, ds) = \nu(\mathrm{supp}(\mu) \times U)$. Therefore, when $\mathrm{supp}(\mu)$ is a compact set, the support of measure ν is also compact and by Lemma 7v it is easy to deduce that for all measures, $\nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S} \times U)$ with compact support, for all $y \in \mathbb{R}^{d_2}$, $f_y^{(l)}$ is ν-integrable for all $l \geq 1$.

The proof of part i of the above lemma is exactly same as that in Yaji and Bhatnagar [41, lemma 4.2]. The proof of part ii is similar but with minor technical modifications and is presented below.

Proof. ii. Fix $y \in \mathbb{R}^{d_2}$, $l \geq 1$ and $\mu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S})$ with compact support.

Consider $z \in \int_{\mathbb{R}^{d_1} \times \mathcal{S}} F_y^{(l)}(x, s)\mu(dx, ds)$. Then there exists $f \in \mathscr{S}(F_y^{(l)})$ such that $z = \int_{\mathbb{R}^{d_1} \times \mathcal{S}} f(x, s)\mu(dx, ds)$. Let $G : \mathbb{R}^{d_1} \times \mathcal{S} \to \{\text{subsets of } U\}$ be such that for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $G(x, s) = \{u \in U : f(x, s) = f_y^{(l)}(x, s, u)\}$. By the fact that $f_y^{(l)}(x, s, U) = F_y^{(l)}(x, s)$ and because $f(x, s) \in F_y^{(l)}(x, s)$ for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, we have that $G(x, s)$ is nonempty. By the continuity of $f_y^{(l)}(x, s, \cdot)$, we have that $G(x, s)$ is closed for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$. For any $C \subseteq U$ closed, $G^{-1}(C) \in \mathscr{B}(\mathbb{R}^{d_1} \times \mathcal{S})$ (for a proof, see Yaji and Bhatnagar [41, appendix B]), and hence G is measurable. Because G is measurable, by Lemma 4i, we have that $\mathscr{S}(G) \neq \emptyset$. Let $g \in \mathscr{S}(G)$ and let $\hat{g} : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathbb{R}^{d_1} \times \mathcal{S} \times U$ be such that for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $\hat{g}(x, s) := (x, s, g(x, s))$. Let $\nu = \mu\hat{g}^{-1}$ (push-forward measure). Clearly, $\nu_{\mathbb{R}^{d_1} \times U} = \mu$ and $\int_{\mathbb{R}^{d_1} \times \mathcal{S} \times U} f_y^{(l)}(x, s, u)\nu(dx, ds, du) = \int_{\mathbb{R}^{d_1} \times \mathcal{S} \times U} f_y^{(l)}(x, s, u)\mu\hat{g}^{-1}(dx, ds, du) = \int_{\mathbb{R}^{d_1} \times \mathcal{S}} f_y^{(l)}(x, s, g(x, s))\mu(dx, ds) = \int_{\mathbb{R}^{d_1} \times \mathcal{S}} f(x, s) \cdot$ $\mu(dx, ds) = z$. Therefore, the left-hand side (L.H.S.) is contained in the right-hand side (R.H.S.).

Let $\nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S} \times U)$ with $\nu_{\mathbb{R}^{d_1} \times \mathcal{S}} = \mu$. By Borkar [12, corollary 3.1.2], there exists a μ a.s. unique measurable map $q : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathcal{P}(U)$ such that $\nu(dx, ds, du) = q(x, s, du)\mu(dx, ds)$. Because μ has compact support, ν has compact support, and hence $f_y^{(l)}$ is ν-integrable (see remark following Lemma 8). Therefore, $\int_{\mathbb{R}^{d_1} \times \mathcal{S} \times U} f_y^{(l)}(x, s, u)\nu(dx, ds, du) =$ $\int_{\mathbb{R}^{d_1} \times \mathcal{S}} [\int_U f_y^{(l)}(x, s, u)q(x, s, du)]\mu(dx, ds)$. By Lemma 7v, we know that for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}, f_y^{(l)}(x, s, U) = F_y^{(l)}(x, s)$, and hence $f^{(l)}(x, s, U)$ is convex and compact subset of $\mathbb{R}^k$. Therefore, for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $\int_U f_y^{(l)}(x, s, u) \cdot$ $q(x, s, du) \in F_y^{(l)}(x, s)$. Let $f : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathbb{R}^k$ be such that for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $f(x, s) := \int_U f^{(l)}(x, s, u)q(x, s, du) \cdot$ Then clearly, f is measurable and $f \in \mathscr{S}(F_y^{(l)})$. Therefore, $\int_{\mathbb{R}^{d_1} \times \mathcal{S} \times U} f_y^{(l)}(x, s, u)\nu(dx, ds, du) = \int_{\mathbb{R}^{d_1} \times \mathcal{S}} [\int_U f_y^{(l)}(x, s, u)$ $q(x, s, du)]\mu(dx, ds) = \int_{\mathbb{R}^{d_1} \times \mathcal{S}} f(x, s)\mu(dx, ds) \in \int_{\mathbb{R}^{d_1} \times \mathcal{S}} F_y^{(l)}(x, s)\mu(dx, ds)$. The above gives us that the R.H.S. is contained in the L.H.S. □

2.3. Differential Inclusions and Their Limit Sets

Here, we will review the results of Benaïm et al. [4] on the theory of differential inclusions and give definitions of limit sets associated with such dynamical systems that are used later in the paper.

First, we define a set-valued map whose associated DI is known to admit at least one solution through every initial condition. Such set-valued maps are called Marchaud maps, and the definition of such a map is stated below.

Definition 8 (Marchaud map). $F : \mathbb{R}^k \to \{\text{subsets of } \mathbb{R}^k\}$ is a Marchaud map if,

i. for every $z \in \mathbb{R}^k$, $F(z)$ is a convex and compact subset of $\mathbb{R}^k$,

ii. there exists $K > 0$ such that for every $z \in \mathbb{R}^k$, $\sup_{z' \in F(z)} \|z'\| \leq K(1 + \|z\|)$, and

iii. for every $z \in \mathbb{R}^k$, for every $\mathbb{R}^k$-valued sequence, $\{z_n\}_{n\geq 0}$ converging to $z \in \mathbb{R}^k$, for every sequence $\{z'_n \in F(z_n)\}_{n\geq 0}$ converging to $z' \in \mathbb{R}^k$, we have that $z' \in F(z)$.

Let F be a Marchaud map. Then the DI associated with the map F is given by

$$\frac{dz}{dt} \in F(z). \tag{11}$$

Because F is a Marchaud map, it is known that the DI (11) admits at least one solution through every initial condition (see Benaïm et al. [4, section 1.2]). By a solution of DI (11) with initial condition $z \in \mathbb{R}^k$, we mean a function $\mathbf{z} : \mathbb{R} \to \mathbb{R}^k$ such that $\mathbf{z}(\cdot)$ is absolutely continuous, $\mathbf{z}(0) = z$ and for *a.e.* $t \in \mathbb{R}$, $\frac{d\mathbf{z}(t)}{dt} \in F(\mathbf{z}(t))$.

Now, we will recall the notions of flow, invariant sets, attracting sets, attractors, basin of attraction, and internally chain transitive sets. All these notions are taken from Benaïm et al. [4].

The *flow* of DI (11) is given by the set-valued map $\Phi : \mathbb{R}^k \times \mathbb{R} \to \{\text{subsets of } \mathbb{R}^k\}$ such that for every $(z,t) \in \mathbb{R}^k \times \mathbb{R}$,

$$\Phi(z,t) := \{\mathbf{z}(t) : \mathbf{z} \textit{ is a solution of DI } (11) \textit{ with } \mathbf{z}(0) = z\}.$$

For any set $A \subseteq \mathbb{R}^k$, let $\Phi(A,t) := \cup_{z \in A} \Phi(z,t)$.

A closed set $A \subseteq \mathbb{R}^k$ is *invariant* for the flow Φ of DI (11) if for every $z \in A$, there exists a solution $\mathbf{z}(\cdot)$ of DI (11) such that, $\mathbf{z}(0) = z$ and for every $t \in \mathbb{R}$, $\mathbf{z}(t) \in A$.

A compact set $A \subseteq \mathbb{R}^k$ is an *attracting* set for the flow Φ of DI (11) if there exists an open neighborhood of A, say $\mathbb{O}$, with the property that for every $\epsilon > 0$, there exists $T > 0$ (depending on $\epsilon > 0$ only) such that for every $t \geq T$, $\Phi(\mathbb{O}, t) \subseteq N^\epsilon(A)$, where $N^\epsilon(A)$ stands for the ϵ-neighborhood of A.

A compact set $A \subseteq \mathbb{R}^k$ is an attractor for the flow Φ of DI (11) if A is an attracting set and is invariant for the flow Φ of DI (11).

For any $z \in \mathbb{R}^k$, $\omega_\Phi(z) := \cap_{t \geq 0} \overline{\Phi(z, [t, \infty))}$, where $\Phi(z, [t,\infty)) := \cup_{q \geq t} \Phi(z,q)$. For any set $A \subseteq \mathbb{R}^k$, the *basin of attraction* of set A is denoted by $B(A)$ and is defined as

$$B(A) := \{z \in \mathbb{R}^k : \omega_\Phi(z) \subseteq A\}.$$

If $A \subseteq \mathbb{R}^k$ is an attractor whose basin of attraction is the whole of $\mathbb{R}^k$ (i.e., $B(A) = \mathbb{R}^k$), then A is called a *global attractor*.

Given a set $A \subseteq \mathbb{R}^k$ and $z, z' \in A$, for any $\epsilon > 0$ and $T > 0$, there exists an (ϵ, T) chain from z to z' for DI (11) if there exists an integer $n \in \mathbb{N}$, solutions $\mathbf{z}_1, \ldots, \mathbf{z}_n$ to DI (11) and real numbers $t_1, \ldots, t_n$ greater than T such that

- for all $i \in \{1, \ldots, n\}$ and for all $q \in [0, t_i]$, and $\mathbf{z}_i(q) \in A$,
- for all $i \in \{1, \ldots, n\}$, $\| \mathbf{z}_i(t_i) - \mathbf{z}_{i+1}(0) \| \leq \epsilon$,
- $\| \mathbf{z}_1(0) - z \| \leq \epsilon$ and $\| \mathbf{z}_n(t_n) - z' \| \leq \epsilon$.

A compact set $A \subseteq \mathbb{R}^d$ is said to be *internally chain transitive* if for every $z, z' \in A$, for every $\epsilon > 0$, and for every $T > 0$, there exists (ϵ, T) chain from z to z' for DI (11).

Suppose $L \subseteq \mathbb{R}^k$ is an invariant set. Then, the flow of DI (11) restricted to the invariant set L is a set-valued map, $\Phi^L : L \times \mathbb{R} \to \{\text{subsets of } \mathbb{R}^k\}$ such that for every $(z,t) \in L \times \mathbb{R}$,

$$\Phi^L(z,t) := \{\mathbf{z}(t) : \mathbf{z}(\cdot) \text{ is a solution of DI (11) with } \mathbf{z}(0) = z, \text{ and for every } t \in \mathbb{R},\ \mathbf{z}(t) \in L\}. \tag{12}$$

2.4. Single Timescale Stochastic Recursive Inclusions with Nonadditive Iterate-Dependent Markov Noise

In this section, we review results by Yaji and Bhatnagar [41] on the analysis of single timescale stochastic recursive inclusions with nonadditive iterate-dependent Markov noise.

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $\{Z_n\}_{n \geq 0}$ be a sequence of $\mathbb{R}^d$-valued random variables satisfying

$$Z_{n+1} - Z_n - a(n) M_{n+1} \in a(n) F(Z_n, S_n), \tag{13}$$

where the following assumptions hold:

Assumption S(A1). The map $F : \mathbb{R}^d \times \mathcal{S} \to \{\text{subsets of } \mathbb{R}^d\}$ where $(\mathcal{S}, d_{\mathcal{S}})$ is a compact metric space is such that

i. for every $(z,s) \in \mathbb{R}^d \times \mathcal{S}$, $F(z,s)$ is a convex and compact subset of $\mathbb{R}^d$,
ii. there exists $K > 0$ such that for every $(z,s) \in \mathbb{R}^d \times \mathcal{S}$, $\sup_{z' \in F(z,s)} \|z'\| \leq K(1 + \|z\|)$, and
iii. for every $z \in \mathbb{R}^d$, for every $\mathbb{R}^d \times \mathcal{S}$ valued sequence, say $\{(z_n, s_n)\}_{n \geq 1}$ converging to (z,s), and for any sequence $\{z'_n \in F(z_n, s_n)\}_{n \geq 1}$ converging to z', we have $z' \in F(z,s)$.

Assumption S(A2). $\{S_n\}_{n \geq 0}$ is a sequence of $\mathcal{S}$-valued measurable functions on Ω such that for every $n \geq 0$, for every $A \in \mathscr{B}(\mathcal{S})$, $\mathbb{P}(S_{n+1} \in A | S_m, Z_m, m \leq n) = \mathbb{P}(S_{n+1} \in A | S_n, Z_n) = \Pi(Z_n, S_n)(A)$ *a.s*, where $\Pi : \mathbb{R}^d \times \mathcal{S} \to \mathcal{P}(\mathcal{S})$ is continuous.

Assumption S(A3). $\{a(n)\}_{n \geq 0}$ is a sequence of positive real numbers satisfying

i. $a(0) \leq 1$ and for every $n \geq 0$, $a(n) \geq a(n+1)$,
ii. $\sum_{n=0}^{\infty} a(n) = \infty$ and $\sum_{n=0}^{\infty} (a(n))^2 < \infty$.

Assumption S(A4). $\{M_n\}_{n\geq 1}$ is a sequence of $\mathbb{R}^d$-valued random variables on Ω such that for $a.s.(\omega)$, for any $T > 0$, $\lim_{n\to\infty}\sup_{n\leq k\leq \tau(n,T)}\|\sum_{m=n}^{k} a(m)M_{m+1}(\omega)\| = 0$, where $\tau(n,T) := \min\{m > n : \sum_{k=n}^{m-1} a(k) \geq T\}$.

Assumption S(A5). $\mathbb{P}(\sup_{n\geq 0}\|Z_n\| < \infty) = 1$.

A detailed motivation for each of these assumptions can be found in Yaji and Bhatnagar [41]. We will briefly explain them and their consequences.

Assumption S(A1) ensures that the set-valued map F is a SAM, and Assumption S(A2) is the iterate-dependent Markov noise assumption. As a consequence of Assumption S(A2), for every $z \in \mathbb{R}^d$, we know that the Markov chain defined by the transition kernel $\Pi(z,\cdot)(\cdot)$ possesses the weak Feller property (see Meyn and Tweedie [29]). In addition to the above, because the state space is compact, the set of stationary distributions for the Markov chain whose transition probability is given by $\Pi(z,\cdot)(\cdot)$ is nonempty for every $z \in \mathbb{R}^d$. Let $D(z) \subseteq \mathcal{P}(\mathcal{S})$ denote the set of stationary distributions of the Markov chain whose transition kernel is $\Pi(z,\cdot)(\cdot)$ (for any $z \in \mathbb{R}^d$, $\mu \in D(z)$ if and only if for every $A \in \mathscr{B}(\mathcal{S})$, $\mu(A) = \int_{\mathcal{S}} \Pi(z,s)(A)\mu(ds)$). We also know that for every $z \in \mathbb{R}^d$, $D(z)$ is a convex and compact subset of $\mathcal{P}(\mathcal{S})$, and the map $z \to D(z)$ has a closed graph (see Yaji and Bhatnagar [41] and the references therein). Assumption S(A3) is the standard step-size assumption, and Assumption S(A4) is the general additive noise assumption, which ensures that the contribution of the additive noise is eventually negligible (for various noise models satisfying Assumption S(A4), see Benaïm et al. [4]). Assumption S(A5) is the stability requirement on the iterate sequence.

The set-valued map, $\hat{F} : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^d\}$ that serves as the vector field for the differential inclusion (DI) that the iterates are expected to track is defined as,

$$\hat{F}(z) := \cup_{\mu\in D(z)} \int_{\mathcal{S}} F_z(s)\mu(ds)$$

for every $z \in \mathbb{R}^d$, where for every $z \in \mathbb{R}^d$, F_z denotes the slice as in Definition 7i of the set-valued map F appearing in Recursion (13). The set-valued map $\hat{F}$ is a Marchaud map (see Yaji and Bhatnagar [41, lemma 4.7]) and the associated DI given by

$$\frac{dz}{dt} \in \hat{F}(z) \tag{14}$$

admits at least one solution through every initial condition (see Benaïm et al. [4, section 1.2]). Let $\Sigma(z_0)$ denote the set of solutions of DI (14) with initial condition $z_0 \in \mathbb{R}^d$ and $\Sigma := \cup_{z_0\in\mathbb{R}^d}\Sigma(z_0)$ (the set of all possible solutions). For every $z_0 \in \mathbb{R}^d$, $\Sigma(z_0)$ is a subset of $\mathscr{C}(\mathbb{R},\mathbb{R}^d)$, the set of all $\mathbb{R}^d$-valued continuous functions on $\mathbb{R}$. The set $\mathscr{C}(\mathbb{R},\mathbb{R}^d)$ is a complete metric space for the metric $\mathbf{D}$ defined by

$$\mathbf{D}(\mathbf{z},\mathbf{z}') := \sum_{k=1}^{\infty} \frac{1}{2^k}\min\left(\|\mathbf{z}-\mathbf{z}'\|_{[-k,k]}, 1\right),$$

where $\|\mathbf{z}-\mathbf{z}'\|_{[-k,k]} := \sup_{t\in[-k,k]}\|\mathbf{z}(t)-\mathbf{z}'(t)\|$. As a consequence of Benaïm et al. [4, lemma 3.1], we have that Σ and for every $z_0 \in \mathbb{R}^d$, $\Sigma(z_0)$ are closed and compact subsets of $\mathscr{C}(\mathbb{R},\mathbb{R}^d)$, respectively.

Let $t_0 := 0$ and for every $n \geq 1$, $t(n) := \sum_{k=0}^{n-1} a(k)$. Define the stochastic process with continuous sample paths $\bar{Z} : \Omega \times \mathbb{R} \to \mathbb{R}^d$ as

$$\bar{Z}(\omega,t) := \left(\frac{t-t(n)}{t(n+1)-t(n)}\right)Z_{n+1}(\omega) + \left(\frac{t(n+1)-t}{t(n+1)-t(n)}\right)Z_n(\omega)$$

for every $(\omega,t) \in \Omega \times [0,\infty)$, where n is such that $t \in [t(n),t(n+1))$ and for every $(\omega,t) \in \Omega \times (-\infty,0]$, let $\bar{Z}(\omega,t) := Z_0(\omega)$. Then the main result from the analysis of recursion (13) in Yaji and Bhatnagar [41] is as follows.

Theorem 2. *Under Assumptions S(A1)-S(A5), for almost every $\omega \in \Omega$,*

i. *the family of functions $\{\bar{Z}(\omega,\cdot+t)\}_{t\geq 0}$ is relatively compact in $\mathscr{C}(\mathbb{R},\mathbb{R}^d)$,*

ii. *every limit point of $\{\bar{Z}(\omega,\cdot+t)\}_{t\geq 0}$ in $\mathscr{C}(\mathbb{R},\mathbb{R}^d)$ is a solution of DI (14), more formally,*

$$\lim_{t\to\infty}\mathbf{D}\left(\bar{Z}(\omega,\cdot+t),\Sigma\right) = 0, \text{ and}$$

iii. *the limit set denoted by* $L(\bar{Z}(\omega,\cdot))$, *defined as*

$$L(\bar{Z}(\omega,\cdot)) := \cap_{t\geq 0}\overline{\{\bar{Z}(\omega,q+t) : q\geq 0\}},$$

is nonempty, compact, and internally chain transitive for the flow of DI (14).

For a proof of the above theorem, see Yaji and Bhatnagar [41, theorems 6.6 and 6.7].

2.5. Space of Probability Measure-Valued Functions

In this section, we will define the space of probability measure-valued measurable functions on $[0,\infty)$. We will introduce an appropriate topology on this space and show that such a space is compact metrizable. These spaces are used in the theory of optimal control of diffusions (see Borkar [8]) and also in analyzing stochastic approximation schemes (see Yaji and Bhatnagar [41], Borkar [10]). Quantities defined in this section will serve as tools in analyzing the stochastic recursions later.

Throughout this section, U will denote the closed unit ball in $\mathbb{R}^{d_2}$, and for any $r>0$, B_r denotes the closed ball of radius r in $\mathbb{R}^{d_1}$ centered on the origin. For every $r>0$, let $\mathcal{M}(U\times B_r\times\mathcal{S})$ denote the set of all functions $\gamma(\cdot)$ on $[0,\infty)$ taking values in $\mathcal{P}(U\times B_r\times\mathcal{S})$ (space of probability measures on $U\times B_r\times\mathcal{S}$ equipped with the Prohorov topology) such that $\gamma(\cdot)$ is measurable. Formally,

$$\mathcal{M}(U\times B_r\times\mathcal{S}) := \{\gamma : [0,\infty)\to\mathcal{P}(U\times B_r\times\mathcal{S}) : \gamma(\cdot) \text{ is measurable}\}.$$

Similarly, for every $r>0$, $\mathcal{M}(B_r\times\mathcal{S})$ (or $\mathcal{M}(B_r)$) denotes the set of all functions $\gamma(\cdot)$ on $[0,\infty)$ taking values in $\mathcal{P}(B_r\times\mathcal{S})$ (or $\mathcal{P}(B_r)$) such that $\gamma(\cdot)$ is measurable. Formally,

$$\mathcal{M}(B_r\times\mathcal{S}) := \{\gamma : [0,\infty)\to\mathcal{P}(B_r\times\mathcal{S}) : \gamma(\cdot) \text{ is measurable}\},$$
$$\mathcal{M}(B_r) := \{\gamma : [0,\infty)\to\mathcal{P}(B_r) : \gamma(\cdot) \text{ is measurable}\}.$$

For every $r>0$, let $\tau_{U\times B_r\times\mathcal{S}}$ denote the coarsest topology on $\mathcal{M}(U\times B_r\times\mathcal{S})$, which renders continuous the functions $\gamma(\cdot)\to\int_0^T g(t)[\int_{U\times B_r\times\mathcal{S}} f(u,x,s)\gamma(t)(du,dx,ds)]dt$ for every $f\in\mathcal{C}(U\times B_r\times\mathcal{S},\mathbb{R})$, for every $g\in L_2([0,T],\mathbb{R})$, and for every $T>0$.

Similarly, for every $r>0$, let $\tau_{B_r\times\mathcal{S}}$ denote the coarsest topology on $\mathcal{M}(B_r\times\mathcal{S})$, which renders continuous the functions $\gamma(\cdot)\to\int_0^T g(t)[\int_{B_r\times\mathcal{S}} f(x,s)\gamma(t)(dx,ds)]dt$ for every $f\in\mathcal{C}(B_r\times\mathcal{S},\mathbb{R})$, for every $g\in L_2([0,T],\mathbb{R})$ and for every $T>0$.

Finally, for every $r>0$, let τ_{B_r} denote the coarsest topology on $\mathcal{M}(B_r)$, which renders continuous the functions, $\gamma(\cdot)\to\int_0^T g(t)[\int_{B_r} f(x)\gamma(t)(dx)]dt$ for every $f\in\mathcal{C}(B_r,\mathbb{R})$, for every $g\in L_2([0,T],\mathbb{R})$, and for every $T>0$.

The following is a well-known metrization lemma for the topological spaces defined above.

Lemma 9 (Metrization)**.**

i. *For every* $r>0$, *the topological space* $(\mathcal{M}(U\times B_r\times\mathcal{S}),\tau_{U\times B_r\times\mathcal{S}})$ *is compact metrizable.*
ii. *For every* $r>0$, *the topological space* $(\mathcal{M}(B_r\times\mathcal{S}),\tau_{B_r\times\mathcal{S}})$ *is compact metrizable.*
iii. *For every* $r>0$, *the topological space* $(\mathcal{M}(B_r),\tau_{B_r})$ *is compact metrizable.*

We refer the reader to Borkar [10, lemma 2.1] for the proof of the above metrization lemma. The next lemma provides continuous functions between the above defined metric spaces, which are used later. The proof of the lemma below is an extension of Yaji and Bhatnagar [41, lemma 5.2] to the above defined metric spaces. Recall that for any probability measure $\nu\in\mathcal{P}(U\times B_r\times\mathcal{S})$, $\nu_{B_r\times\mathcal{S}}\in\mathcal{P}(B_r\times\mathcal{S})$ denotes the image of the measure ν under the projection $U\times B_r\times\mathcal{S}\to B_r\times\mathcal{S}$ (i.e., for every $A\in\mathcal{B}(B_r\times\mathcal{S})$, $\nu_{B_r\times\mathcal{S}}(A)=\int_{U\times A}\nu(du,dx,ds)$). Similarly, $\nu_{B_r}\in\mathcal{P}(B_r)$ denotes the image of measure ν under the projection $U\times B_r\times\mathcal{S}\to B_r$ (i.e., for every $A\in\mathcal{B}(B_r)$, $\nu_{B_r}(A)=\int_{U\times A\times\mathcal{S}}\nu(du,dx,ds)$). It is easy to see that ν_{B_r} is also the image of $\nu_{B_r\times\mathcal{S}}$ under the projection $B_r\times\mathcal{S}\to B_r$.

Lemma 10. *For every* $r>0$,

i. *the map* $\theta_1 : \mathcal{P}(U\times B_r\times\mathcal{S})\to\mathcal{P}(B_r\times\mathcal{S})$ *such that for every* $\nu\in\mathcal{P}(U\times B_r\times\mathcal{S})$, $\theta_1(\nu):=\nu_{B_r\times\mathcal{S}}$ *is continuous.*
ii. *the map* $\theta_2 : \mathcal{P}(B_r\times\mathcal{S})\to\mathcal{P}(B_r)$ *such that for every* $\nu\in\mathcal{P}(B_r\times\mathcal{S})$, $\theta_2(\nu):=\nu_{B_r}$ *is continuous.*
iii. *for any* $\gamma\in\mathcal{M}(U\times B_r\times\mathcal{S})$, *we have that* $\theta_1\circ\gamma\in\mathcal{M}(B_r\times\mathcal{S})$ *where for every* $t\geq 0$, $(\theta_1\circ\gamma)(t)=\theta_1(\gamma(t))$.
iv. *for any* $\gamma\in\mathcal{M}(B_r\times\mathcal{S})$, *we have that* $\theta_2\circ\gamma\in\mathcal{M}(B_r)$ *where for every* $t\geq 0$, $(\theta_2\circ\gamma)(t)=\theta_2(\gamma(t))$.
v. *the map* $\Theta_1 : \mathcal{M}(U\times B_r\times\mathcal{S})\to\mathcal{M}(B_r\times\mathcal{S})$ *such that for every* $\gamma\in\mathcal{M}(U\times B_r\times\mathcal{S})$, $\Theta_1(\gamma):=\theta_1\circ\gamma$ *is continuous.*
vi. *the map* $\Theta_2 : \mathcal{M}(B_r\times\mathcal{S})\to\mathcal{M}(B_r)$ *such that for every* $\gamma\in\mathcal{M}(B_r\times\mathcal{S})$, $\Theta_2(\gamma):=\theta_2\circ\gamma$ *is continuous.*

Proof. Fix $r > 0$.

i. Let $\{\nu^n\}_{n\ge1}$ be a sequence in $\mathscr{P}(U\times B_r\times\mathscr{S})$ converging to $\nu\in\mathscr{P}(U\times B_r\times\mathscr{S})$ as $n\to\infty$, and let $\pi : U\times B_r\times\mathscr{S}\to B_r\times\mathscr{S}$ denote the projection map such that for every $(u,x,s)\in U\times B_r\times\mathscr{S}$, $\pi(u,x,s)=(x,s)$. Clearly, π is continuous and for any continuous function $f\in\mathscr{C}(B_r\times\mathscr{S},\mathbb{R})$, $f\circ\pi$ is continuous. Because $U\times B_r\times\mathscr{S}$ is a compact metric space, from Borkar [12, theorem 2.1.1(ii)], we get that for every $f\in\mathscr{C}(B_r\times\mathscr{S},\mathbb{R})$, $\int_{U\times B_r\times\mathscr{S}}(f\circ\pi)\cdot(u,x,s)\nu^n(du,dx,ds)\to\int_{U\times B_r\times\mathscr{S}}(f\circ\pi)\nu(du,dx,ds)$ as $n\to\infty$. By definition, we have that for every $n\ge0$, $\nu^n_{B_r\times\mathscr{S}}=\nu^n\pi^{-1}$ (the push-forward measure) and $\nu_{B_r\times\mathscr{S}}=\nu\pi^{-1}$. Therefore, for every $f\in\mathscr{C}(B_r\times\mathscr{S},\mathbb{R})$, $\int_{B_r\times\mathscr{S}}f(x,s)\nu^n_{B_r\times\mathscr{S}}\cdot(dx,ds)\to\int_{B_r\times\mathscr{S}}f(x,s)\nu_{B_r\times\mathscr{S}}(dx,ds)$. Hence, by Borkar [12, theorem 2.1.1], we get that $\nu^n_{B_r\times\mathscr{S}}\to\nu_{B_r\times\mathscr{S}}$ as $n\to\infty$ in $\mathscr{P}(B_r\times\mathscr{S})$, which gives us continuity of $\theta_1(\cdot)$.

ii. Similar to part i of this lemma.

iii and iv. Composition of measurable functions is measurable.

v. Let $\{\gamma_n\}_{n\ge1}$ be a sequence in $\mathscr{M}(U\times\ B_r\times\mathscr{S})$ converging to $\gamma\in\mathscr{M}(U\times B_r\times\mathscr{S})$ as $n\to\infty$. Then we know that for every $f\in\mathscr{C}(U\times B_r\times\mathscr{S},\mathbb{R})$, for every $T>0$, and for every $g\in L_2([0,T],\mathbb{R})$, $\int_0^T g(t)[\int_{U\times B_r\times\mathscr{S}}f(u,x,s)\cdot\gamma_n(t)(du,dx,ds)]dt\to\int_0^T g(t)[\int_{U\times B_r\times\mathscr{S}}f(u,x,s)\gamma(t)(du,dx,ds)]dt$ as $n\to\infty$. Let π denote the projection map as in part i of this lemma, and we know that for any $f\in\mathscr{C}(B_r\times\mathscr{S},\mathbb{R})$, $f\circ\pi\in\mathscr{C}(U\times B_r\times\mathscr{S},\mathbb{R})$. Then we have that for every $f\in\mathscr{C}(B_r\times\mathscr{S},\mathbb{R})$, for every $T>0$, and for every $g\in L_2([0,T],\mathbb{R})$, $\int_0^T g(t)[\int_{U\times B_r\times\mathscr{S}}(f\circ\pi)(u,x,s)\gamma_n(t)\cdot(du,dx,ds)]dt\to\int_0^T g(t)[\int_{U\times B_r\times\mathscr{S}}(f\circ\pi)(u,x,s)\gamma(t)(du,dx,ds)]dt$ as $n\to\infty$. By arguments similar to part i of this lemma, we have that for every $f\in\mathscr{C}(B_r\times\mathscr{S},\mathbb{R})$, for every $T>0$, and for every $g\in L_2([0,T],\mathbb{R})$, $\int_0^T g(t)\cdot[\int_{U\times B_r\times\mathscr{S}}f(x,s)(\theta_1\circ\gamma_n)(t)(dx,ds)]dt\to\int_0^T g(t)[\int_{U\times B_r\times\mathscr{S}}f(x,s)(\theta_1\circ\gamma)(t)(dx,ds)]dt$ as $n\to\infty$. Therefore, $\Theta_1(\gamma_n)\to\Theta_1(\gamma)$ in $\mathscr{M}(B_r\times\mathscr{S})$ as $n\to\infty$, which gives us continuity of $\Theta_1(\cdot)$.

vi. Similar to part v. of this lemma. □

3. Two-Timescale Scheme and Preliminary Results

In this section, we will formally define the two-timescale recursion as well as state and motivate the assumptions imposed (Assumptions A1–A10).

3.1. Recursion and Assumptions

Let $(\Omega,\mathscr{F},\mathbb{P})$ denote a probability space, $\{X_n\}_{n\ge0}$ be a sequence of $\mathbb{R}^{d_1}$-valued random variables on Ω, and $\{Y_n\}_{n\ge0}$ be a sequence of $\mathbb{R}^{d_2}$-valued random variables on Ω, which satisfy for every $n\ge0$,

$$Y_{n+1}-Y_n-b(n)M^{(2)}_{n+1}\in b(n)H_2\left(X_n,Y_n,S^{(2)}_n\right),\tag{15a}$$

$$X_{n+1}-X_n-a(n)M^{(1)}_{n+1}\in a(n)H_1\left(X_n,Y_n,S^{(1)}_n\right),\tag{15b}$$

where the following assumptions hold.

Assumption A1. The map $H_1:\mathbb{R}^d\times\mathscr{S}^{(1)}\to\{\text{subsets of }\mathbb{R}^{d_1}\}$ with $\mathscr{S}^{(1)}$ a compact metric space with metric $d_{\mathscr{S}^{(1)}}$, is such that

i. for every $(x,y,s^{(1)})\in\mathbb{R}^d\times\mathscr{S}^{(1)}$, $H_1(x,y,s^{(1)})$ is a convex and compact subset of $\mathbb{R}^{d_1}$,

ii. there exists $K>0$ such that for every $(x,y,s^{(1)})\in\mathbb{R}^d\times\mathscr{S}^{(1)}$, $\sup_{x'\in H_1(x,y,s^{(1)})}\|x'\|\le K(1+\|x\|+\|y\|)$, and

iii. for every $(x,y,s^{(1)})\in\mathbb{R}^d\times\mathscr{S}^{(1)}$, for every $(\mathbb{R}^d\times\mathscr{S}^{(1)})$-valued sequence, $\{(x_n,y_n,s^{(1)}_n)\}_{n\ge1}$ converging to $(x,y,s^{(1)})\in\mathbb{R}^d\times\mathscr{S}^{(1)}$, and for every sequence $\{x'_n\in H_1(x_n,y_n,s^{(1)}_n)\}_{n\ge1}$ converging to $x'\in\mathbb{R}^{d_1}$, we have that $x'\in H_1(x,y,s^{(1)})$.

Assumption A2. The map $H_2:\mathbb{R}^d\times\mathscr{S}^{(2)}\to\{\text{subsets of }\mathbb{R}^{d_2}\}$ with $\mathscr{S}^{(2)}$ a compact metric space with metric $d_{\mathscr{S}^{(2)}}$ is such that

i. for every $(x,y,s^{(2)})\in\mathbb{R}^d\times\mathscr{S}^{(2)}$, $H_2(x,y,s^{(2)})$ is a convex and compact subset of $\mathbb{R}^{d_2}$,

ii. there exists $K>0$ such that for every $(x,y,s^{(2)})\in\mathbb{R}^d\times\mathscr{S}^{(2)}$, $\sup_{y'\in H_2(x,y,s^{(2)})}\|y'\|\le K(1+\|x\|+\|y\|)$,

iii. for every $(x,y,s^{(2)})\in\mathbb{R}^d\times\mathscr{S}^{(2)}$, for every $(\mathbb{R}^d\times\mathscr{S}^{(2)})$-valued sequence, $\{(x_n,y_n,s^{(2)}_n)\}_{n\ge1}$ converging to $(x,y,s^{(2)})\in\mathbb{R}^d\times\mathscr{S}^{(2)}$, and for every sequence $\{y'_n\in H_2(x_n,y_n,s^{(2)}_n)\}_{n\ge1}$ converging to $y'\in\mathbb{R}^{d_2}$, we have that $y'\in H_2(x,y,s^{(2)})$.

Assumption A3. $\{S^{(1)}_n\}_{n\ge0}$ is a sequence of $\mathscr{S}^{(1)}$-valued random variables on Ω such that for every $n\ge0$, for every $A\in\mathscr{B}(\mathscr{S}^{(1)})$, $\mathbb{P}(S^{(1)}_{n+1}\in A|S^1_m,X_m,Y_m,\ 0\le m\le n)=\mathbb{P}(S^{(1)}_{n+1}\in A|S^{(1)}_n,X_n,Y_n)=\Pi^{(1)}(X_n,Y_n,S^{(1)}_n)(A)$ *a.s*, where $\Pi^{(1)}:\mathbb{R}^d\times\mathscr{S}^{(1)}\to\mathscr{P}(\mathscr{S}^{(1)})$ is continuous.

Assumption A4. $\{S^{(2)}_n\}_{n\geq 0}$ is a sequence of $\mathcal{S}^{(2)}$-valued random variables on Ω, such that for every $n \geq 0$, for every $A \in \mathscr{B}(\mathcal{S}^{(2)})$, $\mathbb{P}(S^{(2)}_{n+1} \in A | S^2_m, X_m, Y_m, 0 \leq m \leq n) = \mathbb{P}(S^{(2)}_{n+1} \in A | S^{(2)}_n, X_n, Y_n) = \Pi^{(2)}(X_n, Y_n, S^{(2)}_n)(A)\, a.s,$ where $\Pi^{(2)} : \mathbb{R}^d \times \mathcal{S}^{(2)} \longrightarrow \mathcal{P}(\mathcal{S}^{(2)})$ is continuous.

Assumption A5. $\{a(n)\}_{n\geq 0}$ and $\{b(n)\}_{n\geq 0}$ are two sequences of positive real numbers satisfying

i. $a(0) \leq 1$ and for every $n \geq 0$, $a(n) \geq a(n+1)$,
ii. $b(0) \leq 1$ and for every $n \geq 0$, $b(n) \geq b(n+1)$,
iii. $\lim_{n\to\infty} \frac{b(n)}{a(n)} = 0$, and
iv. $\sum_{n=0}^{\infty} a(n) = \sum_{n=0}^{\infty} b(n) = \infty$ and $\sum_{n=0}^{\infty}((a(n))^2 + (b(n))^2) < \infty$.

Assumption A6. $\{M^{(1)}_n\}_{n\geq 1}$ is a sequence of $\mathbb{R}^{d_1}$-valued random variables on Ω such that for $a.e.(\omega)$, for any $T > 0$, $\lim_{n\to\infty} \sup_{n\leq k\leq \tau^1(n,T)} \| \sum_{m=n}^{k} a(m) M^{(1)}_{m+1}(\omega) \| = 0$, where $\tau^1(n,T) := \min\{m > n : \sum_{k=n}^{m-1} a(k) \geq T\}$.

Assumption A7. $\{M^{(2)}_n\}_{n\geq 1}$ is a sequence of $\mathbb{R}^{d_2}$-valued random variables on Ω such that for $a.e.(\omega)$, for any $T > 0$, $\lim_{n\to\infty} \sup_{n\leq k\leq \tau^2(n,T)} \| \sum_{m=n}^{k} b(m) M^{(2)}_{m+1}(\omega) \| = 0$, where $\tau^2(n,T) := \min\{m > n : \sum_{k=n}^{m-1} b(k) \geq T\}$.

Assumption A8. $\mathbb{P}(\sup_{n\geq 0}(\|X_n\| + \|Y_n\|) < \infty) = 1$.

3.2. Justification for the Assumptions

Assumptions A1 and A2 ensure that H_1 and H_2 are SAMs. Assumptions A3 and A4 are the iterate-dependent Markov noise assumptions. Under Assumption A3, for every $(x, y) \in \mathbb{R}^d$, the Markov chain associated with the transition kernel given by $\Pi^{(1)}(x, y, \cdot)(\cdot)$ possesses the weak Feller property (see Meyn and Tweedie [29]). In addition to the above, because $\mathcal{S}^{(1)}$ is a compact metric space, the Markov chain associated with the transition kernel $\Pi^{(1)}(x, y, \cdot)(\cdot)$ has at least one stationary distribution for every $(x, y) \in \mathbb{R}^d$ ($\mu \in \mathcal{P}(\mathcal{S}^{(1)})$ is stationary for the Markov chain associated with the transition kernel $\Pi^{(1)}(x, y, \cdot)(\cdot)$ if for every $A \in \mathscr{B}(\mathcal{S}^{(1)})$, $\mu(A) = \int_{\mathcal{S}^{(1)}} \Pi^{(1)}(x, y, s^{(1)})(A)\mu(ds^{(1)})$). For every $(x, y) \in \mathbb{R}^d$, let $D^{(1)}(x, y) \subseteq \mathcal{P}(\mathcal{S}^{(1)})$ denote the set of stationary distributions of the Markov chain associated with the transition kernel $\Pi^{(1)}(x, y, \cdot)(\cdot)$. It can easily be shown that

i. for every $(x, y) \in \mathbb{R}^d$, $D^{(1)}(x, y)$ is a convex and compact subset of $\mathcal{P}(\mathcal{S}^{(1)})$, and
ii. the graph of the map $(x, y) \rightarrow D^{(1)}(x, y)$ is closed, that is, the set

$$\mathcal{G}\left(D^{(1)}\right) := \left\{(x, y, \mu) \in \mathbb{R}^d \times \mathcal{P}\left(\mathcal{S}^{(1)}\right) : (x, y) \in \mathbb{R}^d,\ \mu \in D^{(1)}(x, y)\right\}$$

is a closed subset of $\mathbb{R}^d \times \mathcal{P}(\mathcal{S}^{(1)})$.

The proofs of the above two statements are similar to those in Borkar [11, p. 69]. Similarly, under Assumption A4, for every $(x, y) \in \mathbb{R}^d$, the set of stationary distributions (denoted by $D^{(2)}(x, y)$) associated with the Markov chain defined by the transition kernel $\Pi^{(2)}(x, y, \cdot)(\cdot)$ is a nonempty, convex, and compact subset of $\mathcal{P}(\mathcal{S}^{(2)})$, and the map $(x, y) \longrightarrow D^{(2)}(x, y)$ has a closed graph (i.e., the set $\mathcal{G}(D^{(2)})$ defined in an analogous manner as $\mathcal{G}(D^{(1)})$ is a closed subset of $\mathbb{R}^d \times \mathcal{P}(\mathcal{S}^{(2)})$).

Assumption A5 is the standard two-timescale step-size assumption. Assumption A5iii tells that eventually, the time step taken by Recursion (15a) is smaller than the time step taken by Recursion (15b). Hence, Recursion (15a) is called the slower timescale recursion, and Recursion (15b) is called the faster timescale recursion. Assumptions A6 and A7 are the conditions that the additive noise terms satisfy. These guarantee that the contribution of additive noise terms is eventually negligible. For various noise models where these additive noise assumptions are satisfied, we refer the reader to Benaïm et al. [4].

Assumption A8 is the stability assumption that ensures that the iterates remain within a bounded set. Although this is a standard requirement in the study of such recursions, it is highly nontrivial. An important future direction would be to provide sufficient conditions for verification of Assumption A8. In several applications, stability is ensured by projecting the iterates into a convex compact set, where the convex, compact set is chosen such that desired attractors (i.e., set of points to which the iterates converge to) lie within. In such cases, the recursion studied takes the form

$$Y_{n+1} = P_Y\left[Y_n + b(n)\left(h_2\left(X_n, Y_n, S^{(2)}_n\right) + M^{(2)}_{n+1}\right)\right],$$
$$X_{n+1} = P_X\left[X_n + a(n)\left(h_1\left(X_n, Y_n, S^{(1)}_n\right) + M^{(1)}_{n+1}\right)\right],$$

where the maps h_1 and h_2 denote single-valued Lipschitz continuous maps as in Karmakar and Bhatnagar [21], and P_Y and P_X denote projection maps into appropriate convex and compact sets. The above recursion can be rewritten as

$$Y_{n+1} = Y_n + b(n)\left(\gamma_Y\left(Y_n; h_2\left(X_n, Y_n, S_n^{(2)}\right) + M_{n+1}^{(2)}\right) + o(b(n))\right),$$
$$X_{n+1} = X_n + a(n)\left(\gamma_X\left(X_n; h_1\left(X_n, Y_n, S_n^{(1)}\right) + M_{n+1}^{(1)}\right) + o(a(n))\right),$$

where $\gamma_X(X_n; h_1(X_n, Y_n, S_n^{(1)}) + M_{n+1}^{(1)})$ denotes the directional derivative of the projection map $P_X(\cdot)$ along the direction $h_1(X_n, Y_n, S_n^{(1)}) + M_{n+1}^{(1)}$ (similarly, $\gamma_Y(Y_n; h_2(X_n, Y_n, S_n^{(2)}) + M_{n+1}^{(2)})$ denotes the directional derivative of the projection map $P_Y(\cdot)$ along the direction $h_2(X_n, Y_n, S_n^{(2)}) + M_{n+1}^{(2)}$). In most cases, these directional derivatives are not continuous, and in order to analyze the above recursion, we define the underlying set-valued drift functions as

$$H_2(x, y, s^{(2)}) := \cap_{\epsilon>0}\bar{co}\left(\cup_{\|(x',y')-(x,y)\|<\epsilon}\left\{\gamma_Y\left(y'; h_2\left(x', y', s^{(2)}\right) + z\right) : z \in A^{(2)}(x, y)\right\}\right),$$
$$H_1(x, y, s^{(1)}) := \cap_{\epsilon>0}\bar{co}\left(\cup_{\|(x',y')-(x,y)\|<\epsilon}\left\{\gamma_X\left(x'; h_1\left(x', y', s^{(1)}\right) + z\right) : z \in A^{(1)}(x, y)\right\}\right),$$

where $\bar{co}(\cdot)$ denotes the convex closure of a set, and $A^{(i)}(x, y)$ denotes the support of the conditional distribution of $M_{n+1}^{(i)}$ given a sigma algebra $\mathscr{F}_n$ ($\{\mathscr{F}_n\}_{n\geq 0}$ is a filtration with respect to which $\{M_{n+1}^{(i)}\}_{n\geq 0}$ is a martingale difference array), which is assumed to be compact. The above set-valued drift functions can be shown to be SAMs, and hence the projected scheme can be rewritten in the form of the recursion studied in the paper.

Viewing stochastic approximations with projections as schemes with set-valued maps have been used to analyze recursions arising in variational inequalities by Nagurney and Zhang [31]. Several other schemes used to solve optimization problems, which rely on the martingale-based method for the analysis of the underlying stochastic approximation schemes, obtain stability as a consequence of stronger assumptions on noise and drift functions, which are easily verifiable under that particular application setting (see Jiang and Xu [20], Koshal et al. [23]). Further, adaptive projection-based methods, which are rooted in Chen et al. [14] and Chen and Yunmin [15], guarantee stability by truncating iterates when found to be lying outside a prescribed compact set and have been extended to the case with Markov noise by Andrieu et al. [1] and Fort et al. [18]. Yaji and Bhatnagar [42] extend these adaptive projection schemes to stochastic approximations with set-valued maps without Markov noise.

3.3. Some Additional Assumptions and Preliminary Results

The Markov noise terms in the faster timescale, in limit will average the drift function H_1 w.r.t. the stationary distributions given by the map $(x, y) \to D^{(1)}(x, y)$. The appropriate set-valued map whose associated DI the faster timescale recursion is expected to track is given by

$$\hat{H}_1(x, y) := \cup_{\mu\in D^{(1)}(x,y)} \int_{\mathcal{S}^{(1)}} H_{1,(x,y)}\left(s^{(1)}\right)\mu\left(ds^{(1)}\right) \tag{16}$$

for every $(x, y) \in \mathbb{R}^d$m where for every $(x, y) \in \mathbb{R}^d$, the integrand $H_{1,(x,y)}$ denotes the slice as in Definition 7i of the set-valued map H_1 in Recursion (15b). As a consequence of the step-size Assumption A5 with respect to the faster timescale (15b), the slower timescale recursion (15a) appears to be static and one would expect that the family of DIs,

$$\frac{dx}{dt} \in \hat{H}_1(x, y_0), \tag{17}$$

obtained by fixing some $y_0 \in \mathbb{R}^{d_2}$ to describe the behavior of the faster timescale recursion (15b). Before we proceed, we need to ensure that for every $y_0 \in \mathbb{R}^{d_2}$, the DI (17) has solutions through every initial condition. The next lemma states the map $\hat{H}_1(\cdot, y_0)$ is a Marchaud map for every $y_0 \in \mathbb{R}^{d_2}$, which ensures that the DI (17) has solutions.

Lemma 11. *For every $y_0 \in \mathbb{R}^{d_2}$, the set-valued map $\hat{H}_1(\cdot, y_0) : \mathbb{R}^{d_1} \to \{$ subsets of $\mathbb{R}^{d_1}\}$ is a Marchaud map.*

Proof of the above lemma is given in Section 4. The next assumption will ensure that for every $y_0 \in \mathbb{R}^{d_2}$, the DI (17) has a global attractor to which one expects the faster timescale iterates $\{X_n\}_{n\geq 0}$ to converge.

Assumption A9. For every $y_0 \in \mathbb{R}^{d_2}$, the DI (17) admits a globally attracting set, A_{y_0}. The map $\lambda : \mathbb{R}_{d_2} \to$ {subsets of $\mathbb{R}^{d_1}$}, where for every $y \in \mathbb{R}^{d_2}$, $\lambda(y) := A_y$ is such that

i. for every $y \in \mathbb{R}^{d_2}$, $\sup_{x \in \lambda(y)} \|x\| \leq K(1 + \|y\|)$, and

ii. for every $y \in \mathbb{R}^{d_2}$, for every $\mathbb{R}^{d_2}$-valued sequence, $\{y_n\}_{n\geq 1}$ converging to $y \in \mathbb{R}^{d_2}$, and for every $\{x_n \in \lambda(y_n)\}_{n\geq 0}$ converging to $x \in \mathbb{R}^{d_1}$, we have $x \in \lambda(y)$.

With respect to the slower timescale recursion (15a), the faster timescale recursion will appear to have equilibrated. Further, the Markov noise terms average the set-valued drift function H_2 with respect to the stationary distributions. In what follows, we construct the set-valued map that the slower timescale recursion is expected to track and which captures both the equilibration of the faster timescale and the averaging by the Markov noise terms.

Before we proceed, recall that $\mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$ denotes the set of probability measures on $\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}$ with the Prohorov topology. For any $\mu \in \mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$, $\mu_{\mathbb{R}^{d_1}} \in \mathscr{P}(\mathbb{R}^{d_1})$ and $\mu_{\mathscr{S}^{(2)}} \in \mathscr{P}(\mathscr{S}^{(2)})$ denote the images of the probability measure μ under projections $\mathbb{R}^{d_1} \times \mathscr{S}^{(2)} \to \mathbb{R}^{d_1}$ and $\mathbb{R}^{d_1} \times \mathscr{S}^{(2)} \to \mathscr{S}^{(2)}$, respectively (for any $A \in \mathscr{B}(\mathbb{R}^{d_1})$, $\mu_{\mathbb{R}^{d_1}}(A) := \int_{A \times \mathscr{S}^{(2)}} \mu(dx, ds^{(2)})$, and similarly for every $A \in \mathscr{B}(\mathscr{S}^{(2)})$, $\mu_{\mathscr{S}^{(2)}}(A) := \int_{\mathbb{R}^{d_1} \times A} \mu(dx, ds^{(2)})$).

Define the map $D : \mathbb{R}^{d_2} \to$ {subsets of $\mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S})$} such that for every $y \in \mathbb{R}^{d_2}$,

$$D(y) := \left\{\mu \in \mathscr{P}\left(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}\right) : \operatorname{supp}(\mu_{\mathbb{R}^{d_1}}) \subseteq \lambda(y) \text{ and for every } A \in \mathscr{B}\left(\mathscr{S}^{(2)}\right), \mu_{\mathscr{S}^{(2)}}(A) = \int_{\mathscr{S}^{(2)}} \Pi^{(2)}\left(x, y, s^{(2)}\right)(A)\mu\left(dx, ds^{(2)}\right)\right\}, \tag{18}$$

where $\operatorname{supp}(\mu_{\mathbb{R}^{d_1}})$ denotes the support of measure $\mu_{\mathbb{R}^{d_1}}$ (i.e., $\operatorname{supp}(\mu_{\mathbb{R}^{d_1}}) \subseteq \mathbb{R}^{d_1}$ is a closed set such that $\mu_{\mathbb{R}^{d_1}}(\operatorname{supp}(\mu_{\mathbb{R}^{d_1}})) = 1$, and for every closed set $A \subseteq \mathbb{R}^{d_1}$, with $\mu_{\mathbb{R}^{d_1}}(A) = 1$, we have $\operatorname{supp}(\mu_{\mathbb{R}^{d_1}}) \subseteq A$). A natural question to ask is whether $D(y)$ is nonempty for every $y \in \mathbb{R}^{d_2}$, and if it is nonempty, what properties the map $D(\cdot)$ possesses and its relation to the stationary distributions of the Markov noise terms $\{S_n^{(2)}\}_{n\geq 0}$. The lemma below answers these questions.

Lemma 12. *The map $D(\cdot)$ defined in (18) satisfies*

i. *for every $y \in \mathbb{R}^{d_2}$, $D(y)$ is nonempty, convex, and compact subset of $\mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$,*

ii. *for every $y \in \mathbb{R}^{d_2}$, for every $\mathbb{R}^{d_2}$-valued sequence, $\{y_n\}_{n\geq 1}$ converging to $y \in \mathbb{R}^{d_2}$, and for every $\mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$-valued sequence $\{\mu^n \in D(y_n)\}_{n\geq 1}$ converging to $\mu \in \mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$, we have $\mu \in D(y)$.*

iii. *for every $y \in \mathbb{R}^{d_2}$, $\bar{co}(\{\delta_{x^*} \otimes \nu \in \mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}) : x^* \in \lambda(y),\ \nu \in D^{(2)}(x^*, y)\}) \subseteq D(y)$, where for any $x \in \mathbb{R}^{d_1}$, δ_x denotes the Dirac measure.*

Proof.

i. Fix $y \in \mathbb{R}^{d_2}$. Consider the product measure $\mu := \delta_{x^*} \otimes \nu \in \mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$, where $\delta_{x^*} \in \mathscr{P}(\mathbb{R}^{d_1})$ denotes the Dirac measure on some $x^* \in \lambda(y)$ (i.e., for every $A \in \mathscr{B}(\mathbb{R}^{d_1})$, $\delta_{x^*}(A) = 1$ if $x^* \in A$, $\delta_{x^*}(A) = 0$ otherwise) and $\nu \in \mathscr{P}(\mathscr{S}^{(2)})$ is such that $\nu \in D^{(2)}(x^*, y)$ (i.e., ν is a stationary measure of the Markov chain whose transition kernel is given by $\Pi^{(2)}(x^*, y, \cdot)(\cdot)$). Then, $\mu_{\mathbb{R}^{d_1}} = \delta_{x^*}$ and because $x^* \in \lambda(y)$, $\operatorname{supp}(\mu_{\mathbb{R}^{d_1}}) = \{x^*\} \subseteq \lambda(y)$. Further, $\mu_{\mathscr{S}^{(2)}} = \nu$ and for every $A \in \mathscr{B}(\mathscr{S}^{(2)})$, $\int_{\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}} \Pi^{(2)}(x,y,s)(A)\mu(dx,ds) = \int_{\mathscr{S}^{(2)}} [\int_{\mathbb{R}^{d_1}} \Pi^{(2)}(x,y,s^{(2)})(A)\delta_{x^*}(dx)]\nu(ds^{(2)}) = \int_{\mathscr{S}^{(2)}} \Pi^{(2)}(x^*,y,s^{(2)}) \cdot (A)\nu(ds^{(2)}) = \nu(A)$, where the last equality follows from the fact that $\nu \in D^{(2)}(x^*, y)$. Therefore, $\delta_{x^*} \otimes \nu \in D(y)$, and hence $D(y) \neq \emptyset$.

Let $\mu^1, \mu^2 \in D(y)$ and $\alpha \in (0, 1)$. Consider the measure $\mu := \alpha\mu^1 + (1 - \alpha)\mu^2$ (i.e., for any $A \in \mathscr{B}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$, $\mu(A) = \alpha\mu^1(A) + (1 - \alpha)\mu^2(A)$). Clearly, $\mu \in \mathscr{P}(\mathbb{R}^{d_1} \times \mathscr{S}^{(2)})$, $\mu_{\mathbb{R}^{d_1}} = \alpha\mu^1_{\mathbb{R}^{d_1}} + (1 - \alpha)\mu^2_{\mathbb{R}^{d_1}}$, and $\mu_{\mathscr{S}^{(2)}} = \alpha\mu^1_{\mathscr{S}^{(2)}} + (1 - \alpha)\mu^2_{\mathscr{S}^{(2)}}$. For $i \in 1, 2$, $\operatorname{supp}(\mu^i_{\mathbb{R}^{d_1}}) \subseteq \lambda(y)$, from which we have $\mu^i_{\mathbb{R}^{d_1}}(\lambda(y)) = 1$, and hence $\mu_{\mathbb{R}^{d_1}}(\lambda(y)) = \alpha\mu^1_{\mathbb{R}^{d_1}}(\lambda(y)) + (1 - \alpha) \cdot \mu^2_{\mathbb{R}^{d_1}}(\lambda(y)) = 1$. Therefore, $\operatorname{supp}(\mu_{\mathbb{R}^{d_1}}) \subseteq \lambda(y)$. For every $A \in \mathscr{B}(\mathscr{S}^{(2)})$, $\int_{\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}} \Pi^{(2)}(x,y,s^{(2)})(A)\mu(dx,ds^{(2)}) = \alpha \int_{\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}} \cdot \Pi^{(2)}(x,y,s^{(2)})(A)\mu^1(dx,ds^{(2)}) + (1-\alpha) \int_{\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}} \Pi^{(2)}(x,y,s^{(2)})(A)\mu^2(dx,ds^{(2)}) = \alpha\mu^1_{\mathscr{S}^{(2)}}(A) + (1-\alpha)\mu^2_{\mathscr{S}^{(2)}}(A) = \mu_{\mathscr{S}^{(2)}}(A)$. Therefore, $\mu := \alpha\mu^1 + (1 - \alpha)\mu^2 \in D(y)$, which gives us the convexity of $D(y)$.

In order to show that $D(y)$ is compact, we will first show that the set $D(y)$ is a closed set. Consider $\{\mu^n\}_{n\geq 1}$ such that for every $n \geq 1$, $\mu^n \in D(y)$ converging to $\mu \in \mathscr{P}(\mathbb{R}^d \times \mathscr{S}^{(2)})$. Clearly, $\{\mu^n_{\mathbb{R}^{d_1}}\}_{n\geq 1}$ converges to $\mu_{\mathbb{R}^{d_1}}$ in $\mathscr{P}(\mathbb{R}^{d_1})$. Because for every $n \geq 1$ and because $\operatorname{supp}(\mu^n_{\mathbb{R}^{d_1}}) \subseteq \lambda(y)$, we have $\mu^n(\lambda(y)) = 1$ for every $n \geq 1$. By Assumption A9, $\lambda(y)$ is a compact subset of $\mathbb{R}^{d_1}$ and by Borkar [12, theorem 2.1.1(iv)], we have $\limsup_{n\to\infty} \mu^n_{\mathbb{R}^{d_1}}(\lambda(y)) \leq \mu_{\mathbb{R}^{d_1}}(\lambda(y))$. Therefore, $\mu_{\mathbb{R}^{d_1}}(\lambda(y)) = 1$, which gives us that $\operatorname{supp}(\mu_{\mathbb{R}^{d_1}}) \subseteq \lambda(y)$. Clearly, $\{\mu^n_{\mathscr{S}^{(2)}}\}_{n\geq 1}$ converges to $\mu_{\mathscr{S}^{(2)}}$ in $\mathscr{P}(\mathscr{S}^{(2)})$. Because $\mathscr{S}^{(2)}$ is a compact metric space, by Borkar [12, theorem 2.1.1(ii)], we know that for every $f \in \mathscr{C}(\mathscr{S}^{(2)}, \mathbb{R})$, $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\mu^n_{\mathscr{S}^{(2)}}(d\tilde{s}^{(2)}) \to \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\mu_{\mathscr{S}^{(2)}}(d\tilde{s}^{(2)})$ as $n \to \infty$. Let $\nu^n(d\tilde{s}^2) := \int_{\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}} \Pi^{(2)}(x, y, s^{(2)})(d\tilde{s}^{(2)})\mu(dx, ds^{(2)}) \in \mathscr{P}(\mathscr{S}^{(2)})$ for every $n \geq 1$ and $\nu(d\tilde{s}^{(2)}) := \int_{\mathbb{R}^{d_1} \times \mathscr{S}^{(2)}} \Pi^{(2)}(x, y, s^{(2)})(d\tilde{s}^{(2)})\mu(dx, ds^{(2)})$. It is easy to see that for any $f \in \mathscr{C}(\mathscr{S}^{(2)}, \mathbb{R})$,

$\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu^n(d\tilde{s}^{(2)}) = \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}}[\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})]\mu^n(dx,ds^{(2)})$. By Assumption A4, $(x,s^{(2)}) \to \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})$ $\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})$ is continuous for any $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$. Therefore, as $\mu^n \to \mu$ in $\mathscr{P}(\mathbb{R}^{d_1}\times\mathscr{S}^{(2)})$, we have $\int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}}[\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})]\mu^n(dx,ds^{(2)}) \to \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}}[\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})]\mu(dx,ds^{(2)})$ or $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\cdot$ $\nu^n(d\tilde{s}^{(2)}) \to \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu(d\tilde{s}^{(2)})$. Because for every $n \ge 1$, $\mu^n \in D(y)$, we have $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\mu^n_{\mathscr{S}^{(2)}}(d\tilde{s}^{(2)}) = \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu^n(d\tilde{s}^{(2)})$ for every $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$. Thus, for every $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$, we have $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\mu_{\mathscr{S}^{(2)}}(d\tilde{s}^{(2)}) = \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu(d\tilde{s}^{(2)})$. Therefore, $\mu_{\mathscr{S}^{(2)}} = \nu$, which establishes that $\mu \in D(y)$, and hence $D(y)$ is closed. To establish the compactness of $D(y)$, we show that the set $D(y)$ is relatively compact in $\mathscr{P}(\mathbb{R}^{d_1}\times\mathscr{S}^{(2)})$. For any measure $\mu \in D(y)$, the support of the measure μ, denoted by supp(μ) is contained in $\lambda(y)\times\mathscr{S}^{(2)}$, which is a compact set independent of $\mu \in D(y)$. Thus, the family of measures $\{\mu : \mu \in D(y)\}$ is tight, and, by Prohorov's theorem (see Borkar [12, theorem 2.3.1]), we have a set of measures $D(y)$ that is relatively compact in $\mathscr{P}(\mathbb{R}^{d_1}\times\mathscr{S}^{(2)})$. Therefore, $D(y)$ is closed and relatively compact.

ii. Let $y_n \to y$ in $\mathbb{R}^{d_2}$ and $\mu^n \in D(y_n) \to \mu$ in $\mathscr{P}(\mathbb{R}^{d_1}\times\mathscr{S}^{(2)})$ as $n \to \infty$. Let B_1 denote the closed unit ball in $\mathbb{R}^{d_1}$. By Assumption A9, we have that the set-valued map $y \to \lambda(y)$ is u.s.c. Therefore, for every $\epsilon > 0$, there exists $\delta > 0$(depending on ϵ and y) such that for every $y' \in \mathbb{R}^{d_1}$, satisfying $\|y' - y\| < \delta$, we have $\lambda(y') \subseteq \lambda(y) + \epsilon B_1$. Because $\lambda(y)$ is compact, $\lambda(y) + \epsilon B_1$ is compact. Because $y_n \to y$, there exists N such that for every $n \ge N$, $\|y_n - y\| < \delta$. Then, for all $n \ge N$, $\lambda(y_n) \subseteq \lambda(y) + \epsilon B_1$. By the above, we have that $\limsup_{n\to\infty} \mu^n_{\mathbb{R}^{d_1}}(\lambda(y) + \epsilon(B_1)) = 1$ for every $\epsilon > 0$. Because $\mu^n \to \mu$, we have that $\mu^n_{\mathscr{R}^{d_1}} \to \mu_{\mathscr{R}^{d_1}}$ in $\mathscr{P}(\mathscr{R}^{d_1})$ and by Borkar [12, theorem 2.1.1(iv)], we have that for every $\epsilon > 0$, $\mu_{\mathbb{R}^{d_1}}(\lambda(y) + \epsilon B_1) = 1$. Because $\lambda(y)$ is compact, $\lambda(y) = \cap_{n\ge1}(\lambda(y) + \frac{1}{n}B_1)$ and $\mu_{\mathbb{R}^{d_1}}(\lambda(y)) = \lim_{n\to\infty} \mu_{\mathbb{R}^{d_1}}(\lambda(y) + \frac{1}{n}B_1) = 1$. Therefore, supp$(\mu_{\mathbb{R}^{d_1}}) \subseteq \lambda(y)$. Let $\nu^n(d\tilde{s}^{(2)}) := \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} \Pi^{(2)}(x,y_n,s^{(2)})(d\tilde{s}^2)\cdot$ $\mu^n(dx,ds^{(2)}) \in \mathscr{P}(S^2)$ and $\nu(d\tilde{s}^{(2)}) := \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} \Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})\mu(dx,ds^{(2)})$. Then, for any $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$, for any $n \ge 1$, $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu^n(\tilde{s}^{(2)}) = \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}}[\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x, y_n, s^{(2)})(d\tilde{s}^{(2)})]\mu^n(dx, ds^{(2)})$ and $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu(\tilde{s}^{(2)}) = \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}}[\int_{\mathscr{S}^{(2)}}\cdot$ $f(\tilde{s}^{(2)})\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})]\mu(dx,ds^{(2)})$. Because for every $n \ge 1$, $\mu^n \in D(y_n)$, we have that supp$(\mu^n) \subseteq \lambda(y_n)\times S^2$. By using the u.s.c. property of the map $\lambda(\cdot)$ and the fact that $y_n \to y$, we get that for any $\epsilon > 0$, there exists N such that for every $n \ge N$, $\lambda(y_n)\times\mathscr{S}^{(2)} \subseteq (\lambda(y) + \epsilon B_1)\times\mathscr{S}^{(2)}$. Therefore, for every $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$, for every $n \ge N$, $\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu^n(d\tilde{s}^{(2)}) = \int_{(\lambda(y)+\epsilon B_1)\times\mathscr{S}^{(2)}}[\int_{\mathscr{S}^{(2)}} f(s^{(2)})\Pi^{(2)}(x,y_n,s^{(2)})(d\tilde{s}^{(2)})]\mu^n(dx,ds^{(2)})$. By assumption $(A4)$, the map $(x,y,s^{(2)})$ $\to \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})$ is continuous, and hence its restriction to the compact set $(\lambda(y) + \epsilon B_1)\times C\times\mathscr{S}^{(2)}$ is uniformly continuous, where $C \subseteq \mathbb{R}^{d_2}$ is a compact set such that for every $n \ge 1$, $y_n \in C$. By the above, we can conclude that for any $\tilde{\epsilon} > 0$, there exists N_1 such that for every $n \ge N_1$, for every $(x,s^{(2)}) \in (\lambda(y) + \epsilon B_1)\times\mathscr{S}^{(2)}$, $|\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x,y_n,s^{(2)})(d\tilde{s}^{(2)}) - \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\Pi^{(2)}(x,y,s^{(2)})(d\tilde{s}^{(2)})| < \tilde{\epsilon}$. Therefore, for every $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$, there exists $\tilde{N} := \max\{N, N_1\}$ such that for every $n \ge \tilde{N}$,

$$\left|\int_{\mathscr{S}^{(2)}} f\left(\tilde{s}^{(2)}\right)\nu^n\left(d\tilde{s}^{(2)}\right) - \int_{\mathscr{S}^{(2)}} f\left(\tilde{s}^{(2)}\right)\nu\left(d\tilde{s}^{(2)}\right)\right|$$
$$\le$$
$$\tilde{\epsilon} + \left|\int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}}\left[\int_{\mathscr{S}^{(2)}} f\left(\tilde{s}^{(2)}\right)\Pi^{(2)}\left(x,y,s^{(2)}\right)\left(d\tilde{s}^{(2)}\right)\right]\mu^n\left(ds,ds^{(2)}\right) - \int_{\mathscr{S}^{(2)}} f\left(\tilde{s}^{(2)}\right)\nu\left(d\tilde{s}^{(2)}\right)\right|.$$

The second term in the R.H.S. of the above inequality goes to zero as $n \to \infty$ (use the definition of $\nu(d\tilde{s}^{(2)})$, Assumption A4 and Borkar [12, theorem 2.1.1(ii)]). Therefore, taking the limit in the above equation, we get that for any $f \in \mathscr{C}(\mathscr{S}^{(2)},\mathbb{R})$, for every $\tilde{\epsilon} > 0$, $\lim_{n\to\infty}|\int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu^n(d\tilde{s}^{(2)}) - \int_{\mathscr{S}^{(2)}} f(\tilde{s}^{(2)})\nu(d\tilde{s}^{(2)})| \le \tilde{\epsilon}$. Hence, $\nu^n \to \nu$ in $\mathscr{P}(\mathscr{S}^{(2)})$ as $n \to \infty$. Clearly, $\mu^n_{\mathscr{S}^{(2)}} \to \mu_{\mathscr{S}^{(2)}}$ as $n \to \infty$. Therefore, $\nu = \mu_{\mathscr{S}^{(2)}}$, which gives us that $\mu \in D(y)$.

iii. Follows from part i of this lemma. □

Define the set-valued map $\hat{H}_2 : \mathbb{R}^{d_2} \to \{\text{subsets of } \mathbb{R}^{d_1}\}$ such that for every $y \in \mathbb{R}^{d_2}$,

$$\hat{H}_2(y) := \cup_{\mu\in D(y)} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}\left(x,s^{(2)}\right)\mu\left(dx,ds^{(2)}\right), \tag{19}$$

where for every $y \in \mathbb{R}^{d_2}$, $H_{2,y}$ denotes the slice as in Definition 7iv of the set-valued map H_2. Because for every $y \in \mathbb{R}^{d_2}$, for every $\mu \in D(y)$, supp(μ) is compact, by Lemma 7iii, we know that the slices $H_{2,y}$ are μ-integrable for every $\mu \in D(y)$. So the above set-valued map is well defined, and we will show later that the slower timescale iterates track the DI given by

$$\frac{dy}{dt} \in \hat{H}_2(y). \tag{20}$$

The above DI is guaranteed to have solutions as a consequence of the lemma below.

Lemma 13. *The set-valued map $\hat{H}_2 : \mathbb{R}^{d_2} \longrightarrow \{$ subsets of $\mathbb{R}^{d_2}\}$ is a Marchaud map.*

The proof of the above lemma is given in Section 4.

Remark 3. In order to understand the DI (20) better, we consider the cases where the map $\lambda(\cdot)$ is single-valued and the case where Markov noise terms are absent. These special cases also highlight the fact that our results are a significant generalization of the results by Ramaswamy and Bhatnagar [36] and Karmakar and Bhatnagar [21].

1. When the map $\lambda(\cdot)$ is single valued, for any $\mu \in D(y)$, because $\mathrm{supp}(\mu_{\mathbb{R}^{d_1}}) \subseteq \lambda(y)$, we have that $\mu_{\mathbb{R}^{d_1}} = \delta_{\lambda(y)}$, where $\delta_{\lambda(y)} \in \mathcal{P}(\mathbb{R}^{d_1})$ denotes the Dirac measure at $\lambda(y)$. Therefore, the measure $\mu = \delta_{\lambda(y)} \otimes \mu_{\mathcal{S}^{(2)}}$. Because $\mu \in D(y)$, we know that for every $A \in \mathscr{B}(\mathcal{S}^{(2)})$, $\mu_{\mathcal{S}^{(2)}}(A) = \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} \Pi^{(2)}(x,y,s^{(2)})(A)\mu(dx,ds^{(2)}) = \int_{\mathcal{S}^{(2)}}[\int_{\mathbb{R}^{d_1}} \Pi^{(2)}\cdot(x,y,s^{(2)})(A)\delta_{\lambda(y)}(dx)]\ \mu_{\mathcal{S}^{(2)}}(ds^{(2)}) = \int_{\mathcal{S}^{(2)}} \Pi^{(2)}(\lambda(y),y,s^{(2)})(A)\mu_{\mathcal{S}^{(2)}}(ds^{(2)})$. Thus, $\mu_{\mathcal{S}^{(2)}} \in D^{(2)}(\lambda(y),y)$, where $D^{(2)}(\lambda(y),y)$ denotes the set of stationary measures of the Markov chain with transition kernel $\Pi^{(2)}(\lambda(y),y,\cdot)(\cdot)$. Therefore, for every $y \in \mathbb{R}^{d_2}$,

$$\hat{H}_2(y) = \cup_{\mu\in D(y)} \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H_{2,y}\left(x,s^{(2)}\right)\mu\left(dx,ds^{(2)}\right) = \cup_{\nu\in D^{(2)}(\lambda(y),y)} \int_{\mathcal{S}^{(2)}} H_{2,(\lambda(y),y)}\left(s^{(2)}\right)\nu\left(ds^{(2)}\right),$$

where $H_{2,(\lambda(y),y)}$ denotes the slice as in Definition 7i of the set-valued map H_2. Therefore, DI (20) is nothing but the set-valued analog of the slower timescale DI in Karmakar and Bhatnagar [21].

2. Suppose Markov noise terms are absent (for the analysis and definition of such a recursion, see Ramaswamy and Bhatnagar [36]). Then, such a recursion can be rewritten in the form of Recursion (15), with Markov noise terms taking values in a dummy state space $\mathcal{S}^{(1)} = \mathcal{S}^{(2)} = \{s^*\}$ with transition laws $\Pi^{(1)}(x,y,s^*) = \Pi^{(2)}(x,y,s^*) = \delta_{s^*}$ for every $(x,y) \in \mathbb{R}^d$. Then, it is easy to deduce that the stationary distribution maps $D^{(1)}(x,y) = D^{(2)}(x,y) = \delta_{s^*}$ for every $(x,y) \in \mathbb{R}^d$. Then, for every $y \in \mathbb{R}^{d_2}$, any $\mu \in D(y)$ is of the form $\mu = \nu \otimes \delta_{s^*}$, where $\nu \in \mathcal{P}(\mathbb{R}^{d_1})$ with $\mathrm{supp}(\nu) \subseteq \lambda(y)$. Then, for any $y \in \mathbb{R}^{d_2}$,

$$\hat{H}_2(y) = \cup_{\mu\in D(y)} \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H_{2,y}\left(x,s^{(2)}\right)\mu\left(dx,ds^{(2)}\right) = \cup_{\substack{\nu\in\mathcal{P}(\mathbb{R}^{d_1}):\\ \mathrm{supp}(\nu)\subseteq\lambda(y)}} \int_{\mathbb{R}^{d_1}} H_{2,y}(x,s^*)\nu(dx) = \bar{co}\left(\cup_{x\in\lambda(y)}H_2(x,y,s^*)\right),$$

which is exactly the same slower timescale DI as in Ramaswamy and Bhatnagar [36].

Suppose now that the following holds in addition:

Assumption A10. DI (20) has a globally attracting set $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$, then, the main result of this paper states that for almost every ω, as $n \to \infty$,

$$\begin{pmatrix} X_n(\omega) \\ Y_n(\omega) \end{pmatrix} \to \cup_{y\in\mathcal{Y}}\left(\lambda(y) \times \{y\}\right).$$

4. Mean Fields and Their Properties

In this section, we prove that for every $y \in \mathbb{R}^{d_2}$, the set-valued map $\hat{H}_1(\cdot,y)$ and the set-valued map $\hat{H}_2(\cdot)$ defined in Equations (16) and (19), respectively, are Marchaud maps.

Recall that by Assumptions A1 and A2, the set-valued maps H_1 and H_2 are SAMs. For such set-valued maps, by Lemma 2, we know that there exist sequences of continuous set-valued maps, denoted by $\{H_1^{(l)}\}_{l\geq 1}$ and $\{H_2^{(l)}\}_{l\geq 1}$, which approximate H_1 and H_2, respectively. Further, by Lemma 3, these approximating maps admit a continuous parametrization denoted by, $h_1^{(l)}$ and $h_2^{(l)}$. Throughout this section, $\{H_1^{(l)}\}_{l\geq 1}$, $\{H_2^{(l)}\}$, $\{h_1^{(l)}\}_{l\geq 1}$, and $\{h_2^{(l)}\}_{l\geq 1}$ denote the maps as described above.

Similar to the definition of the maps $\hat{H}_1$ and $\hat{H}_2$, we define the maps obtained by averaging the set-valued maps $H_1^{(l)}$ and $H_2^{(l)}$ for every $l \geq 1$ with respect to measures given by the maps $(x,y) \to D^{(1)}(x,y)$ and $y \to D(y)$.

Definition 9. Let the maps $D^{(1)} : \mathbb{R}^d \to \{\text{subsets of } \mathcal{P}(\mathcal{S}^{(1)})\}$ and $D : \mathbb{R}^{d_2} \to \{\text{subsets of } \mathcal{P}(\mathbb{R}^{d_1}\times\mathcal{S}^{(2)})\}$ be as in Section 3. For every $l \geq 1$,

i. for every $(x,y) \in \mathbb{R}^d$, define $\hat{H}_1^{(l)} : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}^{d_1}\}$ such that,

$$\hat{H}_1^{(l)}(x,y) := \cup_{\mu\in D^{(1)}(x,y)} \int_{\mathcal{S}^{(1)}} H_{1,(x,y)}^{(l)}\left(s^{(1)}\right)\mu\left(ds^{(1)}\right),$$

where $H^{(l)}_{1,(x,y)}$ denotes the slice (as in Definition 7ii) of the set-valued map $H^{(l)}_1$, and

ii. for every $y \in \mathbb{R}^{d_2}$, define $\hat{H}^{(l)}_2 : \mathbb{R}^{d_2} \longrightarrow \{\text{subsets of } \mathbb{R}^{d_2}\}$ such that,

$$\hat{H}^{(l)}_2(y) := \cup_{\mu \in D(y)} \int_{\mathcal{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right),$$

where $H^{(l)}_{2,y}$ denotes the slice (as in Definition 7v) of the set-valued map $H^{(l)}_2$.

In the lemma below, we prove that for every $y \in \mathbb{R}^{d_2}$, the maps $\hat{H}^{(l)}_1(\cdot, y)$ and the map $\hat{H}^{(l)}_2(\cdot)$ are Marchaud maps for every $l \geq 1$.

Lemma 14. *For every $l \geq 1$,*

i. *the set-valued map $\hat{H}^{(l)}_1 : \mathbb{R}^d \longrightarrow \{\text{subsets of } \mathbb{R}^{d_1}\}$ is such that,*

a. *for every $(x,y) \in \mathbb{R}^d$, $\hat{H}^{(l)}_1(x,y)$ is a nonempty, convex and compact subset of $\mathbb{R}^{d_1}$, and*

b. *for $K^{(l)} > 0$, where $K^{(l)}$ is as in Lemma 2, for every $(x,y) \in \mathbb{R}^d$, $\sup_{x' \in \hat{H}^{(l)}_1(x,y)} \|x'\| \leq K^{(l)}(1 + \|x\| + \|y\|)$,*

I for every $(x,y) \in \mathbb{R}^d$, for $\mathbb{R}^d$-valued sequence, $\{(x_n, y_n)\}_{n\geq 1}$ converging to $(x,y) \in \mathbb{R}^d$, and for every sequence $\{x'_n \in \hat{H}^{(l)}_1(x_n, y_n)\}$ converging to $x' \in \mathbb{R}^{d_1}$, we have that $x' \in \hat{H}^{(l)}_1(x,y)$.

ii. *for every $y \in \mathbb{R}^{d_2}$, the map $\hat{H}^{(l)}_1(\cdot, y)$ is a Marchaud map,*

iii. *the map $\hat{H}^{(l)}_2(\cdot)$ is a Marchaud map.*

Proof. Fix $l \geq 1$.

i. For every $(x,y) \in \mathbb{R}^d$, by Lemma 6iv, $H^{(l)}_{1,(x,y)}$ is μ-integrable for every $\mu \in D^{(1)}(x,y)$. Hence, for every $(x,y) \in \mathbb{R}^d$, $\hat{H}^{(l)}_1(x,y)$ is nonempty. Let $x^1, x^2 \in \hat{H}^{(l)}_1(x,y)$ and $\alpha \in (0,1)$. Then by Lemma 8i, there exist $\nu^1, \nu^2 \in \mathcal{P}(\mathcal{S}^{(1)} \times U)$ such that for $i \in \{1,2\}$, $\nu^i_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$ and $x^i = \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu^i(ds^{(1)}, du)$, where U denotes the closed unit ball in $\mathbb{R}^{d_1}$. Then, $\alpha x^1 + (1-\alpha)x^2 = \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)(\alpha\nu^1 + (1-\alpha)\nu^2)(ds^{(1)}, du)$. Clearly, $(\alpha\nu^1 + (1-\alpha)\nu^2)_{\mathcal{S}^{(1)}} = \alpha\nu^1_{\mathcal{S}^{(1)}} + (1-\alpha)\nu^2_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$, where the last inclusion follows from the fact that $D^{(1)}(x,y)$ is a convex subset of $\mathcal{P}(\mathcal{S}^{(1)})$. By Lemma 8i, we get that $\alpha x^1 + (1-\alpha)x^2 \in \hat{H}^{(l)}_1(x,y)$. Therefore, $\hat{H}^{(l)}_1(x,y)$ is convex.

By Lemma 6ii, for every $(x,y) \in \mathbb{R}^d$, the set-valued map $H^{(l)}_{1,(x,y)}$ is bounded by $C^{(l)}_{(x,y)} := K^{(l)}(1 + \|x\| + \|y\|)$. Therefore, for every $f \in \mathscr{S}(H_{1,(x,y)})$, for every $s^{(1)} \in \mathcal{S}^{(1)}$, $\|f(s^{(1)})\| \leq C^{(l)}_{(x,y)}$. Thus, for every $x' \in \hat{H}^{(l)}_1(x,y)$, by definition, $x' = \int_{\mathcal{S}^{(1)}} f(s^{(1)})\mu(ds^{(1)})$ for some $f \in \mathscr{S}(H^{(l)}_1)$ and some $\mu \in D^{(1)}(x,y)$. Therefore, for every $x' \in \hat{H}^{(l)}_1(x,y)$, $\|x'\| \leq \int_{\mathcal{S}^{(1)}} \|f(s^{(1)})\|\mu(ds^{(1)}) \leq C^{(l)}_{(x,y)} = K^{(l)}(1 + \|x\| + \|y\|)$.

As a consequence of the arguments in the preceding paragraph, for some $(x,y) \in \mathbb{R}^d$, in order to show that $\hat{H}^{(l)}_1(x,y)$ is compact, it is enough to show that it is closed. Consider a sequence $\{x^n \in \hat{H}^l_1(x,y)\}_{n\geq 1}$ converging to $x^* \in \mathbb{R}^{d_1}$. Then, by the definition of $\hat{H}^{(l)}_1(x,y)$ and by Lemma 8i, for every $n \geq 1$, there exists $\nu^n \in \mathcal{P}(\mathcal{S}^1 \times U)$ such that $\nu^n_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$ and $x^n = \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu^n(ds^{(1)}, du)$. Because $\mathcal{S}^{(1)} \times U$ is a compact metric space, $\mathcal{P}(\mathcal{S}^{(1)} \times U)$ is compact, and hence there exists a subsequence $\{n_k\}_{k\geq 1}$ such that $\{\nu^{n_k}\}_{k\geq 1}$ converges to $\nu \in \mathcal{P}(\mathcal{S}^{(1)} \times U)$. Clearly, $\{\nu^{n_k}_{\mathcal{S}^{(1)}}\}_{k\geq 1}$ converges to $\nu_{\mathcal{S}^{(1)}}$ and by Borkar [12, theorem 2.1.1(ii)], $x^{n_k} = \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu^{n_k}(ds^{(1)}, du) \longrightarrow \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu(ds^{(1)}, du) = x^*$. Because for every k, $\nu^{n_k}_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$ and by the fact that $D^{(1)}(x,y)$ is closed, we get that $\nu_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$. Therefore, $x^* = \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, du)\nu(ds^{(1)}, du)$ and $\nu_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$. Thus, $x^* \in \hat{H}^{(l)}_1(x,y)$, which gives us that $\hat{H}^{(l)}_1(x,y)$, is closed.

Let $\{(x_n, y_n)\}_{n\geq 1}$ be a sequence converging to (x,y) and let $\{x'_n \in \hat{H}^{(l)}_1\}_{n\geq 1}$ be a sequence converging to x'. Then, by Lemma 8i, for every $n \geq 1$, there exists $\nu^n \in \mathcal{P}(\mathcal{S}^{(1)} \times U)$ such that $\nu^n_{\mathcal{S}^{(1)}} \in D^{(1)}(x_n, y_n)$ and $x'_n = \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu^n(ds^{(1)}, du)$. Because $\mathcal{S}^{(1)} \times U$ is a compact metric space, $\mathcal{P}(\mathcal{S}^{(1)} \times U)$ is a compact metric space, and hence there exists a subsequence, say $\{n_k\}_{k\geq 1}$, such that $\{\nu^{n_k}\}_{k\geq 1}$ converges to $\nu \in \mathcal{P}(\mathcal{S}^{(1)} \times U)$. Clearly, $\nu^{n_k}_{\mathcal{S}^{(1)}} \longrightarrow \nu_{\mathcal{S}^{(1)}}$ in $\mathcal{P}(\mathcal{S}^{(1)})$ and by the closed graph property of the map $(x,y) \longrightarrow D^{(1)}(x,y)$, we have that $\nu_{\mathcal{S}^{(1)}} \in D^{(1)}(x,y)$. Using the continuity of the map $h^{(l)}_1(\cdot)$, it is easy to show that $\lim_{k\to\infty} \sup_{(s^{(1)},u)\in\mathcal{S}^{(1)}\times U} \|h^{(l)}_{1,(x_{n_k}, y_{n_k})}(s^{(1)}, u) - h^{(l)}_{1,(x,y)}(s^{(1)}, u)\| = 0$. Then, $\|x' - \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu(ds^{(1)}, du)\| \leq \|x' - \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x_{n_k}, y_{n_k})}(s^{(1)}, u)\nu^{n_k}(ds^{(1)}, du)\| + \int_{\mathcal{S}^{(1)}\times U} \|h^{(l)}_{1,(x_{n_k}, y_{n_k})}(s^{(1)}, u) - h^{(l)}_{1,(x,y)}(s^{(1)}, u)\|\nu^{n_k}(ds^{(1)}, du) + \|\int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu^{n_k}(ds^{(1)}, du) - \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu(ds^{(1)}, du)\|$. Now use Borkar [12, theorem 2.1.1(ii)] in the above inequality to obtain $\lim_{k\to\infty} \|x' - \int_{\mathcal{S}^{(1)}\times U} h^{(l)}_{1,(x,y)}(s^{(1)}, u)\nu(ds^{(1)}, du)\| = 0$. Then Lemma 8i gives us that $x' \in \hat{H}^{(l)}_1(x,y)$.

ii. Follows from part i of this lemma.

iii. The proof′ is similar to part i of this lemma with minor modifications. The first modification is the use of Lemma 8ii instead of Lemma 8i. For example, in order to show that $\hat{H}^{(l)}_2(y)$ is closed for some $y \in \mathbb{R}^{d_2}$, fix sequence $\{y'_n\}_{n\geq 1} \subseteq \hat{H}_2^{(l)}(y)$ converging to y'. Use Lemma 8ii and the definition of $\hat{H}_2^{(l)}(y)$, to obtain $\{\nu^n\}_{n\geq 1} \subseteq \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S}^{(2)} \times U)$, where U denotes the closed unit ball in $\mathbb{R}^{d_2}$, and the sequence $\{\nu^n\}_{n\geq 1}$ is such that for every $n \geq 1$, $\nu^n_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} \in D(y)$ and $y'_n = \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}\times U} h^{(l)}_{2,y}(x, s^{(2)}, u)\nu^n(dx, ds^{(2)}, du)$. By definition of $D(y)$, for every $n \geq 1$, $\text{supp}(\nu^n_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}}) \subseteq \lambda(y) \times \mathcal{S}^{(2)}$ and hence $\text{supp}(\nu^n) \subseteq \lambda(y) \times \mathcal{S}^{(2)} \times U$, which is a compact subset of $\mathbb{R}^{d_1} \times \mathcal{S}^{(2)} \times U$. Now by Prohorov's theorem, the sequence $\{\nu^n\}_{n\geq 1}$ is a relatively compact subset of $\mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S}^{(2)} \times U)$ and hence has a convergent subsequence. By Lemma 12i, $D(y)$ is compact, and hence every limit point of $\{\nu^n_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}}\}$ is in $D(y)$. The rest of the argument is same as the corresponding in part (*i*) of this lemma.

In order to show that $\hat{H}_2^{(l)}(\cdot)$ has a closed graph, fix sequences $\{y_n\}_{n\geq 1}$ converging to y and $\{y'_n \in \hat{H}_2^{(l)}(y_n)\}_{n\geq 1}$ converging to y'. Use Lemma 8ii, to obtain $\{\nu^n\}_{n\geq 1} \subseteq \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S}^{(2)} \times U)$ such that for every $n \geq 1$, $\nu^n_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} \in D(y_n)$ and $y'_n := \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}\times U} h^{(l)}_{2,y}(x, s^{(2)}, u)\nu^n(dx, ds^{(2)}, du)$. Then, for every $n \geq 1$, $\text{supp}(\nu^n) \subseteq \lambda(y_n) \times \mathcal{S}^{(2)} \times U$. By Assumption A9, for any $\delta > 0$, the set $L := \{x \in \lambda(\tilde{y}) : \|\tilde{y} - y\| \leq \delta\}$ is a compact subset of $\mathbb{R}^{d_1}$. Therefore, there exists N large such that for every $n \geq N$, $\text{supp}(\nu^n) \subseteq L \times \mathcal{S}^{(2)} \times U$. By Prohorov's theorem, the sequence of measures $\{\nu^n\}_{n\geq N}$ is tight and has a convergent subsequence. Clearly, by Lemma 12ii, every limit point of $\{\nu^n_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}}\}_{n\geq N}$ is in $D(y)$. Now the rest of the argument is the same as the corresponding argument in part i of this lemma. □

By Lemma 2, we know that for every $l \geq 1$, for every $(x, y) \in \mathbb{R}^d$, $H_1(x, y) \subseteq H_1^{(l+1)}(x, y) \subseteq H_1^{(l)}(x, y)$ (similarly, $H_2(x, y) \subseteq H_2^{(l+1)}(x, y) \subseteq H_2^{(l)}(x, y)$). The next lemma states that the above is true for $\hat{H}_i$ and $\hat{H}_i^{(l)}$ as well for every $i \in \{1, 2\}$.

Lemma 15.

i. *For every* $l \geq 1$, *for every* $(x, y) \in \mathbb{R}^d$, $\hat{H}_1(x, y) \subseteq \hat{H}_1^{(l+1)}(x, y) \subseteq \hat{H}_1^{(l)}(x, y)$.

ii. *For every* $l \geq 1$, *for* $y \in \mathbb{R}^{d_2}$, $\hat{H}_2(y) \subseteq \hat{H}_2^{(l+1)}(y) \subseteq \hat{H}_2^{(l)}(y)$.

iii. *For every* $(x, y) \in \mathbb{R}^d$, $\cap_{l\geq 1}\hat{H}_1^{(l)}(x, y) = \cup_{\mu\in D^{(1)}(x,y)} \cap_{l\geq 1} \int_{\mathcal{S}^{(1)}} H^{(l)}_{1,(x,y)}(s^{(1)})\mu(ds^{(1)})$.

iv. *For every* $y \in \mathbb{R}^{d_2}$, $\cap_{l\geq 1}\hat{H}_2^{(l)}(y) = \cup_{\mu\in D(y)} \cap_{l\geq 1} \int_{\mathbb{R}^{d_2}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$.

v. *For every* $(x, y) \in \mathbb{R}^d$, $\hat{H}_1(x, y) = \cap_{l\geq 1}\hat{H}_1^{(l)}(x, y)$.

vi. *For every* $y \in \mathbb{R}^{d_2}$, $\hat{H}_2(y) = \cap_{l\geq 1}\hat{H}_2^{(l)}(y)$.

Proof. The proofs of parts i and ii follow directly from the definition of $\hat{H}_i$, $\hat{H}_i^{(l)}$ for every $l \geq 1$ and the fact that for every $i \in \{1, 2\}$, $H_i(x, y) \subseteq H_i^{(l+1)}(x, y) \subseteq H_i^{(l)}(x, y)$ for every $(x, y) \in \mathbb{R}^d$. The proof of part iii is similar to part iv, and we present the proof of part iv below (the proof of part iii is in fact the same as that of Yaji and Bhatnagar [41, lemma 4.4(ii)]).

iv. Fix $y \in \mathbb{R}^{d_2}$. Then, by definition of $\hat{H}_2^{(l)}(y)$, we have that for every $l \geq 1$, for any $\mu \in D(y)$, $\int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)}) \cdot \mu(dx, ds^{(2)}) \subseteq \hat{H}_2^{(l)}(y)$. Therefore, $\cup_{\mu\in D(y)} \cap_{l\geq 1} \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)}) \subseteq \cap_{l\geq 1}\hat{H}_2^{(l)}(y)$.

Let $y' \in \cap_{l\geq 1}\hat{H}_2^{(l)}(y)$. Then, for every $l \geq 1$, there exists $\mu^l \in D(y)$ such that $y' \in \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu^l(dx, ds^{(2)})$. Because $\{\mu^l\}_{l\geq 1}$ is a subset of $D(y)$, for every $l \geq 1$, $\text{supp}(\mu^l) \subseteq \lambda(y) \times \mathcal{S}^{(2)}$. Hence, the sequence of probability measures $\{\mu^l\}_{l\geq 1}$ is tight and, by Prohorov's theorem, has a limit, say $\mu^* \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S}^{(2)})$. Let $\{l_k\}_{k\geq 1}$ be a subsequence such that $\mu^{l_k} \to \mu^*$ as $k \to \infty$, and, by Lemma 12i, we know that $D(y)$ is compact, which gives us $\mu^* \in D(y)$. Because for every $l \geq 1$, for every k such that $l_k \geq l$, $\mathcal{S}(H^{(l_k)}_{2,y}) \subseteq \mathcal{S}(H^{(l)}_{2,y})$, we get that for every $l \geq 1$, for every k such that $l_k \geq l$, $\int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu^{l_k}(dx, ds^{(2)}) = y'$. For every $l \geq 1$, by Lemma 8ii, we know that for every k such that $l_k \geq l$, there exists $\nu^{(l,l_k)} \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S}^{(2)} \times U)$ (U denotes the closed unit ball in $\mathbb{R}^{d_2}$) such that $y' = \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}\times U} h^{(l)}_{2,y}(x, s^{(2)}, u) \cdot \nu^{(l,l_k)}(dx, ds^{(2)}, du)$ and $\nu^{(l,l_k)}_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} = \mu^{l_k}$. Further, for every $l \geq 1$, for every k such that $l_k \geq l$, $\text{supp}(\nu^{(l,l_k)}) \subseteq \lambda(y) \times \mathcal{S}^{(2)} \times U$, and hence $\{\nu^{(l,l_k)}\}_{k:l_k\geq l}$ is tight and, by Prohorov's theorem, has a convergent subsequence. For every $l \geq 1$, let $\nu^{(l)}$ denote a limit point of the sequence $\{\nu^{(l,l_k)}\}_{k:l_k\geq l}$. Because for every $l \geq 1$, $\{\nu^{(l,l_k)}_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} = \mu^{l_k}\}_{k:l_k\geq l}$ and $\mu^{l_k} \to \mu^*$ as $k \to \infty$, we have that $\nu^{(l)}_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} = \mu^* \in D(y)$. By Borkar [12, theorem 2.1.1(ii)], for every $l \geq 1$, $y' = \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}\times U} h^{(l)}_{2,y}(x, s^{(2)}, u) \cdot \nu^{(l)}(dx, ds^{(2)}, du)$, and hence by Lemma 8ii, $y' \in \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu^*(dx, ds^{(2)})$, where $\mu^* \in D(y)$. Therefore, there exists $\mu^* \in D(y)$ such that for every $l \geq 1$, $y' \in \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu^*(dx, ds^{(2)})$. Hence, $y' \in \cup_{\mu\in D(y)} \cap_{l\geq 1} \int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} \cdot H^{(l)}_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$.

The proof of part v is similar to the proof of part vi, and we present a proof of part vi below (the proof of part v is exactly the same as that of Yaji and Bhatnagar [41, lemma 4.4(iii)]).

vi. From part ii of this lemma, we have that, for every $y \in \mathbb{R}^{d_2}$, $\hat{H}_2(y) \subseteq \cap_{l\geq 1}\hat{H}_2^{(l)}(y)$.

Fix $y \in \mathbb{R}^{d_2}$ and $\mu \in D(y)$. Let $y' \in \cap_{l\geq 1} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H^{(l)}_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$. Then, for every $l \geq 1$, there exists $f^{(l)} \in \mathscr{S}(H^{(l)}_{2,y})$ such that $y' = \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} f^{(l)}(x, s^{(2)})\mu(dx, ds^{(2)})$. Let $d(\tilde{y}, A) := \inf\{\|\tilde{y} - z\| : z \in A\}$ for every $\tilde{y} \in \mathbb{R}^{d_2}$ and for every $A \subseteq \mathbb{R}^{d_2}$ compact. By Lemma 5, we have that $\int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$ is compact and convex. Then,

$$\begin{aligned} d\left(y', \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right)\right) &= \inf\left\{\|y' - z\| : z \in \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right)\right\} \\ &= \inf_{f\in\mathscr{S}(H_{2,y})} \left\| y' - \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} f\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right)\right\| \\ &= \inf_{f\in\mathscr{S}(H_{2,y})} \left\| \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} \left(f^{(l)}\left(x, s^{(2)}\right) - f\left(x, s^{(2)}\right)\right)\mu\left(dx, ds^{(2)}\right)\right\| \\ &\leq \inf_{f\in\mathscr{S}(H_{2,y})} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} \left\| f^{(l)}\left(x, s^{(2)}\right) - f\left(x, s^{(2)}\right)\right\| \mu\left(dx, ds^{(2)}\right) \\ &= \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} \inf\left\{\left\| f^{(l)}\left(x, s^{(2)}\right) - \tilde{y}\right\| : \tilde{y} \in H_{2,y}\left(x, s^{(2)}\right)\right\}\mu\left(dx, ds^{(2)}\right), \end{aligned}$$

where the last equality follows from Li et al. [27, lemma 1.3.12]. By Yaji and Bhatnagar [41, lemma 3.7], we know that for every $l \geq 1$, the map $(x, s^{(2)}) \rightarrow d(f^{(l)}(x, s^{(2)}), H_{2,y}(x, s^{(2)}))$ is measurable, and from the last equality, it follows that for every $l \geq 1$,

$$d\left(y', \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right)\right) \leq \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} d\left(f^{(l)}\left(x, s^{(2)}\right), H_{2,y}\left(x, s^{(2)}\right)\right)\mu\left(dx, ds^{(2)}\right).$$

By Observation (2) stated after Lemma 2, we have that for every $(x, s^{(2)}) \in \mathbb{R}^{d_2} \times \mathscr{S}^{(2)}$, $\lim_{l\to\infty} d(f^{(l)}(x, s^{(2)}), H_{2,y}(x, s^{(2)})) = 0$. Because $\mu \in D(y)$, $\text{supp}(\mu) \subseteq \lambda(y) \times \mathscr{S}^{(2)}$, for every $l \geq 1$,

$$\int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} d\left(f^{(l)}\left(x, s^{(2)}\right), H_{2,y}\left(x, s^{(2)}\right)\right)\mu\left(dx, ds^{(2)}\right) = \int_{\lambda(y)\times\mathscr{S}^{(2)}} d\left(f^{(l)}\left(x, s^{(2)}\right), H_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right)\right).$$

Because $\lambda(y)$ is compact, there exists $M > 0$ such that for every $x \in \lambda(y)$, $\|x\| \leq M$. By Lemma 7ii, Assumption A2 and Observation (1) stated below Lemma 2, we have that for every $l \geq 1$, for every $(x, s^{(2)}) \in \lambda(y) \times \mathscr{S}^{(2)}$, $d(f^{(l)}(x, s^{(2)}), H_{2,y}(x, s^{(2)})) \leq (K^{(l)}_y + K)(1 + \|x\|) \leq (\max\{\tilde{K}, \tilde{K}\|y\|\} + K)(1 + M)$. By the bounded convergence theorem, we have

$$d\left(y', \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right)\right) \leq \lim_{l\to\infty} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} d\left(f^{(l)}\left(x, s^{(2)}\right), H_{2,y}\left(x, s^{(2)}\right)\right)\mu\left(dx, ds^{(2)}\right) = 0.$$

Therefore, $d(y', \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})) = 0$, and by Lemma 5, we know that $\int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$ is a closed subset of $\mathbb{R}^{d_2}$. Hence, $y' \in \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$.

From the arguments in the preceding paragraph, we have that for every $y \in \mathbb{R}^{d_2}$, for every $\mu \in D(y)$, $\cap_{l\geq 1} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} \cdot H^{(l)}_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)}) \subseteq \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}(x, s^{(2)})\mu(dx, ds^{(2)})$. Thus, for every $y \in \mathbb{R}^{d_2}$,

$$\cup_{\mu\in D(y)} \cap_{l\geq 1} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H^{(l)}_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right) \subseteq \cup_{\mu\in D(y)} \int_{\mathbb{R}^{d_1}\times\mathscr{S}^{(2)}} H_{2,y}\left(x, s^{(2)}\right)\mu\left(dx, ds^{(2)}\right) = \hat{H}_2(y).$$

By part iv of this lemma, we get that for every $y \in \mathbb{R}^{d_2}$, $\cap_{l\geq 1}\hat{H}^{(l)}_2(y) \subseteq \hat{H}_2(y)$. □

Lemma 16. *The set-valued map $\hat{H}_1 : \mathbb{R}^d \rightarrow \{$subsets of $\mathbb{R}^{d_1}\}$ as defined in Equation (16) is such that,*

i. *for every $(x, y) \in \mathbb{R}^d$, $\hat{H}_1(x, y)$ is a nonempty, convex, and compact subset of $\mathbb{R}^{d_1}$,*

ii. *there exists $K > 0$ (same as in Assumption A1ii such that for every $(x, y) \in \mathbb{R}^d$, $\sup_{x'\in\hat{H}_1(x,y)} \|x'\| \leq K(1 + \|x\| + \|y\|)$, and*

iii. *for every $(x, y) \in \mathbb{R}^d$, for every $\mathbb{R}^d$-valued sequence, $\{(x_n, y_n)\}_{n\geq 1}$ converging to (x, y), and for every $\{x'_n \in \hat{H}_1 (x_n, y_n)\}_{n\geq 1}$ converging to x', we have $x' \in \hat{H}_1(x, y)$.*

Proof.

i. Fix $(x, y) \in \mathbb{R}^d$. By Lemma 6iii, $H_{1,(x,y)}$ is μ-integrable for every $\mu \in D^{(1)}(x, y)$. Hence, $\hat{H}_1(x, y)$ is nonempty. For every $l \geq 1$, by Lemma 14i(a), we know that $\hat{H}^{(l)}_1(x, y)$ is convex and compact subset of $\mathbb{R}^{d_1}$. By Lemma 15v, we have that $\hat{H}_1(x, y) = \cap_{l\geq 1}\hat{H}^{(l)}_1(x, y)$, and hence $\hat{H}_1(x, y)$ is convex and compact.

ii. Fix $(x,y) \in \mathbb{R}^d$. For any $x' \in \hat{H}_1(x,y)$, there exists $\mu \in D^{(1)}(x,y)$ and $f \in \mathscr{S}(H_{1,(x,y)})$ such that $x' = \int_{\mathcal{S}^{(1)}} f(s^{(1)}) \cdot \mu(ds^{(1)})$. By Lemma 6i, we know that for every $s^{(1)} \in \mathcal{S}^{(1)}$, $\|f(s^{(1)})\| \leq C_{(x,y)} = K(1 + \|x\| + \|y\|)$. Therefore, $\|x'\| = \| \int_{\mathcal{S}^{(1)}} f(s^{(1)}) \ \mu(ds^{(1)})\| \leq C_{(x,y)} = K(1 + \|x\| + \|y\|)$.

iii. Let $\{(x_n, y_n)\}_{n\geq 1}$ a sequence converging to (x,y) and $\{x'_n \in \hat{H}_1(x_n, y_n)\}_{n\geq 1}$ be a sequence converging to x'. Then, by Lemma 15v, we have that for every $l \geq 1$, for every $n \geq 1$, $x'_n \in \hat{H}_1^{(l)}(x_n, y_n)$. By Lemma 14i(c), we have that for every $l \geq 1$, $x' \in \hat{H}_1^{(l)}(x,y)$. Thus by Lemma 15v, we have $x' \in \hat{H}_1(x,y)$. □

Lemma 11 is an immediate consequence of the above lemma. Similarly, the proof of Lemma 13 follows from the fact that $\{\hat{H}_2^{(l)}\}_{l\geq 1}$ are Marchaud maps (see Lemma 14iii), which approximate $\hat{H}_2$ (see Lemma 15vi and the linear growth property of the map $\lambda(\cdot)$ (i.e., Assumption A9i).

5. Recursion Analysis

In this section, we present the analysis of Recursion (15). The analysis comprises two parts.

The first part deals with the analysis of the faster timescale recursion, where we show that the faster timescale iterates $\{X_n\}_{n\geq 1}$ converge almost surely to $\lambda(y)$ (as in Assumption A9) for some $y \in \mathbb{R}^{d_2}$.

The second part deals with the slower timescale recursion analysis, where we show that the slower timescale iterates $\{Y_n\}_{n\geq 1}$ track the flow of DI (20).

Throughout this section, we assume that Assumptions A1–A9 are satisfied.

5.1. Faster Timescale Recursion Analysis

For every $\omega \in \Omega$, for every $n \geq 0$, the two-timescale recursion (15a and 15b) can be written as

$$Y_{n+1}(\omega) - Y_n(\omega) - b(n)M_{n+1}^{(2)}(\omega) = b(n)V_n^2(\omega), \tag{21a}$$

$$X_{n+1}(\omega) - X_n(\omega) - a(n)M_{n+1}^{(1)}(\omega) = a(n)V_n^1(\omega), \tag{21b}$$

where for every $n \geq 0$, V_n^1 and V_n^2 are such that for every $\omega \in \Omega$,

$$V_n^1(\omega) \in H_1\Big(X_n(\omega), Y_n(\omega), S_n^{(1)}(\omega)\Big),$$

$$V_n^2(\omega) \in H_2\Big(X_n(\omega), Y_n(\omega), S_n^{(2)}(\omega)\Big).$$

Recursion (21) can be rewritten as

$$Y_{n+1}(\omega) - Y_n(\omega) = a(n)\left(\frac{b(n)}{a(n)}V_n^2(\omega) + \frac{b(n)}{a(n)}M_{n+1}^{(2)}(\omega)\right),$$

$$X_{n+1}(\omega) - X_n(\omega) = a(n)\Big(V_n^1(\omega) + M_{n+1}^{(1)}(\omega)\Big)$$

for every $\omega \in \Omega$ and for every $n \geq 0$. The above can be now written in the form of the single timescale recursion (i.e., (13)):

$$Z_{n+1} - Z_n - a(n)M_{n+1} \in a(n)F\Big(Z_n, S_n^{(1)}\Big), \tag{23}$$

where

1. for every $n \geq 0$, $Z_n = (X_n, Y_n)$,
2. for $n \geq 0$, $M_{n+1} = (M_{n+1}^{(1)}, \frac{b(n)}{a(n)}(V_n^2 + M_{n+1}^{(2)}))$,
3. $F : \mathbb{R}^d \times \mathcal{S}^{(1)} \to \{\text{subsets of } \mathbb{R}^d\}$ such that for every $(x,y,s^{(1)}) \in \mathbb{R}^d \times \mathcal{S}^{(1)}$, $F(x,y,s^{(1)}) = (H_1(x,y,s^{(1)}), 0)$.

We now show that the quantities defined above satisfy the assumptions associated with the single timescale recursion as in Section 2.4. Clearly, by Assumption A5, the step-size sequence $\{a(n)\}_{n\geq 0}$ satisfies Assumption S(A3), and by Assumption A3, the Markov noise terms, $\{S_n^{(1)}\}_{n\geq 0}$ satisfy Assumption S(A2). As a consequence of the stability Assumption A8, we have that $\mathbb{P}(\sup_{n\geq 0} \|Z_n := (X_n, Y_n)\| < \infty) = 1$, and hence Assumption S(A5) is satisfied.

Consider the set-valued map F defined above. Clearly, by Assumption A1i, for every $(x,y,s^{(1)}) \in \mathbb{R}^d \times \mathcal{S}^{(1)}$, $F(x,y,s^{(1)})$ is a nonempty, convex, and compact subset of $\mathbb{R}^d$. Further, by Assumption A1ii, we have that for every $(x,y,s^{(1)}) \in \mathbb{R}^d \times \mathcal{S}^{(1)}$, $\sup_{(x',y') \in F(x,y,s^{(1)})} \|(x',y')\| = \sup_{x' \in H_1(x,y,s^{(1)})} \|x'\| \leq K(1 + \|x\| + \|y\|) \leq \max\{K, KC\}(1 + \|(x,y)\|)$, where $C > 0$ is such that $\|x\| + \|y\| \leq C\|(x,y)\|$ for every $(x,y) \in \mathbb{R}^d$ (see Kumaresan [24, theorem 4.3.26]).

By Assumption A1iii, the map H_1 has a closed graph, and hence the map F also has a closed graph. Therefore, the set-valued map satisfies Assumption S(A1).

Recall that for every $T > 0$, for every $n \geq 0$, $\tau^1(n,T) := \min\{m > n : \sum_{k=n}^{m-1} a(k) \geq T\}$. Let

$$\Omega_1 := \{\omega \in \Omega : \text{Assumptions A6, A7, and A8 hold}\}. \tag{24}$$

Clearly, $\mathbb{P}(\Omega_1) = 1$. Let $\omega \in \Omega_1$ and fix $T > 0$. For any $n \geq 0$,

$$\sup_{n \leq k \leq \tau^1(n,T)} \left\| \sum_{m=n}^{k} a(m) M_{m+1}(\omega) \right\| \leq \sup_{n \leq k \leq \tau^1(n,T)} \left\| \sum_{m=n}^{k} a(m) M_{m+1}^{(1)}(\omega) \right\| + \sup_{n \leq k \leq \tau^1(n,T)} \left\| \sum_{m=n}^{k} a(m) \frac{b(m)}{a(m)} \left(V_n^2(\omega) + M_{m+1}^{(2)}(\omega) \right) \right\|.$$

By Assumption A8, for every $\omega \in \Omega_1$, there exists $r > 0$ such that $\sup_{n \geq 0}(\|X_n(\omega)\| + \|Y_n(\omega)\|) \leq r$. Because for every $n \geq 0$, for every $\omega \in \Omega_1$, $V_n^2(\omega) \in H_2(X_n(\omega), Y_n(\omega), S_n^{(2)}(\omega))$, and by Assumption A2ii, we have that $\|V_n^2(\omega)\| \leq K(1 + \|X_n(\omega)\| + \|Y_n(\omega)\|) \leq K(1 + r) =: R < \infty$. Further, by Assumption A5iii, for every $0 < \epsilon < T$, there exists N such that for every $n \geq N$, $b(n) \leq \frac{\epsilon}{T+1} a(n)$. Therefore, for every $n \geq N$, for every $m > n$, $\sum_{k=n}^{m-1} b(k) \leq \frac{\epsilon}{T+1} \sum_{k=n}^{m-1} a(k)$. Thus, for every $n \geq N$, $\sum_{k=n}^{\tau^1(n,T)-1} b(k) \leq \epsilon$ and $\tau^1(n,T) \leq \tau^2(n,T)$. Therefore, for every $0 < \epsilon < T$, for every $n \geq N$,

$$\sup_{n \leq k \leq \tau^1(n,T)} \left\| \sum_{m=n}^{k} a(m) M_{m+1}(\omega) \right\| \leq \sup_{n \leq k \leq \tau^1(n,T)} \left\| \sum_{m=n}^{k} a(m) M_{m+1}^{(1)}(\omega) \right\| + \epsilon R + \sup_{n \leq k \leq \tau^2(n,T)} \left\| \sum_{m=n}^{k} b(m) M_{m+1}^{(2)}(\omega) \right\|.$$

Taking the limit in the above equation and using Assumptions A6 and A7 give us that for every $0 < \epsilon < T$,

$$\lim_{n \to \infty} \sup_{n \leq k \leq \tau^1(n,T)} \left\| \sum_{m=n}^{k} a(m) M_{m+1}(\omega) \right\| \leq R\epsilon.$$

Therefore, for every $\omega \in \Omega_1$, for every $T > 0$, $\lim_{n \to \infty} \sup_{n \leq k \leq \tau^1(n,T)} \|\sum_{m=n}^{k} a(m) M_{m+1}(\omega)\| = 0$. Thus the additive noise terms $\{M_n\}_{n \geq 1}$ satisfy Assumption SA4.

Therefore, quantities in Recursion (23) satisfy Assumptions SA1–SA5, and we apply the main result of the single timescale recursion (see Theorem 2iii) to conclude the following.

Lemma 17. *Under Assumptions A1–A8, for almost every ω, there exists a nonempty compact set $L \subseteq \mathbb{R}^d$ (depending on ω) such that*

i. *$(X_n(\omega), Y_n(\omega)) \to L$ as $n \to \infty$, where $\{(X_n, Y_n)\}_{n \geq 0}$ is as in Recursion (15).*

ii. *the set L is internally chain transitive for the flow of the DI,*

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} \in \begin{pmatrix} \hat{H}_1(x,y) \\ 0 \end{pmatrix}. \tag{25}$$

The set-valued map associated with the DI (25) is clearly a Marchaud map (use Lemma 16). Further, any solution $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ of DI (25) is such that for every $t \in \mathbb{R}$, $\mathbf{y}(t) = \mathbf{y}(0)$ and $\mathbf{x}(\cdot)$ is a solution to DI (17) with $y_0 = \mathbf{y}(0)$.

Fix $\omega \in \Omega$ such that Lemma 17 holds. Let $L \subseteq \mathbb{R}^d$ be as in Lemma 17. Let

$$A := \{(x,y) \in \mathbb{R}^d : x \in \lambda(y),\ y \in \mathbb{R}^{d_2}\}, \tag{26}$$

where for every $y \in \mathbb{R}^{d_2}$, $\lambda(y)$ is as in Assumption A9. Because L is internally chain transitive for the flow of DI (25), by Benaïm et al. [4, lemma 3.5], we know that it is invariant. Let $(x^*, y^*) \in L$ and $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ be a solution to DI (25) with initial condition (x^*, y^*) and for all $t \in \mathbb{R}$, $(\mathbf{x}(t), \mathbf{y}(t)) \in L$. Then, for every $t \in \mathbb{R}$, $\mathbf{y}(t) = y^*$ and $\mathbf{x}(\cdot)$ is a solution of DI (17) with $y_0 = y^*$. By Assumption (A9), there exists a compact subset $\lambda(y^*) \subseteq \mathbb{R}^{d_1}$, which is a globally attracting set for the flow of DI (17) with $y_0 = y^*$. By definition of a globally attracting set, we have that $\cap_{t \geq 0} \overline{\{\mathbf{x}(q+t) : q \geq 0\}} \subseteq \lambda(y^*)$. Therefore, $(\mathbf{x}(t), \mathbf{y}(t)) \to \lambda(y^*) \times \{y^*\} \subseteq A$, and because for every $t \in \mathbb{R}$, $(\mathbf{x}(t), \mathbf{y}(t)) \in L$, we get that $L \cap A \neq \emptyset$. In fact, for any closed set $C \subseteq \mathbb{R}^d$ invariant for the flow of DI (25), the above argument gives us that $C \cap A \neq \emptyset$. If we are able to show that $L \cap A = L$, then, by Lemma 17i, we obtain that $(X_n(\omega), Y_n(\omega)) \to L \subseteq A$ as $n \to \infty$. In this regard, we need to impose the following assumption.

Assumption A11. For any compact set $C \subseteq \mathbb{R}^d$, invariant for the flow of DI (25), for any open neighborhood $\mathcal{O}$ of $C \cap A$, there exists an open neighborhood $\mathcal{O}'$ of $C \cap A$ such that

$$\Phi^C(\mathcal{O}' \cap C, [0,\infty)) \subseteq \mathcal{O} \cap C,$$

where $\Phi^C : C \times \mathbb{R} \to \{\text{subsets of } \mathbb{R}^d\}$ denotes the flow of DI (25) restricted to the invariant set C (see Section 2.3 for definition).

Remark 4. The above assumption is a weaker form of Assumption A1 imposed by Borkar [11, chapter 6] (that implies Assumption A11 above). The above assumption is basically the Lyapunov stability condition (see Benaïm et al. [4, definition IX(ii)]) for the flow restricted to the invariant set C. We will see that in the application studied later, the above assumption is satisfied.

Lemma 18. *Under Assumptions A1–A9 and A11, for almost every ω, $L \subseteq \mathbb{R}^d$ as in Lemma 17 is such that*

$$L \subseteq \left\{(x,y) \in \mathbb{R}^d : x \in \lambda(y),\ y \in \mathbb{R}^{d_2}\right\}.$$

Therefore, $(X_n(\omega), Y_n(\omega)) \to L \subseteq \{(x,y) \in \mathbb{R}^d : x \in \lambda(y),\ y \in \mathbb{R}^{d_2}\}$ as $n \to \infty$, where $\{(X_n, Y_n)\}_{n\geq 0}$ is as in Recursion (15).

Proof. We present a brief outline here to highlight where Assumption A11 is used.

Let $A \subseteq \mathbb{R}^d$ be as in Equation (26). Fix $\omega \in \Omega$ and obtain L as in Lemma 17. We know that L is internally chain transitive for the flow of DI (25), and because it is also invariant, $L \cap A \neq \emptyset$. By Assumption A9, for every $(x^*, y^*) \in L$, $\omega_{\Phi^L}((x^*, y^*)) \subseteq L \cap A$. Thus, for every $(x^*, y^*) \in L$, for every solution of DI (25), $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ such that $(\mathbf{x}(0), \mathbf{y}(0)) = (x^*, y^*)$, and for every $t \in \mathbb{R}$, $(\mathbf{x}(t), \mathbf{y}(t)) \in L$, for every open neighborhood $\mathcal{O}$ of $L \cap A$, there exists $t > 0$ such that $(\mathbf{x}(t), \mathbf{y}(t)) \in \mathcal{O} \cap L$. By Benaïm et al. [4, lemma 3.13], we get that for every open neighborhood $\mathcal{O}$ of $L \cap A$, there exists $T > 0$, for every $(x^*, y^*) \in L$, for every solution of DI (25), $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ such that $(\mathbf{x}(0), \mathbf{y}(0)) = (x^*, y^*)$ and for every $t \in \mathbb{R}$, $(\mathbf{x}(t), \mathbf{y}(t)) \in L$, for some $t \in [0, T]$, $(\mathbf{x}(t), \mathbf{y}(t)) \in \mathcal{O} \cap L$.

Fix $\epsilon > 0$. Then, by Assumption A11, there exists an open neighborhood of $L \cap A$, $\mathcal{O}$ such that $\Phi^L(\mathcal{O} \cap L, [0,\infty)) \subseteq N^\epsilon(L \cap A) \cap L$. From arguments in the previous paragraph, we can find $T > 0$ such that for every $(x^*, y^*) \in L$, for every solution of DI (25) with $(\mathbf{x}(0), \mathbf{y}(0)) = (x^*, y^*)$ and $(\mathbf{x}(t), \mathbf{y}(t)) \in L$ for every $t \in \mathbb{R}$, there exists $t \in [0, T]$ such that $(\mathbf{x}(t), \mathbf{y}(t)) \in \mathcal{O} \cap L$. Therefore, $\Phi^L(L, [T, \infty)) \subseteq N^\epsilon(L \cap A) \cap L$. Thus, $L \cap A$ is an attracting set for Φ^L. Now the claim follows from Benaïm et al. [4, proposition 3.20]. □

5.2. Slower Timescale Recursion Analysis

Before we present the analysis of slower timescale recursion, we present some preliminaries where we will define various quantities needed later. Throughout this section, let $\{H_2^{(l)}\}_{l\geq 1}$ and $\{h_2^{(l)}\}_{l\geq 1}$ denote maps as in Section 4. Further, we will allow Assumptions A1–A9 and A11 to be satisfied. The slower timescale recursion analysis is similar to the analysis of single timescale inclusion by Yaji and Bhatnagar [41] with minor modifications arising because of the presence of faster timescale iterates. Throughout this section, U denotes the closed unit ball in $\mathbb{R}^{d_2}$, and B_r denotes the closed ball of radius $r > 0$ in $\mathbb{R}^{d_1}$ centered on the origin.

5.2.1. Preliminaries. Let $t^s(0) := 0$ and for every $n \geq 1$, $t^s(n) := \sum_{m=0}^{n-1} b(m)$. Define $\bar{Y} : \Omega \times [0,\infty) \to \mathbb{R}^{d_2}$ such that for every $(\omega, t) \in \Omega \times [0, \infty)$,

$$\bar{Y}(\omega, t) := \left(\frac{t - t^s(n)}{t^s(n+1) - t^s(n)}\right) Y_{n+1}(\omega) + \left(\frac{t^s(n+1) - t}{t^s(n+1) - t^s(n)}\right) Y_n(\omega),$$

where n is such that $t \in [t^s(n), t^s(n+1))$.

Consider the slower timescale recursion (15a) given by

$$Y_{n+1} - Y_n - b(n) M_{n+1}^{(2)} \in b(n) H_2\left(X_n, Y_n, S_n^{(2)}\right)$$

for every $n \geq 0$. By Lemma 2, we have that for every $l \geq 1$, for every $n \geq 0$, $H_2(X_n, Y_n, S_n^{(2)}) \subseteq H_2^{(l)}(X_n, Y_n, S_n^{(2)})$. Therefore, for every $l \geq 1$, the following recursion follows from the above (i.e., (15a)):

$$Y_{n+1} - Y_n - b(n) M_{n+1}^{(2)} \in b(n) H_2^{(l)}\left(X_n, Y_n, S_n^{(2)}\right).$$

By Lemma 3, we know that for every $l \geq 1$, the set-valued map $H_2^{(l)}$ admits a continuous single-valued parametrization $h_2^{(l)}$. The next lemma allows us to write the slower timescale inclusion in terms of the parametrization of $H_2^{(l)}$, and the result follows from Yaji and Bhatnagar [41, lemma 6.1].

Lemma 19. *For every $l \geq 1$, for every $n \geq 0$, there exists a U-valued random variable on Ω, $U_n^{(l)}$, such that for every $\omega \in \Omega$,*

$$Y_{n+1}(\omega) - Y_n(\omega) - b(n)M_{n+1}^{(2)}(\omega) = b(n)h_2^{(l)}\left(X_n(\omega), Y_n(\omega), S_n^{(2)}(\omega), U_n^{(l)}(\omega)\right).$$

For every $l \geq 1$, define $\Gamma^{(l)} : \Omega \times [0,\infty) \to \mathcal{P}(U \times \mathbb{R}^{d_1} \times \mathcal{S}^{(2)})$ such that for every $(\omega, t) \in \Omega \times [0,\infty)$,

$$\Gamma^{(l)}(\omega, t) := \delta_{U_n^{(l)}(\omega)} \otimes \delta_{X_n(\omega)} \otimes \delta_{S_n^{(2)}(\omega)}, \tag{27}$$

where $\delta_{U_n^{(l)}(\omega)} \in \mathcal{P}(U)$ denotes the Dirac measure at $U_n^{(l)}(\omega) \in U$ (for every $A \in \mathcal{B}(U)$, $\delta_{U_n^{(l)}(\omega)}(A) = 1$ if $U_n^{(l)}(\omega) \in A$, 0 otherwise), $\delta_{X_n(\omega)} \in \mathcal{P}(\mathbb{R}^{d_1})$ denotes the Dirac measure at $X_n(\omega) \in \mathbb{R}^{d_1}$, and similarly, $\delta_{S_n^{(2)}(\omega)} \in \mathcal{P}(\mathcal{S}^{(2)})$ denotes the Dirac measure at $S_n^{(2)}(\omega) \in \mathcal{S}^{(2)}$.

The lemma below provides an equicontinuity result used later.

Lemma 20. *For every $l \geq 1$, for every $r > 0$, the family of maps*

$$\left\{ y \in rU \to \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\left(x, y, s^{(2)}, u\right) \nu\left(du, dx, ds^{(2)}\right) : \nu \in \mathcal{P}\left(U \times B_r \times \mathcal{S}^{(2)}\right) \right\}$$

is equicontinuous.

Proof. Fix $l \geq 1$. By Lemma 3, we know that the map $h_2^{(l)}(\cdot)$ is continuous. Hence, the map $h_2^{(l)}(\cdot)$ restricted to the compact set $B_r \times rU \times \mathcal{S}^{(2)} \times U$ is uniformly continuous. Therefore, for every $\epsilon > 0$, there exists $\delta > 0$ such that for every $(x,y,s^{(2)},u), (x,y',s^{(2)},u) \in B_r \times rU \times \mathcal{S}^{(2)} \times U$, satisfying $\|y - y'\| < \delta$, $\|h_2^{(l)}(x,y,s^{(2)},u) - h_2^{(l)}(x,y',s^{(2)},u\| < \epsilon$. Therefore, for $\delta > 0$ as above, with $\|y - y'\| < \delta$, for any $\nu \in \mathcal{P}(U \times B_r \times \mathcal{S}^{(2)})$, $\| \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}(x,y,s^{(2)},u) \cdot \nu(du,dx,ds^{(2)}) - \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}(x,y',s^{(2)},u)\nu(du,dx,ds^{(2)})\| \leq \int_{U \times B_r \times \mathcal{S}^{(2)}} \|h_2^{(l)}(x,y,s^{(2)},u) - h_2^{(l)}(x,y',s^{(2)},u)\| \nu(du,dx,ds^{(2)}) \leq \epsilon$. □

For every $l \geq 1$, define $G^{(l)} : \Omega \times [0,\infty) \to \mathbb{R}^{d_2}$ such that for every $(\omega, t) \in \Omega \times [0,\infty)$,

$$G^{(l)}(\omega, t) := h_2^{(l)}\left(X_n(\omega), Y_n(\omega), S_n^{(2)}(\omega), U_n^{(l)}(\omega)\right), \tag{28}$$

where n is such that $t \in [t^s(n), t^s(n+1))$.

In what follows, most of the arguments are sample pathwise, and we use smaller case symbols to denote the above-defined quantities along a particular sample path. For example, x_n, y_n, $u_n^{(l)}$, $m_{n+1}^{(2)}$, $s_n^{(2)}$, $\bar{y}(t)$, $\gamma_n^{(l)}(t)$, and $g^{(l)}(t)$ denote $X_n(\omega)$, $Y_n(\omega)$, $U_n^{(l)}(\omega)$, $M_{n+1}^{(2)}(\omega)$, $S_n^{(2)}(\omega)$,$\bar{Y}(\omega, t)$, $\Gamma_n^{(l)}(\omega, t)$, and $G^{(l)}(\omega, t)$, respectively, for some ω fixed.

5.2.2. Main Result: Asymptotic Pseudotrajectory. For every $\omega \in \Omega$, for every $l \geq 1$, and for every $\tilde{t} \geq 0$, let $\tilde{y}^{(l)}(\cdot; \tilde{t})$ denote the solution of the o.d.e.

$$\dot{\tilde{y}}^{(l)}(t; \tilde{t}) = g^{(l)}\left(t + \tilde{t}\right) \tag{29}$$

for every $t \geq 0$ with initial condition $\tilde{y}^{(l)}(0; \tilde{t}) = \bar{y}(\tilde{t})$.

Let Ω_1 be as in (24). Then by Assumptions A6–A8, we have that $\mathbb{P}(\Omega_1) = 1$. First, we will get rid of the additive noise terms. In this regard, we prove the lemma below that states that for every $\omega \in \Omega_1$, the family of functions $\{\bar{y}(\cdot + t)\}_{t \geq 0}$ and $\{\tilde{y}^{(l)}(\cdot; t)\}_{t \geq 0}$ have the same limit points in $\mathcal{C}([0,\infty), \mathbb{R}^{d_2})$ for every $l \geq 1$. The proof of the lemma below is similar to that of Yaji and Bhatnagar [41, lemma 6.3] and is given in Appendix A.

Lemma 21. *For almost every ω, for every $l \geq 1$, and for every $T > 0$,*

$$\lim_{t \to \infty} \sup_{0 \leq q \leq T} \left\| \bar{y}(q + t) - \tilde{y}^{(l)}(q; t) \right\| = 0.$$

The lemma below guarantees the existence of limit points for $\{\tilde{y}^{(l)}(\cdot; t)\}_{t \geq 0}$ in $\mathcal{C}([0,\infty), \mathbb{R}^{d_2})$. The proof is similar to that of Yaji and Bhatnagar [41, lemma 6.4] and is given in Appendix B.

Lemma 22. *For almost every ω, for every $l \geq 1$, the family of functions $\{\tilde{y}^{(l)}(\cdot; t)\}_{t \geq 0}$ is relatively compact in $\mathcal{C}([0,\infty), \mathbb{R}^{d_2})$.*

As a consequence of Lemmas 21 and 22, we get that for almost every ω,

i. the family of functions $\{\bar{y}(\cdot + t)\}_{t \geq 0}$ is relatively compact in $\mathscr{C}([0,\infty), \mathbb{R}^d)$, and

ii. the linearly interpolated trajectory of the slower timescale iterates, $\bar{y}(\cdot)$, is uniformly continuous on $[0,\infty)$.

The next proposition states that every limit point of $\{\bar{y}(\cdot + t)\}_{t \geq 0}$ is a solution of DI (20) on $[0,\infty)$. The proof is along the lines of that of Yaji and Bhatnagar [41, proposition 6.5], but with modifications arising because of the presence of faster timescale iterates.

Proposition 1. *For almost every ω, every limit point $y^*(\cdot)$ of $\{\bar{y}(\cdot + t)\}_{t \geq 0}$ in $\mathscr{C}([0,\infty), \mathbb{R}^{d_2})$ satisfies the following.*

i. *For some $r > 0$, for every $l \geq 1$, there exists $\tilde{\gamma}^{(l)} \in \mathcal{M}(U \times B_r \times \mathcal{S}^{(2)})$ such that for every $t \geq 0$,*

$$y^*(t) = y^*(0) + \int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\left(x, y^*(q), s^{(2)}, u\right) \tilde{\gamma}^{(l)}(q)\left(du, dx, ds^{(2)}\right) \right] dq.$$

ii. *For every $l \geq 1$, $\tilde{\gamma}^{(l)}$ as in part* i *of this proposition is such that for almost every $t \geq 0$,*

$$\Theta_1\left(\tilde{\gamma}^{(l)}\right)(t) \in D\left(y^*(t)\right).$$

iii. *$y^*(\cdot)$ is absolutely continuous and for almost every $t \in [0,\infty)$,*

$$\frac{dy^*(t)}{dt} \in \hat{H}_2\left(y^*(t)\right).$$

Proof. Let $\Omega_2 := \{\omega \in \Omega$:Lemma 18 holds$\}$. From the proof of Yaji and Bhatnagar [41, theorem 6.6], it is clear that $\Omega_1 \subseteq \Omega_2$ and $\mathbb{P}(\Omega_2) = 1$. Fix $\omega \in \Omega_2$ and let $t_n \to \infty$ be such that $\bar{y}(\cdot + t_n) \to y^*(\cdot)$ in $\mathscr{C}([0,\infty), \mathbb{R}^{d_2})$.

i. Fix $l \geq 1$. By Assumption 8, there exists $r > 0$ such that $\sup_{n \geq 0}(\|x_n\| + \|y_n\|) \leq r$. Then, for any $t \geq 0$, $\gamma^{(l)}(t) := \delta_{u_n^{(l)}} \otimes \delta_{x_n} \otimes \delta_{s_n^{(2)}} \in \mathcal{P}(U \times B_r \times \mathcal{S}^{(2)})$, where n is such that $t \in [t^s(n), t^s(n+1))$. Therefore, $\gamma^{(l)} \in \mathcal{M}(U \times B_r \times \mathcal{S}^{(2)})$, and by Lemma 9i, we get that the sequence $\{\gamma^{(l)}(\cdot + t_n)\}_{n \geq 1}$ has a convergent subsequence in $\mathcal{M}(U \times B_r \times \mathcal{S}^{(2)})$. Let $\tilde{\gamma}^{(l)}$ be a limit point of $\{\gamma^{(l)}(\cdot + t_n)\}_{n \geq 1}$ and, without loss of generality, assume $\gamma^{(l)}(\cdot + t_n) \to \tilde{\gamma}^{(l)}$ as $n \to \infty$. By definition of $\tilde{y}^{(l)}(\cdot; t_n)$, we have that for every $n \geq 1$, for every $t \geq 0$,

$$\tilde{y}^{(l)}(t; t_n) = \bar{y}(t_n) + \int_0^t g^{(l)}(q + t_n) dq = \bar{y}(t_n) + \int_0^t h_2^{(l)}\left(x_{[t_n+q]}, y_{[t_n+q]}, s^{(2)}_{[t_n+q]}, u^{(l)}_{[t_n+q]}\right) dq,$$

where for any $t \geq 0$, $[t] := \max\{n \geq 0 : t \geq t^s(n)\}$. Using the definition of $\gamma^{(l)}(\cdot + t_n)$ (see (27) and recall that $\gamma^{(l)}(\cdot + t_n) := \Gamma^{(l)}(\omega, \cdot + t_n)$) in the above, we get that for every $n \geq 1$, for every $t \geq 0$,

$$\tilde{y}^{(l)}(t; t_n) = \bar{y}(t_n) + \int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\left(x, y_{[t_n+q]}, s^{(2)}, u\right) \gamma^{(l)}(q + t_n)\left(du, dx, ds^{(2)}\right) \right] dq.$$

Because $\bar{y}(\cdot + t_n) \to y^*(\cdot)$ in $\mathscr{C}([0,\infty), \mathbb{R}^{d_2})$, by Lemma 21, we have that $\tilde{y}^{(l)}(\cdot; t_n) \to y^*(\cdot)$ as $n \to \infty$. By taking the limit in the above equation, we get that for every $t \geq 0$,

$$\begin{aligned} \lim_{n \to \infty}\left[\tilde{y}^{(l)}(t; t_n) - \bar{y}(t_n)\right] &= \lim_{n \to \infty} \int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\left(x, y_{[t_n+q]}, s^{(2)}, u\right) \gamma^{(l)}(q + t_n)\left(du, dx, ds^{(2)}\right) \right] dq, \\ y^*(t) - y^*(0) &= \lim_{n \to \infty} \int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\left(x, y_{[t_n+q]}, s^{(2)}, u\right) \gamma^{(l)}(q + t_n)\left(du, dx, ds^{(2)}\right) \right] dq. \end{aligned} \tag{30}$$

Because $\gamma^{(l)}(\cdot + t_n) \to \tilde{\gamma}^{(l)}(\cdot)$ and by our choice of the topology for $\mathcal{M}(U \times B_r \times \mathcal{S}^{(2)})$, we have

$$\int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} \tilde{f}\left(q, u, x, s^{(2)}\right) \gamma^{(l)}(q + t_n)\left(du, dx, ds^{(2)}\right) \right] dq - \int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} \tilde{f}\left(q, u, x, s^{(2)}\right) \tilde{\gamma}^{(l)}(q)\left(du, dx, ds^{(2)}\right) \right] dq \to 0$$

for all bounded continuous $\tilde{f} : [0,t] \times U \times B_r \times \mathcal{S}^{(2)} \to \mathbb{R}$ of the form

$$\tilde{f}\left(q, u, x, s^{(2)}\right) = \sum_{m=1}^{N} a_m g_m(q) f_m\left(u, x, s^{(2)}\right)$$

for some $N \geq 1$, scalars a_m, and bounded continuous functions g_m, f_m on $[0,t]$, $U \times B_r \times \mathcal{S}^{(2)}$, respectively, for $1 \leq m \leq N$. By the Stone-Weierstrass theorem, such functions can uniformly approximate any function in $\mathscr{C}([0,t] \times U \times B_r \times \mathcal{S}^{(2)}, \mathbb{R})$. Thus, the above convergence holds true for all real valued continuous functions on $[0,t] \times U \times B_r \times \mathcal{S}^{(2)}$, implying that $t^{-1}\gamma^{(l)}(q+t_n)(du, dx, ds^{(2)})dq \to t^{-1}\tilde{\gamma}^{(l)}(q)(du, dx, ds^{(2)})dq$ in $\mathcal{P}([0,t] \times U \times B_r \times \mathcal{S}^{(2)})$. Thus,

$$\begin{aligned}\Bigg\| \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \\ - \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \tilde{\gamma}^{(l)}(q)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \Bigg\| \to 0\end{aligned} \tag{31}$$

as $n \to \infty$. Because $\{\bar{y}(\cdot + t_n)|_{[0,t]}\}_{n \geq 1}$ uniformly converges to $y^*(\cdot)|_{[0,t]}$, we have that the function $q \to y_{[t_n+q]}$ uniformly converges to $y^*(\cdot)|_{[0,t]}$ on $[0,t]$. Using the above and Lemma 20, we have that for every $\epsilon > 0$, there exists N(depending on ϵ) such that for every $n \geq N$, for every $q \in [0,t]$,

$$\begin{aligned}\Bigg\| \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y_{[t_n+q]}, s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big) \\ - \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big)\Bigg\| < \epsilon.\end{aligned} \tag{32}$$

Now,

$$\begin{aligned}&\Bigg\| \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y_{[t_n+q]}, s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \\ &\qquad - \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \tilde{\gamma}^{(l)}(q)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \Bigg\| \\ &\leq \Bigg\| \int_0^t \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y_{[t_n+q]}, s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big) dq \\ &\qquad - \int_0^t \int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big) dq \Bigg\| \\ &+ \Bigg\| \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \\ &\qquad - \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \tilde{\gamma}^{(l)}(q)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \Bigg\|.\end{aligned}$$

Taking the limit as $n \to \infty$ in the above equation and using (31) and (32), we get

$$\begin{aligned}\lim_{n \to \infty} \Bigg\| \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y_{[t_n+q]}, s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \\ - \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \tilde{\gamma}^{(l)}(q)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \Bigg\| \leq \epsilon t\end{aligned}$$

for every $\epsilon > 0$. Therefore, for every $t \geq 0$,

$$\begin{aligned}\lim_{n \to \infty} \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y_{[t_n+q]}, s^{(2)}, u\Big) \gamma^{(l)}(q+t_n)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq \\ = \int_0^t \Bigg[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\Big(x, y^*(q), s^{(2)}, u\Big) \tilde{\gamma}^{(l)}(q)\Big(du, dx, ds^{(2)}\Big)\Bigg] dq.\end{aligned}$$

Substituting the above limit in Equation (30), we get that for every $t \geq 0$,

$$y^*(t) - y^*(0) = \int_0^t \left[\int_{U \times B_r \times \mathcal{S}^{(2)}} h_2^{(l)}\left(x, y^*(q), s^{(2)}, u\right) \tilde{\gamma}^{(l)}(q)\left(du, dx, ds^{(2)}\right) \right] dq.$$

ii. Fix $l \geq 1$. Let $\mu^{(l)} := \Theta_1(\gamma^{(l)})$. Then, for every $n \geq 1$, $\mu^{(l)}(\cdot + t_n) = \Theta_1(\gamma^{(l)}(\cdot + t_n))$ and because $\gamma^{(l)}(\cdot + t_n) \longrightarrow \tilde{\gamma}^{(l)}$ in $\mathcal{M}(U \times B_r \times \mathcal{S}^{(2)})$, by Lemma 10v, we get that $\mu^{(l)}(\cdot + t_n) \longrightarrow \tilde{\mu}^{(l)}(\cdot) =: \Theta_1(\tilde{\gamma}^{(l)})$ in $\mathcal{M}(B_r \times \mathcal{S}^{(2)})$ as $n \longrightarrow \infty$. In order to prove that for almost every $t \geq 0$, $\tilde{\mu}^{(l)}(t) \in D(y^*(t))$, we need to show that for almost every $t \geq 0$,

1. $\text{supp}(\tilde{\mu}^{(l)}_{B_r}(t)) \subseteq \lambda(y^*(t))$, and
2. for every $A \in \mathcal{B}(\mathcal{S}^{(2)})$, $\tilde{\mu}^{(l)}_{\mathcal{S}^{(2)}}(t)(A) = \int_{\mathbb{R}^{d_1} \times \mathcal{S}^{(2)}} \Pi^{(2)}(x, y^*(t), s^{(2)})(A) \tilde{\mu}^{(l)}(t)(dx, ds^{(2)})$.

First, we present a proof of the claim in part 1 above. $\mu^{(l)}_{B_r} := \Theta_2(\mu^{(l)}) \in \mathcal{M}(B_r)$ and because as $n \longrightarrow \infty$, $\mu^{(l)}(\cdot + t_n) \longrightarrow \tilde{\mu}^{(l)}$, by Lemma 10vi, we have that $\mu^{(l)}_{B_r}(\cdot + t_n) \longrightarrow \tilde{\mu}^{(l)}_{B_r}$ in $\mathcal{M}(B_r)$. Because as $n \longrightarrow \infty$, $\mu^{(l)}_{B_r}(\cdot + t_n) \longrightarrow \tilde{\mu}^{(l)}_{B_r}$ in $\mathcal{M}(B_r)$, by Borkar [11, chapter 6, lemma 3], we have that for almost every $t \geq 0$, there exists a subsequence $\{n_k\}_{k \geq 1}$ and a subsequence of natural numbers $\{c_p\}_{p \geq 1}$ such that

$$\frac{1}{c_p} \sum_{k=1}^{c_p} \mu^{(l)}_{B_r}(t + t_{n_k}) \longrightarrow \tilde{\mu}^{(l)}_{B_r}(t) \tag{33}$$

in $\mathcal{P}(B_r)$ as $p \longrightarrow \infty$. Fix $t \geq 0$ such that the above holds. By the definition of $\mu^{(l)}_{B_r}$, we have that for every $k \geq 1$, $\mu^{(l)}_{B_r}(t + t_{n_k}) = \delta_{x_{[t+t_{n_k}]}}$, where $[t + t_{n_k}] := \max\{m > n : t + t_{n_k} \geq t^s(m)\}$. Because $\bar{y}(\cdot + t_{n_k}) \longrightarrow y^*(t)$, using the uniform continuity of $\bar{y}(\cdot)$, we have that the function $\tilde{t} \in [0, \infty) \longrightarrow y_{[\tilde{t}+t_{n_k}]}$ uniformly converges on compacts to the function $y^*(\cdot)$. Therefore, $y_{[t+t_{n_k}]} \longrightarrow y^*(t)$ as $k \longrightarrow \infty$, and by Lemma 18, we have that $x_{[t+t_{n_k}]} \longrightarrow \lambda(y^*(t))$. Further, by the definition of $r > 0$, as in part i of this proposition, we get that $\sup_{n \geq 1} \|x_n\| \leq r$. Hence, $\lambda(y^*(t)) \cap B_r \neq \emptyset$ and $x_{[t+t_{n_k}]} \longrightarrow \lambda(y^*(t)) \cap B_r$ as $k \longrightarrow \infty$. For every $\epsilon > 0$, clearly $(\lambda(y^*(t)) + B_\epsilon) \cap B_r$ is compact and there exists K large such that for every $k \geq K$, $x_{[t+t_{n_k}]} \in (\lambda(y^*(t)) + B_\epsilon) \cap B_r$, and hence $\delta_{x_{[t+t_{n_k}]}}((\lambda(y^*(t)) + B_\epsilon) \cap B_r) = 1$ for every $k \geq K$. Because $\frac{1}{c_p} \sum_{k=1}^{c_p} \mu^{(l)}_{B_r}(t + t_{n_k}) \longrightarrow \tilde{\mu}^{(l)}_{B_r}(t)$ in $\mathcal{P}(B_r)$ as $p \longrightarrow \infty$, by Borkar [12, theorem 2.1.1(iv)] we have that, for every $\epsilon > 0$,

$$\begin{aligned}
\limsup_{p \to \infty} \frac{1}{c_p} \sum_{k=1}^{c_p} \mu^{(l)}_{B_r}(t + t_{n_k})\left((\lambda(y^*(t)) + B_\epsilon) \cap B_r\right) &= \limsup_{p \to \infty} \frac{1}{c_p} \sum_{k=1}^{c_p} \delta_{x_{[t+t_{n_k}]}}\left((\lambda(y^*(t)) + B_\epsilon) \cap B_r\right) \\
&= 1 \\
&\leq \tilde{\mu}^{(l)}_{B_r}(t)\left((\lambda(y^*(t)) + B_\epsilon) \cap B_r\right) \leq 1.
\end{aligned}$$

Therefore, for every $\epsilon > 0$, $\tilde{\mu}^{(l)}_{B_r}(t)((\lambda(y^*(t)) + B_\epsilon) \cap B_r) = 1$, which gives us that $\tilde{\mu}^{(l)}_{B_r}(t)(\lambda(y^*(t)) \cap B_r) = 1$, and hence $\text{supp}(\tilde{\mu}^{(l)}_{B_r}(t)) \subseteq \lambda(y^*(t))$. Because (33) holds for almost every $t \geq 0$, we have that for almost every $t \geq 0$, $\text{supp}(\tilde{\mu}^{(l)}_{B_r}(t)) \subseteq \lambda(y^*(t))$.

The proof of the claim in part 2 above is similar to the proof of Yaji and Bhatnagar [41, proposition 6.5(ii)], and we provide a brief outline for the sake of completeness. Let $\{f_i\}_{i \geq 1} \subseteq \mathcal{C}(\mathcal{S}^{(2)}, \mathbb{R})$ be a convergence determining class for $\mathcal{P}(\mathcal{S}^{(2)})$. By an appropriate affine transformation, we can ensure that for every $i \geq 1$, for every $s^{(2)} \in \mathcal{S}^{(2)}$, $0 \leq f_i(s^{(2)}) \leq 1$. Define

$$\zeta^i_n := \sum_{k=0}^{n-1} b(k)\left(f_i(S^2_{k+1}) - \int_{\mathcal{S}^{(2)}} f_i\left(\tilde{s}^{(2)}\right) \Pi^{(2)}\left(X_k, Y_k, S^2_k\right)\left(d\tilde{s}^{(2)}\right)\right)$$

for every $n \geq 1$, for every $i \geq 1$. For every $i \geq 1$, $\{\zeta^i_n\}_{n \geq 1}$ is a square integrable martingale w.r.t. the filtration $\{\mathcal{F}_n := \sigma(X_k, Y_k, S^2_k : 0 \leq k \leq n)\}_{n \geq 1}$ and further $\sum_{n=0}^{\infty} \mathbb{E}[(\zeta^i_{n+1} - \zeta^i_n)^2 | \mathcal{F}_n] \leq 2 \sum_{n=1}^{\infty} (b(n))^2 < \infty$. By the Martingale convergence theorem (see Borkar [11, appendix C, theorem 11]), we get that for almost every ω, for every $i \geq 1$, $\{\zeta^i_n\}_{n \geq 1}$ converges. Let $\Omega_m := \{\omega \in \Omega : \ \forall i \geq 1, \ \{\zeta^i_n\} \text{ converges}\}$. Define $\Omega^* := \Omega_m \cap \Omega_2$ and from the arguments above, we get that $\mathbb{P}(\Omega^*) = 1$. Therefore, for every $\omega \in \Omega^*$, for every $i \geq 1$, and for every $T > 0$,

$$\sum_{k=n}^{\tau^2(n,T)} b(k)\left(f_i\left(s^{(2)}_{k+1}\right) - \int_{\mathcal{S}^{(2)}} f_i\left(\tilde{s}^{(2)}\right) \Pi^{(2)}\left(x_k, y_k, s^{(2)}_k\right)\left(d\tilde{s}^{(2)}\right)\right) \longrightarrow 0$$

as $n \to \infty$. By our choice of $\{f_i\}_{i\geq 1}$, the fact that the step-size sequence $\{b(n)\}_{n\geq 0}$ is nonincreasing and the definition of $\mu^{(l)}$, we get that for every $\omega \in \Omega^*$, for every $i \geq 1$, and for every $T > 0$,

$$\lim_{t\to\infty} \int_0^T \int_{B_r \times \mathcal{S}^{(2)}} \left[f_i\left(s^{(2)}\right) - \int_{\mathcal{S}^{(2)}} f_i\left(\tilde{s}^{(2)}\right) \Pi^{(2)}\left(x, y_{[q+t]}, s^{(2)}\right)\left(d\tilde{s}^{(2)}\right) \right] \mu^{(l)}(q+t)\left(dx, ds^{(2)}\right) dq = 0,$$

where $[q+t] := \max\{n \geq 0 : t \geq t^s(n)\}$. By Assumption A4, we have that for every $i \geq 1$, the function $(x, y, s^{(2)}) \to f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y, s^{(2)})(d\tilde{s}^{(2)})$ is continuous, and hence the restriction of the above function to the compact set $B_r \times rU \times \mathcal{S}^{(2)}$ is uniformly continuous, where $r > 0$ is as in part i of this proposition. Using the uniform continuity above and the fact that $\lim_{t\to\infty} \sup_{0\leq q\leq T} \|\bar{y}(q+t) - y_{[q+t]}\| = 0$ (which follows from definition of $\bar{y}(\cdot)$ and uniform continuity of $\bar{y}(\cdot)$), we get that for every $\omega \in \Omega^*$, for every $i \geq 1$, and for every $T > 0$,

$$\lim_{t\to\infty} \int_0^T \int_{B_r \times \mathcal{S}^{(2)}} \left[f_i\left(s^{(2)}\right) - \int_{\mathcal{S}^{(2)}} f_i\left(\tilde{s}^{(2)}\right) \Pi^{(2)}\left(x, \bar{y}(q+t), s^{(2)}\right)\left(d\tilde{s}^{(2)}\right) \right] \mu^{(l)}(q+t)\left(dx, ds^{(2)}\right) dq = 0. \tag{34}$$

We know that as $n \to \infty$, $\bar{y}(\cdot + t_n) \to y^*(\cdot)$ in $\mathcal{C}([0,\infty), \mathbb{R}^{d_2})$ and $\mu^{(l)}(\cdot + t_n) \to \tilde{\mu}^{(l)}(\cdot)$ in $\mathcal{M}(B_r \times \mathcal{S}^{(2)})$. Further, by arguments similar to Lemma 20, the family of functions

$$\left\{ y \in rU \to \int_{B_r \times \mathcal{S}^{(2)}} \left[f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y, s^{(2)})(d\tilde{s}^{(2)}) \right] \nu(dx, ds^{(2)}) : \nu \in \mathcal{P}(B_r \times \mathcal{S}^{(2)}) \right\}$$

is equicontinuous. Therefore, for every $\omega \in \Omega^*$, for every $T > 0$,

$$\begin{aligned}
\lim_{t\to\infty} \Bigg\| \int_0^T \int_{B_r \times \mathcal{S}^{(2)}} \Bigg[\int_{\mathcal{S}^{(2)}} f_i\left(\tilde{s}^{(2)}\right) \Pi^{(2)}\left(x, y^*(q), s^{(2)}\right)\left(d\tilde{s}^{(2)}\right) \\
- \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, \bar{y}(q+t), s^{(2)})(d\tilde{s}^{(2)}) \Bigg] \mu^{(l)}(q+t)(dx, ds^{(2)}) dq \Bigg\| = 0, \\
\lim_{t\to\infty} \Bigg\| \int_0^T \int_{B_r \times \mathcal{S}^{(2)}} \left[f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y^*(q), s^{(2)})(d\tilde{s}^{(2)}) \right] \mu^{(l)}(q+t)(dx, ds^{(2)}) dq \\
- \int_0^T \int_{B_r \times \mathcal{S}^{(2)}} \left[f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y^*(q), s^{(2)})\left(d\tilde{s}^{(2)}\right) \right] \tilde{\mu}^{(l)}(q)(dx, ds^{(2)}) dq \Bigg\| = 0.
\end{aligned}$$

Using the above and Equation (34), we get that for every $\omega \in \Omega^*$, for every $T > 0$,

$$\int_0^T \int_{B_r \times \mathcal{S}^{(2)}} \left[f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y^*(q), s^{(2)})(d\tilde{s}^{(2)}) \right] \tilde{\mu}^{(l)}(q)\left(dx, ds^{(2)}\right) dq = 0.$$

By applying Lebesgue's differentiation theorem (see Borkar [11, chapter 11.1.3]), we get that for every $\omega \in \Omega^*$, for every $i \geq 1$, and for almost every $t \geq 0$,

$$\int_{B_r \times \mathcal{S}^{(2)}} \left[f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y^*(t), s^{(2)})(d\tilde{s}^{(2)}) \right] \tilde{\mu}^{(l)}(t)(dx, ds^{(2)}) = 0.$$

Because for every $t \geq 0$, $\tilde{\mu}^{(l)}(t) \in \mathcal{P}(B_r \times \mathcal{S}^{(2)})$, it is also an element of $\mathcal{P}(\mathbb{R}^{d_1} \times \mathcal{S}^{(2)})$ with $\text{supp}(\tilde{\mu}^{(l)}(t)) \subseteq B_r \times \mathcal{S}^{(2)}$. Therefore, for every $\omega \in \Omega^*$, for every $i \geq 1$, and for almost every $t \geq 0$,

$$\int_{\mathbb{R}^{d_1} \times \mathcal{S}^{(2)}} \left[f_i(s^{(2)}) - \int_{\mathcal{S}^{(2)}} f_i(\tilde{s}^{(2)}) \Pi^{(2)}(x, y^*(t), s^{(2)})(d\tilde{s}^{(2)}) \right] \tilde{\mu}^{(l)}(t)(dx, ds^{(2)}) = 0.$$

Because $\{f_i\}_{i\geq 1}$ is a convergence determining class for $\mathcal{P}(\mathcal{S}^{(2)})$, from the above, it follows that for every $\omega \in \Omega^*$, for almost every $t \geq 0$,

$$\tilde{\mu}^{(l)}_{\mathcal{S}^{(2)}}(t)(d\tilde{s}^{(2)}) = \int_{\mathbb{R}^{d_1} \times \mathcal{S}^{(2)}} \Pi^{(2)}(x, y^*(t), s^{(2)})(d\tilde{s}^{(2)}) \tilde{\mu}^{(l)}(t)(dx, ds^{(2)}).$$

iii. Fix $l \ge 1$. Then, by part i of this proposition, $y^*(\cdot)$ is clearly absolutely continuous, and for almost every $t \ge 0$,

$$\frac{dy^*(t)}{dt} = \int_{U\times\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} h_2^{(l)}\big(x, y^*(t), s^{(2)}, u\big)\tilde{\gamma}^{(l)}(t)\big(du, dx, ds^{(2)}\big).$$

By part ii of this lemma, we know that for almost every $t \ge 0$, $\Theta_1(\tilde{\gamma}^{(l)})(t) = \tilde{\gamma}^{(l)}_{B_r\times\mathcal{S}^{(2)}}(t) = \tilde{\gamma}^{(l)}_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}}(t) \in D(y^*(t))$. Hence, by Lemma 8ii, for almost every $t \ge 0$,

$$\begin{aligned}\frac{dy^*(t)}{dt} &= \int_{U\times\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} h_2^{(l)}\big(x, y^*(t), s^{(2)}, u\big)\tilde{\gamma}^{(l)}(t)\big(du, dx, ds^{(2)}\big)\\ &\in \cup_{\mu\in D(y^*(t))}\int_{\mathbb{R}^{d_1}\times\mathcal{S}^{(2)}} H^{(l)}_{2,y^*(t)}\big(x, s^{(2)}\big)\mu\big(dx, ds^{(2)}\big)\\ &= \hat{H}_2^{(l)}\big(y^*(t)\big).\end{aligned}$$

Because the above holds for every $l \ge 1$, we get that for almost every $t \ge 0$,

$$\frac{dy^*(t)}{dt} \in \cap_{l\ge1}\hat{H}_2^{(l)}(y^*(t)) = \hat{H}_2(y^*(t)),$$

where the last equality follows from Lemma 15vi. □

A continuous function $\mathbf{y}:\mathbb{R}\to\mathbb{R}^{\mathrm{d}_2}$ is said to be an asymptotic pseudotrajectory for the flow of DI (20) if $\lim_{t\to\infty}\mathbf{D}(\mathbf{y}(\cdot+\mathrm{t}), \Sigma^2) = 0$, where $\Sigma^2 \subseteq \mathcal{C}(\mathbb{R}, \mathbb{R}^{d_2})$ denotes the set of solutions of DI (20). Fix $\omega \in \Omega^*$. Extend $\bar{y}(\cdot)$ to the whole of $\mathbb{R}$ by defining $\bar{y}(t) = \bar{y}(0)$ for every $t < 0$. Then by Assumption A8 and uniform continuity of $\bar{y}(\cdot)$, we have that the family of functions $\{\bar{y}(\cdot + t)\}_{t\ge0}$ is relatively compact in $\mathcal{C}(\mathbb{R}, \mathbb{R}^{d_2})$. Let $y^*(\cdot)$ be a limit point of the above family of functions. Then, by Proposition 1iii, $y^*(\cdot)|_{[0,\infty)}$ is a solution of DI (20) on $[0,\infty)$. Usually, the negative time argument is omitted because it follows from the positive time argument.

Fix $T > 0$. Let $t_n \to \infty$ be such that $\bar{y}(\cdot + t_n) \to y^*(\cdot)$ in $\mathcal{C}(\mathbb{R}, \mathbb{R}^{d_2})$. Then, $\bar{y}(\cdot + t_n - T) \to y^*(\cdot - T)$. By Proposition 1iii, $y^*(\cdot - T)|_{[0,\infty)}$ is a solution of DI (20) on $[0,\infty)$. Therefore, $y^*(\cdot)|_{[-T,0]}$ is absolutely continuous and for *a.e.* $t \in [-T, 0]$, $\frac{dy^*(t)}{dt} \in \hat{H}_2(y^*(t))$. Because T was arbitrary, we have that the $y^*(\cdot)|_{(-\infty,0]}$ is solution of DI (20) on $(-\infty, 0]$. Therefore, $y^*(\cdot) \in \Sigma^2$, and, by Benaïm et al. [4, theorem 4.1], we get the following result.

Theorem 3 (APT). *Under Assumptions A1–A9 and A11, for almost every ω, the linearly interpolated trajectory of the slower timescale recursion (15a), $\bar{y}(\cdot)$, is an asymptotic pseudotrajectory of DI (20).*

5.2.3. Characterization of Limit Sets. As a consequence of Theorem 3, for almost every ω, the limit sets of the slower timescale recursion, $L(\bar{y})$, defined as

$$L(\bar{y}) := \cap_{t\ge0}\overline{\{\bar{y}(q + t) : q \ge 0\}} \tag{35}$$

can be characterized in terms of the dynamics of DI (20). Further, using Lemma 18, we obtain the main result of this paper, which we will state below.

Theorem 4 (Limit set). *Under assumptions A1–A9 and A11, for almost every ω,*

i. *$L(\bar{y})$ is a nonempty, compact subset of $\mathbb{R}^{d_2}$ and is internally chain transitive for the flow of DI (20), and*

ii. *if Assumption 10 is satisfied, then $L(\bar{y}) \subseteq \mathcal{Y}$ and as $n \to \infty$,*

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} \to \cup_{y\in\mathcal{Y}}\big(\lambda\big(y\big) \times \{y\}\big).$$

Proof. Fix $\omega \in \Omega^*$.

i. By Theorem 3, we know that $\bar{y}(\cdot)$ is an asymptotic pseudotrajectory for the flow of DI (20). Now the claim follows from Benaïm et al. [4, theorem 4.3].

ii. By part i of this theorem, we know that $L(\bar{y})$ is internally chain transitive for the flow of DI(20). Because $\mathcal{Y}$ is a globally attracting set for DI(20), by Benaïm et al. [4, corollary 3.24], we get that $L(\bar{y}) \subseteq \mathcal{Y}$. Therefore, $y_n \to \mathcal{Y}$ as $n \to \infty$ and by Lemma 18, we get that, as $n \to \infty$,

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} \to \cup_{y\in\mathcal{Y}}\big(\lambda(y) \times \{y\}\big). \quad \square$$

6. Applications

In this section, we consider applications of the theory to problems in convex optimization. First, we present an application of the theory to solve a constrained convex optimization problem, where the objective function and the functions defining the constraints are averaged with respect to the stationary distribution of a Markov process. Then, we present another application where Newton's method is used to solve an unconstrained optimization problem. Throughout this section, we assume that $\mathcal{S}^{(1)} = \mathcal{S}^{(2)} = \mathcal{S}$ and $|\mathcal{S}| < \infty$.

7. Application: Constrained Convex Optimization

In this section, we consider an application of the theory to a problem of constrained convex optimization. Throughout this section, we assume that $\mathcal{S}^{(1)} = \mathcal{S}^{(2)} = \mathcal{S}$ and $|\mathcal{S}| < \infty$. The finiteness of the state space is an assumption imposed only for this application and not in the analysis of the scheme presented in the paper.

Let the objective function $J : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathbb{R}$ be such that $J(\cdot)$ is continuous and for every $s \in \mathcal{S}$, $J(\cdot, s)$ is convex and coercive (i.e., for any $M > 0$, there exists $r > 0$ such that for any $x \in \mathbb{R}^{d_1}$ with $\|x\| \geq r$, we have that $J(x, s) \geq M$). The functions describing the constraints are given by $C : \mathcal{S} \to \mathbb{R}^{d_2 \times d_1}$ and $w : \mathcal{S} \to \mathbb{R}^{d_2}$. We assume that for any $s \in \mathcal{S}$, the set $\mathcal{X}(s) := \{x \in \mathbb{R}^{d_1} : C(s)x = d(s)\}$ is nonempty. The law of the Markov noise terms is given by $\Pi : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathcal{P}(\mathcal{S})$ such that Π is continuous and let $\mu \in \mathcal{P}(\mathcal{S})$ denote the unique stationary distribution of the Markov chain given by the transition kernel $\Pi(x, \cdot)(\cdot)$ for every $x \in \mathbb{R}^{d_1}$.

Let $\partial J(x, s)$ denote the set of subgradients of the convex function $J(\cdot, s)$ at the point $x \in \mathbb{R}^{d_1}$. Formally,

$$\partial J(x,s) := \left\{ g \in \mathbb{R}^{d_1} : \ \forall x' \in \mathbb{R}^{d_1},\ J(x',s) \geq J(x,s) + \langle g, x' - x\rangle \right\}.$$

Then, it is easy to show that for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $\partial J(x, s)$ is convex and compact. Further, the map $(x, s) \to \partial J(x, s)$ possesses the closed graph property. We assume that the map $(x, s) \to \partial J(x, s)$ satisfies the linear growth property, that is, there exists $K > 0$ such that $\sup_{x' \in \partial J(x,s)} \|x'\| \leq K(1 + \|x\|)$.

Let $J_\mu : \mathbb{R}^{d_1} \to \mathbb{R}$ be defined such that for every $x \in \mathbb{R}^{d_1}$, $J_\mu(x) := \int_{\mathcal{S}} J(x, s)\mu(ds)$. Similarly, define $C_\mu := \int_{\mathcal{S}} C(s)\cdot \mu(ds) \in \mathbb{R}^{d_2 \times d_1}$ and $w_\mu := \int_{\mathcal{S}} w(s)\mu(ds) \in \mathbb{R}^{d_2}$. The optimization problem that we wish to solve is given by

$$\begin{aligned} OP(\mu):\ & \min_{x\in\mathbb{R}^{d_1}} J_\mu(x),\ \text{subject to}: \\ & C_\mu x = w_\mu. \end{aligned}$$

The standard approach in solving the optimization problem $OP(\mu)$ is the projected subgradient descent algorithm whose recursion is given by

$$X_{n+1} = P_\mu\big(X_n - a(n)(g_n + M_{n+1})\big),$$

where $g_n \in \partial J_\mu(X_n)$, M_{n+1} is the subgradient estimation error and P_μ denotes the projection operation onto the affine subspace $\mathcal{X}_\mu := \{x \in \mathbb{R}^{d_1} : C_\mu x = w_\mu\}$. Such a scheme cannot be implemented when μ is not known. This is the case in problems arising in optimal control.

The feasible set of the optimization problem $OP(\mu)$, given by $\mathcal{X}_\mu$, is nonempty because for every $s \in \mathcal{S}$, $\mathcal{X}(s)$ is nonempty. Further, because for every $s \in \mathcal{S}$, $J(\cdot, s)$ is coercive, the function $J_\mu(\cdot)$ is coercive and hence bounded below. Therefore, the optimization problem $OP(\mu)$ has at least one solution. Let the solution set of the optimization problem $OP(\mu)$ be denoted by Z.

For any $r > 0$, let B_r denote the closed ball of radius r in $\mathbb{R}^{d_1}$ centered on the origin. For every $s \in \mathcal{S}$, pick $x_s \in \mathcal{X}(s)$ and compute $M_1 := \max\{J(x_s, s') : s, s' \in \mathcal{S}\}$. Then, $x_\mu := \sum_{s\in\mathcal{S}} \mu(s)x_s \in \mathcal{X}_\mu$ and $J_\mu(x_\mu) \leq M_1$. Because $|\mathcal{S}| < \infty$ and the functions $J(\cdot, s)$ are coercive, for some $M > \max\{0, M_1\}$, there exists $r > \max\{\|x_s\| : s \in \mathcal{S}\}$ such that for every $s \in \mathcal{S}$, for every $x \in B_r^c$, $J(x, s) \geq M$, and for every $s \in \mathcal{S}$, $B_r \cap \mathcal{X}(s) \neq \emptyset$. Then $Z \subseteq B_r$. Instead of $OP(\mu)$, we will solve the following penalized/regularized optimization problem given by

$$\begin{aligned} \tilde{OP}(\mu):\ & \min_{x\in\mathbb{R}^{d_1}} J_\mu(x) + \frac{\epsilon}{2r^2}\|x\|^2 + \frac{K+1}{2}\max\{\|x\|^2 - r^2, 0\}, \\ \text{subject to:}\ & C_\mu x = w_\mu, \end{aligned}$$

where $r > 0$ is as determined above, K is the constant associated with the linear growth property of the subgradient map $\partial J(\cdot, s)$, and $\epsilon > 0$ is an arbitrary constant small in value. Then, it is easy to show that $\tilde{OP}(\mu)$ has at least one solution, and the set of solutions of $\tilde{OP}(\mu)$ denoted by $\tilde{Z}$ is such that $\tilde{Z} \subseteq B_r$. Further, for any $\tilde{x} \in \tilde{Z}$, for any $x^* \in Z$, $J_\mu(\tilde{x}) - J_\mu(x^*) \leq \epsilon$.

Consider the Lagrangian $L : \mathbb{R}^d \to \mathbb{R}$ associated with the optimization problem $\tilde{OP}(\mu)$ defined such that for every $(x, y) \in \mathbb{R}^d$,

$$L(x, y) := J_\mu(x) + \frac{\epsilon}{2r^2}\|x\|^2 + \frac{K+1}{2}\max\{\|x\|^2 - r^2, 0\} + \langle y, C_\mu x - w_\mu \rangle.$$

Let $\hat{J} : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathbb{R}$, be defined such that for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$, $\hat{J}(x, s) := J(x, s) + \frac{\epsilon}{2r^2}\|x\|^2 + \frac{K+1}{2}\max\{\|x\|^2 - r^2, 0\}$. Then, for every $(x, s) \in \mathbb{R}^{d_1} \times \mathcal{S}$,

$$\hat{J}_\mu(x) := J_\mu(x) + \frac{\epsilon}{2r^2}\|x\|^2 + \frac{K+1}{2}\max\{\|x\|^2 - r^2, 0\} = \int_{\mathcal{S}} \hat{J}(x, s)\mu(ds).$$

When the transition law Π (and hence μ) is not known, we propose the following recursion, which performs primal descent along the faster timescale (i.e., minimization of $L(\cdot, y)$ w.r.t. x) and dual ascent on the slower timescale (i.e., maximization of $L(x, \cdot)$ w.r.t. y). The recursion is given by

$$Y_{n+1} - Y_n = b(n)(C(S_n)X_n - w(S_n)), \tag{36a}$$

$$X_{n+1} - X_n - a(n)M^1_{n+1} \in -a(n)\big(\partial\hat{J}(X_n, S_n) + C(S_n)^T Y_n\big), \tag{36b}$$

where the step-size sequences $\{a(n)\}_{n\geq 0}$ and $\{b(n)\}_{n\geq 0}$ are chosen such that they satisfy Assumption A5 and $\{M^1_n\}_{n\geq 1}$ denotes the subgradient estimation error, which is assumed to satisfy Assumption A6 (e.g., when $\{M^1_n\}_{n\geq 1}$ is i.i.d. zero mean with finite variance, Assumption A6 is satisfied; more generally, Assumption A6 is satisfied if $\{M^1_n\}_{n\geq 1}$ is a martingale difference sequence satisfying Assumption A3 in Borkar [11, chapter 2.1]).

It is easy to see that the maps $(x, y, s) \to -(\partial\hat{J}(x, s) + C(s)^T y)$ and $(x, y, s) \to C(s)x - w(s)$ satisfy Assumptions A1 and A2, respectively. The linear growth property of the map $(x, y, s) \to -(\partial\hat{J}(x, s) + C(s)^T y)$ follows from the linear growth property of $x \to \partial J(x, s)$. Further, by Bertsekas [6, proposition 5.4.6], we get that for every $(x, y) \in \mathbb{R}^d$, $-\int_{\mathcal{S}}(\partial\hat{J}(x, s) + C(s)^T y)\mu(ds) = -(\partial\hat{J}_\mu(s) + C^T_\mu y) = -\partial L(x, y)$.

For every $y \in \mathbb{R}^{d_2}$, let $\lambda(y) := \{x \in \mathbb{R}^{d_1} : -C^T_\mu y \in \partial\hat{J}_\mu(x)\}$. Then, for every $y \in \mathbb{R}^{d_2}$, $\lambda(y)$ is nonempty because $L(\cdot, y)$ is convex and coercive. Further, $|\lambda(y)| = 1$; that is, $\lambda(y)$ is a singleton because $L(\cdot, y)$ is strictly convex. For any $y \in \mathbb{R}^{d_2}$, $x' \in \lambda(y)$ if and only if there exists $\tilde{g} \in \mathbb{R}^{d_1}$ in the set of subgradients of the function $J_\mu(\cdot) + \frac{K+1}{2}\max\{\|\cdot\|^2 - r^2, 0\}$ at x' such that $\tilde{g} + \frac{\epsilon}{r^2}x' + C^T_\mu y = 0$. So either $\|x'\| \leq r$ or if $\|x'\| > r$, then $\max\{\|x'\|^2 - r^2, 0\} = \|x'\|^2 - r^2$, and hence for some $g \in \partial J_\mu(x')$, $\tilde{g} = g + (K+1)x'$, from which we get that,

$$\left(K + 1 + \frac{\epsilon}{2r^2}\right)\|x'\| = \|g + C^T_\mu y\| \leq K + K\|x'\| + \|C^T_\mu\|\,\|y\|.$$

Thus, for $K' := \max\{K, r, \|C^T_\mu\|\}$, we get that for every $y \in \mathbb{R}^{d_2}$, $\|\lambda(y)\| \leq K'(1 + \|y\|)$. The set $\lambda(y)$ is clearly globally attracting for the flow of DI $\frac{dx}{dt} \in -(\partial\hat{J}_\mu(x) + C^T_\mu y)$, and by Aubin and Cellina [2, theorem 6], the map $y \in \mathbb{R}^{d_2} \to \lambda(y)$ is u.s.c. (because $\lambda(\cdot)$ is also single valued, it is continuous). Hence, the map $\lambda(\cdot)$ satisfies Assumption 9.

If the iterates are stable for a.e. ω (i.e., Assumption 8 is satisfied), the result in Section 5.1 gives us that for almost every ω, there exists a nonempty compact set $A \subseteq \mathbb{R}^d$ such that $(X_n(\omega), Y_n(\omega)) \to A$ as $n \to \infty$ and A is internally chain transitive for the flow of DI,

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} \in \begin{pmatrix} -\partial\hat{J}_\mu(x) - C^T_\mu y \\ 0 \end{pmatrix}. \tag{37}$$

By the arguments in Section 5.1, we have that $A \cap \mathcal{G}(\lambda) \neq \emptyset$, where $\mathcal{G}(\lambda) := \{(\lambda(y), y) : y \in \mathbb{R}^{d_2}\}$. Let $\mathcal{O} \subseteq \mathbb{R}^d$ be an open neighborhood of $\mathcal{G}(\lambda)$. Let $\mathcal{O}'(\delta) := \{(x, y) \in \mathbb{R}^d : L(x, y) - L(\lambda(y), y) < \delta\}$. By Aubin and Cellina [2, chapter 1.2, theorem 6], the map $y \in \mathbb{R}^{d_2} \to L(\lambda(y), y)$ is continuous, and hence $\mathcal{O}'(\delta)$ is an open neighborhood of $\mathcal{G}(\lambda)$. Further, it is easy to show that $\cap_{\delta>0}\mathcal{O}'(\delta) = \mathcal{G}(\lambda)$, and hence $\cap_{\delta>0}(\mathcal{O}'(\delta) \cap A) = \mathcal{G}(\lambda) \cap A$. Because $A \subseteq \mathbb{R}^d$ is compact, there exists $\delta^* > 0$ such that $\mathcal{O}'(\delta^*) \cap A \subseteq \mathcal{O} \cap A$. Consider any solution of DI (37), $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ starting at $(x^*, y^*) \in \mathcal{O}'(\delta^*) \cap A$ and satisfying for every $t \in \mathbb{R}$, $(\mathbf{x}(t), \mathbf{y}(t)) \in A$. Recall from Section 5.1 that $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ as above is such that for every $t \geq 0$, $\mathbf{y}(t) = y^*$ and $\mathbf{x}(\cdot)$ is a solution of DI, $\frac{dx}{dt} \in -(\partial\hat{J}_\mu(x) + C^T_\mu y^*) = -\partial L(x, y^*)$ and hence descends along the potential $L(x, y^*)$. Therefore, the solution $(\mathbf{x}(\cdot), \mathbf{y}(\cdot))$ remains within $\mathcal{O}'(\delta^*) \cap A$, which gives us that $\Phi^A(\mathcal{O}'(\delta^*) \cap A, [0, \infty)) \subseteq \mathcal{O} \cap A$, where Φ^A denotes the flow of DI (37) restricted to the set A. Thus, Assumption A11 is satisfied, and, from Lemma 18, we get the following result.

Lemma 23 (Faster Timescale Convergence). *For almost every ω, $(X_n(\omega), Y_n(\omega)) \to \mathcal{G}(\lambda)$ as $n \to \infty$.*

Theorem 3 gives us that the iterates $\{Y_n\}_{n\geq 0}$ in Recursion (36a) track the flow of DI,

$$\frac{dy}{dt} \in \cup_{\nu \in D(y)} \int_{\mathbb{R}^{d_1}\times\mathcal{S}} (C(s)x - w(s))\nu(dx, ds), \tag{38}$$

where, for every $y \in \mathbb{R}^{d_2}$, $D(y)$ is as in Equation (18). Because for every $y \in \mathbb{R}^{d_2}$, $\lambda(y)$ is a singleton and because μ is the unique stationary distribution of the Markov chain given by transition kernel $\Pi(\cdot)(\cdot)$, we get that for every $y \in \mathbb{R}^{d_2}$, $D(y) = \delta_{\lambda(y)} \otimes \mu$. Therefore, DI (38) takes the following form:

$$\frac{dy}{dt} = C_\mu \lambda(y) - w_\mu. \tag{39}$$

In order to analyze the asymptotic behavior of o.d.e. (39), we need the following version of the envelope theorem. The proof of the envelope theorem below is similar to that by Milgrom and Segal [30].

Lemma 24 (Envelope Theorem)**.** *Let $\mathbf{y} : [0,T] \longrightarrow \mathbb{R}^{d_2}$ be an absolutely continuous function. Let $\tilde{L} : \mathbb{R}^{d_1} \times [0,T] \longrightarrow \mathbb{R}$ be defined such that for every $(x,t) \in \mathbb{R}^{d_1} \times [0,T]$, $\tilde{L}(x,t) := L(x, \mathbf{y}(t))$. Then,*

i. *for every $x \in \mathbb{R}^{d_1}$, $\tilde{L}(x,\cdot)$ is absolutely continuous and there exists $\mathfrak{D} \subseteq [0,T]$ with Lebesgue measure T such that for every $t \in \mathfrak{D}$, for every $x \in \mathbb{R}^{d_1}$, $\frac{\partial \tilde{L}(x,t)}{\partial t}$ exists and $\frac{\partial \tilde{L}(x,t)}{\partial t} = \langle \frac{d\mathbf{y}(t)}{dt}, C_\mu x - w_\mu \rangle$.*

ii. *the function $V : [0,T] \longrightarrow \mathbb{R}$, where for every $0 \leq t \leq T$, $V(t) := \inf_{x\in\mathbb{R}^{d_1}} \tilde{L}(x,t)$ is absolutely continuous. Further, for any $0 < t \leq T$,*

$$V(t) = V(0) + \int_0^t \frac{\partial \tilde{L}(x,q)}{\partial q}\bigg|_{x=\lambda(\mathbf{y}(q))} dq.$$

Proof.

i. Because $\mathbf{y}(\cdot)$ is absolutely continuous, it is differentiable almost everywhere, and let $\mathfrak{D} \subseteq [0,T]$ be the set of $t \in [0,T]$ such that $\frac{d\mathbf{y}(t)}{dt}$ exists. Then clearly, the Lebesgue measure of $\mathfrak{D}$ is T.

Fix $x \in \mathbb{R}^{d_1}$. Then, $L(x,\cdot)$ is a Lipschitz continuous function because for any $y', y'' \in \mathbb{R}^{d_2}$,

$$\begin{aligned} |L(x,y') - L(x,y'')| &= |\langle y' - y'', C_\mu x - w_\mu \rangle| \\ &\leq \|y' - y''\| \, \|C_\mu x - w_\mu\| \\ &= \beta_x \|y' - y''\|, \end{aligned}$$

where $\beta_x := \|C_\mu x - w_\mu\|$. Further, $L(x,\cdot)$ is differentiable (i.e., totally differentiable because it is linear in y), and the total derivative is given by $\nabla_y L(x,y') = (C_\mu x - w_\mu)$ for every $y' \in \mathbb{R}^{d_2}$. Because $\tilde{L}(x,\cdot)$ is the composition of an absolutely continuous function $\mathbf{y}(\cdot)$ and a Lipschitz continuous function $L(x,\cdot)$, we have that $\tilde{L}(x,\cdot)$ is absolutely continuous. By Rudin [38, theorem 9.15], we have that for every $t \in \mathfrak{D}$, $\frac{\partial \tilde{L}(x,t)}{dt}$ exists and $\frac{\partial \tilde{L}(x,t)}{dt} = \langle \frac{d\mathbf{y}(t)}{dt}, C_\mu x - w_\mu \rangle$.

ii. Because $\mathbf{y}(\cdot)$ is absolutely continuous, there exists $\alpha > 0$ such that $\sup_{t\in[0,T]} \|\mathbf{y}(t)\| \leq \alpha$. Further, by Assumption A9, for every $t \in [0,T]$,

$$V(t) = \inf_{\substack{x\in\mathbb{R}^{d_1}:\\ \|x\|\leq K'(1+\|\mathbf{y}(t)\|)}} \tilde{L}(x,t) = \inf_{\substack{x\in\mathbb{R}^{d_1}:\\ \|x\|\leq K'(1+\alpha)}} \tilde{L}(x,t).$$

Therefore, for every $0 \leq t < t' \leq T$,

$$\begin{aligned} |V(t') - V(t)| &\leq \sup_{\substack{x\in\mathbb{R}^{d_1}:\\ \|x\|\leq K'(1+\alpha)}} \left|\tilde{L}(x,t') - \tilde{L}(x,t)\right| \\ &\leq \sup_{\substack{x\in\mathbb{R}^{d_1}:\\ \|x\|\leq K'(1+\alpha)}} \left|\int_t^{t'} \frac{\partial \tilde{L}(x,q)}{\partial q} dq\right| \\ &= \sup_{\substack{x\in\mathbb{R}^{d_1}:\\ \|x\|\leq K'(1+\alpha)}} \left|\langle \mathbf{y}(t') - \mathbf{y}(t), C_\mu x - w_\mu \rangle\right| \\ &\leq \left(\sup_{\substack{x\in\mathbb{R}^{d_1}:\\ \|x\|\leq K'(1+\alpha)}} \|C_\mu x - w_\mu\| \right) \|\mathbf{y}(t') - \mathbf{y}(t)\|. \end{aligned}$$

Now the absolute continuity of $V(\cdot)$ follows from absolute continuity of $\mathbf{y}(\cdot)$. Because $V(\cdot)$ is absolutely continuous, $\frac{dV(q)}{dq}$ exists for a.e. $q \in [0, T]$, and for any $0 < t \le T$, $V(t) = V(0) + \int_0^t \frac{dV(q)}{dq} dq$. Let $q \in (0, T)$ be such that $\frac{dV(q)}{dq}$ exists and $q \in \mathcal{D}$. Then, for $q' > q$, $V(q') - V(q) \le \tilde{L}(\lambda(\mathbf{y}(q)), q') - \tilde{L}(\lambda(y(q)), q)$. The right derivative of $V(\cdot)$ at q, which is the same as $\frac{dV(q)}{dq}$ satisfies $\frac{dV(q)}{dq} \le \frac{\partial \tilde{L}(x,q)}{\partial q}\big|_{x=\lambda(\mathbf{y}(q))}$. Considering $q < q'$ and repeating the above argument gives us $\frac{\partial \tilde{L}(x,q)}{\partial q}\big|_{x=\lambda(\mathbf{y}(q))} \le \frac{dV(q)}{dq}$. Thus, for a.e. $q \in [0, T]$, $\frac{\partial \tilde{L}(x,q)}{\partial q}\big|_{x=\lambda(\mathbf{y}(q))} = \frac{dV(q)}{dq}$ and because $V(\cdot)$ is absolutely continuous, for any $0 < t \le T$,

$$V(t) = V(0) + \int_0^t \frac{\partial \tilde{L}(x,q)}{\partial q}\bigg|_{x=\lambda(\mathbf{y}(q))} dq. \quad \square$$

Let $Q_\mu : \mathbb{R}^{d_2} \to \mathbb{R}$ be defined such that for $y \in \mathbb{R}^{d_2}$, $Q_\mu(y) := \inf_{x \in \mathbb{R}^{d_1}} L(x, y) = L(\lambda(y), y)$. The function $Q_\mu(\cdot)$ is the objective function of the dual of the optimization problem $\tilde{OP}(\mu)$ and is a concave function. By the strong duality theorem (see Bertsekas [6, proposition 5.3.3]), the dual optimization problem given by $\max_{y \in \mathbb{R}^{d_2}} Q_\mu(y)$ has at least one solution and let the set of solutions of the dual optimization problem be denoted by $\mathcal{Y}$. Further, the strong duality theorem also gives us that for any $y \in \mathcal{Y}$ and for any $x \in \tilde{Z}$, $Q_\mu(y) = \hat{J}_\mu(x)$.

Let $\mathbf{y} : \mathbb{R} \to \mathbb{R}^{d_2}$ be a solution of the o.d.e. (39) with initial condition $y \in \mathcal{Y}$. Then, $\mathbf{y}(\cdot)$ is absolutely continuous and for a.e. $t \in [0, \infty)$, $\frac{d\mathbf{y}(t)}{dt} = C_\mu \lambda(\mathbf{y}(t)) - w_\mu$. By Lemma 24ii, we have that for any $t \ge 0$,

$$\begin{aligned}
V(t) &= V(0) + \int_0^t \frac{\partial \tilde{L}(x,q)}{\partial q}\bigg|_{x=\lambda(\mathbf{y}(q))} dq, \\
&= V(0) + \int_0^t \left\langle \frac{d\mathbf{y}(q)}{dq}, C_\mu \lambda(\mathbf{y}(q)) - w_\mu \right\rangle dq, \\
&= V(0) + \int_0^t \|C_\mu \lambda(\mathbf{y}(q)) - w_\mu\|^2 dq. \qquad (40)
\end{aligned}$$

Because $V(t) = Q_\mu(\mathbf{y}(t))$ and $V(0) = Q_\mu(y)$, where $y \in \mathcal{Y}$, we get that $V(t) - V(0) \le 0$. Hence, for every $t \ge 0$, $\int_0^t \|C_\mu \lambda(\mathbf{y}(q)) - w_\mu\|^2 dq \le 0$, which gives us that $\|C_\mu \lambda(\mathbf{y}(t)) - w_\mu\| = 0$ for a.e. $t \in [0, \infty)$. Thus, for any solution of o.d.e. (39), $\mathbf{y}(\cdot)$, with initial condition $y \in \mathcal{Y}$, we have that $C_\mu \lambda(y) - w_\mu = 0$ and for every $t \ge 0$, $\mathbf{y}(t) = y$. Therefore, $\mathcal{Y} \subseteq \{y \in \mathbb{R}^{d_2} : C_\mu \lambda(y) - w_\mu = 0\}$. Further, by Bertsekas [6, proposition 5.3.3(ii)], any $y \in \mathbb{R}^{d_2}$ such that $C_\mu \lambda(y) - w_\mu = 0$ is a solution of the dual optimization problem, and hence $\mathcal{Y} = \{y \in \mathbb{R}^{d_2} : C_\mu \lambda(y) - w_\mu = 0\}$ (from this, it also follows that $\mathcal{Y}$ is closed).

In the theorem below, we summarize the main convergence result associated with Recursion (36).

Theorem 5 (Convergence to Lagrangian Saddle Points)**.**

i. *For any solution* $\mathbf{y}(\cdot)$ *of the o.d.e.* (39) *with any initial condition* $y_0 \in \mathbb{R}^{d_2}$*, which is bounded for* $t \ge 0$ *(i.e.,* $\sup_{t \ge 0} \|\mathbf{y}(t)\| < \infty$*), we have that as* $t \to \infty$, $\inf_{y \in \mathcal{Y}} \|\mathbf{y}(t) - y\| \to 0$.

ii. *For any* $y \in \mathcal{Y}$, $\lambda(y)$ *is a solution of the optimization problem* $\tilde{OP}(\mu)$ *(i.e.,* $\lambda(y) \in \tilde{Z}$*).*

iii. *If the iterates remain stable for almost every* ω *(i.e., Assumption* A8 *is satisfied), then, for almost every* ω,

a. $Y_n(\omega) \to \mathcal{Y}$ *as* $n \to \infty$*, and*

b. $\begin{pmatrix} X_n(\omega) \\ Y_n(\omega) \end{pmatrix} \to \cup_{y \in \mathcal{Y}} \left\{ \begin{pmatrix} \lambda(y) \\ y \end{pmatrix} \right\} \subseteq \mathbb{R}^d$.

Proof.

i. Let $\mathbf{y}(\cdot)$ be a solution of the o.d.e. (39) with initial condition $y_0 \in \mathbb{R}^{d_2}$ (assume $y_0 \notin \mathcal{Y}$; otherwise, we know that for every $t \ge 0$, $\mathbf{y}(t) = y_0$ and hence the claim follows) such that $\sup_{t \ge 0} \|\mathbf{y}(t)\| \le \mathrm{M}$ for some $M > 0$. Then, $\mathbf{y}(\cdot)|_{[0,\infty)}$ is uniformly continuous because for any $0 \le t < t' < \infty$,

$$\begin{aligned}
\|\mathbf{y}(t') - \mathbf{y}(t)\| &= \left\| \int_t^{t'} (C_\mu \mathbf{y}(q) - w_\mu) dq \right\| \\
&\le \int_t^{t'} (\|C_\mu\| \, \|\mathbf{y}(q)\| + \|w_\mu\|) dq \\
&\le (\|C_\mu\| M + \|w_\mu\|)(t' - t).
\end{aligned}$$

The function $y \in \mathbb{R}^{d_2} \to \|C_\mu y - w_\mu\|$ is uniformly continuous, and hence the function $t \in [0,\infty) \to \|C_\mu \mathbf{y}(t) - w_\mu\|$ is uniformly continuous. Further by Lemma 24ii, for any $t > 0$, $0 \le V(t) - V(0) \le Q_\mu(y) - V(0) < \infty$, where $y \in \mathcal{Y}$. The claim that as $t \to \infty$, $\mathbf{y}(t) \to \mathcal{Y}$ is equivalent to the claim that as $t \to \infty$, $\|C_\mu \mathbf{y}(t) - w_\mu\| \to 0$.

Suppose there exists $\epsilon > 0$ such that for every $T > 0$, there exists $t \ge T$ such that $\|C_\mu \mathbf{y}(t) - w_\mu\| > \epsilon$. From the uniform continuity of $t \to \|C_\mu \mathbf{y}(t) - w_\mu\|$, there exists $\delta > 0$ such that for every $t, t' \in [0,\infty)$ satisfying $|t - t'| < \delta$, $|\|C_\mu \mathbf{y}(t) - w_\mu\| - \|C_\mu \mathbf{y}(t') - w_\mu\|| < \frac{\epsilon}{2}$. Therefore, we can obtain a sequence $\{t_n\}_{n\ge1}$ such that for every $n \ge 1$, $\delta < t_n < t_{n+1} - 2\delta$ and for every $t \in (t_n - \delta, t_n + \delta)$, $\|C_\mu \mathbf{y}(t) - w_\mu\| > \frac{\epsilon}{2}$. Let N be such that $\frac{2(Q_\mu(y) - V(0))}{\epsilon^2\delta} < N$, where $y \in \mathcal{Y}$. Then, by Lemma 24ii, we get that

$$
\begin{aligned}
Q_\mu(\mathbf{y}(t_{N+1})) - V(0) &= V(t_{N+1}) - V(0) \\
&= \int_0^{t_{N+1}} \left\langle \frac{d\mathbf{y}(q)}{dq}, C_\mu \lambda(\mathbf{y}(q)) - w_\mu \right\rangle dq \\
&= \int_0^{t_{N+1}} \|C_\mu \mathbf{y}(q) - w_\mu\|^2 dq \\
&\ge \sum_{n=1}^{N} \int_{t_n - \delta}^{t_n + \delta} \|C_\mu \mathbf{y}(q) - w_\mu\|^2 dq \\
&> N\left(\frac{\epsilon^2\delta}{2}\right) \\
&> Q_\mu(y) - V(0),
\end{aligned}
$$

which contradicts the fact that $V(t) - V(0) \le Q_\mu(y) - V(0)$. Therefore, $\lim_{t\to\infty} \|C_\mu \mathbf{y}(t) - w_\mu\| = 0$.

ii. Let $y \in \mathcal{Y}$. Then, we know that $C_\mu \lambda(y) - w_\mu = 0$, and hence $\lambda(y)$ is feasible for $\tilde{OP}(\mu)$. By the definition of $\lambda(y)$, we have that for every $x \in \mathbb{R}^{d_1}$, $L(\lambda(y), y) \le L(x, y)$. Now the claim follows from Bertsekas [6, proposition 5.3.3(ii)].

iii. Let ω be such that Theorem 4 holds.

a. Then, by Theorem 4i, we know that there exists a nonempty, compact set $A \subseteq \mathbb{R}^{d_2}$ such that as $n \to \infty$, $Y_n(\omega) \to A$. Further, A is internally chain transitive for the flow of o.d.e. (39) and hence is invariant. Let $\mathbf{y}(\cdot)$ be a solution to o.d.e. (39) with initial condition in A and for every $t \in \mathbb{R}$, $\mathbf{y}(t) \in A$. Because A is compact, $\sup_{t\ge0} \|\mathbf{y}(t)\| < \infty$, and hence by part i of this theorem, we get that $\mathbf{y}(t) \to \mathcal{Y}$ as $t \to \infty$. Because for every $t \ge 0$, $\mathbf{y}(t) \in A$, we get that $\mathcal{Y} \cap A \ne \emptyset$. Further, for some $y \in \mathcal{Y}$, $(\cap_{\delta>0}\{y' \in \mathbb{R}^{d_2} : Q_\mu(y) - Q_\mu(y') < \delta\}) \cap A = \mathcal{Y} \cap A$. For any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $\{y' \in \mathbb{R}^{d_2} : Q_\mu(y) - Q_\mu(y') < \delta\} \cap A \subseteq N^\epsilon(\mathcal{Y} \cap A)$, where $N^\epsilon(\cdot)$ denotes the ϵ-neighborhood of a set. By using Lemma 24ii, it is easy to show that $\Phi^A(\{y' \in \mathbb{R}^{d_2} : Q_\mu(y) - Q_\mu(y') < \delta\} \cap A, [0,\infty)) \subseteq N^\epsilon(\mathcal{Y} \cap A)$, where Φ^A denotes the flow of o.d.e. (39) restricted to set A (see Section 2.3). Therefore, $\mathcal{Y} \cap A$ is an attracting set for the flow Φ^A. From Benaïm et al. [4, proposition 3.20], we get that $\mathcal{Y} \cap A = A$. Therefore, as $n \to \infty$, $Y_n(\omega) \to \mathcal{Y}$.

b. Follows from part iii(a) of this theorem and Lemma 23. □

7.1. Newton's Method with Markov Sampling

Let the objective function $J : \mathbb{R}^{d_1} \times \mathcal{S} \to \mathbb{R}$ be such that $J(\cdot)$ is continuous and for every $s \in \mathcal{S}$, $J(\cdot, s)$ is twice continuously differentiable, convex, and coercive (i.e., for any $M > 0$, there exists $r > 0$ such that for any $x \in \mathbb{R}^{d_1}$ with $\|x\| \ge r$, we have that $J(x,s) \ge M$). Let $\nabla J(x,s)$ and $H(x,s)$ denote the gradient and hessian, respectively, of the convex function $J(\cdot,s)$ at the point $x \in \mathbb{R}^{d_1}$. We assume that the map $(x,s) \to \nabla J(x,s)$ satisfies the linear growth property, that is, there exists $K > 0$ such that $\nabla J(x,s) \le K(1 + \|x\|)$ and that there exists $r > 0$ such that the maximum eigenvalue of the hessian $H(x,s)$ is bounded above by r.

Let $J_\mu : \mathbb{R}^{d_1} \to \mathbb{R}$ be defined such that for every $x \in \mathbb{R}^{d_1}$, $J_\mu(x) := \int_{\mathcal{S}} J(x,s)\mu(ds)$. The optimization problem that we wish to solve is given by

$$\min_{x \in \mathbb{R}^{d_1}} J_\mu(x).$$

Clearly, the function J_μ is twice continuously differentiable. While for every $s \in \mathcal{S}$ the hessian of the function $J_\mu(\cdot, s)$ need not be invertible, we assume that the hessian of the averaged function $J_\mu(\cdot)$ at x, denoted by $H_\mu(x)$, is invertible for any x and that there exists an $\eta > 0$ such that the minimum eigenvalue of the matrix $H_\mu(x)$ is bounded below by η for every x. Further, because for every $s \in \mathcal{S}$, $J(\cdot, s)$ is coercive, the function $J_\mu(\cdot)$ is coercive and hence bounded below. Therefore, the optimization problem $OP(\mu)$ has at least one solution. Let the solution set of the optimization problem $OP(\mu)$ be denoted by Z and because $H_\mu(\cdot)$ is invertible, $|Z| = 1$.

The standard approach in solving the optimization problem $OP(\mu)$ is the Newton's algorithm whose recursion is given by

$$X_{n+1} = X_n - a(n)\left(H_\mu^{-1}(X_n)\nabla J_\mu(X_n) + M_{n+1}\right),$$

where M_{n+1} is the error in gradient and hessian estimation. When the cardinality of the set $\mathcal{S}$ is large, computation of expectations of the gradient and hessian for every iteration is infeasible. Hence, one resorts to sampling-based approaches in which the expectation of the gradient and hessian are replaced by the gradient and hessian of the function $J(\cdot, S_n)$, respectively, where S_n is sampled from a distribution such that the desired expectation is obtained in the limit.

We consider a popular sampling approach known as the Markov sampling method where samples are obtained from a Markov chain with a unique stationary distribution μ. Let the law of the Markov sampling terms be given by $\Pi : \mathbb{R}^d \times \mathcal{S} \to \mathcal{P}(\mathcal{S})$. We assume that Π is continuous and $\mu \in \mathcal{P}(\mathcal{S})$ is the unique stationary distribution of the Markov chain given by the transition kernel $\Pi(x, y, \cdot)(\cdot)$ for every $(x, y) \in \mathbb{R}^{d_1+d_1}$. As in the previous section, for any $r > 0$, let B_r denote the closed ball of radius r in $\mathbb{R}^{d_1}$ centered on the origin.

We propose the following recursion, which performs Newton update estimation on the fast timescale and the Newton update on the slow timescale. The recursion is given by

$$X_{n+1} - X_n = -b(n)Y_n, \tag{41a}$$

$$Y_{n+1} - Y_n - a(n)M_{n+1} \in -a(n)\left(H(X_n, S'_n)^T\left[H(X_n, S''_n)Y_n - \nabla J(X_n, S_n)\right] \oplus B_\delta\right), \tag{41b}$$

where the step-size sequences $\{a(n)\}_{n\geq 0}$ and $\{b(n)\}_{n\geq 0}$ are chosen such that they satisfy Assumption A5 and B_δ, the closed ball of radius δ centered on the origin in $\mathbb{R}^{d_1}$, denotes the nondiminishing, bounded error component in the estimation of the gradient/hessian, whereas $\{M_n\}$ denotes the diminishing error component in the estimation and is assumed to satisfy Assumption A6. Further, $\{S'_n\}$, $\{S''_n\}$, and $\{S_n\}$ are chosen such that they are independent sampling processes with the same sampling law as $\{S_n\}$, by which we mean that at iteration n, if X_n, Y_n are the solution estimate and the current update estimate, respectively, and S_n, S'_n, and S''_n are the current samples, the samples S_{n+1}, S'_n, and S''_n are sampled independently from the distributions $\Pi(X_n, Y_n, S_n)$, $\Pi(X_n, Y_n, S'_n)$, and $\Pi(X_n, Y_n, S''_n)$, respectively.

One can easily show that the set-valued drift function in Recursion (41b) is upper semicontinuous with convex and compact set values. The linear growth property of the drift function follows from the maximum eigenvalue bound and the linear growth property of the map $(x, s) \to \nabla J(x, s)$. If the iterates remain stable for a.e. ω, by Lemma 18, we have that for a.e. ω, there exists a nonempty compact set $A \subseteq \mathbb{R}^{d_1}$ such that $(Y_n(\omega), X_n(\omega)) \to A$ as $n \to \infty$ and A is internally chain transitive for the flow of DI,

$$\begin{pmatrix} \frac{dy}{dt} \\ \frac{dx}{dt} \end{pmatrix} \in \begin{pmatrix} -\nabla_y\left(\|H_\mu(x)y - \nabla J_\mu(x)\|^2\right) \oplus B_\delta \\ 0 \end{pmatrix}, \tag{42}$$

where ∇_y denotes the gradient with respect to y. For every $x \in \mathbb{R}^{d_1}$, the differential inclusion given by

$$\frac{dy}{dt} \in -\nabla_y\left(\|H_\mu(x)y - \nabla J_\mu(x)\|^2\right) \oplus B_\delta \tag{43}$$

is a δ-inflated system (see Kloeden and Kozyakin [22] for a definition) of the o.d.e.,

$$\frac{dy}{dt} = -\nabla_y\left(\|H_\mu(x)y - \nabla J_\mu(x)\|^2\right),$$

and clearly the point $H_\mu(x)^{-1}\nabla J_\mu(x)$ is a global attractor of the above o.d.e..

From the result of Kloeden and Kozyakin [22] on continuity of attractors, we have that for any (x, δ), the δ-inflated dynamical system as in Equation (43) has a global attractor (denoted by $\lambda_\delta(x) \subseteq \mathbb{R}^{d_1}$), and for any sequence $\{(x_n, \delta_n)\}$ converging to (x, δ), $\mathbf{H}(\lambda_{\delta_n}(x_n), \lambda_\delta(x))$ converges to zero, where $\mathbf{H}(\cdot)$ denotes the Hausdorff metric on the family of compact and convex subsets of $\mathbb{R}^d$ (see equation (13) in Yaji and Bhatnagar [42] for a definition). Clearly, for any x, $\lambda_0(x) = H_\mu(x)^{-1}\nabla J_\mu(x)$. Further, we assume that for any $\delta_1 > 0$, there exists a δ small enough such that the Hausdorff distance between $\lambda_\delta(x)$ and $\lambda_0(x)$ is less than δ_1 uniformly for all x. (Although we state this as an assumption, under the assumptions on the hessian matrix H_μ imposed above, we feel that the analysis in Kloeden and Kozyakin [22] can be modified to obtain the same.) From the above and the linear growth property of the gradient $\nabla J_\mu(\cdot)$, it now follows that the map $x \to \lambda_\delta(x)$ satisfies Assumption A9.

Theorem 4 now gives us that for a.e. ω, the iterates $X_n(\omega)$ converge to an internally chain transitive set of the DI,

$$\begin{aligned}\frac{dx}{dt} &\in \bar{co}(\lambda_\delta(x)) \\ &\subseteq H_\mu(x)^{-1}\nabla J_\mu(x) \oplus B_{\delta_1},\end{aligned} \tag{44}$$

where the last inclusion follows from choosing δ small enough. The lower bound on the minimum eigenvalue of the hessian matrix $H_\mu(\cdot)$ gives us that the mean field of the DI (44) is a Marchaud map. From the result on continuity of attractors in Li and Zhang [26], we have that for any $\epsilon > 0$, there exists δ_1 small enough such that the DI (44) has a global attractor A_{δ_1} such that $A_{\delta_1} \subseteq A_0 \oplus B_\epsilon$, where $A_0 = \{x \in \mathbb{R}^{d_1} : \nabla J_\mu(x) = 0\}$, the global attractor of the o.d.e.,

$$\frac{dx}{dt} = H_\mu(x)^{-1}\nabla J_\mu(x).$$

Thus, by choosing $\delta > 0$ small enough, the iterates X_n of Recursion (41) converge to the ϵ-perturbed solution of the optimization problem given by

$$\min_{x\in\mathbb{R}^{d_1}} J_\mu(x).$$

8. Conclusions and Directions for Future Work

We have presented a detailed analysis of a two-timescale stochastic recursive inclusion with set-valued drift functions and in the presence of nonadditive iterate-dependent Markov noise with nonunique stationary distributions. The analysis in Section 5 shows us that the asymptotic behavior of the two-timescale recursion (15) is such that the faster timescale iterates in recursion (15b) track the flow of DI (17) for some fixed value of the slower timescale variable, and the slower timescale iterates track the flow of DI (20). The assumptions under which the two-timescale recursion is studied in this paper are weaker than those in current literature. Recursions with such behavior are often required to solve nested minimization problems that arise in machine learning and optimization. A special case of constrained convex optimization with linear constraints is considered as an application where the objective function is not assumed to be differentiable, and further, the objective function and constraints are averaged with respect to stationary distribution of an underlying Markov chain. When the transition law and hence the stationary distribution is not known in advance, a primal descent-dual ascent algorithm as in Recursion (36) can be implemented with the knowledge of the sample paths of the underlying Markov chain, and the analysis presented in this paper guarantees convergence to an ϵ-optimal solution for a user specified choice of ϵ. Further, another application is presented where we show that the Newton's method to compute a solution to an unconstrained convex optimization problem with nondiminishing bounded noise model for the errors in gradient and hessian estimation converges to an ϵ-optimal solution if the bound on the nondiminishing error is sufficiently small, while not requiring an explicit inversion of the hessian.

We outline a few important directions for future work.

1. For two-timescale stochastic approximation schemes with set-valued mean fields, to the best of our knowledge, there are no sufficient conditions for stability in the current literature. We believe extensions of the stability result for single-timescale stochastic approximation by Borkar and Meyn [13] and Ramaswamy and Bhatnagar [35] can be made to the case of two-timescale recursions. Another approach to stability could be along the lines of Andrieu et al. [1].

2. In many applications, the iterates are projected at each time step and are ensured to remain within a compact, convex set. Such projections often arise due to the inherent need of the application at hand or are used to ensure stability. Such projected schemes tend to introduce spurious equilibrium points at the boundary of the feasible set. Further complications arise due to the presence of Markov noise terms; most of the time, the projection map is not differentiable, and only directional derivatives are known to exist. Such projected stochastic approximation schemes for a single-valued case without Markov noise component are analyzed by Nagurney and Zhang [31] and should serve as a basis for analyzing more general frameworks with projection.

3. In some applications arising in reinforcement learning, the noise terms are not Markov by themselves, but their lack of Markov property comes through the dependence on a control sequence in addition to the

iterate sequence. Under such controlled Markov noise assumption, the two-timescale stochastic approximation scheme has been analyzed by Karmakar and Bhatnagar [21] but with single valued, Lipschitz continuous drift functions. Extending the analysis presented in this paper to the case with set-valued drift functions and controlled Markov noise is straightforward and requires no major change in the overall flow of the analysis. This extension allows one to analyze the asymptotic behavior of a larger class of reinforcement learning algorithms (see Perkins and Leslie [34]).

4. Several other applications, such as two-timescale controlled stochastic approximation and two-timescale approximate drift problem also can be analyzed with the help of the results presented in this paper (see Borkar [11, chapter 5.3] for definitions of the above).

Appendix A. Proof of Lemma 21

Fix $\omega \in \Omega_1$, $l \geq 1$, and $T > 0$. We prove the claim along the sequence $\{t^s(n)\}_{n\geq 1}$ from which the claim of Lemma 21 easily follows.

Fix $n \geq 0$. Let $\tau^2(n,T) := \min\{m > n : t^s(m) \geq t^s(n) + T\}$. Let $q \in [0,T]$. Then there exists k such that $t^s(n) + q \in [t^s(k), t^s(k+1))$ and $n \leq k \leq \tau^2(n,T) - 1$. By the definition of $\bar{y}(\cdot)$ and $\tilde{y}^{(l)}(\cdot; t^s(n))$, we have that $\bar{y}(t^s(n)+q) = \alpha y_k + (1-\alpha) y_{k+1}$ and $\tilde{y}^{(l)}(q; t^s(n)) = \alpha \tilde{y}^{(l)}(t^s(k) - t^s(n); t^s(n)) + (1-\alpha)\tilde{y}^{(l)}(t^s(k+1) - t^s(n); t^s(n))$, where $\alpha = \frac{t^s(k+1) - t^s(n) - q}{t^s(k+1) - t^s(k)}$. Because $\tilde{y}^{(l)}(\cdot; t^s(n))$ is a solution of the o.d.e. (29), we have that for every $k \geq n$, $\tilde{y}^{(l)}(t^s(k) - t^s(n); t^s(n)) = y_n + \sum_{j=n}^{k-1} b(j) h_2^{(l)}(x_j, y_j, s_j^{(2)}, u_j^{(l)})$, and by Lemma 19, we have that for every $k \geq n$, $y_k = \bar{y}(t^s(k)) = y_n + \sum_{j=n}^{k-1} b(j) h_2^{(l)}(x_j, y_j, s_j^{(2)}, u_j^{(l)}) + \sum_{j=n}^{k-1} b(j) m_{j+1}^{(2)}$. Thus,

$$\begin{aligned}
\left\|\bar{y}(t^s(n)+q) - \tilde{y}^{(l)}(q; t^s(n))\right\| &\leq \left\|\alpha \sum_{j=n}^{k-1} b(j) m_{j+1}^{(2)} + (1-\alpha) \sum_{j=n}^{k} b(j) m_{j+1}^{(2)}\right\| \\
&\leq \alpha \left\|\sum_{j=n}^{k-1} b(j) m_{j+1}^{(2)}\right\| + (1-\alpha)\left\|\sum_{j=n}^{k} b(j) m_{j+1}^{(2)}\right\| \\
&\leq \sup_{n \leq k \leq \tau(n,T)} \left\|\sum_{j=n}^{k} b(j) m_{j+1}^{(2)}\right\|.
\end{aligned}$$

Because the R.H.S. of the above inequality is independent of $q \in [0,T]$, we have $\sup_{0\leq q\leq T} \|\bar{y}(t^s(n)+q) - \tilde{y}^{(l)}(q; t^s(n))\| \leq \sup_{n\leq k\leq \tau^2(n,T)} \|\sum_{j=n}^{k} b(j) m_{j+1}^{(2)}\|$. Therefore, $\lim_{n\to\infty} \sup_{0\leq q\leq T} \|\bar{y}(t^s(n)+q) - \tilde{y}^{(l)}(q; t^s(n))\| \leq \lim_{n\to\infty} \sup_{n\leq k\leq \tau^2(n,T)} \|\sum_{j=n}^{k} b(j) m_{j+1}^{(2)}\|$. Now the claim follows from Assumption A7.

Appendix B. Proof of Lemma 22

Fix $l \geq 1$, $\omega \in \Omega_1$. By Assumption A8, we know that there exists $r > 0$ such that $\sup_{n\geq 0}(\|x_n\| + \|y_n\|) \leq r$, and hence $\sup_{t\geq 0} \tilde{y}^{(l)}(0; t) = \sup_{t\geq 0} \bar{y}(t) \leq r$.

For any $t \geq 0$, let $[t] := \max\{n \geq 0 : t^s(n) \leq t\}$. For every $t \geq 0$ and $q_1, q_2 \in [0, \infty)$ (w.l.o.g. assume $q_1 < q_2$), we have

$$\begin{aligned}
\left\|\tilde{y}^{(l)}(q_1; t) - \tilde{y}^{(l)}(q_2; t)\right\| &= \left\|\int_{q_1}^{q_2} h_2^{(l)}\left(x_{[t+q]}, y_{[t+q]}, s_{[t+q]}^{(2)}, u_{[t+q]}^{(l)}\right) dq\right\| \\
&\leq \int_{q_1}^{q_2} \left\|h_2^{(l)}\left(x_{[t+q]}, y_{[t+q]} s_{[t+q]}^{(2)}, u_{[t+q]}^{(l)}\right)\right\| dq \\
&\leq \int_{q_1}^{q_2} K^{(l)}(1 + \|x_{[t+q]}\| + \|y_{[t+q]}\|) dq \\
&\leq C^{(l)}(q_2 - q_1),
\end{aligned}$$

where $C^{(l)} := K^{(l)}(1+r)$ and $r > 0$ is such that $\sup_{n\geq 0}(\|x_n\| + \|y_n\|) \leq r$. Thus, $\{\tilde{y}^{(l)}(\cdot; t)\}_{t\geq 0}$ is an equicontinuous family. Now the claim follows from the Arzella-Ascoli theorem.

References

[1] Andrieu C, Moulines É, Priouret P (2005) Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* 44(1): 283–312.

[2] Aubin JP, Cellina A (2012) *Differential Inclusions: Set-Valued Maps and Viability Theory*, vol. 264 (Springer Science & Business Media, New York).

[3] Benaim M (1996) A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.* 34(2):437–472.

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

[5] Benzi M, Golub GH, Liesen J (2005) Numerical solution of saddle point problems. *Acta Numerica* 14:1–137.

[6] Bertsekas DP (2009) *Convex Optimization Theory* (Athena Scientific, Belmont, MA).

[7] Bhatnagar S, Prashanth L (2015) Simultaneous perturbation newton algorithms for simulation optimization. *J. Optim. Theory Appl.* 164(2):621–643.

[8] Borkar VS (1989) *Optimal Control of Diffusion Processes.* Pitman Lecture Notes in Mathematics, vol. 203 (Longman Scientific and Technical, Harlow, UK).

[9] Borkar VS (1997) Stochastic approximation with two time scales. *Systems Control Lett.* 29(5):291–294.

[10] Borkar VS (2006) Stochastic approximation with controlled Markov noise. *Systems Control Lett.* 55(2):139–145.

[11] Borkar VS (2008) *Stochastic Approximation: A Dynamical Systems Viewpoint* (Cambridge University Press, Cambridge, UK).

[12] Borkar VS (2012) *Probability Theory: An Advanced Course* (Springer Science & Business Media, New York).

[13] Borkar VS, Meyn SP (2000) The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.* 38(2):447–469.

[14] Chen HF, Guo L, Gao AJ (1987) Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes Their Appl.* 27:217–231.

[15] Chen HF, Yunmin Z (1986) Stochastic approximation procedures with randomly varying truncations. *Sci. China Ser. A: Math., Phys., Astron., Tech. Sci.* 29(9):914–926.

[16] Derevitskii D, Fradkov AL (1974) Two models analyzing the dynamics of adaptation algorithms. *Avtomatika i Telemekhanika* (1):67–75.

[17] Duflo M (2013) *Random Iterative Models*, vol. 34 (Springer Science & Business Media, New York).

[18] Fort G, Moulines E, Schreck A, Vihola M (2016) Convergence of markovian stochastic approximation with discontinuous dynamics. *SIAM J. Control Optim.* 54(2):866–893.

[19] Iusem A, Jofré A, Oliveira RI, Thompson P (2017) Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.* 27(2):686–724.

[20] Jiang H, Xu H (2008) Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Trans. Automatic Control* 53(6):1462–1475.

[21] Karmakar P, Bhatnagar S (2017) Two time-scale stochastic approximation with controlled markov noise and off-policy temporal-difference learning. *Math. Oper. Res.* 43(1):130–151.

[22] Kloeden P, Kozyakin V (2000) The inflation of attractors and their discretization: The autonomous case. *Nonlinear Anal.: Theory Methods Appl.* 40(1–8):333–344.

[23] Koshal J, Nedic A, Shanbhag UV (2013) Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Trans. Automatic Control* 58(3):594–609.

[24] Kumaresan S (2005) *Topology of Metric Spaces* (Alpha Science International, Oxford, UK).

[25] Kushner H, Yin G (2003) *Stochastic approximation and recursive algorithms and applications* (Springer, New York).

[26] Li D, Zhang X (2002) On dynamical properties of general dynamical systems and differential inclusions. *J. Math. Anal. Appl.* 274(2):705–724.

[27] Li S, Ogura Y, Kreinovich V (2013) *Limit Theorems and Applications of Set-Valued and Fuzzy Set-Valued Random Variables*, vol. 43 (Springer Science & Business Media, New York).

[28] Metivier M, Priouret P (1984) Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Trans. Inform. Theory* 30(2):140–151.

[29] Meyn SP, Tweedie RL (2012) *Markov Chains and Stochastic Stability* (Springer Science & Business Media, New York).

[30] Milgrom P, Segal I (2002) Envelope theorems for arbitrary choice sets. *Econometrica* 70(2):583–601.

[31] Nagurney A, Zhang D (2012) *Projected Dynamical Systems and Variational Inequalities with Applications*, vol. 2 (Kluwer Academic Publishers, Norwell, MA).

[32] Nedić A, Ozdaglar A (2009) Subgradient methods for saddle-point problems. *J. Optim. Theory Appl.* 142(1):205–228.

[33] Parthasarathy KR (1967) *Probability Measures on Metric Spaces*, vol. 352 (American Mathematical Society, Providence, RI).

[34] Perkins S, Leslie DS (2012) Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems* 2(2):409–446.

[35] Ramaswamy A, Bhatnagar S (2016) A generalization of the borkar-meyn theorem for stochastic recursive inclusions. *Math. Oper. Res.* 42(3):648–661.

[36] Ramaswamy A, Bhatnagar S (2016) Stochastic recursive inclusion in two timescales with an application to the Lagrangian dual problem. *Stochastics* 88(8):1173–1187.

[37] Robbins H, Monro S (1951) A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

[38] Rudin W (1976) *Principles of mathematical analysis*. International Series in Pure and Applied Mathematics (McGraw-Hill, New York).

[39] Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control* 37(3):332–341.

[40] Sutton RS, Maei HR, Precup D, Bhatnagar S, Silver D, Szepesvári C, Wiewiora E (2009) Fast gradient-descent methods for temporal-difference learning with linear function approximation. *Proc. 26th Annual Internat. Conf. Machine Learning* (ACM, New York), 993–1000.

[41] Yaji VG, Bhatnagar S (2018) Stochastic recursive inclusions with non-additive iterate-dependent markov noise. *Stochastics* 90(3):330–363.

[42] Yaji VG, Bhatnagar S (2020) Analysis of stochastic approximation schemes with set-valued maps in the absence of a stability guarantee and their stabilization. *IEEE Trans. Automatic Control* 65(3):1100–1115.