

STOCHASTIC APPROXIMATION WITH DISCONTINUOUS DYNAMICS, DIFFERENTIAL INCLUSIONS, AND APPLICATIONS

BY NHU NGUYEN^{1,a} AND GEORGE YIN^{2,b}

¹*Department of Mathematics and Applied Mathematical Sciences, University of Rhode Island, ^anhu.math.2611@gmail.com*

²*Department of Mathematics, University of Connecticut, ^bgyin@uconn.edu*

This work develops new results for stochastic approximation algorithms. The emphases are on treating algorithms and limits with discontinuities. The main ingredients include the use of differential inclusions, set-valued analysis, nonsmooth analysis, and stochastic differential inclusions. Under broad conditions, it is shown that a suitably scaled sequence of the iterates has a differential inclusion limit. In addition, it is shown for the first time that a centered and scaled sequence of the iterates converges weakly to a stochastic differential inclusion limit. The results are then used to treat several application examples including Markov decision processes, Lasso algorithms, Pegasos algorithms, support vector machine classification, and learning. Some numerical demonstrations are also provided.

1. Introduction. This paper examines stochastic approximation from new angles. One of the main motivations stems from the minimization of a nondifferentiable function or finding the zeros of a set-valued mapping corrupted with random disturbances. In contrast to the existing literature, this paper focuses on stochastic approximation with discontinuous dynamics and set-valued mappings, and develops new techniques for analyzing algorithms involving set-valued analysis and stochastic differential inclusions.

Let us begin with a stochastic approximation algorithm of the form

$$(1.1) \quad \mathbf{X}_{n+1} = \mathbf{X}_n + a_n \mathbf{b}_n(\mathbf{X}_n, \xi_n),$$

and the corresponding projection algorithm

$$(1.2) \quad \mathbf{X}_{n+1} = \Pi_H(\mathbf{X}_n + a_n \mathbf{b}_n(\mathbf{X}_n, \xi_n)),$$

where H is a constraint set and Π_H is the projection onto the set H . Introduced by Robbins–Monro in [47] in 1951, stochastic approximation algorithms have been studied extensively with a wide range of applications [3, 7, 36, 37, 39, 40]. In addition to the traditional areas, recent applications also include cooperative dynamics and games [4] and multilevel Monte Carlo methods [19, 25]. When the sequence $\mathbf{b}_n(\cdot, \xi_n)$ and the associated “average” satisfy some smoothness conditions, the asymptotic properties of the algorithms are relatively well-understood [36, 37, 39]. We refer to such cases as “stochastic approximation with continuous dynamics and continuous limits.” When $\mathbf{b}_n(\cdot, \xi_n)$ is not necessarily continuous but its average is, the analysis can be found in [35, 37, 40]. We refer to such cases as “stochastic approximation with discontinuous dynamics and continuous limits.” In [37], the case that $\mathbf{b}_n(\cdot, \xi_n)$ is continuous while the limits belong to a set-valued mapping is considered and is referred to as “stochastic approximation with continuous dynamics but discontinuous limits.” In many applications, we need to handle discontinuous functions and/or set-valued mappings for both

Received May 2021; revised February 2022.

MSC2020 subject classifications. Primary 62L20, 60H10, 60J60; secondary 34A60.

Key words and phrases. Stochastic approximation, stochastic subgradient descent, differential inclusion, stochastic differential inclusion, convergence, rate of convergence.

$\mathbf{b}_n(\cdot, \xi_n)$ and its average. We refer to such cases as “stochastic approximation with discontinuous dynamics and discontinuous limits.”

Numerous systems and practical problems require optimizing nonsmooth functions such as dimension reduction in high-dimensional statistics (the L^1 -norm regularized term) [20], support vector machines classification (the hinge loss function), neural networks (the rectified linear unit), collaborative filtering and recommender systems (various types of matrix regularizers) [30], complementarity problems [11], compressed modes in physics, and partial consensus problems [23], among others. Because the objective functions are not continuously differentiable, the gradient-based methods are often replaced by subgradient-based counterparts. As can be seen, discontinuous dynamics and set-valued mappings are ubiquitous in optimization problems. There are also numerous problems and algorithms in control engineering, economics, and operations research that require the treatment of discontinuous dynamics and/or set-valued mappings; see learning algorithms in Markov decision processes [45], algorithms in approachability theory and the study of fictitious play in game theory [5, 6], and adaptive filtering [7, 18, 37, 56] arising from signal processing. As a concrete example, we consider the sign-error algorithm [58] frequently used in adaptive filtering. Let $\varphi_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}$ be measured output and reference signals, respectively, with $\{(y_n, \varphi_n)\}$ being a stationary process. We adjust a system parameter $\theta \in \mathbb{R}^d$ adaptively so that the weighted output $\theta^\top \varphi_n$ best matches the reference signal y_n in the sense that a L^1 cost function is minimized. The algorithm reads

$$\theta_{n+1} = \theta_n + a_n \varphi_n \operatorname{sign}(y_n - \varphi_n^\top \theta_n),$$

where $\operatorname{sign}(y) = \mathbf{1}_{\{y>0\}} - \mathbf{1}_{\{y<0\}}$ is the sign operator used for reducing the computational complexity.

Note that randomness can affect samplings, mini-batching computations, partial observations, noisy measurements, and many other sources. As was mentioned, various functions involved in applications could possibly be nonsmooth or even not continuous. Thus, it is necessary to study stochastic approximation algorithms (1.1) and (1.2) with both $\mathbf{b}_n(\cdot, \cdot)$ and its averages being discontinuous functions and/or set-valued mappings.

With the aforementioned motivations, this paper formulates the problem by using a general and unified setting, introduces new techniques, proves convergence under mild conditions, and establishes rates of convergence of stochastic approximations with possibly discontinuous dynamics and discontinuous limits. Both constrained, unconstrained, and biased algorithms are considered. To be more specific, using appropriate piecewise linear and piecewise constant interpolations, we prove the boundedness and equicontinuity of the sequences in a functional space. The compactness enables us to extract a convergent subsequence. To characterize the limits, most existing works in the literature use continuity for either the dynamics or the limit systems or both. If the dynamics are not continuous but the limit systems have enough regularity, Kushner used an “averaging method” in [35] to handle this problem under some conditions on the existence of certain Lyapunov functions. In contrast, Métivier and Priouret in [40] used a probabilistic approach by averaging out the noise with respect to the invariant measure. To analyze algorithms with both the dynamics and limits being discontinuous, we need a new approach. In this paper, we use ordinary differential equations (ODEs) with discontinuous right-hand sides, differential inclusions, set-valued dynamical systems, and convex analysis to characterize the asymptotic behavior of the algorithms. To obtain the stability, we use results of stability for differential inclusions together with novel concepts and techniques from nonsmooth analysis. Moreover, we examine biased stochastic approximation using continuation of chain recurrent sets in set-valued dynamic systems. Furthermore, we ascertain the rates of convergence by using the theory of stochastic differential inclusions and the newly developed theory of variational analysis.

Related works and our contributions. To proceed, we highlight our contributions and novelties of the paper in contrast to the existing literature.

- Although the algorithms involving discontinuous dynamics and set-valued mappings were considered in [36], continuity in an appropriate sense of set-valued mappings was needed. The continuity, however, may fail in applications. Except [5], there has been no general approach in the literature for studying convergence of stochastic approximation schemes involving set-valued mappings without continuity. Although the aforementioned works deal with differential inclusions, the setup and results of the current paper are different than that of [5]. Using our approach, it is possible to recover the setting in [5]; see Remark 3. Moreover, the limit processes in [5] were shown to be perturbed solutions of differential inclusions, whereas in the current paper, the limit processes are solutions (rather than perturbed solutions) of the limit differential inclusions. Our convergence analysis is done partially by examining the closure of the set of solutions of a family of differential inclusions for general set-valued mappings, which is a crucial point in the development. Both constrained and unconstrained algorithms are considered in this paper.
- To prove the convergence to the equilibrium point, the stability of differential inclusions corresponding to stochastic approximation schemes is carefully investigated using a Lyapunov functional method that is novel and not considered in the existing literature of stochastic approximation. To be more specific, we use a \mathcal{U} -generalized Lyapunov functional. Our approaches and results appear to be more effective and easily applicable; see the examples in Section 4. The idea behind this approach is that one can ignore some “less important” points that do not affect the stability of the dynamics.
- We consider biased stochastic approximation with discontinuous dynamics and set-valued mappings. Although biased stochastic approximation counterpart with smooth dynamics was dealt with in [53], to the best of our knowledge, this is the first time biased stochastic approximation in conjunction with set-valued mappings without continuity is treated.
- In addition, this work provides rates of convergence for the algorithms with discontinuities and set-valued mappings. Stochastic differential inclusions are used for the first time to ascertain the rates of convergence of stochastic approximation algorithms.
- With a wide range of applications in mind, we provide a unified framework and new approaches to analyze convergence, rates of convergence, robustness for stochastic and nonsmooth optimization problems, and algorithms involving discontinuous dynamics and set-valued mappings. The applications considered in the paper include algorithms in machine learning and Markov decision processes. For applications to machine learning algorithms, we provide new insights in analyzing these algorithms by characterizing the limit behavior and rates of convergence using the dynamic systems generated by differential inclusions and stochastic differential inclusions. Our work fills in the gap for studying convergence of algorithms with nonsmooth loss functions. Treating Markov decision processes, we demonstrate how to apply our results for multistage decision making with partial observations.

Outline of the paper. The rest of paper is arranged as follows. Section 2 obtains convergence of stochastic approximation algorithms with the emphasis on discontinuity and set-valued mappings. Section 3 ascertains rates of convergence leading to the stochastic differential inclusion limits. Section 4 examines a number of applications together with the corresponding numerical results. Section 5 summarizes our findings and provides further remarks. Finally, mathematical background on ODEs with discontinuous right-hand sides, differential inclusions, nonsmooth analysis, set-valued dynamic systems, and stochastic differential inclusions are summarized in Appendix 5.

2. Convergence. Denote by \mathbb{R}^d the d -dimensional Euclidean space with the usual Euclidean norm $|\cdot|$, and let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$ be a complete filtered probability space satisfying the usual conditions. Consider the following general stochastic approximation algorithm:

$$(2.1) \quad \mathbf{X}_{n+1} = \mathbf{X}_n + a_n \mathbf{b}_n(\mathbf{X}_n, \xi_n) + a_n \mathbf{h}(\mathbf{X}_n, \zeta_n) + a_n \mathbf{h}_0(\tilde{\zeta}_n) + a_n \boldsymbol{\beta}_n,$$

and the associated projection algorithm

$$(2.2) \quad \begin{cases} \tilde{\mathbf{X}}_{n+1} = \mathbf{X}_n + a_n \mathbf{b}_n(\mathbf{X}_n, \xi_n) + a_n \mathbf{h}(\mathbf{X}_n, \zeta_n) + a_n \mathbf{h}_0(\tilde{\zeta}_n) + a_n \boldsymbol{\beta}_n, \\ \mathbf{X}_{n+1} = \Pi_H(\tilde{\mathbf{X}}_{n+1}), \end{cases}$$

where Π_H is the projection operator (orthogonal projection into the set H), and H is a closed and nonempty subset of \mathbb{R}^d . The $\{a_n\}$ is a sequence of step sizes (a sequence of positive real numbers) satisfying $a_n \rightarrow 0$ and $\sum_{n=1}^{\infty} a_n = \infty$. The sequences $\{\xi_n\}$, $\{\zeta_n\}$, and $\{\tilde{\zeta}_n\}$ noise processes that are correlated in time but independent of each other, and $\{\boldsymbol{\beta}_n\}$ represents the bias; see [36, 37]. In the literature, $\boldsymbol{\beta}_n$ is often formulated as a diminishing bias so that it tends to 0 w.p.1. However, there are cases that one has to face asymptotically nonzero bias in the sense $\lim_{n \rightarrow \infty} \|\boldsymbol{\beta}_n\| > 0$.

Motivated by many applications, the functions $\mathbf{b}_n(\cdot, \cdot)$ are allowed to be discontinuous and/or set-valued mappings, which can be used to represent subgradients of nondifferentiable components in the loss function, whereas $\mathbf{h}(\cdot, \cdot)$ is a continuous function (in \mathbf{x}) representing the gradient of the smooth parts in the loss function. The discontinuity of $\mathbf{b}_n(\cdot, \cdot)$ and/or set-valued mappings appear frequently in applications. Dealing with such functions and mappings is one of the main objectives of this paper.

Notation. Similar to [36, 37], define $t_0 = 0$ and for $n \geq 1$, $t_n := \sum_{i=0}^{n-1} a_i$, $m(t) := \max\{n : t_n \leq t\}$ if $t \geq 0$ and $m(t) := 0$ if $t < 0$; and define the piecewise constant interpolation $\bar{\mathbf{X}}^0(t)$ and the piecewise linear interpolation $\mathbf{X}^0(t)$ of \mathbf{X}_n with interpolation intervals $\{a_n\}$ as

$$\begin{aligned} \bar{\mathbf{X}}^0(t) &:= \mathbf{X}_n \quad \text{in } [t_n, t_{n+1}), \\ \mathbf{X}^0(t_n) &:= \mathbf{X}_n \quad \text{and} \quad \mathbf{X}^0(t) := \frac{t_{n+1} - t}{a_n} \mathbf{X}_n + \frac{t - t_n}{a_n} \mathbf{X}_{n+1} \quad \text{in } (t_n, t_{n+1}), \end{aligned}$$

respectively, and define the shift sequence $\mathbf{X}^n(\cdot)$ on $(-\infty, \infty)$ as

$$\mathbf{X}^n(t) := \begin{cases} \mathbf{X}^0(t + t_n) & \text{if } t \geq -t_n, \\ \mathbf{X}_0 & \text{if } t \leq -t_n. \end{cases}$$

For two sets S , S_1 , and either a set-valued or vector-valued mapping F , and a real number k , we define $S + S_1 := \{\mathbf{x} + \mathbf{y} : \mathbf{x} \in S, \mathbf{y} \in S_1\}$, and $F(S) := \bigcup_{\mathbf{x} \in S} F(\mathbf{x})$, and $kS := \{k\mathbf{x} : \mathbf{x} \in S\}$. Throughout the paper, B denotes the unit open ball $B = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| < 1\}$ and \bar{B} is its closure; “co” is the convex hull and “ $\overline{\text{co}}$ ” is the convex closure; $2^{\mathbb{R}^d}$ is the collection of all subsets of \mathbb{R}^d . To analyze the convergence, we present the following standard assumptions first.

REMARK 1. Reference label convention. Throughout the paper, we use several sets of assumptions. To facilitate the reading, we shall use the following conventions. Conditions headed by (A) corresponds to standard assumptions; conditions headed by (K), (G), and (P) are assumptions involving *Krasovskii* operator, *general* set-valued mapping, and *projection*, respectively; conditions headed by (KS), (GS), and (PS) are stability assumptions corresponding to that of (K), (G), and (P), respectively; conditions headed by (R) are for the rates of convergence study.

- (A) (i) $\mathbf{h}(\cdot, \zeta)$ is continuous in \mathbf{x} , uniformly in ζ on bounded sets of \mathbf{x} .
 (ii) Either $\mathbf{h}(\cdot, \cdot)$ is a bounded measurable function or there are nonnegative measurable functions $g_1(\cdot)$ of \mathbf{x} , and $g_2(\cdot)$ and $g_3(\cdot)$ of ζ such that $g_1(\cdot)$ is bounded on bounded sets (of \mathbf{x}) and

$$(2.3) \quad |\mathbf{h}(\mathbf{x}, \zeta)| \leq g_1(\mathbf{x})g_2(\zeta) + g_3(\zeta),$$

and for each $\varepsilon > 0$,

$$(2.4) \quad \lim_{\Delta \rightarrow 0} \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{j \geq n} \max_{t \leq \Delta} \sum_{i=m(j\Delta)}^{m(j\Delta+t)-1} a_i [g_2(\zeta_i) + g_3(\zeta_i)] \geq \varepsilon \right\} = 0.$$

- (iii) There exists a continuous function $\bar{\mathbf{h}}(\cdot)$ such that for some $T > 0$, each $\varepsilon > 0$, and each \mathbf{x} ,

$$(2.5) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} a_i (\mathbf{h}(\mathbf{x}, \zeta_i) - \bar{\mathbf{h}}(\mathbf{x})) \right| \geq \varepsilon \right\} = 0.$$

- (iv) The $\{\xi_n\}$, $\{\zeta_n\}$, $\{\tilde{\zeta}_n\}$ are sequences of independent and exogenous noises, and the function $\mathbf{h}_0(\cdot)$ is measurable such that for some $T > 0$ and each $\varepsilon > 0$,

$$(2.6) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} a_i \mathbf{h}_0(\tilde{\zeta}_i) \right| \geq \varepsilon \right\} = 0.$$

By exogenous noises, we mean that the distribution of $\{\xi_i, i > n\}$ conditioned on $\{\xi_i, \mathbf{X}_n : i \leq n\}$ is the same as that of $\{\xi_i, i > n\}$ conditioned on $\{\xi_i : i \leq n\}$ and similar assumptions for ζ_n and $\tilde{\zeta}_n$.

- (v) The $\{\beta_n\}$ is a sequence of bounded random variables satisfying $|\beta_n| \rightarrow 0$ w.p.1.

REMARK 2. Assumption (A) together with the boundedness of the iterates $\{\mathbf{X}_n\}$ or a projection algorithm (e.g., Assumption (P) given later) presents broad conditions, which guarantee the boundedness and equicontinuity of $\{\mathbf{X}^n(\cdot)\}$. Sufficient conditions guaranteeing the boundedness can be provided; see [36], Section 4.7 and Theorem 4.7.4, (see also [39]) or using a projection algorithm [36, 37]. To handle nonexogenous noise, the reader can consult [37], Section 6.6, for the treatment of state-dependent noise. In this paper, for simplicity, we will not deal with such cases. The noise processes $\{\xi_n\}$, $\{\zeta_n\}$, $\{\tilde{\zeta}_n\}$ take values in some measurable spaces. However, because we do not assume any regularity of functions \mathbf{b} , \mathbf{h} , \mathbf{h}_0 on these variables, we often do not specify these spaces. Moreover, one can combine $\mathbf{h}_0(\tilde{\zeta}_n)$ and β_n , however, due to their motivations in application (one presents the noise and the other presents the bias), we still keep these two different terms. Assumption (A)(v) (as well as (2.4), (2.5), (2.6)) can be relaxed, which will be considered later.

Convergence. Now, we state our main convergence results; some preliminary results and concepts are relegated to Appendix 5. We use $\mathcal{C}^d(-\infty, \infty)$ to denote the space of \mathbb{R}^d -valued continuous functions defined on $(-\infty, \infty)$, and $D(-\infty, \infty)$ and $D[0, \infty)$ to denote the spaces of real-valued functions defined on $(-\infty, \infty)$ and $[0, \infty)$, respectively, which are right continuous and have left limits, endowed with the Skorohod topology. We use $D^d(-\infty, \infty)$ (resp., $D^d[0, \infty)$) to denote the corresponding D spaces taking values in \mathbb{R}^d . The convergence of sequence of functions in $\mathcal{C}^d(-\infty, \infty)$ or $D^d(-\infty, \infty)$ (resp., $D^d[0, \infty)$) is in the sense of weak topology (uniform convergence on bounded intervals).

As was mentioned, the functions $\mathbf{b}_n(\cdot, \cdot)$ are possibly discontinuous and belong to some set-valued mapping so that they can be used to represent subgradients of nondifferentiable

components of the loss function. To illustrate, we first consider the case that this set-valued mapping can be expressed as the Krasovskii operator of some vector-valued function. [For example, subgradient of $|\cdot|$ can be expressed as the Krasovskii operator of the $\text{sign}(\cdot)$ function.] In fact, we allow perturbations of this set-valued mapping, which is presented in the following Assumption (K).

(K) There are a locally bounded function $\bar{\mathbf{b}}(\cdot)$ and a sequence of (positive real-valued) continuous (in \mathbf{x} , uniformly in ξ) functions $\{m_n(\mathbf{x}, \xi)\}$ such that $\forall n, \mathbf{x}, \xi$,

$$\mathbf{b}_n(\mathbf{x}, \xi) \in \mathcal{K}[\bar{\mathbf{b}}](\mathbf{x}) + m_n(\mathbf{x}, \xi)\bar{B},$$

and for some $T > 0$, each $\varepsilon > 0$, and each \mathbf{x} ,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} a_i m_i(\mathbf{x}, \xi_i) \right| \geq \varepsilon \right\} = 0.$$

In the above, $\mathcal{K}[\bar{\mathbf{b}}]$ is the Krasovskii operator of $\bar{\mathbf{b}}$, that is, $\mathcal{K}[\bar{\mathbf{b}}] : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ is defined by

$$\mathcal{K}[\bar{\mathbf{b}}](\mathbf{y}) := \bigcap_{\delta > 0} \text{co} \bar{\mathbf{b}}(B(\mathbf{y}, \delta)),$$

where $B(\mathbf{y}, \delta)$ is the open ball in \mathbb{R}^d with center \mathbf{y} and radius δ . More details on the Krasovskii operator and related results are provided in Section A.1.

THEOREM 2.1. *Consider algorithm (2.1). Assume that (A) and (K) hold and that $\{\mathbf{X}_n\}$ is bounded w.p.1.*

- *Then there is a null set Ω_0 such that $\forall \omega \notin \Omega_0$, $\{\mathbf{X}^n(\cdot)\}$ is bounded and equicontinuous on bounded intervals.*
- *Let $\mathbf{X}(\cdot)$ be the limit of a convergent subsequence of $\{\mathbf{X}^n(\cdot)\}$. Then $\mathbf{X}(t)$ is a Krasovskii solution of*

$$(2.7) \quad \dot{\mathbf{X}}(t) \in \bar{\mathbf{b}}(\mathbf{X}(t)) + \bar{\mathbf{h}}(\mathbf{X}(t)),$$

that is, $\mathbf{X}(\cdot)$ is a solution of the differential inclusion (see Section A.1 for detailed definitions)

$$(2.8) \quad \dot{\mathbf{X}}(t) \in \mathcal{K}[\bar{\mathbf{b}} + \bar{\mathbf{h}}](\mathbf{X}(t)).$$

- *The limit set of $\mathbf{X}(\cdot)$ is internally chain transitive (with respect to (2.8)) and the limit points of $\{\mathbf{X}_n\}$ are contained in \mathcal{R} , the set of chain-recurrent points of (2.8). (A limit point of $\{\mathbf{X}_n\}$ is the limit of some convergent subsequence of $\{\mathbf{X}_n\}$; see Appendix A.4 for other definitions.)*
- *Moreover, let Λ be a locally asymptotically stable set (in the sense of Lyapunov) of all Krasovskii solutions of (2.7) and $\text{DA}(\Lambda)$ be its domain of attraction. If $\{\mathbf{X}_n\}$ visits the compact subset of $\text{DA}(\Lambda)$ infinitely often with probability 1 (resp., with probability $\geq \rho$), then $\mathbf{X}_n \rightarrow \Lambda$ when $n \rightarrow \infty$ with probability 1 (resp., with probability $\geq \rho$).*

PROOF OF THEOREM 2.1. To help the reading, we divide the proof into four parts.

Part 1: Boundedness and equi-continuity. The argument in proving $\{\mathbf{X}^n(\cdot)\}$ being bounded and equicontinuous is similar to [36], Proof of Theorem 2.4.1 and Theorem 2.4.2, or [37], Proof of Theorem 6.1.1. Hence, we only outline the main points and highlight the differences. Let \mathcal{H}_0 be a countable dense subset of \mathbb{R}^d and Ω_0 be the null set that contains all paths, in which $\{\mathbf{X}_n\}$ is unbounded and the exceptional sets in Assumption (A)(ii)–(v), (K), union over \mathcal{H}_0 . In the assumptions, the null or exceptional sets are the sets in which the boundedness or

convergence does not hold. For example, the exceptional set (at \mathbf{x}) in (A)(iii) is the set of all ω , in which

$$\limsup_{n \rightarrow \infty} \max_{t \leq T} \left| \sum_{i=m(nT)}^{m(nT+t)-1} a_i (\mathbf{h}(\mathbf{x}, \zeta_i) - \bar{\mathbf{h}}(\mathbf{x})) \right| \neq 0.$$

We refer to [36], Proof of Lemma 2.2.1, for the proof of the above exceptional sets being null sets. Since \mathcal{H}_0 is countable, Ω_0 is still a null set. Now, we work with a fixed $\omega \notin \Omega_0$. We write $\mathbf{X}^n(\cdot)$ as

$$(2.9) \quad \begin{aligned} \mathbf{X}^n(t) = & \mathbf{X}_n + \int_0^t \mathbf{b}_n(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) ds + \int_0^t \mathbf{h}(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) ds \\ & + \mathbf{\Gamma}^n(t) + \mathbf{\Psi}^n(t), \end{aligned}$$

if $t \geq -t_n$, otherwise $\mathbf{X}^n(t) = \mathbf{X}_0$, where $\mathbf{\Gamma}^n(t)$ and $\mathbf{\Psi}^n(t)$ are the piecewise linear interpolations of $\sum_{i=0}^{n-1} a_i \beta_i$ and $\sum_{i=0}^{n-1} a_i \mathbf{h}_0(\xi_i)$, respectively. That is,

$$\begin{aligned} \mathbf{\Gamma}^0(t_n) &= \sum_{i=0}^{n-1} a_i \beta_i; \\ \mathbf{\Gamma}^0(t) &= \frac{t_{n+1} - t}{a_n} \mathbf{\Gamma}^0(t_n) + \frac{t - t_n}{a_n} \mathbf{\Gamma}^0(t_{n+1}) \quad \text{for } t \in (t_n, t_{n+1}); \\ \mathbf{\Gamma}^n(t) &= \begin{cases} \mathbf{\Gamma}^0(t + t_n) - \mathbf{\Gamma}^0(t_n) & \text{if } t \geq -t_n, \\ -\mathbf{\Gamma}^0(t_n) & \text{if } t \leq -t_n; \end{cases} \\ \mathbf{\Psi}^0(t_n) &= \sum_{i=0}^{n-1} a_i \mathbf{h}_0(\xi_i); \\ \mathbf{\Psi}^0(t) &= \frac{t_{n+1} - t}{a_n} \mathbf{\Psi}^0(t_n) + \frac{t - t_n}{a_n} \mathbf{\Psi}^0(t_{n+1}) \quad \text{for } t \in (t_n, t_{n+1}); \\ \mathbf{\Psi}^n(t) &= \begin{cases} \mathbf{\Psi}^0(t + t_n) - \mathbf{\Psi}^0(t_n) & \text{if } t \geq -t_n, \\ -\mathbf{\Psi}^0(t_n) & \text{if } t \leq -t_n; \end{cases} \end{aligned}$$

and $\bar{\beta}^0(\cdot)$ and $\bar{\xi}^0(\cdot)$ are the piecewise constant interpolations of $\{\beta_n\}$ and $\{\xi_n\}$, that is, $\bar{\beta}^0(t) = \beta_n$ and $\bar{\xi}^0(t) = \xi_n$ for $t \in [t_n, t_{n+1})$. Note that we have three different noise processes, $\{\xi_n\}$, $\{\zeta_n\}$, and $\{\tilde{\zeta}_n\}$, but we write them as $\{\xi_n\}$ (and $\bar{\xi}^0(t)$ for the interpolations) to simplify the notation. For simplicity again, we will always write the algorithm as (2.9), whether $t \geq -t_n$ or $t \leq -t_n$ with the understanding that $\mathbf{X}^n(t) = \mathbf{X}_0$ if $t \leq -t_n$.

First, by (A)(iv) and (A)(v), $\{\mathbf{\Gamma}^n(\cdot)$ and $\mathbf{\Psi}^n(\cdot)\}$ are equicontinuous and bounded, and any convergent subsequence converges uniformly to a zero process on bounded intervals (see, e.g., [36], Lemma 2.2.1). Second, note that $\{\mathbf{b}_n(\cdot, \cdot)\}$ is (uniformly in the variable ξ) bounded due to Assumption (K) and the boundedness of $\{\mathbf{X}_n\}$. By (A)(ii), if $\mathbf{h}(\cdot, \cdot)$ is uniformly bounded, combining with boundedness of $\{\mathbf{b}_n(\cdot, \cdot)\}$, $\{\mathbf{X}^n(\cdot)\}$ is equicontinuous. Otherwise, by (2.3) and (2.4), we obtain

$$\int_t^{t+s} |\mathbf{h}(\bar{\mathbf{X}}^0(r), \bar{\xi}^0(r))| dr \leq K \int_t^{t+s} g_2(\bar{\xi}^0(r)) dr + \int_t^{t+s} g_3(\bar{\xi}^0(r)) dr,$$

where K is some finite constant; such a K always exists due to the boundedness of $\{\mathbf{X}_n\}$ and local boundedness of $g_1(\cdot)$ in (A)(ii). Thus, by (2.4), we get that $\int_t^{t+s} |\mathbf{h}(\bar{\mathbf{X}}^0(r), \bar{\xi}^0(r))| dr$ is

uniformly continuous in t, s in $[0, \infty)$. Therefore, it is easy to show that $\mathbf{X}^0(\cdot)$ is uniformly continuous, so $\{\mathbf{X}^n(\cdot)\}$ is equicontinuous. As a consequence, we obtain boundedness and equicontinuity of $\{\mathbf{X}^n(\cdot)\}$.

Part 2: Characterize the limit. Take a convergent subsequence of $\{\mathbf{X}^n(\cdot)\}$ and still denote it by $\{\mathbf{X}^n(\cdot)\}$ for simplicity of notation and denote its limit by $\mathbf{X}(\cdot)$. From the integral form (2.9), we have that

$$(2.10) \quad \begin{aligned} \mathbf{X}^n(t) = & \mathbf{X}_n + \int_0^t \mathbf{b}_n(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) ds + \int_0^t \bar{\mathbf{h}}(\mathbf{X}(s)) ds \\ & + \int_0^t [\mathbf{h}(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) - \bar{\mathbf{h}}(\mathbf{X}(s))] ds + \mathbf{\Gamma}^n(t) + \mathbf{\Psi}^n(t). \end{aligned}$$

Hence, we obtain that

$$(2.11) \quad \mathbf{Q}^n(t) = \mathbf{Q}^n(0) + \int_0^t \mathbf{b}_n(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) ds + \int_0^t \bar{\mathbf{h}}(\mathbf{X}(s)) ds,$$

where

$$\mathbf{Q}^n(t) := \mathbf{X}^n(t) - \mathbf{\Gamma}^n(t) - \mathbf{\Psi}^n(t) - \int_0^t [\mathbf{h}(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) - \bar{\mathbf{h}}(\mathbf{X}(s))] ds.$$

Because of Assumption (K), we get

$$(2.12) \quad \begin{aligned} \mathbf{b}_n(\bar{\mathbf{X}}^0(t_n + t), \bar{\xi}^0(t_n + t)) & \in \mathcal{K}[\bar{\mathbf{b}}](\bar{\mathbf{X}}^0(t_n + t)) + m_n(\bar{\mathbf{X}}^0(t_n + t), \bar{\xi}^0(t_n + t))\bar{B} \\ & = \mathcal{K}[\bar{\mathbf{b}}](\mathbf{X}(t) + \mathbf{p}_n(t)) + m_n(\bar{\mathbf{X}}^0(t_n + t), \bar{\xi}^0(t_n + t))\bar{B}, \end{aligned}$$

where $m_n(\mathbf{x}, \xi)$ is as in Assumption (K) and $\mathbf{p}_n(t) := \bar{\mathbf{X}}^0(t_n + t) - \mathbf{X}(t)$.

Next, we prove $\mathbf{p}_n(t)$ converges to $\mathbf{0}$ and $\mathbf{Q}^n(t)$ converges to $\mathbf{X}(t)$ uniformly on bounded t -intervals. First, it is easy to see that $\bar{\mathbf{X}}^0(t_n + \cdot) - \mathbf{X}(\cdot)$ converges to $\mathbf{0}$ uniformly on bounded intervals, which leads to that $\{\mathbf{p}_n(\cdot)\}$ converges to $\mathbf{0}$ uniformly on bounded intervals. Second, by the continuity of $\mathbf{h}(\cdot, \xi)$ in Assumptions (A)(i), and the fact that $\bar{\mathbf{X}}^0(t_n + \cdot) - \mathbf{X}(\cdot)$ converges to $\mathbf{0}$ uniformly on bounded intervals, we obtain that (see, e.g., [36], Proof of Theorem 2.4.1)

$$(2.13) \quad \int_0^t (\mathbf{h}(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) - \mathbf{h}(\mathbf{X}(s), \bar{\xi}^0(t_n + s))) ds \rightarrow \mathbf{0},$$

uniformly on bounded intervals. On the other hand, we also have that

$$(2.14) \quad \lim_{n \rightarrow \infty} \int_0^t (\mathbf{h}(\mathbf{x}, \bar{\xi}^0(t_n + s)) - \bar{\mathbf{h}}(\mathbf{x})) ds = \mathbf{0},$$

uniformly in (t, \mathbf{x}) on bounded sets. In fact, by (A)(iii) we first only get the convergence (2.14) being uniform on bounded t -intervals for \mathbf{x} being in countable dense set \mathcal{H}_0 . However, because of the assumptions on regularity of $\mathbf{h}(\cdot, \cdot)$ and $\bar{\mathbf{h}}(\cdot)$, we obtain the uniform convergence on bounded sets. Combining (2.13) and (2.14) implies that

$$\int_0^t [\mathbf{h}(\bar{\mathbf{X}}^0(t_n + s), \bar{\xi}^0(t_n + s)) - \bar{\mathbf{h}}(\mathbf{X}(s))] ds \rightarrow \mathbf{0} \quad \text{uniformly on bounded intervals.}$$

The uniform convergences to $\mathbf{0}$ of $\mathbf{\Gamma}^n(\cdot)$ and $\mathbf{\Psi}^n(\cdot)$ follow from Assumptions (A)(iv) and (A)(v). Hence, $\{\mathbf{Q}^n(\cdot)\}$ converges to $\mathbf{X}(\cdot)$ uniformly on bounded intervals. To proceed, we have the following proposition, whose proof can be found in [22], Lemma 4.1 and Lemma 4.2.

PROPOSITION 2.1. *We have the following results.*

(a) *Let $C(t) : \mathbb{R} \rightarrow 2^{\mathbb{R}^d}$ be a set-valued mapping, whose values are compact, convex, and all contained in a common ball, that is, there is a finite ball $B_C \subset \mathbb{R}^d$ such that $C(t) \subset B_C$ for all t . Then $\int_0^1 C(t) dt$ is compact and convex.*

(b) *Let $S(t) : \mathbb{R} \rightarrow 2^{\mathbb{R}^d}$ be a set-valued mapping, whose values are all contained in a common ball. If $\mathbf{X}(\cdot) : [0, 1] \rightarrow \mathbb{R}^d$ satisfies that*

$$\mathbf{X}(t) - \mathbf{X}(s) \in \int_s^t S(r) dr, \quad \text{for all } s < t \in [0, 1],$$

then $\mathbf{X}(\cdot)$ is absolutely continuous and satisfies that $\dot{\mathbf{X}}(t) \in \overline{\text{co}}S(t)$ almost everywhere in $[0, 1]$.

Now, let $\varepsilon, \delta > 0$ be arbitrary. On bounded intervals, for n large enough, $|\mathbf{p}_n(\cdot)| < \varepsilon/2$. Moreover, because of Assumption (K), the average of the “radius of neighbor” $m_n(\mathbf{x}, \xi_n)$ tends to 0, thus on bounded intervals, for n large enough, we have from (2.12) that

$$\int_s^t \mathbf{b}_n(\bar{\mathbf{X}}^0(t_n + r), \bar{\xi}^0(t_n + r)) dr \in \int_s^t \mathcal{K}[\bar{\mathbf{b}}](\mathbf{X}(r) + \mathbf{p}_n(r)) dr + \delta \bar{B}.$$

Hence, for all t, s in bounded intervals, for n large enough, one obtains from (2.11) and (2.12) that

$$\mathbf{Q}^n(t) - \mathbf{Q}^n(s) \in \int_s^t \bar{\mathbf{h}}(\mathbf{X}(r)) dr + \int_s^t \overline{\text{co}}(\bar{\mathbf{b}}(\mathbf{X}(r) + \varepsilon B)) dr + \delta \bar{B}.$$

By part (a) of Proposition 2.1, letting $n \rightarrow \infty$, we obtain that

$$\mathbf{X}(t) - \mathbf{X}(s) \in \int_s^t \bar{\mathbf{h}}(\mathbf{X}(r)) dr + \int_s^t \overline{\text{co}}(\bar{\mathbf{b}}(\mathbf{X}(r) + \varepsilon B)) dr + \delta \bar{B}.$$

Letting $\delta \rightarrow 0$ combined with part (b) implies that $\mathbf{X}(t)$ is absolutely continuous and for almost t in bounded intervals,

$$\dot{\mathbf{X}}(t) \in \overline{\text{co}}(\bar{\mathbf{b}}(\mathbf{X}(t) + \varepsilon B)) + \bar{\mathbf{h}}(\mathbf{X}(t)), \quad \forall \varepsilon > 0.$$

Taking $\varepsilon \rightarrow 0$, we obtain that for almost t in bounded intervals

$$\dot{\mathbf{X}}(t) \in \bigcap_{\varepsilon > 0} \overline{\text{co}}(\bar{\mathbf{b}}(\mathbf{X}(t) + \varepsilon B)) + \bar{\mathbf{h}}(\mathbf{X}(t)) = \mathcal{K}[\bar{\mathbf{b}}](\mathbf{X}(t)) + \bar{\mathbf{h}}(\mathbf{X}(t)).$$

Hence, combined with Lemma A.1, we obtain that $\mathbf{X}(t)$ satisfies the differential inclusion

$$\dot{\mathbf{X}}(t) \in \mathcal{K}[\bar{\mathbf{b}} + \bar{\mathbf{h}}](\mathbf{X}(t)).$$

Part 3: Stability. The proof of the limit set of $\mathbf{X}(\cdot)$ being internally-chain transitive can be found in [5], Theorem 3.6. Hence, the limit points of $\{\mathbf{X}_n\}$ are contained in \mathcal{R} , the set of chain-recurrent points. Since we still use the definition of stability in the sense of Lyapunov, the argument for obtaining stability is the same as that of [36], Proof of Theorem 2.3.1, or [37], Proof of Theorem 5.2.1. We will study the stability (in the sense of Lyapunov) for differential inclusions later. \square

Theorem 2.1 can be generalized when we replace the Krasovskii operator by arbitrary set-valued mappings. We proceed with the conditions needed and the assertions.

(G) There is a set-valued mapping $G : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ satisfying:

(i) $G(\cdot)$ has nonempty, compact, convex values, and all values are contained in a finite common ball, that is, there is a finite ball $B_G \subset \mathbb{R}^d$ such that $G(\mathbf{x}) \subset B_G$ for all \mathbf{x} ;

(ii) G has a closed graph, that is, $\text{Graph}(G) := \{(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in G(\mathbf{x})\}$ is a closed subset of $\mathbb{R}^d \times \mathbb{R}^d$;

(iii) there is a sequence of (positive real-valued) continuous (in \mathbf{x} , uniformly in ξ) functions $\{m_n(\mathbf{x}, \xi)\}$ such that for all n, \mathbf{x}, ξ ,

$$\mathbf{b}_n(\mathbf{x}, \xi) \in G(\mathbf{x}) + m_n(\mathbf{x}, \xi)\bar{B},$$

and that for some $T > 0$, each $\varepsilon > 0$, and each \mathbf{x} ,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} a_i m_i(\mathbf{x}, \xi_i) \right| \geq \varepsilon \right\} = 0.$$

THEOREM 2.2. *If we replace Assumption (K) by (G) in Theorem 2.1, then the conclusions in Theorem 2.1 continue to hold with the limit differential inclusion (2.8) replaced by*

$$(2.15) \quad \dot{\mathbf{X}}(t) \in \bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t)).$$

PROOF OF THEOREM 2.2. To prove Theorem 2.2, we need to generalize the results on closure of the set of Krasovskii solutions for the set of solutions of the classes of differential inclusions that satisfy a “nice” property (property 2.16) like the Krasovskii operator. Then we will prove this property holds for the set-valued mappings in our setting (having compact, convex values, contained in a finite common ball and having close graph). The results are shown in the following two propositions.

PROPOSITION 2.2. *Let $\mathbf{X}_k(\cdot)$ be satisfied by the following for all t, s in $[0, 1]$:*

$$\mathbf{X}_k(t) - \mathbf{X}_k(s) \in \int_s^t (F(\mathbf{X}_k(r) + \mathbf{p}_k(r, \mathbf{X}_k(r))) + \mathbf{q}_k(r, \mathbf{X}_k(r))) dr,$$

for some sequences of functions $\{\mathbf{p}_k(\cdot)\}$ and $\{\mathbf{q}_k(\cdot)\}$ satisfying $\mathbf{p}_k \rightarrow \mathbf{0}$ $\mathbf{q}_k \rightarrow \mathbf{0}$ uniformly (in $[0, 1]$). Assume that $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ is a set-valued mapping, whose values are nonempty, compact, convex, and in a common ball, and that

$$(2.16) \quad \bigcap_{\varepsilon > 0} \overline{\text{co}} F(\mathbf{x} + \varepsilon B) = F(\mathbf{x}), \quad \forall \mathbf{x}.$$

If $\mathbf{X}_k(\cdot)$ converges (uniformly) to $\mathbf{X}(\cdot)$, then the limit $\mathbf{X}(\cdot)$ is a solution of the following differential inclusion:

$$\dot{\mathbf{X}}(t) \in F(\mathbf{X}(t)).$$

PROOF. For arbitrary $\varepsilon, \delta > 0$, there is a large number N such that $\forall n \geq N$,

$$|\mathbf{p}_n(\cdot)| < \varepsilon, \quad |\mathbf{q}_n(\cdot)| < \delta, \quad |\mathbf{X}_n(\cdot) - \mathbf{X}(\cdot)| < \varepsilon.$$

Hence, we have

$$\mathbf{X}_k(t) - \mathbf{X}_k(s) \in \int_s^t \overline{\text{co}}(F(\mathbf{X}_k(r) + \varepsilon B) + \delta B) dr \in \int_s^t \overline{\text{co}}(F(\mathbf{X}(r) + 2\varepsilon B) + \delta B) dr.$$

Letting $k \rightarrow \infty$, it follows from Proposition 2.1 that $\mathbf{X}(t)$ is absolutely continuous and for almost all t

$$\dot{\mathbf{X}}(t) \in \overline{\text{co}}(F(\mathbf{X}(t) + 2\varepsilon B) + \delta B).$$

Taking $\delta \rightarrow 0$, we obtain $\dot{\mathbf{X}}(t) \in \overline{\text{co}} F(\mathbf{X}(t) + 2\varepsilon B) \forall \varepsilon > 0$. As a consequence, $\dot{\mathbf{X}}(t) \in \bigcap_{\varepsilon > 0} \overline{\text{co}} F(\mathbf{X}(t) + 2\varepsilon B)$. Using (2.16), we complete the proof. \square

PROPOSITION 2.3. Assume $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ is a set-valued mapping, whose values are nonempty, convex, compact subsets, and contained in a finite common ball, and whose graph is closed. Then, one has

$$\bigcap_{\varepsilon > 0} \overline{\text{co}}F(\mathbf{x} + \varepsilon B) = F(\mathbf{x}), \quad \forall \mathbf{x}.$$

PROOF. Let \mathbf{x} be fixed but otherwise arbitrary. By Lemma A.3, F is upper semicontinuous. Hence, by [1], Proposition 3, Chapter 1, we have $\overline{\text{co}}F(\mathbf{x} + \varepsilon B) \subset \overline{\text{co}}F(\mathbf{x} + \varepsilon \overline{B}) = \text{co}F(\mathbf{x} + \varepsilon \overline{B})$. Therefore, $\bigcap_{\varepsilon > 0} \overline{\text{co}}F(\mathbf{x} + \varepsilon B) \subset \bigcap_{\varepsilon > 0} \text{co}F(\mathbf{x} + \varepsilon \overline{B})$. On the other hand, by [48], Theorem 5.7, from closed graph property of F , we obtain that

$$(2.17) \quad \{\mathbf{u} : \text{there exist } \mathbf{x}_n \rightarrow \mathbf{x} \text{ and } \mathbf{y}_n \in F(\mathbf{x}_n) \text{ such that } \mathbf{y}_n \rightarrow \mathbf{u}\} \subset F(\mathbf{x}).$$

Now, let $\mathbf{u} \in \bigcap_{\varepsilon > 0} \overline{\text{co}}F(\mathbf{x} + \varepsilon B)$. Then $\mathbf{u} \in \bigcap_{\varepsilon > 0} \text{co}F(\mathbf{x} + \varepsilon \overline{B})$. As a consequence, $\mathbf{u} \in \text{co}F(\mathbf{x} + \frac{1}{n}\overline{B})$ for all $n \in \mathbb{N}$. By Carathéodory's theorem for convex hulls of sets in a Euclidean space [48], Theorem 2.29, for each n , there are $d + 1$ points $\mathbf{y}_n^0, \dots, \mathbf{y}_n^d$ and $d + 1$ points $\mathbf{x}_n^0, \dots, \mathbf{x}_n^d$, $|\mathbf{x} - \mathbf{x}_n^i| \leq \frac{1}{n}$, $\forall i = 0, \dots, d$ and real numbers $a_n^0, \dots, a_n^d \in [0, 1]$, $\sum_i a_n^i = 1$ such that $\mathbf{u} = \sum_{i=0}^d a_n^i \mathbf{y}_n^i$, $\mathbf{y}_n^i \in F(\mathbf{x}_n^i)$. Since $2d + 2$ sequences $\{a_n^i\}_{n=0}^\infty$, $\{\mathbf{y}_n^i\}_{n=0}^\infty$ for $i = 0, \dots, d$ are bounded, we can extract subsequences (still index the sequences by n for simplicity) such that all of them are convergent. As a result, $\mathbf{u} = \sum_{i=0}^d a^i \lim_n \mathbf{y}_n^i$, $\mathbf{y}_n^i \in F(\mathbf{x}_n^i)$, where $a^i = \lim_n a_n^i$, and $a^i \in [0, 1]$, $\sum_{i=0}^d a^i = 1$. Since $\mathbf{x}_n^i \rightarrow \mathbf{x}$, by (2.17), $\lim_n \mathbf{y}_n^i \in F(\mathbf{x})$. Combined with the convexity of $F(\mathbf{x})$, we obtain $\mathbf{u} \in F(\mathbf{x})$. So, $\bigcap_{\varepsilon > 0} \overline{\text{co}}F(\mathbf{x} + \varepsilon B) \subset F(\mathbf{x})$. The proof is complete. \square

It is noted that we need only take care of the “characterization of the limit” part since the other parts are the same as that of Theorem 2.1. With the help of Propositions 2.2 and 2.3, the arguments of “characterization of the limit” part are similar to that of Theorem 2.1; the details are thus omitted. \square

REMARK 3. The difficulty in our setting is that we impose neither continuity to the dynamics of the discrete iterations nor the limit systems. As a result, although we obtain the boundedness and equicontinuity of $\{\mathbf{X}^n(\cdot)\}$ and can extract a convergent subsequence with the limit $\mathbf{X}(\cdot)$, it is impossible to characterize the limit using continuity and compactness. To illustrate, we mention some related works and methods in the literature. In [36, 37], the continuous dynamics with the limits being a set-valued mapping were treated. In this case, it is still possible to pass to the limit after extracting convergent subsequence to characterize the limit. In [35, 37, 40], possibly discontinuous $\mathbf{b}_n(\cdot, \cdot)$ were considered, but the limits have some regularities. Under the regularities of the limits and some assumptions on existence of a Lyapunov function, certain average takes place; see [35] for more details. Along another line, Métivier and Priouret in [40] express the limit function in term of integration of $\mathbf{b}_n(\cdot, \xi)$ over invariant measure of the noise process and use a Poisson equation approach; and thus, under some suitable conditions, the corresponding limit (continuous) differential equation may be obtained. In [36], the case of that $\mathbf{b}_n(\cdot, \cdot)$ allowing to be discontinuous and the limit being a set-valued mapping $G(\cdot)$ is considered. But the continuity of $G(\cdot)$ in the Hausdorff metric defined as

$$d(S_1, S_2) := \sup_{\mathbf{y} \in S_2} \inf_{\mathbf{x} \in S_1} |\mathbf{y} - \mathbf{x}| + \sup_{\mathbf{x} \in S_1} \inf_{\mathbf{y} \in S_2} |\mathbf{y} - \mathbf{x}|, \quad \forall S_i \subset \mathbb{R}^d, i = 1, 2,$$

is needed. However, these assumptions may not be satisfied when we do not have the desired continuity in applications. For instant, in the example of Lasso algorithm, to be illustrated

later, we need to consider set-valued mapping representing the subgradient of the function $|\mathbf{x}|$. For example, in a one-dimensional example, one may need to consider

$$G(x) = \begin{cases} \{-1\} & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x < 0. \end{cases}$$

This set-valued mapping is not continuous at 0 in the Hausdorff metric. Except [5], there has been no general approach in the literature for studying convergence of stochastic approximation schemes involving set-valued mappings without continuity. However, the setup and results of the current paper are different than that of [5]. If we let $\mathbf{b}_n(\mathbf{x}, \xi)$ be independent of ξ , $\mathbf{h}(\mathbf{x}, \xi) = 0$, $\beta_n = 0$, and $m_n(\mathbf{x}, \xi) = 0$, $\forall n, \mathbf{x}, \xi$, where $m_n(\mathbf{x}, \xi)$ is as in Assumption (K) or (G), we recover the setting and results in [5]. In this paper, $m_n(\mathbf{x}, \xi)$ is not required tending to 0, which makes the setting more general and applicable in real applications. In addition, in [5], the limit processes are perturbed solutions of the corresponding differential inclusions, whereas we characterize the limit processes by differential inclusions rather than perturbed differential inclusions. That is done by examining the closure of the set of solutions of a family of differential inclusions for general set-valued mappings.

Convergence to equilibrium point. The following results are concerned with globally asymptotic stability of the limit differential inclusions. It also establishes the convergence to the equilibria of stochastic approximation algorithm (2.1). We introduce the following stability condition for Krasovskii solutions of ODEs with discontinuous right-hand sides, which is similar to the Lyapunov condition in classical stability theory.

(KS) There is a unique equilibrium \mathbf{x}^* of $\bar{\mathbf{b}}(\cdot) + \bar{\mathbf{h}}(\cdot)$, that is, $\bar{\mathbf{b}}(\mathbf{x}^*) + \bar{\mathbf{h}}(\mathbf{x}^*) = \mathbf{0}$ (where, $\bar{\mathbf{h}}(\cdot)$ is as in Assumption (A)(iii) and $\bar{\mathbf{b}}(\cdot)$ is as in Assumption (K)); and there exists a C^∞ -smooth pair of functions (V, W) satisfying that $V(\mathbf{x}) > 0$ and $W(\mathbf{x}) > 0$, $\forall \mathbf{x} \neq \mathbf{0}$, $V(\mathbf{0}) = 0$, and the sublevel sets $\{\mathbf{x} \in \mathbb{R}^d : V(\mathbf{x}) \leq l\}$ are bounded for every $l \geq 0$, and

$$\limsup_{\mathbf{y} \rightarrow \mathbf{x}} \langle \nabla V(\mathbf{x}), \bar{\mathbf{b}}(\mathbf{y} + \mathbf{x}^*) + \bar{\mathbf{h}}(\mathbf{y} + \mathbf{x}^*) \rangle \leq -W(\mathbf{x}), \quad \forall \mathbf{x} \neq \mathbf{0}.$$

THEOREM 2.3. *Consider algorithm (2.1). Under Assumptions (A), (K), (KS), and boundedness of $\{\mathbf{X}_n\}$, there exists a null set Ω_0 such that if $\omega \notin \Omega_0$, then \mathbf{X}_n converges to the unique equilibrium \mathbf{x}^* .*

For the general case, where the Krasovskii operator is replaced by set-valued mappings, we introduce a stability condition (GS) as follows. Our approach is based on a novel method, namely, \mathcal{U} -generalized Lyapunov functional method for differential inclusions.

(GS) There is a unique \mathbf{x}^* such that $\mathbf{0} \in \bar{\mathbf{h}}(\mathbf{x}^*) + G(\mathbf{x}^*)$ (where, $\bar{\mathbf{h}}(\cdot)$ is as in Assumption (A)(iii) and $G(\cdot)$ is as in Assumption (G)); and there exists a \mathcal{U} -generalized Lyapunov function $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ such that the sublevel sets $\{\mathbf{x} \in \mathbb{R}^n : V(\mathbf{x}) \leq l\}$ are compact for every $l > 0$ and the \mathcal{U} -generalized derivative $\dot{\bar{V}}_{\mathcal{U}}^{G^*}(\mathbf{x})$ satisfies $\dot{\bar{V}}_{\mathcal{U}}^{G^*}(\mathbf{x}) \leq -\hat{V}_0(\mathbf{x})$, $\forall \mathbf{x} \neq \mathbf{0}$, for some positive definite function \hat{V}_0 , where $G^*(\mathbf{x}) := \bar{\mathbf{h}}(\mathbf{x} + \mathbf{x}^*) + G(\mathbf{x} + \mathbf{x}^*)$; see Section A.2 (Definition A.5, Definition A.4(iv)) for these concepts.

THEOREM 2.4. *If we replace Assumptions (K) and (KS) by (G) and (GS), then the conclusion of Theorem 2.3 continue to hold.*

PROOF OF THEOREMS 2.3 AND 2.4. The stability of differential inclusions is carefully studied in Section A.3. The proof of Theorem 2.3 follows from Theorem 2.1 and Theorem A.1 in Section A.3. First, under Assumption (KS), the Krasovskii solutions of (2.7) are strongly asymptotically stable (in Clarke's sense) at $\mathbf{x} = \mathbf{x}^*$. Therefore, every Krasovskii solutions of (2.7) is globally asymptotically stable at $\mathbf{x} = \mathbf{x}^*$ in the Lyapunov sense. As the last part of Theorem 2.1, $\{\mathbf{X}_n\}$ must converge to the equilibrium point \mathbf{x}^* w.p.1. Similarly, Theorem 2.4 is obtained by combining Theorem 2.2 and Theorem A.2. \square

Projection algorithms. As was mentioned before, the assumption on boundedness of $\{\mathbf{X}_n\}$ is not restrictive. Since the boundedness is not our main focus, we often assume it in our main results so as to make the argument simpler. Further conditions and/or various projection algorithms may be used; see Remark 2. We proceed to state the results for constrained algorithm (2.2).

(P) The projection space H is a hyper-rectangle, that is, $H = \{\mathbf{x} \in \mathbb{R}^d : b_i \leq x_i \leq c_i\}$ for simplifying arguments. In general, H can be compact and convex; and $H = \{\mathbf{x} \in \mathbb{R}^d : q_i(\mathbf{x}) \leq 0, i = 1, \dots, N\}$, the constrained functions $q_i(\cdot), i = 1, \dots, N$ are continuously differentiable and at $\mathbf{x} \in \partial H$, the gradients $q_{i,\mathbf{x}}(\cdot)$ are linearly independent.

(PS) There is a unique $\mathbf{x}^* \in H$ such that $\mathbf{0} \in \overline{\text{co}}\Pi_H[\bar{\mathbf{h}}(\mathbf{x}^*) + G(\mathbf{x}^*)]$; and there exists a \mathcal{U} -generalized Lyapunov function $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ such that the sublevel sets $\{\mathbf{x} \in \mathbb{R}^d : V(\mathbf{x}) \leq l\}$ are compact for every $l > 0$ and $\dot{\bar{V}}_{\mathcal{U}}^{G_H^*}(\mathbf{x}) \leq -\hat{V}_0(\mathbf{x}), \forall \mathbf{0} \neq \mathbf{x} \in H$, for some positive definite function \hat{V}_0 , where $G_H^*(\mathbf{x}) := \overline{\text{co}}\Pi_H[\bar{\mathbf{h}}(\mathbf{x} + \mathbf{x}^*) + G(\mathbf{x} + \mathbf{x}^*)]$.

THEOREM 2.5. Consider algorithm (2.2). Assume (G), (P), and (A) with (A)(ii) replaced by $\mathbf{h}(\mathbf{x}, \xi)$ being (uniformly in ξ) locally bounded in \mathbf{x} (i.e., $|\mathbf{h}(\mathbf{x}, \xi)| \leq K(\mathbf{x})$ for some locally bounded function K). Then, there is a null set Ω_0 such that $\forall \omega \notin \Omega_0, \{\mathbf{X}^n(\cdot)\}$ is bounded and equicontinuous. Let $\mathbf{X}(\cdot)$ be the limit of a convergent subsequence of $\{\mathbf{X}^n(\cdot)\}$. Then $\mathbf{X}(t)$ is a solution of the differential inclusion

$$(2.18) \quad \dot{\mathbf{X}}(t) \in \overline{\text{co}}\Pi_H[\bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t))].$$

The limit set of $\{\mathbf{X}(\cdot)\}$ is internally chain transitive and as a consequence, the limit points of $\{\mathbf{X}_n\}$ are contained in \mathcal{R} , the set of chain recurrent points of (2.18) (see Section A.4 for the definitions). In addition, if we assume further (PS), then $\{\mathbf{X}_n\}$ converges to \mathbf{x}^* w.p.1.

PROOF OF THEOREM 2.5. First, to use Assumption (A)(iv) in the projection algorithm, let Y_n be a sequence of positive real numbers such that $Y_n \rightarrow 0$ and $|a_n \mathbf{h}_0(\tilde{\zeta}_n)| \leq Y_n/2$ excepting a finite number of n w.p.1 (such a sequence Y_n exists owing to Assumption (A)(iv), Borel–Cantelli lemma [36], Section 5), and let I_n be the indicator of the set where $|a_n \mathbf{h}_0(\tilde{\zeta}_n)| \leq Y_n/2$. To proceed, we write algorithm (2.2) as

$$(2.19) \quad \mathbf{X}_{n+1} = \mathbf{X}_n + a_n[\mathbf{b}_n(\mathbf{X}_n, \xi_n) + \mathbf{h}(\mathbf{X}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \boldsymbol{\beta}_n] + \boldsymbol{\tau}_n + \boldsymbol{\psi}_n,$$

where

$$\boldsymbol{\tau}_n = \Pi_H(\mathbf{M}_n^Y) - \mathbf{M}_n^Y, \quad \boldsymbol{\psi}_n = (\mathbf{M}_n^Y - \mathbf{M}_n) + [\Pi_H(\mathbf{M}_n) - \mathbf{X}_n](1 - I_n),$$

$$\mathbf{M}_n = \mathbf{X}_n + a_n[\mathbf{b}_n(\mathbf{X}_n, \xi_n) + \mathbf{h}(\mathbf{X}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \boldsymbol{\beta}_n],$$

and

$$\mathbf{M}_n^Y = \mathbf{X}_n + a_n[\mathbf{b}_n(\mathbf{X}_n, \xi_n) + \mathbf{h}(\mathbf{X}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \boldsymbol{\beta}_n]I_n.$$

The purpose of partitioning (2.19) is to enable us to apply directly our assumptions (which is assumed without any constraints).

Part 1: Boundedness and equicontinuity. Similar to (2.9), we have that

$$(2.20) \quad \begin{aligned} \mathbf{X}^n(t) = & \mathbf{X}_n + \int_0^t \mathbf{b}_n(\bar{\mathbf{X}}^0(t_n+s), \bar{\xi}^0(t_n+s)) ds + \int_0^t \mathbf{h}(\bar{\mathbf{X}}^0(t_n+s), \bar{\xi}^0(t_n+s)) ds \\ & + \mathbf{\Gamma}^n(t) + \mathbf{\Psi}^n(t) + \boldsymbol{\tau}^n(t) + \boldsymbol{\psi}^n(t). \end{aligned}$$

In the above, $\boldsymbol{\tau}^n(t) := \boldsymbol{\tau}^0(t_n+t) - \boldsymbol{\tau}^0(t_n)$, $\boldsymbol{\psi}^n(t) := \boldsymbol{\psi}^0(t_n+t) - \boldsymbol{\psi}^0(t_n)$, where, $\boldsymbol{\tau}^0(\cdot)$ and $\boldsymbol{\psi}^0(\cdot)$ are the piecewise linear interpolations of $\{\sum_{i=0}^{n-1} a_i \boldsymbol{\tau}_i\}$ and $\{\sum_{i=0}^{n-1} a_i \boldsymbol{\psi}_i\}$, respectively; and the $\mathbf{\Gamma}^n(\cdot)$, $\mathbf{\Psi}^n(\cdot)$ are as in the proof of Theorem 2.1.

Let Ω_0 be the union of sets in which $|a_n \mathbf{h}_0(\tilde{\zeta}_n)| \geq Y_n/2$ infinitely often and the exceptional sets in (A)(iii)–(v), (G) (the union being taken over countable dense set \mathcal{H}_0). [As we mentioned before, for each $\mathbf{x} \in \mathcal{H}_0$, there are (null) exceptional sets (in which, the convergence assumptions do not hold) corresponding to (A)(iii)–(v), and (G); Ω_0 is taken to contain all these sets. Since \mathcal{H}_0 is countable, Ω_0 still has measure zero.] Therefore, we work with a fixed $\omega \notin \Omega_0$.

As in the proof of Theorem 2.1, we proved that $\mathbf{\Gamma}^n(\cdot)$ and $\mathbf{\Psi}^n(\cdot)$ converge uniformly to $\mathbf{0}$ on finite t -intervals. Moreover, because I_n is 0 only for a finite number of n , only a finite number of the terms of $\{(1 - I_n)\}$ are nonzero. Since

$$\boldsymbol{\psi}_n \leq a_n |h(\mathbf{X}_n, \xi_n) + b_n(\mathbf{X}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \boldsymbol{\beta}_n|(1 - I_n) + |\Pi_H(\mathbf{M}_n) - \mathbf{X}_n|(1 - I_n),$$

it is readily seen that $\boldsymbol{\psi}^n(\cdot)$ converges to $\mathbf{0}$ uniformly on finite intervals as $n \rightarrow \infty$. The boundedness of $\{\mathbf{X}^n(\cdot)\}$ is clear because of the use of the projection algorithm.

Next, we prove the equicontinuity of $\{\mathbf{X}^n(\cdot), \boldsymbol{\tau}^n(\cdot)\}$. It suffices to prove the equicontinuity for $\boldsymbol{\tau}^n(\cdot)$; see the proof of Theorem 2.1. By the definition of $\boldsymbol{\tau}_n$, we have following observations (see, e.g., [36], Proof of Theorem 5.3.1): $\boldsymbol{\tau}_n$ is orthogonal to H at the point $\Pi_H(\mathbf{M}_n^Y)$; $|\boldsymbol{\tau}_n| \leq a_n(K_1 + Y_n)$ for some constant K_1 ; there is a constant K_2 such that $\boldsymbol{\tau}_n = \mathbf{0}$ if $\text{distance}(\partial H, \mathbf{X}_n) \geq K_2(Y_n + a_n)$. Because of these observations and the fact that $\mathbf{X}^0(\cdot) - \boldsymbol{\tau}^0(\cdot)$ is uniformly continuous (due to this difference is in fact the process in nonprojected case and is proved before), $\boldsymbol{\tau}^0(\cdot)$ must be uniformly continuous on $[0, \infty)$. Otherwise, there would be $s_k \rightarrow \infty$, $\delta_k \rightarrow 0$ and $\varepsilon > 0$ such that

$$|\mathbf{X}^0(s_k + \delta_k) - \mathbf{X}^0(s_k)| \geq \varepsilon, \quad \text{for all } k,$$

with $\text{distance}(\mathbf{X}^0(s_k), \partial H) \rightarrow 0$ as $k \rightarrow \infty$ and $\text{distance}(\mathbf{X}^0(s_k + \delta_k), \partial H) \geq \varepsilon/2$. However, this contradicts the observations of $\boldsymbol{\tau}_n$ and the uniform continuity of $\mathbf{X}^0(\cdot) - \boldsymbol{\tau}^0(\cdot)$. The uniform continuity of $\boldsymbol{\tau}^0(\cdot)$ implies the equicontinuity of $\{\boldsymbol{\tau}^n(\cdot)\}$.

Part 2: Characterization of the limit. Now, we extract a convergent subsequence of $\{(\mathbf{X}^n(\cdot), \boldsymbol{\tau}^n(\cdot))\}$, and index it again by n with the limit $(\mathbf{X}(\cdot), \boldsymbol{\tau}(\cdot))$. Using the fact that $\mathbf{\Gamma}^n(\cdot)$, $\mathbf{\Psi}^n(\cdot)$, $\boldsymbol{\psi}^n(\cdot)$ converge to $\mathbf{0}$ uniformly and letting $n \rightarrow \infty$ in (2.20), by a similar argument as in the unconstrained case, one has that on bounded intervals

$$(2.21) \quad \mathbf{X}(t+s) - \mathbf{X}(t) \in \int_t^{t+s} G(\mathbf{X}(r)) dr + \int_t^{t+s} \bar{\mathbf{h}}(\mathbf{X}(r)) dr + \boldsymbol{\tau}(t+s) - \boldsymbol{\tau}(s).$$

As in [36], Proof of Theorem 5.3.1, or [37], Proof of Theorem 6.8.1, we have

$$(2.22) \quad \boldsymbol{\tau}(t+s) - \boldsymbol{\tau}(s) = \int_t^{t+s} \mathbf{z}(\mathbf{X}(r)) dr,$$

where $\mathbf{z}(\mathbf{X}(t))$ is the minimal force needed to keep $\mathbf{X}(t)$ in H . A consequence of (2.21) and (2.22) is that

$$(2.23) \quad \mathbf{X}(t+s) - \mathbf{X}(s) \in \int_t^{t+s} \overline{\text{co}} \Pi_H[\bar{\mathbf{h}}(\mathbf{X}(r)) + G(\mathbf{X}(r))] dr.$$

Combining (2.23) and Proposition 2.1, one has that $\mathbf{X}(t)$ is absolutely continuous and for almost all t ,

$$\dot{\mathbf{X}}(t) \in \overline{\text{co}}\Pi_H[\bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t))].$$

Part 3: Asymptotic stability. This part is the same as that of the unconstrained case and is thus omitted. \square

REMARK 4. Recall that we often wish to find roots of some functions and/or set-valued mappings. [For the roots of set-valued mappings, we mean that at these points (roots), the value of these mappings (being a set) contains $\mathbf{0}$]. These points are often called “stationary points” of the corresponding differential equations or inclusions. In the set-valued and differential inclusion cases, the roots may not be (strongly) stationary, where “strongly” means the statement is true for all solutions. If the function is vector-valued and is sufficiently smooth (namely, d -time continuously differentiable, where d is the dimension), then the set of stationary points is equal to \mathcal{R} , the set of chain recurrent points (of the corresponding differential equation). Otherwise, by a sophisticated process, Hurley in [27] shows that if the function is smooth but not smooth enough, the set of chain-recurrent points maybe strictly larger than the set of stationary points. Hence, $\{\mathbf{X}_n\}$ may not converge to the desired stationary points. In our setting, if there is no condition to guarantee the stability of stationary points (termed roots for simplicity), the algorithm may not converge to a set of roots, even if the algorithm starts at one of the roots. It is easy to give an example; see Example 4.4 in Section 4.7.

REMARK 5. The stability of the systems of interest can be characterized by means of the stability in set-valued dynamical systems [5], Section 3 and 4, and references therein. However, the conditions in the aforementioned reference is relatively abstract and difficult to verify in applications. We use criteria on \mathcal{U} -generalized Lyapunov functions instead. Moreover, we give an example in Section 4.4 to show that convergence can be proved by applying our results, but cannot be done otherwise.

REMARK 6. It is worth noting that the Clarke subdifferential of Lipschitz continuous function has the important property (2.16). In addition, the stability assumptions (KS), (GS), and (PS) are not restrictive. They are similar to Lyapunov conditions in classical stability analysis excepting that we need to compute a new type of derivative (namely, \mathcal{U} -generalized derivative) for \mathcal{U} -generalized Lyapunov functional. The examples of computing these new functions are given in Section 4 and Appendix A.2. Moreover, Theorems 2.1 and 2.3 are less general than Theorems 2.2 and 2.4. However, if we can express a set-valued mapping as the Karasovskii operator of some vector-valued function, condition (KS) is more convenient to verify than that of (GS). In the projection algorithm, since H is convex, $\Pi_H(\mathbf{x})$ is uniquely defined and $\Pi_H(\cdot)$ is continuous. However, the convex closure in (2.18) cannot be relaxed since a continuous projection operator may not preserve the convexity.

Biased stochastic approximation. Next, we study biased stochastic approximation. With the term β_n representing a bias, by “biased stochastic approximation,” we mean the bias is not “asymptotically negligible.” To proceed, let $\eta = \limsup_{n \rightarrow \infty} \|\beta_n\|$ be a random variable that is bounded w.p.1. We study stochastic approximation schemes (2.1) and (2.2) with the dependence on η . For a set $S \subset \mathbb{R}^d$, an ε -neighborhood of S denoted by $N_\varepsilon(S)$ is defined as

$$N_\varepsilon(S) = \{\mathbf{x} \in \mathbb{R}^d : \text{distance}(\mathbf{x}, S) \leq \varepsilon\}, \quad \text{distance}(\mathbf{x}, S) := \inf_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|.$$

THEOREM 2.6. Consider algorithm (2.1), assume that (A)(i)–(iv) and (G) hold, and that $\{\mathbf{X}_n\}$ is bounded w.p.1 (resp., consider algorithm (2.2) and assume that (A)(i)–(iv), (G), and (P) hold).

- Then, there is a null set Ω_0 such that $\forall \omega \notin \Omega_0$, $\{\mathbf{X}^n(\cdot)\}$ is bounded and equicontinuous.
- Let $\mathbf{X}(\cdot)$ be the limit of a convergent subsequence of $\{\mathbf{X}^n(\cdot)\}$. Then $\mathbf{X}(\cdot)$ is a solution of the differential inclusion

$$(2.24) \quad \dot{\mathbf{X}}(t) \in N_{2\eta}(\bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t))) \quad (\text{resp., } \dot{\mathbf{X}}(t) \in N_{2\eta}(\overline{\text{co}}\Pi_H(\bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t)))).$$

- There exists a (deterministic) positive function $\phi(\cdot) : [0, \infty) \rightarrow [0, \infty)$ depending on $\limsup_n |\mathbf{X}_n|$ (resp., the projection space H) such that $\lim_{t \rightarrow 0} \phi(t) = \phi(0) = 0$ and

$$(2.25) \quad \limsup_{n \rightarrow \infty} \text{distance}(\mathbf{X}_n, \mathcal{R}) \leq \phi(\eta),$$

where \mathcal{R} is the set of chain recurrent points of differential inclusion

$$\dot{\mathbf{X}}(t) \in \bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t)) \quad (\text{resp., } \dot{\mathbf{X}}(t) \in \overline{\text{co}}\Pi_H(\bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t)))).$$

- Assume further that there is a unique \mathbf{x}^* such that $\mathbf{0} \in \bar{\mathbf{h}}(\mathbf{x}^*) + G(\mathbf{x}^*)$; and that there exists a \mathcal{U} -generalized Lyapunov function $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ such that the sublevel sets $\{x \in \mathbb{R}^n : V(\mathbf{x}) \leq l\}$ are compact for every $l > 0$ and the \mathcal{U} -generalized derivative $\dot{V}_{\mathcal{U}}^{G^*}(\mathbf{x})$ satisfies the “decay condition” in the sense of Assumption (GS) (resp., (PS)) with $G^*(\mathbf{x})$ being replaced by $G_\eta^*(\mathbf{x}) := N_{2\eta}(\bar{\mathbf{h}}(\mathbf{x} + \mathbf{x}^*) + G(\mathbf{x} + \mathbf{x}^*))$. Then, $\{\mathbf{X}_n\}$ converges to \mathbf{x}^* w.p.1.

REMARK 7. If the dynamics and the limits of stochastic approximations are smooth enough (namely, real analytic or k -times continuously differentiable with $k > d$ and d being the dimension of the space), a more precise characterization of the asymptotic bias ($\phi(\eta)$ in (2.25)) is obtained by Tadić and Doucet in [53] using the Yomdin theorem (a qualitative version of the Morse–Sard theorem) and the Łojasiewicz inequality. This paper deals with systems with discontinuity.

PROOF OF THEOREM 2.6. We prove the assertion for unconstrained case only; the constrained case can be handled similarly.

Part 1: Boundedness and equicontinuity. This part is the same as that of the unbiased case. In fact, we can treat β_n as a (uniformly) bounded term and hence, boundedness and equicontinuity of sequence of the piecewise linear interpolated processes still hold.

Part 2: Characterization of the limit. The process of obtaining the limit system is almost the same as that of the unbiased case, excepting for that the limit differential inclusion should be relaxed. To be more specific, in Proposition 2.2, if we relax the condition that $\mathbf{q}_k(\cdot) \rightarrow \mathbf{0}$ uniformly to be that $|\mathbf{q}_k(\cdot)| < \eta$ uniformly for k large enough, we will obtain similar results as in the unbiased case with $G(\cdot)$ being replaced by its 2η -neighborhood due to the bias term β_n .

Part 3: Proof of (2.25) and stability assertion. Let Q be a compact set such that $\{\mathbf{X}_n\}_{n=1}^\infty \subset Q$ and \mathcal{M}_Q be the largest invariant set contained in Q ; and let $\mathcal{R}_{2\eta}(Q)$ be the set of chain recurrence points of following differential inclusion restricted in Q :

$$(2.26) \quad \dot{\mathbf{X}}(t) \in N_{2\eta}(\bar{\mathbf{h}}(\mathbf{X}(t)) + G(\mathbf{X}(t))),$$

that is, $\mathcal{R}_{2\eta}(Q)$ contains all θ satisfying that for any $\varepsilon > 0$, $T > 0$, there are an integer n , real numbers $t_1, \dots, t_n > T$, and solutions $\mathbf{x}_1(\cdot), \dots, \mathbf{x}_n(\cdot)$ of (2.26) such that $\forall k = 1, \dots, n$,

$$\mathbf{x}_k(0) \in \mathcal{M}_Q, \quad |\mathbf{x}_1(0) - \theta| < \varepsilon, \quad |\mathbf{x}_k(t_k) - \mathbf{x}_{k+1}(0)| \leq \varepsilon, \quad |\mathbf{x}_n(t_n) - \theta| \leq \varepsilon.$$

To proceed, we need the following lemma, whose proof can be found in [53], Lemma 5.1, which used the continuation of chain-recurrent set developed in [6], Theorem 3.1.

LEMMA 2.1 (Continuation of chain recurrent set). *There exists a function $\phi(\cdot) : [0, \infty) \rightarrow [0, \infty)$ (depending on Q and \mathcal{R}) such that $\phi(\cdot)$ is nondecreasing with $\lim_{t \rightarrow 0} \phi(t) = \phi(0) = 0$ and $\mathcal{R}_{2\eta}(Q) \subset N_{\phi(\eta)}(\mathcal{R})$, where $N_{\phi(\eta)}(\mathcal{R})$ is a $\phi(\eta)$ -neighborhood of \mathcal{R} .*

It is similar to the arguments in the unbiased case, we obtain that the limit points of $\{\mathbf{X}_n\}$ are contained in $\mathcal{R}_{2\eta}(Q)$. On the other hand, applying Lemma 2.1, we obtain that $\mathcal{R}_{2\eta}(Q) \subset N_{\phi(\eta)}(\mathcal{R})$. Therefore, we conclude our results. Finally, the stability assertion is similar to that of the unbiased case, which turns out to be the study of the stability of the limit differential inclusion. \square

3. Rates of convergence. This section is concerned with the rates of convergence of the stochastic approximation algorithms. One of the new features of our work is that stochastic differential inclusions are used in the rate of convergence study for the first time.

For simplicity, we consider the following algorithm:

$$(3.1) \quad \mathbf{X}_{n+1} = \mathbf{X}_n + a_n \mathbf{h}(\mathbf{X}_n, \xi_n) + a_n \mathbf{b}_n(\mathbf{X}_n), \quad \mathbf{b}_n(\mathbf{X}_n) \in G(\mathbf{X}_n).$$

We assume that the limit dynamical system has a global stable limit point \mathbf{x}^* . The rate of convergence is focused on the asymptotic behavior of $\mathbf{U}_n := \frac{\mathbf{X}_n - \mathbf{x}^*}{\sqrt{a_n}}$. Let $\mathbf{U}^0(\cdot)$ be the piecewise constant interpolation of $\{\mathbf{U}_n\}$, and $\mathbf{U}^n(\cdot)$ be its shifted process, that is,

$$\mathbf{U}^0(t) := \mathbf{U}_n \quad \text{if } t \in [t_n, t_{n+1}); \quad \text{and} \quad \mathbf{U}^n(t) := \mathbf{U}^0(t_n + t), \quad t \geq 0.$$

We state only the results for the unconstrained case with assumption on boundedness of $\{\mathbf{X}_n\}$. The projection case is similar with a slight modification. We assume the following assumption.

(R) (i) The sequence of step sizes $\{a_n\}_{n \geq 0}$ satisfies $0 < a_n \rightarrow 0$ as $n \rightarrow \infty$ and $(a_n/a_{n+1})^{1/2} = 1 + \varepsilon_n$ where (a) $\varepsilon_n = \frac{1}{2n} + o(\varepsilon_n)$ if $a_n = 1/n$, or (b) $\varepsilon_n = o(a_n)$.

(ii) There is a limit point \mathbf{x}^* satisfying the following conditions: (a) $\mathbf{X}_n \rightarrow \mathbf{x}^*$ w.p.1 and $\bar{\mathbf{h}}(\mathbf{x}^*) + G(\mathbf{x}^*) = \{\mathbf{0}\}$; (b) $\{(\mathbf{X}_n - \mathbf{x}^*)/\sqrt{a_n}\}$ is tight.

(iii) The functions $\mathbf{h}(\cdot, \cdot)$ and $\mathbf{h}_{\mathbf{x}}(\cdot, \cdot)$ (gradient with respect to \mathbf{x}) are continuous in (\mathbf{x}, ξ) and bounded on bounded \mathbf{x} -sets. The second partial derivative (with respect to \mathbf{x}) $\mathbf{h}_{\mathbf{xx}}(\cdot, \xi)$ exists and is bounded uniformly in ξ , and $\mathbf{h}_{\mathbf{xx}}(\cdot, \xi)$ is continuous in a neighborhood of \mathbf{x}^* . The $\{\xi_n\}$ is a sequence of uniformly bounded and stationary uniform mixing process satisfying that: $\mathbb{E}\mathbf{h}(\mathbf{x}, \xi_n) = \bar{\mathbf{h}}(\mathbf{x})$ and $\mathbb{E}\mathbf{h}_{\mathbf{x}}(\mathbf{x}, \xi_n) = \bar{\mathbf{h}}_{\mathbf{x}}(\mathbf{x})$. Let

$$\begin{aligned} \psi_n &= \mathbf{h}(\mathbf{x}^*, \xi_n) - \bar{\mathbf{h}}(\mathbf{x}^*), \\ \tilde{\psi}_n &= \mathbf{h}_{\mathbf{x}}(\mathbf{x}^*, \xi_n) - \bar{\mathbf{h}}_{\mathbf{x}}(\mathbf{x}^*), \\ \mathcal{G}_n &= \sigma\{\psi_j; j \leq n\}, \quad \mathcal{G}^n = \sigma\{\psi_j; j \geq n\}, \\ \mathcal{H}_n &= \sigma\{\psi_j; j \leq n\}, \quad \mathcal{H}^n = \sigma\{\psi_j; j \geq n\}, \\ \phi(m) &= \sup_{\mathcal{A} \in \mathcal{G}^{n+m}} |\mathbb{P}(\mathcal{A}|\mathcal{G}_n) - \mathbb{P}(\mathcal{A})|_{\infty}, \\ \tilde{\phi}(m) &= \sup_{\mathcal{A} \in \mathcal{H}^{n+m}} |\mathbb{P}(\mathcal{A}|\mathcal{H}_n) - \mathbb{P}(\mathcal{A})|_{\infty}, \end{aligned}$$

for some $\Delta > 0$, $\sum_j \phi^{\frac{\Delta}{1+\Delta}}(j) < \infty$, $\sum_j \tilde{\phi}^{\frac{\Delta}{1+\Delta}}(j) < \infty$.

(iv) The set-valued mapping $G(\cdot)$ has nonempty, convex, and compact values, which are contained in a finite common ball such that $\mathbf{b}_n(\mathbf{x}) \in G(\mathbf{x}) \forall n$. Moreover, there is a continuous and positively homogeneous set-valued mapping T , whose values are nonempty, convex, compact, and contained in a finite common ball such that G is outer T -differentiable at \mathbf{x}^* (see Section A.5 for these concepts).

REMARK 8. Condition (R)(i) covers commonly used step sizes $\{a_n\}$. Because our main interest here is on the rate of convergence, we simply assume the convergence of \mathbf{X}_n to \mathbf{x}^* . For simplicity of presentation and as a division of labor, we assume the tightness of $\{\frac{\mathbf{X}_n - \mathbf{x}^*}{\sqrt{a_n}}\}$ in (R)(ii). Sufficient conditions ensuring the tightness are given at the end of this section and presented as Proposition 3.1. Regarding (R)(iii), we use the notation as in [17], Chapter 7, pp. 345–346. That is, $\|\cdot\|_p$ denotes the p -norm for $L^p(\Omega, \mathcal{F}, \mathbb{P})$ with $1 \leq p \leq \infty$. It can be shown (see [57]) that:

- (a) $\sum_{i=n}^{m(t_n+t)-1} \sqrt{a_i} [\mathbf{h}(\mathbf{x}^*, \xi_i) - \bar{\mathbf{h}}(\mathbf{x}^*)]$ converges weakly to a Brownian motion $\mathbf{W}(\cdot)$ with covariance $\Sigma_1 t$ as $n \rightarrow \infty$, and
- (b) $\sum_{i=n}^{m(t_n+t)-1} a_i \mathbf{h}_x(\mathbf{x}^*, \xi_i)$ converges in probability to $\bar{\mathbf{h}}_x(\mathbf{x}^*) := A$ as $n \rightarrow \infty$.

THEOREM 3.1. Consider algorithm (3.1) and assume Assumption (R) holds. Then $\{\mathbf{U}^n(\cdot)\}$ converges weakly to the solutions of the following stochastic differential inclusion (see Section A.6 for the definitions)

$$(3.2) \quad d\mathbf{U}(t) \in [A\mathbf{U}(t) + T(\mathbf{U}(t))]dt + \Sigma_1^{1/2} d\bar{\mathbf{W}}(t),$$

if (R)(i)(a) holds, and

$$(3.3) \quad d\mathbf{U}(t) \in [(A + I/2)\mathbf{U}(t) + T(\mathbf{U}(t))]dt + \Sigma_1^{1/2} d\bar{\mathbf{W}}(t),$$

if (R)(i)(b) holds, where $\bar{\mathbf{W}}(t)$ is a d -dimensional standard Brownian motion.

REMARK 9. The main difficulties in deriving the result come from the lack of continuity of $\mathbf{b}_n(\cdot)$ and the handling of the set-valued mappings, provided the normalized noise terms converge (in distribution) to a Wiener process. Although we only state and prove the rate of convergence results for a simple algorithm, similar results for general algorithms can also be obtained with modifications.

PROOF OF THEOREM 3.1. Define $\mathbf{W}^n(\cdot)$ on $(-\infty, \infty)$ by

$$\mathbf{W}^n(t) = \begin{cases} \sum_{i=n}^{m(t_n+t)-1} \sqrt{a_i} [\mathbf{h}(\mathbf{x}^*, \xi_i) - \bar{\mathbf{h}}(\mathbf{x}^*)] & \text{if } t \geq 0, \\ - \sum_{i=m(t_n+t)}^{n-1} \sqrt{a_i} [\mathbf{h}(\mathbf{x}^*, \xi_i) - \bar{\mathbf{h}}(\mathbf{x}^*)] & \text{if } t \leq 0. \end{cases}$$

It is similar to [37] (see also [57], Theorem 10.2.1) that $\{(\mathbf{U}^n(\cdot), \mathbf{W}^n(\cdot))\}$ is tight in $D^d[0, \infty) \times D^d(-\infty, \infty)$. Hence, we can extract a convergent subsequence (still denoted by $\{(\mathbf{U}^n(\cdot), \mathbf{W}^n(\cdot))\}$) such that $\{(\mathbf{U}^n(\cdot), \mathbf{W}^n(\cdot))\}$ converges weakly to a limit, denoted by $(\mathbf{U}(\cdot), \mathbf{W}(\cdot))$. First, Remark 8 yields that $\mathbf{W}(t)$ is a Wiener process with covariance matrix $\Sigma_1 t$. For simplicity of notation, we also assume that the sequence $\{\mathbf{U}_n\}$ is bounded and suppress the truncation step (see, e.g., [37], Theorem 10.2.1). Note that the difficulty in proving Theorem 3.1 comes from the discontinuity of $\mathbf{b}_n(\cdot)$ and the appearance of set-valued mapping $G(\cdot)$. However, this term is assumed to be bounded. Thus, we only need to use the truncated process to handle the smooth term $\mathbf{h}(\cdot)$, which is the reason that the similar truncation step in [37], Theorem 10.2.1, is valid here. To proceed, we work with the case (R)(i)(b); the case (R)(i)(a) can be handled similarly and is thus omitted.

To proceed, we have that

$$\begin{aligned}
 \mathbf{U}_{n+1} &= \left(\frac{a_n}{a_{n+1}} \right)^{1/2} \{ \mathbf{U}_n + \sqrt{a_n}(\bar{\mathbf{h}}(\mathbf{x}^*) + \mathbf{b}_n(\mathbf{X}_n) \\
 &\quad + \mathbf{h}_{\mathbf{x}}(\mathbf{x}^*, \xi_n)(\mathbf{X}_n - \mathbf{x}^*) + a_n O(|\mathbf{U}_n|^2)) \\
 &\quad + \sqrt{a_n}[\mathbf{h}(\mathbf{x}^*, \xi_n) - \bar{\mathbf{h}}(\mathbf{x}^*)] \} \\
 &= \mathbf{U}_n + \left(\left(\frac{a_n}{a_{n+1}} \right)^{1/2} - 1 \right) \mathbf{U}_n + \left(\frac{a_n}{a_{n+1}} \right)^{1/2} \{ a_n \mathbf{h}_{\mathbf{x}}(\mathbf{x}^*, \xi_n) \mathbf{U}_n + a_n \mathbf{v}_n(\mathbf{U}_n) \\
 &\quad + a_n^{3/2} O(|\mathbf{U}_n|^2) + \sqrt{a_n}[\mathbf{h}(\mathbf{x}^*, \xi_n) - \bar{\mathbf{h}}(\mathbf{x}^*)] \},
 \end{aligned}
 \tag{3.4}$$

where

$$\mathbf{v}_n(\mathbf{U}_n) := \frac{\bar{\mathbf{h}}(\mathbf{x}^*) + \mathbf{b}_n(\mathbf{X}_n)}{\sqrt{a_n}}.$$

Let $\delta \in (0, 1)$ be fixed and otherwise arbitrary. Since $G(\cdot)$ is outer T -differentiable at \mathbf{x}^* , there is a neighborhood V of \mathbf{x}^* such that (A.7) holds, that is,

$$G(\mathbf{x}) \subset G(\mathbf{x}^*) + T(\mathbf{x} - \mathbf{x}^*) + \delta |\mathbf{x} - \mathbf{x}^*| B \quad \text{for all } \mathbf{x} \in V.$$

Since \mathbf{X}_n tends to \mathbf{x}^* w.p.1 (Assumption (R)(ii)(a)), for n large enough, we have that

$$\begin{aligned}
 \mathbf{h}(\mathbf{x}^*) + \mathbf{b}_n(\mathbf{X}_n) &\in \mathbf{h}(\mathbf{x}^*) + G(\mathbf{X}_n) \\
 &\subset \mathbf{h}(\mathbf{x}^*) + G(\mathbf{x}^*) + T(\mathbf{X}_n - \mathbf{x}^*) + \delta |\mathbf{X}_n - \mathbf{x}^*| B \\
 &\subset T(\mathbf{X}_n - \mathbf{x}^*) + \delta |\mathbf{X}_n - \mathbf{x}^*| B.
 \end{aligned}
 \tag{3.5}$$

As a consequence,

$$\mathbf{v}_n(\mathbf{U}_n) \in \frac{1}{\sqrt{a_n}} T(\mathbf{X}_n - \mathbf{x}^*) + \delta \frac{|\mathbf{X}_n - \mathbf{x}^*|}{\sqrt{a_n}} B = T(\mathbf{U}_n) + \delta |\mathbf{U}_n| B,$$

where we have used the fact that T is positively homogeneous (in Assumption (R)(iv)). Let

$$M_\delta(\mathbf{x}) := T(\mathbf{x}) + \delta |\mathbf{x}| \bar{B}.$$

Then one has

$$\mathbf{v}_n(\mathbf{U}_n) \in M_\delta(\mathbf{U}_n).$$

Hence, from (3.4), (3.6), $(a_n/a_{n+1})^{1/2} = 1 + o(a_n)$, and $\sum_{i=n}^{m(t_n+t)-1} a_i \mathbf{h}_{\mathbf{x}}(\mathbf{x}^*, \xi_i)$ converges in probability to $\bar{\mathbf{h}}_{\mathbf{x}}(\mathbf{x}^*) := A$ in Remark 8, by the same argument as in the proofs of previous theorems, we obtain that for n large enough

$$\mathbf{U}^n(t) - \mathbf{U}^n(s) \in \int_s^t (A \mathbf{U}^n(r) + M_\delta(\mathbf{U}^n(r))) dr + \mathbf{y}_n(t) - \mathbf{y}_n(s) + \mathbf{W}^n(t) - \mathbf{W}^n(s),$$

where $\mathbf{y}_n(\cdot)$ is some process converging to zero and $\mathbf{W}^n(\cdot)$ converges to $\mathbf{W}(\cdot)$ weakly. Using the Skorohod representation theorem [17], Chapter 3, Theorem 1.8, but without changing notation, we can assume $\mathbf{y}_n(\cdot) + \mathbf{W}^n(\cdot)$ converges to $\mathbf{W}(\cdot)$ w.p.1. Let $\delta_1 \in (0, \delta)$ (depending on δ) be such that

$$T(\mathbf{x} + \delta_1 B) \subset T(\mathbf{x}) + \delta B.$$

Such δ_1 always exists since T is continuous. Because of the convergence of $\mathbf{y}_n(\cdot) + \mathbf{W}^n(\cdot)$ to $\mathbf{W}(\cdot)$, we have that on bounded intervals, for n large, $|\mathbf{y}_n(\cdot) + \mathbf{W}^n(\cdot) - \mathbf{W}(\cdot)| \leq \delta_1/2$. As a consequence, if we let

$$\bar{\mathbf{U}}^n(\cdot) := \mathbf{U}^n(\cdot) - \mathbf{y}_n(\cdot) - \mathbf{W}^n(\cdot) + \mathbf{W}(\cdot),$$

then on bounded intervals, for n large, $|\bar{\mathbf{U}}^n(\cdot) - \mathbf{U}^n(\cdot)| \leq \delta_1/2$, which together with (3.7) implies that for s, t in bounded intervals, for n large,

$$(3.9) \quad \begin{aligned} \bar{\mathbf{U}}^n(t) - \bar{\mathbf{U}}^n(s) &\in \int_s^t (A\mathbf{U}^n(r) + M_\delta(\mathbf{U}^n(r))) dr + \mathbf{W}(t) - \mathbf{W}(s) \\ &\subset \int_s^t (A\bar{\mathbf{U}}^n(r) + \bar{M}_\delta(\bar{\mathbf{U}}^n(r))) dr + \mathbf{W}(t) - \mathbf{W}(s), \end{aligned}$$

where

$$(3.10) \quad \bar{M}_\delta(\mathbf{x}) := T(\mathbf{x}) + \delta(2 + \|A\| + |\mathbf{x}|)\bar{B},$$

and in (3.9), we have used the following facts:

$$|A\mathbf{x} - A\mathbf{y}| \leq \|A\|\|\mathbf{x} - \mathbf{y}\|, \quad \|A\| \text{ is the sup-norm of } A,$$

and if $|\mathbf{x} - \mathbf{y}| < \delta_1$ then $T(\mathbf{y}) \subset T(\mathbf{x}) + \delta B$, due to (3.8). It indicates that on bounded intervals, $\bar{\mathbf{U}}^n(\cdot)$ (for n large) is a solution of

$$(3.11) \quad d\bar{\mathbf{U}}^n(t) \in [A\bar{\mathbf{U}}^n(t) + \bar{M}_\delta(\bar{\mathbf{U}}^n(t))]dt + \Sigma_1^{1/2} d\bar{\mathbf{W}}(t),$$

where $\bar{\mathbf{W}}(t)$ is a d -dimensional standard Wiener process.

To proceed, we state in Lemma 3.1 sufficient conditions for the weak compactness of the set of solutions of stochastic differential inclusions by Kisielewicz in [31] (see also [32, 33]). Then in Lemma 3.2, we verify these conditions.

LEMMA 3.1 (See [31], Theorem 12). *Consider the stochastic differential inclusion*

$$(3.12) \quad d\mathbf{X}(t) \in F_1(\mathbf{X}(t))dt + F_2(\mathbf{X}(t))d\bar{\mathbf{W}}(t).$$

Assume that set-valued mappings $F_1 : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$, $F_2 : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^{d \times d}}$ are measurable and bounded, and have convex values, where F_2 has convex values in the sense that $\{gg^\top : g \in F_2(\mathbf{x})\}$ is convex for each $\mathbf{x} \in \mathbb{R}^d$; and that F_1, F_2 are continuous (see Section A.5 for the definition). Then, for any initial distribution, the set of solutions to (3.12) is sequentially (weakly) closed with respect to the convergence in distribution.

LEMMA 3.2. *For each $\delta > 0$, $\bar{M}_\delta(\cdot)$ is continuous, where $\bar{M}_\delta(\cdot)$ was defined in (3.10).*

PROOF. We prove this lemma by using Lemma A.6. Let $\mathbf{p} \in \mathbb{R}^d$ be arbitrary and consider the map $\sigma(\mathbf{p}, \bar{M}_\delta(\cdot))$, defined by $\sigma(\mathbf{p}, \bar{M}_\delta(\mathbf{x})) := \sup_{\mathbf{a} \in \bar{M}_\delta(\mathbf{x})} \mathbf{p}^\top \mathbf{a}$. We have

$$\begin{aligned} \sigma(\mathbf{p}, \bar{M}_\delta(\mathbf{x})) &= \sup_{\mathbf{a}_1 \in T(\mathbf{x}), \delta_2 \in [0, \delta], \mathbf{e} \text{ is the unit vector in } \mathbb{R}^d} \mathbf{p}^\top (\mathbf{a}_1 + \delta_2(2 + \|A\| + |\mathbf{x}|)\mathbf{e}) \\ &= \sup_{\mathbf{a}_1 \in T(\mathbf{x})} \mathbf{p}^\top \mathbf{a}_1 + \delta(2 + \|A\| + |\mathbf{x}|)|\mathbf{p}| \\ &= \sigma(\mathbf{p}, T(\mathbf{x})) + \delta(2 + \|A\| + |\mathbf{x}|)|\mathbf{p}|. \end{aligned}$$

Since $T(\cdot)$ is continuous, $\sigma(\mathbf{p}, T(\cdot))$ is continuous. As a result, $\sigma(\mathbf{p}, \bar{M}_\delta(\cdot))$ is continuous and then, $\bar{M}_\delta(\cdot)$ is continuous. \square

Since $\mathbf{U}(\cdot)$ is the limit of $\mathbf{U}^n(\cdot)$, it is also the limit of $\bar{\mathbf{U}}^n(\cdot)$. Hence, by Lemmas 3.1 and 3.2, on bounded intervals, $\mathbf{U}(\cdot)$ is such that

$$(3.13) \quad d\mathbf{U}(t) \in [A\mathbf{U}(t) + \bar{M}_\delta(\mathbf{U}(t))]dt + \Sigma_1^{1/2} d\bar{\mathbf{W}}(t), \quad \text{for all } \delta > 0.$$

Because $\mathbf{U}(\cdot)$ is a solution to (3.13), by [31], Lemma 1, we deduce from (3.13) that for any bounded interval $[0, T_0]$ and for all $k \in \mathbb{N}$, there exists $f^k(\cdot)$ such that $f^k(\mathbf{x}) \in \overline{M}_{1/k}(\mathbf{x}) \forall \mathbf{x}$ and for all $s < t \in [0, T_0]$, $\mathbf{U}(t) - \mathbf{U}(s) + \mathbf{W}(t) - \mathbf{W}(s) = \int_s^t A\mathbf{U}(r) dr + \int_s^t f^k(\mathbf{U}(r)) dr$, w.p.1. This yields that

$$(3.14) \quad \mathbf{U}(t) - \mathbf{U}(s) + \mathbf{W}(t) - \mathbf{W}(s) = \int_s^t A\mathbf{U}(r) dr + \int_s^t f^k(\mathbf{U}(r)) dr, \quad \forall k \in \mathbb{N}, \text{ w.p.1.}$$

A consequence of (3.14) is that

$$(3.15) \quad \mathbf{U}(t) - \mathbf{U}(s) + \mathbf{W}(t) - \mathbf{W}(s) - \int_s^t A\mathbf{U}(r) dr \in \int_s^t \overline{M}_{1/k}(\mathbf{U}(r)) dr, \quad \forall k \in \mathbb{N}, \text{ w.p.1.}$$

The $T(\mathbf{x})$ is nonempty, compact, and convex, so is $\overline{M}_{1/k}(\mathbf{x})$. It is readily seen that $\bigcap_{k \in \mathbb{N}} \overline{M}_{1/k}(\mathbf{x}) = T(\mathbf{x}), \forall \mathbf{x}$. Combining this fact together with (3.15) and Proposition 2.1, we have that for all $s, t \in [0, T_0]$

$$\mathbf{U}(t) - \mathbf{U}(s) + \mathbf{W}(t) - \mathbf{W}(s) - \int_s^t A\mathbf{U}(r) dr \in \int_s^t T(\mathbf{U}(r)) dr, \quad \text{w.p.1.}$$

Therefore, we have

$$\mathbf{U}(t) - \mathbf{U}(s) \in \int_s^t [A\mathbf{U}(r) + T(\mathbf{U}(r))] dr + \int_s^t \Sigma_1^{1/2} d\overline{\mathbf{W}}(r), \quad \text{w.p.1.}$$

Equivalently, $\mathbf{U}(\cdot)$ is a solution to

$$d\mathbf{U}(t) \in [A\mathbf{U}(t) + T(\mathbf{U}(t))] dt + \Sigma_1^{\frac{1}{2}} d\overline{\mathbf{W}}(t).$$

The proof is complete. \square

Tightness of normalized sequence. In Assumption (R)(ii), we assumed the tightness of the normalized sequence as a division of labor. To end this section, we provide sufficient conditions for the tightness of sequence $\{\frac{\mathbf{X}_n - \mathbf{x}^*}{\sqrt{a_n}}\}$ for large n . These conditions are essentially concerned with the stability of the limit point \mathbf{x}^* . We will obtain the tightness by adapting and modifying the perturbed Lyapunov functional method for differential inclusions. Such a method was first used in the treatment of partial differential equations and stochastic analysis, and later on used for many different stochastic systems in [37]. Here, we modify this idea to treat our cases. [The assumptions given below are not restrictive. In fact, in many applications, $V(\mathbf{x}) = |\mathbf{x}|^2$ can be used as a simple but promising candidate, which is shown in Section 4. In addition, locally quadratic Lyapunov functions (see [37]) can also be considered.] We state a proposition below. A sketch of the proof is relegated to the Appendix A.7.

PROPOSITION 3.1. *Consider algorithm (3.1) with $\mathbf{b}_n(\mathbf{x}) \in G(\mathbf{x}), \forall n$, $G(\mathbf{x})$ is a set-valued mapping; and suppose \mathbf{X}_n is bounded and converges to \mathbf{x}^* w.p.1 and Assumption (R)(iii) holds. Assume that there is a function $V : \mathbb{R}^d \rightarrow \mathbb{R}$ such that:*

- $V(\mathbf{x}^*) = 0, V(\mathbf{x}) > 0$ for each $\mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{x}^*, V(\cdot)$ together with its partial derivatives up to the second order in \mathbf{x} is continuous; $|V_{\mathbf{x}}(\mathbf{x})|^2 \leq K(1 + V(\mathbf{x})), V_{\mathbf{xx}}(\cdot)$ is uniformly bounded; and $V(\mathbf{x}) \geq c_0|\mathbf{x} - \mathbf{x}^*|^2 + o(|\mathbf{x} - \mathbf{x}^*|^2)$ as $\mathbf{x} \rightarrow \mathbf{x}^*$ for some positive constant c_0 ; and
- there is a $\lambda > 0$ such that $\max \overline{\dot{V}}^{G+\overline{\mathbf{h}}}(\mathbf{x}) \leq -\lambda V(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{x}^*$ (where $\overline{\dot{V}}^{G+\overline{\mathbf{h}}}$ is the set-valued derivative of V with respect to the set $G + \overline{\mathbf{h}}$, see definition A.4(iii)); and
- for each n , each $\mathbf{x} \in \mathbb{R}^d$, and each $\xi, |\mathbf{b}_n(\mathbf{x})|^2 + |\mathbf{h}(\mathbf{x}, \xi)|^2 \leq K(1 + V(\mathbf{x}))$.

Moreover, assume the sequence of step sizes $\{a_n\}$ satisfies either $a_n = \frac{1}{n}$ and $\lambda > 1$, or $a_n \rightarrow 0$, and for each $T > 0$

$$\liminf_n \min_{n \geq i \geq m(t_n - T)} \frac{a_n}{a_i} = 1,$$

where $t_n, m(t)$ are defined at the beginning. Then, there is an N such that $\{\frac{\mathbf{x}_n - \mathbf{x}^*}{\sqrt{a_n}}; n \geq N\}$ is tight in \mathbb{R}^d .

4. Applications. In this section, we apply our results developed in previous sections to a number of application examples.

4.1. Stochastic subgradient descent. We begin with the description under a deterministic setup. Suppose that we aim to find the minimizers of a loss function $L(\mathbf{w})$, that is, $\arg\min_{\mathbf{w} \in \mathbb{R}^d} L(\mathbf{w})$. If $L(\mathbf{w})$ is continuously differentiable with respect to \mathbf{w} , the minimizer \mathbf{w}^* is a solution of the equation $\nabla_{\mathbf{w}} L(\mathbf{w}) = 0$. In this case, we can find the optimum by gradient descent algorithms as usual. However, if $L(\mathbf{w})$ is only strictly convex and not differentiable, we cannot define the gradient $\nabla_{\mathbf{w}} L(\mathbf{w})$. Rather, we define its subgradient $\partial L(\mathbf{w})$ as $\partial L(\mathbf{w}) := \{\mathbf{m} \in \mathbb{R}^d : L(\mathbf{y}) \geq L(\mathbf{w}) + \mathbf{m}^\top (\mathbf{y} - \mathbf{w}), \forall \mathbf{y} \in \mathbb{R}^d\}$. Hence, the minimizer \mathbf{w}^* satisfies $\mathbf{0} \in \partial L(\mathbf{w}^*)$. The algorithm for the minimization is of the form $\mathbf{w}_{n+1} = \mathbf{w}_n - a_n \mathbf{g}_n$, for some $\mathbf{g}_n \in \partial L(\mathbf{w}_n)$. Assume that $L(\cdot)$ can be decomposed into two components, one satisfies certain smooth conditions and the other verifies convexity. Then we can assume $\partial L(\mathbf{w}) = \bar{\mathbf{h}}(\mathbf{w}) + G(\mathbf{w})$, where $\bar{\mathbf{h}}$ is a continuous function and $G(\mathbf{w})$ is a set-valued mapping. Our objective is to find the minimizer \mathbf{w}^* satisfying $\mathbf{0} \in \partial L(\mathbf{w}^*)$.

When noisy observations or measurements are involved, $\partial L(\mathbf{w}_n)$ is often not available. As a result, we use $\tilde{\mathbf{g}}_n$, which is an unbiased or biased estimator of $\partial L(\mathbf{w}_n)$. With noisy observations or measurements, we can write the estimator of $\tilde{\mathbf{g}}_n$ as

$$(4.1) \quad \tilde{\mathbf{g}}_n = \mathbf{b}_n(\mathbf{w}_n, \xi_n) + \mathbf{h}(\mathbf{w}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \beta_n,$$

where $\mathbf{h}(\cdot, \cdot)$ is a smooth (w.r.t. \mathbf{w}) function that will be averaged out to $\bar{\mathbf{h}}$ (or a neighborhood of $\bar{\mathbf{h}}$ if it involves some bias term that is not asymptotically negligible), $\mathbf{b}_n(\cdot, \cdot)$ is bounded with values belonging to a set-valued function $G(\cdot)$, $\xi_n, \zeta_n, \tilde{\zeta}_n$ are the noises, and $\mathbf{h}_0(\tilde{\zeta}_n)$ and β_n can be either averaged out or asymptotically bounded by η when the bias cannot be ignored.

Using (4.1), we construct the algorithm

$$(4.2) \quad \mathbf{w}_{n+1} = \mathbf{w}_n - a_n [\mathbf{b}_n(\mathbf{w}_n, \xi_n) + \mathbf{h}(\mathbf{w}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \beta_n],$$

or its projected version

$$(4.3) \quad \begin{cases} \tilde{\mathbf{w}}_{n+1} = \mathbf{w}_n - a_n [\mathbf{b}_n(\mathbf{w}_n, \xi_n) + \mathbf{h}(\mathbf{w}_n, \zeta_n) + \mathbf{h}_0(\tilde{\zeta}_n) + \beta_n], \\ \mathbf{w}_{n+1} = \Pi_H(\tilde{\mathbf{w}}_{n+1}). \end{cases}$$

Then, under our conditions, in algorithms (4.2) or (4.3), \mathbf{w}_n converges w.p.1 to the minimizer \mathbf{w}^* . We can also obtain robustness and rates of convergence of these algorithms by applying Theorems 2.6 and 3.1. Our proposed conditions are mild and can be verified. The assumptions in the noises are mild and can be verified by many common noise sequences such as i.i.d. sequences, martingale difference sequences, mixing noise, etc. Note also that the boundedness of nonsmooth term \mathbf{b} and local boundedness of smooth term \mathbf{h} are often clear if we use projection algorithms and/or the noise does not make the iterates blow-up. Only conditions for stability (such as (KS), (GS), (PS)) need to be verified carefully. However, it is shown later that many algorithms in the literature satisfy these conditions. Some specific examples (e.g., Lasso algorithm for high-dimensional statistics, and Pegasos algorithm in support vector machine (SVM) classification) will be studied next and some numerical results will be given in Section 4.7.

REMARK 10. Note that stochastic subgradient descent algorithms are used often in machine learning to minimize a loss function for online learning in which the loss function can often be nonsmooth. When the number of data in the training set is large, because computational cost using exact subgradient is expensive, sampling or mini-batching computations are needed. In the literature of machine learning, some existing studies are based purely on establishing a kind of “contraction estimate” (in the sense of in expectation); see for example, [21, 46, 51, 52] and references therein. For example, convergence in expectation was proved in [21, 46] and references therein, or the convergence in probability and almost surely of the sequence $\{\min_{1 \leq k \leq n} \|\mathbf{w}_k\|\}_{n=1}^{\infty}$ were obtained in [41]. Recently, convergence of stochastic subgradient descent algorithms using differential inclusions was also obtained in [8, 9, 16]. In this paper, our effort is to provide a new approach for analyzing the convergence, rates of convergence, robustness of stochastic subgradient algorithms, and other algorithms in nonsmooth optimization by characterizing their behaviors using dynamical systems generated from differential inclusions and stochastic differential inclusions. As a direct application of our results, if the corresponding differential inclusion has the minimizer as a globally asymptotically stable point, then we can obtain the almost sure convergence of the algorithm to the minimizer. Our convergence analysis recovers and generalizes convergence results in stochastic subgradient descent algorithms in the literature. Moreover, we also use a novel Lyapunov functional method to verify the globally asymptotic stability, which was presented in Section 2. The approach appears to be effective. In addition, the rates of convergence and robustness can also be deduced from our results.

REMARK 11. Recently, some other variants of stochastic subgradient or gradient descent algorithms for nonsmooth and/or nonconvex optimization are studied widely, including incremental subgradient descent [24, 34], proximal algorithms and stochastic proximal algorithms [42], perturbed proximal primal dual algorithm [23], smoothing methods [11], gradient sampling methods [10], among others. Nevertheless, the central issue is the handling of set-valued mappings and nonsmooth loss functions. Although we plan not to dwell on each of such algorithms, using our results we can treat such algorithms and obtain respective convergence results.

REMARK 12. In the next two sections, we present how our results can be applied to study algorithms in L^1 -norm penalized (regularized) minimization and support vector machine (SVM) classification. We will only focus on verifying the stability conditions since assumptions in the noises are mild and can be verified by many common noise sequences such as i.i.d. sequences, martingale difference sequences, mixing noise, etc. It is worth noting that although we will not state explicitly the results for algorithms in L^1 -norm regularized minimization in Section 4.2 and SVM classification problem in Section 4.3, our results on convergence (Theorems 2.3, 2.4, and 2.5), robustness (Theorem 2.6), and rates of convergence (Theorem 3.1) hold for these algorithms. These results recover, improve, and further the state-of-art development for Lasso and SVM algorithms.

4.2. L^1 -norm penalized (regularized) minimization: Lasso algorithms, least absolute deviation (LDA) estimators. We consider stochastic algorithms for minimizing loss functions containing L^1 -norm, by providing explicit computations for Lasso algorithm since other cases are similar. Let us start with the following optimization problem. Given a sequence of i.i.d. random variables $\{\mathbf{x}_n, y_n\}$, with $\mathbf{x}_n \in \mathbb{R}^d$, $y \in \mathbb{R}$, we wish to find the weight vector \mathbf{w} so that $\mathbf{x}_n^\top \mathbf{w}$ best matches y_n in the sense $\mathbb{E}\|\mathbf{x}_n^\top \mathbf{w} - y_n\|^2$ is minimized with the constraint $\sum_{i=1}^d |w_i| = 0$, which can be recast into the following problem:

$$(4.4) \quad \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} L(\mathbf{w}), \quad L(\mathbf{w}) = \frac{1}{2} \mathbb{E} \|\mathbf{x}_n^\top \mathbf{w} - y_n\|^2 + \lambda \sum_{i=1}^d |w_i|.$$

Alternatively, we are trying to find \mathbf{w}^* such that $\mathbf{0} \in \bar{\mathbf{h}}(\mathbf{w}^*) + G(\mathbf{w}^*)$, where

$$(4.5) \quad \begin{aligned} \bar{\mathbf{h}}(\mathbf{w}) &= -\frac{1}{2} \nabla_{\mathbf{w}} \mathbb{E} \|\mathbf{x}_n^\top \mathbf{w} - y_n\|^2, \\ G(\mathbf{w}) &= \mathcal{K}(w_1) \times \cdots \times \mathcal{K}(w_d), \quad \text{with } \mathbf{w} = (w_1, \dots, w_d)^\top, \end{aligned}$$

and

$$\mathcal{K}(w_i) = \begin{cases} \{-\lambda\} & \text{if } w_i > 0, \\ [-\lambda, \lambda] & \text{if } w_i = 0, \\ \{\lambda\} & \text{if } w_i < 0. \end{cases}$$

A stochastic algorithm can be constructed as

$$(4.6) \quad \mathbf{w}_{n+1} = \mathbf{w}_n + a_n(y_n - \mathbf{w}_n^\top \mathbf{x}_n) \mathbf{x}_n + a_n g_n(\mathbf{w}_n),$$

where $g_n(\mathbf{w}_n) \in G(\mathbf{w}_n)$ with $G(\cdot)$ defined in (4.5); and the projection algorithm can be written as

$$(4.7) \quad \begin{cases} \tilde{\mathbf{w}}_{n+1} = \mathbf{w}_n + a_n(y_n - \mathbf{w}_n^\top \mathbf{x}_n) \mathbf{x}_n + a_n g_n(\mathbf{w}_n), \\ \mathbf{w}_{n+1} = \Pi_H(\tilde{\mathbf{w}}_{n+1}), \end{cases}$$

with H being a compact and convex set. While the other assumptions are easily verified, the stability assumption needs to be checked carefully. We verify condition (GS) for algorithm (4.6) later in Proposition 4.1, whose proof is postponed to Section 4.6.

Note that loss functions defined as the sum of the errors of prediction and the L^1 -norm regularization are often used in dimension reduction problem in high-dimensional statistics [20], in which, the L^1 -norm is used to penalize the dimension of subspace that we are trying to project onto. Roughly, $\sum_{i=1}^d |w_i|$ cannot be large causing w_i to be small for all $i \in \{1, \dots, d\}$. If we use the squared norm, all w_i would bare the same weight. If we use the absolute deviation, some “less-informative” coordinates will be highlighted and leads to $w_i = 0$ for such coordinates. More intuitively, in a two-dimensional case, from a geometric point of view, the unit ball in L^1 -norm is of diamond shape with four vertices instead of a circle in L^2 -norm so that the optimal value will often be obtained on some axis. For more intuition on the Lasso algorithm as well as L^1 -norm penalization, we refer to the work by Tibshirani in [54].

REMARK 13. In practice, the above algorithms may need to be modified such as stochastic coordinate descent (SCD), truncated gradient (TruncGrad), etc., to be more effective in real data and/or in the problem of inducing sparsity [38]. The convergence of these modified algorithms can be obtained by applying our results with modifications. Here, we only discuss a simple version of the algorithm. There are other algorithms, which minimize loss functions containing absolute norm such as robust regression and least absolute deviation (LAD) with/without Lasso [26, 55]. Algorithms (4.6) and (4.7) and their variants are widely applied by the machine learning community in applications with a large-scale data set [26, 38, 55]. Our main effort in this section is to verify the stability condition that is the most important one for the analysis. It is noted that in many practical algorithms minimizing L^1 -norm loss functions, different types of stepsizes may be used. The approach and results in this paper can be adapted for these settings by using suitably scaled interpolated sequences.

PROPOSITION 4.1. Assume that $\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]$ is a positive definite matrix. Let $G^*(\mathbf{w}) = \bar{\mathbf{h}}(\mathbf{w} + \mathbf{w}^*) + G(\mathbf{w} + \mathbf{w}^*)$, where $\bar{\mathbf{h}}(\cdot)$, $G(\cdot)$ are as in (4.5) and $V(\mathbf{w}) = |\mathbf{w}|^2$, $U(\mathbf{w}) = \sum_{i=1}^d w_i$. Then $\dot{\bar{V}}_{\{U\}}^{G^*}(\mathbf{w}) \leq -c_1 |\mathbf{w}|^2$, where $c_1 > 0$ is the smallest eigenvalue of $\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]$.

4.3. Support vector machine (SVM) classification. We first consider a stochastic optimization problem and then treat the problem of support vector machine (SVM) classification problem. The Pegasos algorithm will be introduced next. Consider the following problem: minimize $L(\mathbf{w}) := \frac{\lambda}{2} \|\mathbf{w}\|^2 + \max\{0, 1 - \mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n]\}$, where (\mathbf{x}_n, y_n) is a sequence of i.i.d. random variables. The stochastic version of subgradient descent algorithm for this problem is as follows

$$(4.8) \quad \mathbf{w}_{n+1} = \mathbf{w}_n - a_n \lambda \mathbf{w}_n + a_n g_n(\mathbf{w}_n, \mathbf{x}_n, y_n),$$

where $g_n(\mathbf{w}_n, \mathbf{x}_n, y_n) \in \partial(-\max\{0, 1 - y_n \mathbf{w}_n^\top \mathbf{x}_n\})$, that is,

$$g_n(\mathbf{w}_n, \mathbf{x}_n, y_n) \in \begin{cases} \{\mathbf{0}\} & \text{if } y_n \mathbf{w}_n^\top \mathbf{x}_n > 1, \\ \overline{\text{co}}\{\mathbf{0}, y_n \mathbf{x}_n\} & \text{if } y_n \mathbf{w}_n^\top \mathbf{x}_n = 1, \\ \{y_n \mathbf{x}_n\} & \text{if } y_n \mathbf{w}_n^\top \mathbf{x}_n < 1; \end{cases}$$

or as the following projection algorithm with the set H being a compact and convex set,

$$(4.9) \quad \begin{cases} \tilde{\mathbf{w}}_{n+1} = \mathbf{w}_n - a_n \lambda \mathbf{w}_n + a_n g_n(\mathbf{w}_n, \mathbf{x}_n, y_n), \\ \mathbf{w}_{n+1} = \Pi_H(\tilde{\mathbf{w}}_{n+1}). \end{cases}$$

Applying our results, the convergence to the optimal point, robustness, rates of convergence of algorithm (4.8) (as well as algorithm (4.9)) can be obtained under conditions in our setup. We will verify the stability condition (GS) for algorithm (4.8) later in Proposition 4.2, whose proof is postponed to Section 4.6. The corresponding numerical example is given in Example 4.2 in Section 4.7.

Algorithms (4.8) and (4.9) can be recast into a form known as Pegasos algorithms and widely applied to support vector machine (SVM) classification problem; SVM is an effective and a popular classification learning tool [15]. More intuition, motivation, and details of the hinge loss function as well as the above loss function in SVM classification can be found in [15, 49, 50] and references therein. Algorithms (4.8) and (4.9) as well as their modified versions were studied in [49, 50] and references therein. However, the convergence was only given in high probability, not w.p.1. By applying our results, the convergence w.p.1 is obtained. The applications of algorithms (4.8) and (4.9) to classification problem in large-scale data can be found in [49, 50] and references therein.

PROPOSITION 4.2. *Let $\bar{\mathbf{h}}(\mathbf{w}) = -\lambda \mathbf{w}$, and*

$$G_1(\mathbf{w}) = \begin{cases} \{\mathbf{0}\} & \text{if } \mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] > 1, \\ \overline{\text{co}}\{\mathbf{0}, \mathbb{E}[y_n \mathbf{x}_n]\} & \text{if } \mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] = 1, \\ \{\mathbb{E}[y_n \mathbf{x}_n]\} & \text{if } \mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] < 1, \end{cases}$$

and $G^(\mathbf{w}) = \bar{\mathbf{h}}(\mathbf{w} + \mathbf{w}^*) + G_1(\mathbf{w} + \mathbf{w}^*)$, and $V(\mathbf{w}) = \|\mathbf{w}\|^2$, $U(\mathbf{w}) = \sum_{i=1}^d w_i$. Then $\frac{\dot{G}^*}{V_{\{U\}}}(\mathbf{w}) \leq -\lambda \|\mathbf{w}\|^2$.*

4.4. Root finding for set-valued mappings. In this section, we demonstrate the effectiveness of our results in proving convergence of a stochastic approximation algorithm for set-valued mappings. Assume that we need to find zero points of a set-valued mapping $G(\cdot)$, that is, find \mathbf{w}^* such that $\mathbf{0} \in G(\mathbf{w}^*)$, where $G: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is as follow $G(\mathbf{w}) = (-w_1 + w_2 + h(w_2), -w_1 - w_2 + h(w_1))$ with $\mathbf{w} = (w_1, w_2)^\top$, and $h(\cdot): \mathbb{R} \rightarrow 2^{\mathbb{R}}$ is defined as

$$h(w) = \begin{cases} 0 & \text{if } w \neq 1, \\ [-1, 1] & \text{if } w = 1. \end{cases}$$

When only noisy observations or measurements are available, a stochastic approximation algorithm for root finding takes the form

$$(4.10) \quad \mathbf{w}_{n+1} = \mathbf{w}_n + a_n(f(\mathbf{w}_n, \xi_n) + \beta_n), \quad f(\mathbf{w}_n, \xi_n) \in G(\mathbf{w}_n),$$

where $\{a_n\}$ is a sequence of step sizes and $\{\beta_n\}$ is a sequence of two-dimensional i.i.d. random variables that are normally distributed with mean $\mathbf{0}$ and identity covariance matrix.

We compare our results with the results in [5] as well as other approaches in studying stability of differential inclusions for applications to stochastic approximations. Under boundedness assumption of $\{\mathbf{w}_n\}$ (or using projection algorithm to convex and compact set), we obtain the limit points of $\{\mathbf{w}_n\}$ are contained in the set of chain-recurrent points of the limit system $\dot{\mathbf{w}}(t) \in G(\mathbf{w}(t))$. Using results in [5], Section 3 and 4, to prove $\{\mathbf{w}_n\}$ converges to $\mathbf{0}$, we need to construct a Lyapunov function V such that $\nabla V(\mathbf{w})g(\mathbf{w}) < 0$, $\forall g(\mathbf{w}) \in G(\mathbf{w})$ for all $\mathbf{w} \neq \mathbf{0}$. Consider a candidate Lyapunov function $V(\mathbf{w}) = \|\mathbf{w}\|^2$. Then

$$\nabla V(\mathbf{w})g(\mathbf{w}) = -\|\mathbf{w}\|^2 + w_1g_1(w_2) + w_2g_2(w_1),$$

$$\text{where } g_1(w_2) \in h(w_2), g_2(w_1) \in h(w_1).$$

At $\mathbf{w} = (1, 1)^\top$, one possibility is that $\nabla V((1, 1))g((1, 1)) = (-\|\mathbf{w}\|^2 + w_1 + w_2)|_{\mathbf{w}=(1,1)^\top} = 0$. So, we cannot guarantee the set $\{\mathbf{0}\}$ to be a globally stable and attracting set. That is, we cannot prove that $\{\mathbf{w}_n\}$ converges to $\mathbf{0}$ using this Lyapunov function. However, using our results, we can prove that $\{\mathbf{w}_n\}$ tends to $\mathbf{0}$ w.p.1 by using the \mathcal{U} -generalized Lyapunov function corresponding to such a candidate function. Condition (GS) needs to be verified, and it is stated in the following proposition, whose proof is in Section 4.6. Roughly speaking, compared with the existing results in the literature, our approach allows one to ignore some “less important” points (e.g., the point $(1, 1)$ above), that may make a (promising) candidate Lyapunov function not satisfy the conditions for the stability in the literature though they generally do not affect the stability of the systems. Moreover, in fact, our setting even allows $f(\mathbf{w}_n, \xi_n)$ to be in a neighbor of $G(\mathbf{w}_n)$ with (random) radius averaged out to 0. A numerical example is given in Example 4.3 in Section 4.7.

PROPOSITION 4.3. *Let $V(\mathbf{w}) = \|\mathbf{w}\|^2$ and*

$$U(\mathbf{w}) = \max\{w_1 - 1, 0\} - \min\{w_1 + 1, 0\} + \max\{w_2 - 1, 0\} - \min\{w_2 + 1, 0\}.$$

Then, one has $\dot{\bar{V}}_{\{U\}}^G(\mathbf{w}) \leq -\|\mathbf{w}\|^2$, $\forall \mathbf{w}$.

4.5. Multistage decision making with partial observations. Let \mathcal{E} and \mathcal{B} be measurable spaces denoting the action space and the state space, respectively. Suppose that $\mathcal{O} \subset \mathbb{R}^d$ is a convex and compact set denoting the outcome space. At discrete times $n = 1, 2, \dots$, a decision maker chooses an action e_n from \mathcal{E} and observes an outcome $M(e_n, b_n)$, where $M : \mathcal{E} \times \mathcal{B} \rightarrow \mathcal{O}$ is a (measurable) function. However, it is worth noting that the outcome is not always observable in application but is only partially observed with noise. So, the exact outcome $M(e_n, b_n)$ is not available for the decision maker, but only noise corrupted outcome $\tilde{M}(e_n, b_n, \xi_n)$ is available, where ξ_n represents the noise.

Thus, we consider the following multistage decision making model with partial observations: (1) the sequence $\{(e_n, b_n)\}_{n \geq 0}$ and $\{\xi_n\}_{n \geq 0}$ are random processes defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and adapted to the filtration $\{\mathcal{F}_n\}$ and the noise sequence $\{\xi_n\}$ satisfies that for some $T > 0$, each $\varepsilon > 0$, and each $(e, b) \in \mathcal{E} \times \mathcal{B}$,

$$(4.11) \quad \lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(t)}^{m(jT+t)-1} \frac{1}{i+1} (M(e, b) - \tilde{M}(e, b; \xi_i)) \right| \geq \varepsilon \right\} = 0,$$

where $m(t) := \max\{n \in \mathbb{N} : \sum_{i=1}^n \frac{1}{i} \leq t\}$; (2) the action of the decision maker is independent of the environment if provided the past information $\{(e_1, b_1), \dots, (e_n, b_n)\}$, that is, $\mathbb{P}((e_{n+1}, b_{n+1}) \in de \times db | \mathcal{F}_n) = \mathbb{P}(e_{n+1} \in de | \mathcal{F}_n) \mathbb{P}(b_{n+1} \in db | \mathcal{F}_n)$; (3) the decision maker records only the cumulative average of the past (partially observed) outcomes,

$$(4.12) \quad \mathbf{X}_n = \frac{1}{n} \sum_{i=1}^n \tilde{M}(e_i, b_i; \xi_i);$$

(4) her/his decisions are based on this average, that is, $\mathbb{P}(e_{n+1} \in de | \mathcal{F}_n) = Q_{\mathbf{X}_n}(de)$, where for each $\mathbf{x} \in \mathcal{O}$, $Q_{\mathbf{x}}(\cdot)$ is a probability measure (in \mathcal{E}), and for each measurable set $A \subset \mathcal{E}$, the map: $\mathbf{x} \in \mathcal{O} \rightarrow Q_{\mathbf{x}}(A) \in [0, 1]$ is measurable. The family $Q = \{Q_{\mathbf{x}} : \mathbf{x} \in \mathcal{O}\}$ is termed a strategy for the decision maker.

DEFINITION 4.1 (Blackwell's approachability). A set $E \subset \mathcal{O}$ is said to be approachable if there exists a strategy Q such that $\mathbf{X}_n \rightarrow E$ w.p.1.

Directed calculations show that

$$(4.13) \quad \mathbf{X}_{n+1} = \mathbf{X}_n + \frac{1}{n+1} (-\mathbf{X}_n + \tilde{M}(e_{n+1}, b_{n+1}; \xi_{n+1})).$$

For each $\mathbf{x} \in \mathcal{O}$, let $G_1(\mathbf{x}) = \{\int_{\mathcal{E} \times \mathcal{B}} M(e, b) Q_{\mathbf{x}}(de) \nu(db) : \nu \in \mathcal{P}(\mathcal{B})\}$, where $\mathcal{P}(\mathcal{B})$ is the set of probability measures over \mathcal{B} . Define $G(\mathbf{x}) = -\mathbf{x} + \text{co} G_1(\Pi_{\mathcal{O}}(\mathbf{x}))$, where $\Pi_{\mathcal{O}}(\cdot)$ is the (orthogonal) projection (onto \mathcal{O}) operator. Applying our results (Theorems 2.2, 2.6, and 3.1), we obtain following results.

THEOREM 4.1. *Under the above settings, the following claims hold.*

(1) *The limit of any convergent subsequence of the shifted sequence of linear continuous time interpolated processes of (4.12) is a solution of the following differential inclusion w.p.1*

$$(4.14) \quad \dot{\mathbf{X}}(t) \in G(\mathbf{X}(t)).$$

(2) *If there is a strategy Q such that E is a globally asymptotically stable set of differential inclusion (4.14), then E is approachable.*

(3) *If there exists a strategy Q such that $E = \{\mathbf{x}^*\}$ is a unique approachable set, then under further technical conditions (as in Theorem 3.1), the limit processes of convergent subsequences of shifted interpolated processes generated by normalized sequence $\frac{\mathbf{X}_n - \mathbf{x}^*}{\sqrt{n}}$ converges weakly to solutions of a stochastic differential inclusion.*

(4) *If the "convergence to 0" condition (4.11) is relaxed as $|M(e, b) - \tilde{M}(e, b, \xi)| < \eta, \forall e, b, \xi$, w.p.1, then the conclusions (1) and (2) still hold with G in (4.14) being replaced by its neighbor with radius η . Moreover, if E_{η} is a globally asymptotically stable set of the corresponding (limit) differential inclusions (and thus, is a approachable set), then there is a (deterministic) nondecreasing function $\phi(\cdot)$ satisfying $\lim_{t \rightarrow 0} \phi(t) = 0$ such that $\text{distance}(E_{\eta}, E) \leq \phi(\eta)$.*

REMARK 14. The studies on the stability of differential inclusions can be found in Appendix A.3 (see also [5] and references therein). [As was noted, in some cases (e.g., as in Section 4.4), the \mathcal{U} -generalized Lyapunov condition presented in this work is more effective than the stability conditions counterpart in existing results.] Multistage decision making models (without partial observations) was considered in [5]. In this application, we allow the outcome to be partially observed under noise by the decision maker. In addition, we characterize the limit processes as solutions rather than perturbed solutions of the limit differential

inclusion, and we also obtain results of rates of convergence and robustness. This example can be further generalized to treat other criteria such as overtaking, bias, and other so-called advanced criteria of optimality, as well as other systems such as switching dynamical systems. We refer the reader to [28] and references therein. Some other applications to Markov decision process using stochastic approximation can be found in [45] and references therein.

4.6. Proof of theorems in Section 4.

PROOF OF PROPOSITION 4.1. The Clarke gradient of $U(\mathbf{w})$ is given by $\partial U(\mathbf{w}) = (1, \dots, 1)^\top$. As a consequence,

$$\tilde{G}_{\{U\}}^*(\mathbf{w}) = G^*(\mathbf{w}).$$

Moreover, V is continuously differentiable, $\partial V(\mathbf{w}) = (2w_1, \dots, 2w_2)^\top$. Therefore, the $\{U\}$ -generalized derivative of V in the direction G is given by

$$(4.15) \quad \dot{\tilde{V}}_{\{U\}}^{G^*}(\mathbf{w}) = \max_{\mathbf{q} \in \tilde{G}_{\{U\}}^*(\mathbf{w})} (\partial V(\mathbf{w}))^\top \mathbf{q} = \max_{\mathbf{q} \in \tilde{G}_{\{U\}}^*(\mathbf{w})} 2\mathbf{w}^\top \mathbf{q}.$$

Noting that for any $\mathbf{q} \in \tilde{G}_{\{U\}}^*(\mathbf{w})$,

$$\mathbf{q} = \bar{\mathbf{h}}(\mathbf{w} + \mathbf{w}^*) + \bar{\mathbf{q}}, \quad \text{for } \bar{\mathbf{q}} \in G(\mathbf{w} + \mathbf{w}^*),$$

and hence,

$$(4.16) \quad \mathbf{w}^\top \mathbf{q} = -\mathbf{w}^\top \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top] \mathbf{w} + \mathbf{w}^\top [\mathbb{E} \mathbf{x}_n (\mathbf{x}_n^\top \mathbf{w}^* - y) + \bar{\mathbf{q}}].$$

Since \mathbf{w}^* is the minimizer, one has

$$\mathbf{0} \in \bar{\mathbf{h}}(\mathbf{w}^*) + G(\mathbf{w}^*).$$

In particular,

$$\mathbf{0} \in -\mathbb{E} \mathbf{x}_n (\mathbf{x}_n^\top \mathbf{w}^* - y) + G(\mathbf{w}^*).$$

Hence, it is equivalent to

$$(4.17) \quad \mathbb{E} \mathbf{x}_n (\mathbf{x}_n^\top \mathbf{w}^* - y) \in -G(\mathbf{w}^*).$$

LEMMA 4.1. For any $w_i, k \in -\mathcal{K}(w_i^*), \bar{q}_i \in \mathcal{K}(w_i + w_i^*)$,

$$(4.18) \quad w_i(k + \bar{q}_i) \leq 0.$$

As a consequence, for all $\mathbf{w} \in \mathbb{R}^d$, one has

$$(4.19) \quad \mathbf{w}^\top (-G(\mathbf{w}^*) + G(\mathbf{w} + \mathbf{w}^*)) \leq 0,$$

where, (4.19) is understood as

$$\mathbf{w}^\top (\mathbf{k} + \bar{\mathbf{q}}) \leq 0 \quad \text{for all } \mathbf{k} \in -G(\mathbf{w}^*), \bar{\mathbf{q}} \in G(\mathbf{w} + \mathbf{w}^*).$$

PROOF. Three cases are considered.

Case 1: $w_i^* = 0$. (4.18) is equivalent to

$$w_i(k + \bar{q}_i) \leq 0, \quad \text{for all } k \in [-\lambda, \lambda], \bar{q}_i \in \mathcal{K}(w_i).$$

If $w_i = 0$, it is obvious. If $w_i > 0$ then $\bar{q}_i = -\lambda$ and $k + \bar{q}_i \leq 0$ for all $k \in [-\lambda, \lambda]$. If $w_i < 0$ then $\bar{q}_i = \lambda$ and $k + \bar{q}_i \geq 0$ for all $k \in [-\lambda, \lambda]$.

Case 2: $w_i^* > 0$. (4.18) is equivalent to

$$(4.20) \quad w_i(\lambda + \bar{q}_i) \leq 0, \quad \text{for all } \bar{q}_i \in \mathcal{K}(w_i + w_i^*).$$

Since $\lambda + \bar{q}_i \geq 0$ for all $\bar{q}_i \in \mathcal{K}(w_i + w_i^*)$, if $w_i \leq 0$, (4.20) is clear. If $w_i > 0$ then $\lambda + \bar{q}_i = 0$ (due to $\mathcal{K}(w_i + w_i^*) = \{-\lambda\}$) and (4.20) holds.

Case 3: $w_i^* < 0$. This case is similar to case 2. The proof of the lemma is complete. \square

Combining (4.15), (4.16), (4.17), and Lemma 4.1, we obtain that

$$\dot{\bar{V}}_{\{U\}}^{G^*}(\mathbf{w}) \leq -\mathbf{w}^\top \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top] \mathbf{w}.$$

Since $\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]$ is positive definite and by Rayleigh's inequality (see, e.g., [12], Chapter 3), one has

$$\mathbf{w}^\top \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top] \mathbf{w} \geq c_1 \|\mathbf{w}\|^2,$$

where $c_1 > 0$ is the smallest eigenvalue of $\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top]$. Therefore, the proposition is proved. \square

REMARK 15. In practice, to guarantee the boundedness of \mathbf{w}_n , we can use a projection algorithm with a hyper-rectangle $H := \{\mathbf{w} \in \mathbb{R}^d : -h \leq w_i \leq h, \forall i\}$ with $h > \lambda$ being sufficiently large. In this case, the proof of Proposition 4.1 for the projection case is similar. Moreover, the above proof can be simplified by applying Theorem 2.3 (in which, we only need to verify condition (KS) instead of (GS)) and using $G(\mathbf{w}) = \mathcal{K}[-\lambda \text{sign}](\mathbf{w})$, where \mathcal{K} is the Krasovskii operator and $-\lambda \text{sign}(\mathbf{w}) = (-\lambda \text{sign}(w_1), \dots, -\lambda \text{sign}(w_d))$. However, in general, given a set-valued mapping $G(\cdot)$, we may not know explicitly $f(\cdot)$ (if it exists) satisfying $G(\cdot) = \mathcal{K}[f](\cdot)$. That is the reason in the proof, we only treat $G(\cdot)$ as a general set-valued mapping, not the Krasovskii operator of some vector-valued function.

PROOF OF PROPOSITION 4.2. Similar to the proof of Proposition 4.1, the Clarke gradient of $U(\mathbf{w})$ is given by $\partial U(\mathbf{w}) = (1, \dots, 1)^\top$ and then $\tilde{G}_{\{U\}}^*(\mathbf{w}) = G^*(\mathbf{w})$. Moreover, V is continuously differentiable, $\partial V(\mathbf{w}) = (2w_1, \dots, 2w_d)^\top = 2\mathbf{w}$. Hence, the $\{U\}$ -generalized derivative of V in direction F is given by

$$(4.21) \quad \dot{\bar{V}}_{\{U\}}^{G^*}(\mathbf{w}) = \max_{\mathbf{q} \in \tilde{G}_{\{U\}}^*(\mathbf{w})} (\partial V(\mathbf{w}))^\top \mathbf{q}.$$

Let $\mathbf{q} \in G^*(\mathbf{w})$ be arbitrary, then

$$(4.22) \quad \mathbf{q} = -\lambda \mathbf{w} - \lambda \mathbf{w}^* + \bar{\mathbf{q}}, \quad \bar{\mathbf{q}} \in G_1(\mathbf{w} + \mathbf{w}^*).$$

Since $\mathbf{0} \in -\lambda \mathbf{w}^* + g_1(\mathbf{w}^*)$, $-\lambda \mathbf{w}^* \in -g_1(\mathbf{w}^*)$. Therefore, we obtain from (4.22) that

$$(4.23) \quad \mathbf{q} = -\lambda \mathbf{w} + \mathbf{m} + \bar{\mathbf{q}}, \quad \bar{\mathbf{q}} \in G_1(\mathbf{w} + \mathbf{w}^*), \text{ (for some) } \mathbf{m} \in -G_1(\mathbf{w}^*).$$

If we can prove

$$(4.24) \quad \mathbf{w}^\top [\mathbf{m} + \bar{\mathbf{q}}] \leq 0 \quad \text{for all } \mathbf{m} \in -G_1(\mathbf{w}^*), \bar{\mathbf{q}} \in G_1(\mathbf{w} + \mathbf{w}^*),$$

then combining (4.21), (4.23), and (4.24), one has $\dot{\bar{V}}_{\{U\}}^{G^*}(\mathbf{w}) < -\lambda \|\mathbf{w}\|^2$. Now we prove (4.24). Three cases are considered next.

Case 1: $\mathbb{E}[y_n(\mathbf{w}^*)^\top \mathbf{x}_n] = 1$. So, $\mathbf{m} \in -G_1(\mathbf{w}^*) = -\overline{\text{co}}\{\mathbb{E}[y_n \mathbf{x}_n], \mathbf{0}\}$ and then $\mathbf{m} = -m \mathbb{E}[y_n \mathbf{x}_n]$ for some $m \in [0, 1]$. If $\mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] = 0$ then $G_1(\mathbf{w} + \mathbf{w}^*) = \overline{\text{co}}\{\mathbb{E}[y_n \mathbf{x}_n], \mathbf{0}\}$. As a consequence, $\bar{\mathbf{q}} = \bar{q} \mathbb{E}[y_n \mathbf{x}_n]$, for some $\bar{q} \in [0, 1]$. Therefore, $\mathbf{w}^\top [\mathbf{m}^k + \bar{\mathbf{q}}^k] = (-m + \bar{q}) \mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] = 0$ and (4.24) is clear. If $\mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] > 0$, then $G_1(\mathbf{w} + \mathbf{w}^*) = \{\mathbf{0}\}$ and thus, $\mathbf{w}^\top [\mathbf{m}^k + \bar{\mathbf{q}}^k] = -m \mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] \leq 0$ and (4.24) holds. On the other hand, if $\mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] < 0$,

$G_1(\mathbf{w} + \mathbf{w}^*) = \{\mathbb{E}[y_n \mathbf{x}_n]\}$ and thus, $\mathbf{w}^\top [\mathbf{m}^k + \bar{\mathbf{q}}^k] = (-m + 1)\mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] \leq 0$ and (4.23) is satisfied.

Case 2: $\mathbb{E}[y_n(\mathbf{w}^*)^\top \mathbf{x}_n] > 1$. Then $\mathbf{m} = \mathbf{0}$ and so, for all $\bar{\mathbf{q}} \in G_1(\mathbf{w} + \mathbf{w}^*)$, $\mathbf{m} + \bar{\mathbf{q}} = \bar{q}\mathbb{E}[y\mathbf{x}]$, for some $\bar{q} \in [0, 1]$. As a result, (4.23) holds if $\mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] \leq 0$. Otherwise, if $\mathbb{E}[y_n \mathbf{w}^\top \mathbf{x}_n] > 0$, then $G_1(\mathbf{w} + \mathbf{w}^*) = \{\mathbf{0}\}$, so $\mathbf{m} + \bar{\mathbf{q}} = \mathbf{0}$ and (4.23) still holds.

Case 3: $\mathbb{E}[y_n(\mathbf{w}^*)^\top \mathbf{x}_n] < 1$. This case is similar to case 2. \square

PROOF OF PROPOSITION 4.3. Since U is convex, it is regular. The Clarke gradient ∂U of U is given by

$$\partial U(\mathbf{x}) = \begin{cases} \{(s(w_1), s(w_2))\} & \text{if } |w_1| \neq 1 \text{ and } |w_2| \neq 1, \\ \overline{\text{co}}\{0, \text{sign}(w_1)\} \times \{s(w_2)\} & \text{if } |w_1| = 1 \text{ and } |w_2| \neq 1, \\ \{s(w_1)\} \times \overline{\text{co}}\{0, \text{sign}(w_2)\} & \text{if } |w_1| \neq 1 \text{ and } |w_2| = 1, \\ \overline{\text{co}}\{0, \text{sign}(w_1)\} \times \overline{\text{co}}\{0, \text{sign}(w_2)\} & \text{if } |w_1| = 1 \text{ and } |w_2| = 1, \end{cases}$$

where

$$s(w) := \begin{cases} 0 & \text{if } -1 < w < 1, \\ \text{sign}(w) & \text{otherwise.} \end{cases}$$

It is noted that

$$G(\mathbf{w}) = \begin{cases} \{(-w_1 + w_2, -w_1 - w_2)\} & \text{if } w_1 \neq 1 \text{ and } w_2 \neq 1, \\ \{(-w_1 + w_2, -w_1 - w_2)\} + [-1, 1] \times \{0\} & \text{if } w_1 = 1 \text{ and } w_2 \neq 1, \\ \{(-w_1 + w_2, -w_1 - w_2)\} + \{0\} \times [-1, 1] & \text{if } w_1 \neq 1 \text{ and } w_2 = 1, \\ \{(-w_1 + w_2, -w_1 - w_2)\} + [-1, 1] \times [-1, 1] & \text{if } w_1 = 1 \text{ and } w_2 = 1. \end{cases}$$

Therefore, direct calculation yields that

$$M_{\{U\}}^G(\mathbf{w}) = \begin{cases} G(\mathbf{w}) & \text{if } |w_1| \neq 1 \text{ and } |w_2| \neq 1, \\ \emptyset & \text{otherwise.} \end{cases}$$

Equivalently, one has

$$M_{\{U\}}^G(\mathbf{w}) = \begin{cases} \{(-w_1 + w_2, -w_1 - w_2)\} & \text{if } |w_1| \neq 1 \text{ and } |w_2| \neq 1, \\ \emptyset & \text{otherwise.} \end{cases}$$

We have $\tilde{G}_{\{U\}}(\mathbf{w}) = M_{\{U\}}^G(\mathbf{w})$; and $\partial V(\mathbf{w}) = 2(w_1, w_2)^\top$. Hence, the $\{U\}$ -generalized derivative of V in direction G is given by

$$\begin{aligned} \dot{\tilde{V}}_{\{U\}}^G(\mathbf{w}) &= \max_{\mathbf{q} \in \tilde{G}_{\{U\}}(\mathbf{x})} \partial V(\mathbf{w})^\top \mathbf{q} \\ &= \begin{cases} -2\|\mathbf{w}\|^2 & \text{if } |w_1| \neq 1 \text{ and } |w_2| \neq 1, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

As a result, the proposition is proved. \square

4.7. Numerical examples. In this section, we provide some numerical examples to illustrate our findings.

EXAMPLE 4.1. This example demonstrates the results in Section 4.2 as well as Theorem 2.6. We are concerned with the following optimization problem:

$$\text{Find } w^* \text{ to minimize } \mathbb{E}(h(w, \xi_n) + \beta_n) + \lambda \|w\|.$$

TABLE 1
Numerical results of algorithm 4.25

Examples		ex1	ex2	ex3	ex4	ex5	ex6
Num. of iterations	\hat{n}	10^3	10^3	10^3	10^3	10^3	10^3
Num. of repeat		10^3	10^3	10^3	10^3	10^3	10^3
Initial value	w_0	5	50	5	5	5	5
Variance c_n of the bias	c_n	$1/n$	$1/n$	$n^{-0.5}$	1	10	10
Step sizes	a_n	$1/\sqrt{n}$	$1/\sqrt{n}$	$1/\sqrt{n}$	$1/\sqrt{n}$	$1/\sqrt{n}$	$1/n$
Error	$ \hat{w} - w^* $	10^{-3}	10^{-3}	10^{-3}	0.01	0.37	0.12

For simplicity, we consider a real-valued function with $h(w, \xi_n) = \frac{1}{2}(w + \xi_n - 1)^2$, $\lambda = 0.7$, $\{\xi_n\}$ is a sequence of random variables with mean 0 and finite variance, and $\{\beta_n\}$ is a sequence of random variables (assumed to be independent for simplicity) satisfying variance of $\beta_n \leq c_n$. We vary c_n to see the effect of the bias on the convergence of the algorithm. The problem becomes: find minimizer w^* of $\mathbb{E}(h(w, \xi_n) + \beta_n) + \lambda|w| = \frac{1}{2}(w - 1)^2 + 0.7|w|$. Direct calculation shows that the true value is $w^* = 0.3$.

Suppose that only the noisy observations or measurements $h(w_n, \xi_n) + \beta_n$ are available, we can construct a recursive algorithm

(4.25)
$$w_{n+1} = w_n + a_n[(1 + \xi_n - w_n) + \beta_n + g(w_n)].$$

In each iteration, we choose

$$g(w) \in \begin{cases} \{-1\} & \text{if } w > 0, \\ [-1, 1] & \text{if } w = 0, \\ \{1\} & \text{if } w < 0. \end{cases}$$

The numerical results are given in Table 1.

In Table 1, columns “ex1” and “ex2” show the minimizer is globally attractive. Columns “ex1,” “ex3,” and “ex4” show the dependence of the convergence rate on how fast the bias going to 0. If c_n is large, the algorithm may not converge fast enough to the true minimizer, but just in its neighborhood, which is shown in columns “ex5,” and “ex6.”

The relation between c_n and the mean of the error (of approximated value) after repeating algorithm 4.25 (with $\hat{n} = 1000$ iterations for each) is shown in Figure 1 (the left one). This shows the numerical results for the theoretical one in Theorem 2.6, that is, the difference between the approximated value and the true value tends to 0 when $\eta := \limsup_n \|\beta_n\| \rightarrow 0$. It is worth noting that the graph depends on β_n through η in two ways. First, η and the upper bound of errors inherit the behavior of normal distributions β_n . Second, they also depend on the magnitude of β_n . As a result, the graph describes the relationship between errors and bias η varies like normal distributions (at each fixed η) with nonzero means (but tending to 0 as $\eta \rightarrow 0$). The graph on the right in Figure 1 shows the convergence rate to 0 of c_n affects the convergence of the algorithm.

EXAMPLE 4.2. This example is concerned with using results in Section 4.3. Consider the following problem (for better visualization, we consider $\mathbf{w} \in \mathbb{R}^2$):

Find minimizer \mathbf{w}^* of $\lambda\|\mathbf{w}\|^2 + \max\{0, 1 - \mathbb{E}\mathbf{w}^\top \mathbf{h}(\xi_n)\}$.

Assume that $\{\mathbf{h}(\xi_n)\}$ is a sequence of independent two-dimensional Gaussian vectors with mean $(1, 2)^\top$ and covariance matrix $I_{2 \times 2}$ (two-dimensional identity matrix), and $\lambda = 1$. A closed-form solution is $\mathbf{w}^* = (0.2, 0.4)^\top$. We will design an algorithm to locate the

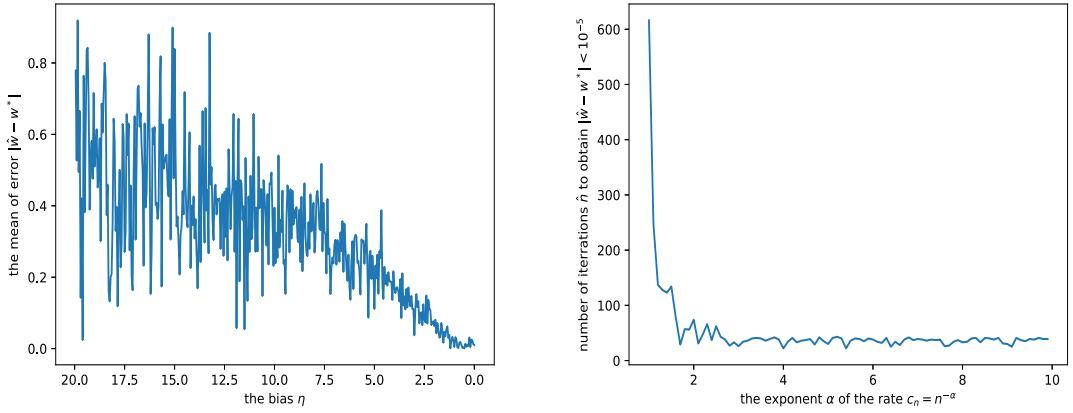


FIG. 1. Numerical results for Example 4.1. Left: relation between bias η and $|\hat{w} - w^*|$. Right: relation between number of iterations to obtain $|\hat{w} - w^*| \leq 10^{-5}$ and the exponent α of $c_n = n^{-\alpha}$ (describing the convergence rate to 0 of unbiased term β_n).

optimum with noise corrupted measurements or observations $\mathbf{h}(\xi_n)$ and bias β_n . Denote $\mathbf{x}_n = \mathbf{h}(\xi_n)$. Consider the algorithm with step sizes $a_n = 1/\sqrt{n}$,

$$(4.26) \quad \mathbf{w}_{n+1} = \mathbf{w}_n + a_n [-2\mathbf{w}_n + g(\mathbf{w}_n, \mathbf{x}_n)],$$

where

$$g(\mathbf{w}, \mathbf{x}) \in \begin{cases} \{\mathbf{0}\} & \text{if } \mathbf{w}^\top \mathbf{x} > 1, \\ \overline{\text{co}}\{\mathbf{0}, \mathbf{x}\} & \text{if } \mathbf{w}^\top \mathbf{x} = 1, \\ \{\mathbf{x}\} & \text{if } \mathbf{w}^\top \mathbf{x} < 1. \end{cases}$$

Let $\mathbf{w}_0 = (3, 5)^\top$, with 1000 replications (i.e., run algorithm 4.26 1000 times), the numerical results are given in Figure 2. Moreover, the mean of $\hat{\mathbf{w}}$ is $(0.2 + 10^{-3}, 0.4 + 10^{-3})^\top$.

EXAMPLE 4.3. This example is concerned with the results in Section 4.4. We wish to find \mathbf{w}^* such that $\mathbf{0} \in G(\mathbf{w}^*)$, where $G(\mathbf{w}) := (-w_1 + w_2 + h(w_2), -w_1 - w_2 + h(w_1))$ with $\mathbf{w} = (w_1, w_2)^\top$ and

$$h(w) = \begin{cases} 0 & \text{if } w \neq 1, \\ [-1, 1] & \text{if } w = 1. \end{cases}$$

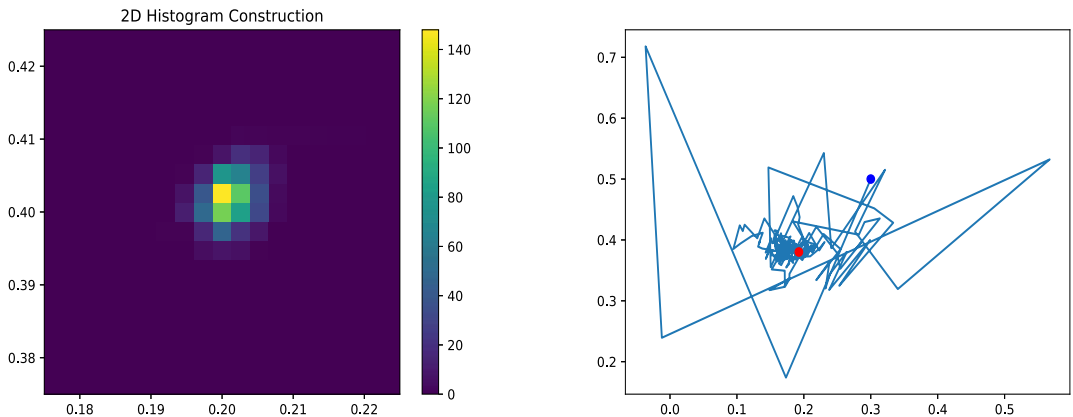


FIG. 2. Numerical results for Example 4.2: Left: 2D histogram of $\hat{\mathbf{w}}$. Right: a trajectory of \mathbf{w}_n (the blue and red points are the starting and ending points).

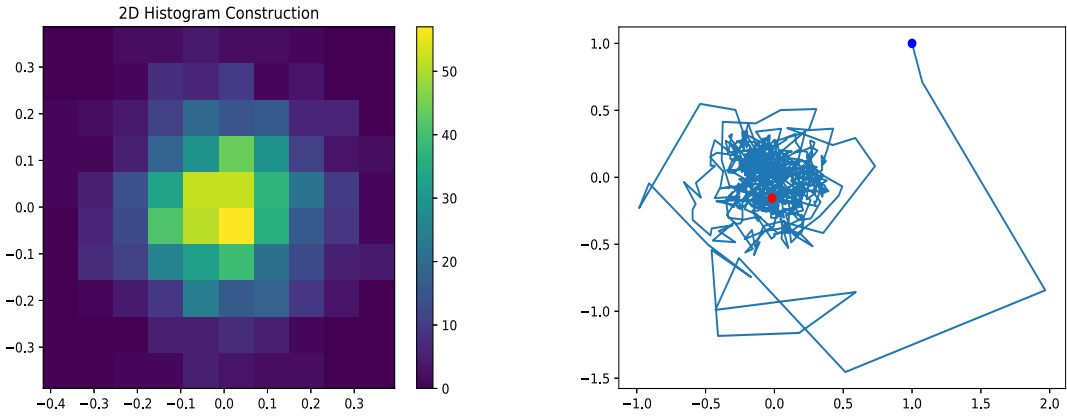


FIG. 3. Numerical results in Example 4.3. Left: 2D histogram of $\hat{\mathbf{w}}$. Right: a trajectory of \mathbf{w}_n (the blue and red points are the starting and ending points).

The true value is $\mathbf{w}^* = (0, 0)^\top$. Consider the stochastic algorithm for the above problem when the observations are corrupted by random disturbances with step sizes $a_n = 1/\sqrt{n}$ and

$$(4.27) \quad \mathbf{w}_{n+1} = \mathbf{w}_n + a_n[g(\mathbf{w}_n) + \boldsymbol{\beta}_n], \quad g(\mathbf{w}_n) \in G(\mathbf{w}_n),$$

and $\boldsymbol{\beta}_n = (\beta_n^1, \beta_n^2)^\top$ so that $\{\boldsymbol{\beta}_n\}$ is a sequence of i.i.d. normal random variables with mean $(0, 0)^\top$ and covariance being the identity matrix. We consider two initial points $\mathbf{w}_0 = (1, 1)^\top$ (near the minimizer) and $\mathbf{w}_0 = (10, -20)^\top$ (far away from the minimizer). Running algorithm 4.27 1000 times, we obtain the mean of $\hat{\mathbf{w}}$ to be $(10^{-3}, 10^{-2})^\top$ for both cases. A histogram and a trajectory of $\{\mathbf{w}_n\}$ (in the case of $\mathbf{w}_0 = (1, 1)^\top$) are shown in Figure 3.

EXAMPLE 4.4. This example considers the comments in Remark 4. We will give an example to show that if conditions on stability of the zero points are violated, the sequence obtained by stochastic approximation may not converge to the right points even if the algorithm starts from one of the optima. Assume that

$$\mathbf{h}(\mathbf{w}) = \begin{cases} [0, 1] \times [-2, 1] & \text{if } \mathbf{w} = (2, 2)^\top, \\ \{0\} \times [-2, -1] & \text{if } 1 \leq w_1 \leq 2, -1 < w_2 \leq 2, \mathbf{w} \neq (2, 2)^\top, \\ [-2, -1] \times \{0\} & \text{if } -1 < w_1 \leq 2, -2 < w_2 \leq -1, \\ \{0\} \times [1, 2] & \text{if } -2 < w_1 \leq -1, -2 \leq w_2 < -1, \\ [1, 2] \times \{0\} & \text{if } -2 \leq w_1 < 1, 1 \leq w_2 \leq 2, \\ \{(-0.005w_1, -0.005w_2)^\top\} & \text{otherwise,} \end{cases}$$

and consider the problem: find \mathbf{w}^* such that $\mathbf{0} \in \mathbb{E}(\mathbf{h}(\mathbf{w}^*) + \boldsymbol{\beta}_n)$, where $\{\boldsymbol{\beta}_n\}$ is a sequence of i.i.d. normal random variables with mean $(0, 0)^\top$ and covariance being the identity matrix. The optimum is given by $\mathbf{w}^* \in \{(0, 0)^\top, (2, 2)^\top\}$. Consider a stochastic approximation algorithm for this problem as follow

$$(4.28) \quad \mathbf{w}_{n+1} = \mathbf{w}_n + \frac{1}{\sqrt{n}}[g(\mathbf{w}_n) + \boldsymbol{\beta}_n], \quad g(\mathbf{w}_n) \in \mathbf{h}(\mathbf{w}_n).$$

We run the algorithm with $\mathbf{w}_0 = (2, 2)^\top$ for 10 millions iterations and note the points at 1 million, 2 million, ..., 10 million iterations. The algorithm does not converge even if the number of iterations is large. In fact, $\{\mathbf{w}_n\}$ tends to be close to some subset of chain-recurrent points, which are strictly larger than the set of the roots. The numerical results are shown in Figure 4.

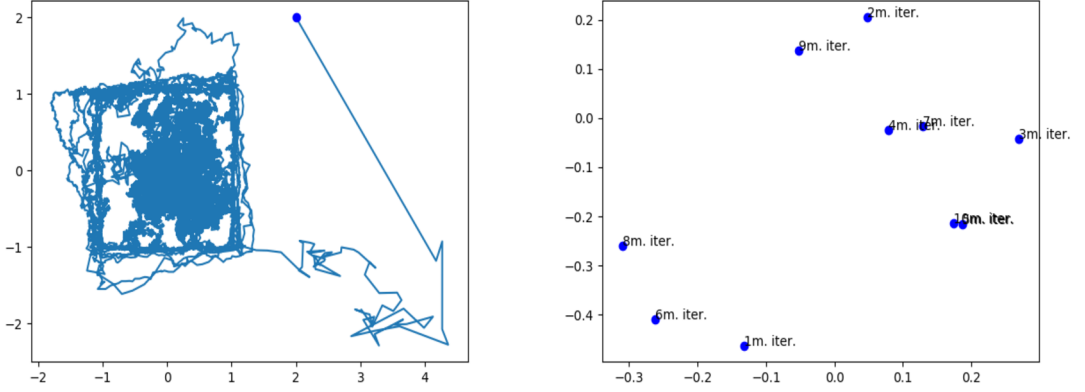


FIG. 4. Numerical results for Example 4.4. Left: a trajectory of $\{w_n\}$, starting from $(2, 2)$ (the blue point). Right: The points at 1 million, 2 million, ..., 10 million iterations.

5. Concluding remarks. Motivated by a wide variety of applications, we considered stochastic approximation with discontinuous dynamics and set-valued mappings. Unconstrained, constrained, and biased algorithms are considered. The traditional approach in the existing literature cannot be used due to the discontinuity. Another main challenge is that we have to deal with set-valued mappings.

Under broad conditions, we use the theory of ODEs with discontinuous right-hand side, differential inclusions, and set-valued analysis, to overcome the difficulties of lack of continuity. Concepts in nonsmooth analysis, set-valued dynamic systems, and novel results in stability of differential inclusions enable us to obtain the convergence to the desired optimal points. The continuation of chain recurrent set of the limit differential inclusions enables us to obtain desired bounds in biased stochastic approximation. The rates of convergence are obtained by using the newly developed concepts in set-valued analysis (T -differentiability) and stochastic differential inclusions (weak compactness of the set of solutions).

Then we make use of our results in applications including Markov decision processes, stochastic subgradient descent algorithms, minimizing L^1 regularized loss functions (online Lasso algorithms, among others), and Pegasos algorithms (in SVMs classification). It is shown that convergence w.p.1 of these stochastic algorithms can be obtained using our results. It is also demonstrated that our results can be used to prove convergence in certain cases, which cannot be done otherwise in the existing literature. New insights for analyzing convergence, rates of convergence, and robustness of these algorithms are also obtained.

APPENDIX: MATHEMATICS PREPARATION

A.1. ODEs with discontinuous right-hand sides and differential inclusions. This section is devoted to ODEs with discontinuous right-hand sides and differential inclusions. Consider the differential equation

$$(A.1) \quad \dot{\mathbf{X}}(t) = f(\mathbf{X}(t)).$$

Given a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, define the set-valued function $\mathcal{K}[f] : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$, known as the Krasovskii operator, as follows

$$\mathcal{K}[f](\mathbf{y}) = \bigcap_{\delta > 0} \overline{\text{co}} f(B(\mathbf{y}, \delta)).$$

LEMMA A.1. *If f is continuous, then $\mathcal{K}[f](\mathbf{x}) = \{f(\mathbf{x})\}$. If f, g are locally bounded and either f or g is continuous then $\mathcal{K}[f + g](\mathbf{x}) = \mathcal{K}[f](\mathbf{x}) + \mathcal{K}[g](\mathbf{x})$.*

PROOF. The first assertion is obvious. By [43], Theorem 1, we have that $\mathcal{K}[g](\mathbf{x}) = \text{co}\{\lim g(\mathbf{x}_i) | \mathbf{x}_i \rightarrow \mathbf{x}\}$. Using this fact, the lemma can be proved; some details are omitted. \square

DEFINITION A.1 (See [22]). A function $\varphi : J \rightarrow \mathbb{R}^d$ (J is an interval in \mathbb{R}) is said to be a Krasovskii solution to (A.1) if it is absolutely continuous on each compact subinterval of J and is a solution of the differential inclusion

$$(A.2) \quad \dot{\mathbf{X}}(t) \in \mathcal{K}[f](\mathbf{X}(t)),$$

that is, φ satisfies (A.2) almost every $t \in J$. Moreover, φ is said to be a Carathéodory solution if it satisfies the (A.1) for almost every $t \in J$, or equivalently, it satisfies the corresponding integral equation.

DEFINITION A.2. A set-valued mapping F is upper semicontinuous at a given $\bar{\mathbf{x}}$, if for every open set U , $F(\bar{\mathbf{x}}) \subset U$, there is an open set V such that $\bar{\mathbf{x}} \in V$ and $F(\mathbf{x}) \subset V$ for every $\mathbf{x} \in V$.

Note that if f is a locally bounded function, then $\mathcal{K}[f](\cdot)$ is upper semicontinuous, nonempty, compact, and convex. The following theorem of the existence of the Krasovskii solution can be found in [22].

LEMMA A.2. *If $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a locally bounded function, there exists at least a Krasovskii solution of (A.2) starting from any initial condition.*

REMARK 16. Some remarks are in order; for more details, we refer to [22].

(i) For the uniqueness of the Krasovskii solution, we need further conditions for $f(\cdot)$, which can be found in [22]. The Carathéodory solutions are always Krasovskii solutions (if both of them exist), but the converse is not true. If f is continuous, they are the same.

(ii) Consider an example with $f(\cdot) = -\text{sign}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$, that is,

$$f(y) = \begin{cases} -1 & \text{if } y > 0, \\ 0 & \text{if } y = 0, \\ 1 & \text{if } y < 0. \end{cases}$$

In this case,

$$\mathcal{K}[f](y) = \begin{cases} \{-1\} & \text{if } y > 0, \\ [-1, 1] & \text{if } y = 0, \\ \{1\} & \text{if } y < 0. \end{cases}$$

Next, ODEs with discontinuous right-hand sides are generalized to differential inclusions.

DEFINITION A.3. Let $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ be a set-valued mapping. A solution to the differential inclusion

$$(A.3) \quad \dot{\mathbf{X}}(t) \in F(\mathbf{X}(t))$$

with initial point $\mathbf{x} \in \mathbb{R}^d$ is an absolutely continuous function $\mathbf{X}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^d$ such that $\mathbf{X}(0) = \mathbf{x}$ and satisfies (A.3) for almost every $t \in \mathbb{R}$.

The following lemma shows that under Assumption (G) in our paper, the solutions of differential inclusion exists.

LEMMA A.3 (See [1], Chapter 1 and Chapter 2). *Let $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ be a set-valued map with values contained in a finite common ball and whose graph is closed. Then F is upper semicontinuous, and (A.3) admits at least one solution with any initial point.*

A.2. Nonsmooth analysis: Set-valued derivative and \mathcal{U} -generalized derivative. In this section, we provide some definitions of generalized derivatives in nonsmooth analysis, which will be key in studying stability of solutions of differential inclusions.

DEFINITION A.4. We introduce the following definitions.

(i) ([2] or [13], p. 39) A function $V(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be regular at $\mathbf{x} \in \mathbb{R}^d$ if for all $\mathbf{v} \in \mathbb{R}^d$, there exists the usual right directional derivative $V'_+(\mathbf{x}, \mathbf{v})$ and $V'_+(\mathbf{x}, \mathbf{v}) = V^\circ(\mathbf{x}, \mathbf{v})$; where

$$V'_+(\mathbf{x}, \mathbf{v}) := \lim_{t \downarrow 0} \frac{V(\mathbf{x} + t\mathbf{v}) - V(\mathbf{x})}{t},$$

and $V^\circ(\mathbf{x}, \mathbf{v})$ is the generalized directional derivative defined as

$$V^\circ(\mathbf{x}, \mathbf{v}) := \limsup_{\mathbf{y} \rightarrow \mathbf{x}, t \downarrow 0} \frac{V(\mathbf{y} + t\mathbf{v}) - V(\mathbf{y})}{t}.$$

V is said to be regular if it is regular at every $\mathbf{x} \in \mathbb{R}^d$. Note that a convex function is not only Lipschitz continuous (in suitable domain), but also regular.

(ii) (see [13]) The Clarke gradient ∂V of V is defined as $\partial V(\mathbf{x}) := \overline{\text{co}}\{\lim \nabla V(\mathbf{x}_i) | \mathbf{x}_i \rightarrow \mathbf{x}, \mathbf{x} \notin \Omega_V\}$, where Ω_V is the set of measure zero with ∇V being not defined.

(iii) (see [2]) The set-valued derivative of a regular function V with respect to F is defined as

$$\dot{\bar{V}}^F(\mathbf{x}) = \{a \in \mathbb{R} | \text{there is } \mathbf{q} \in F(\mathbf{x}) \text{ such that } \mathbf{p}^\top \mathbf{q} = a, \forall \mathbf{p} \in \partial V(\mathbf{x})\}.$$

(iv) A function $V : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be positive definite if it is continuous, $V(\mathbf{0}) = 0$ and there are continuous increasing functions α_1 and $\alpha_2 : \mathbb{R}_+ \rightarrow \mathbb{R}$ with $\alpha_1(0) = \alpha_2(0) = 0$ such that $\alpha_1(|\mathbf{x}|) \leq V(\mathbf{x}) \leq \alpha_2(|\mathbf{x}|)$, $\forall \mathbf{x} \in \mathbb{R}^d$.

The following lemma provides a view of the relationship between the above definitions and the dynamics of solutions of differential inclusions.

LEMMA A.4 (See [2], Lemma 1). *Let $\mathbf{X}(\cdot)$ be a solution of $\dot{\mathbf{X}}(t) \in F(\mathbf{X}(t))$, and $V : \mathbb{R}^d \rightarrow \mathbb{R}$ be a locally Lipschitz continuous and regular function. Then, $\frac{d}{dt}V(\mathbf{X}(t))$ exists almost everywhere and $\frac{d}{dt}V(\mathbf{X}(t)) \in \dot{\bar{V}}^F(\mathbf{X}(t))$ almost everywhere.*

Finally, we recall the following definitions introduced in [29], which are used in this paper.

DEFINITION A.5.

(i) Let $\mathcal{U} := \{U_i\}_{i=1}^\infty$ be a collection of real-valued Lipschitz regular functions. We define $\tilde{F}_{\mathcal{U}} := \bigcap_{i=1}^\infty M_{U_i}^F(\mathbf{x})$, where $M_{U_i}^F := \{\mathbf{q} \in F(\mathbf{x}) | \text{there exists } a \in \mathbb{R} \text{ such that } \mathbf{p}^\top \mathbf{q} = a, \forall \mathbf{p} \in \partial U_i(\mathbf{x})\}$. If \mathcal{U} is empty, we define $\tilde{F}_{\mathcal{U}} = F(\mathbf{x})$. $\tilde{F}_{\mathcal{U}}$ is called the \mathcal{U} -reduced differential inclusion.

(ii) The \mathcal{U} -generalized derivative of locally Lipschitz function $V : \mathbb{R}^d \rightarrow \mathbb{R}$ with direction F , denoted by $\dot{\bar{V}}_{\mathcal{U}}^F$ is defined as

$$\dot{\bar{V}}_{\mathcal{U}}^F(\mathbf{x}) := \begin{cases} \min_{\mathbf{p} \in \partial V(\mathbf{x})} \max_{\mathbf{q} \in \tilde{F}_{\mathcal{U}}(\mathbf{x})} \mathbf{p}^\top \mathbf{q} & \text{if } V \text{ is regular,} \\ \max_{\mathbf{p} \in \partial V(\mathbf{x})} \max_{\mathbf{q} \in \tilde{F}_{\mathcal{U}}(\mathbf{x})} \mathbf{p}^\top \mathbf{q} & \text{if } V \text{ is not regular.} \end{cases}$$

The \mathcal{U} -generalized derivative is understood to be $-\infty$ if $\tilde{F}_{\mathcal{U}}$ is empty. Such a Lyapunov function V with $\dot{\bar{V}}_{\mathcal{U}}^F(\mathbf{x}) \leq 0, \forall \mathbf{x}$ is called a \mathcal{U} -generalized Lyapunov function.

EXAMPLE A.1. To illustrate, let $F(x) = \mathcal{K}[f](x) : \mathbb{R} \rightarrow 2^{\mathbb{R}}, f(x) = -\text{sign}(x)$, that is,

$$F(x) = \begin{cases} -1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ 1 & \text{if } x < 0, \end{cases}$$

$U : \mathbb{R} \rightarrow \mathbb{R}, U(x) = \max\{x, 0\}, \mathcal{U} = \{U\}$ and $V(x) = x^2$. Since U is convex, it is regular. The Clarke gradient of U is given by

$$\partial U(x) = \begin{cases} 1 & \text{if } x > 0, \\ [0, 1] & \text{if } x = 0, \\ 0 & \text{if } x < 0. \end{cases}$$

The reduced inclusion $M_{\mathcal{U}}^F$ is given by

$$M_{\mathcal{U}}^F(x) = \begin{cases} -1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x < 0. \end{cases}$$

Moreover, V is continuously differentiable, $\partial V(x) = 2x$. Hence, the \mathcal{U} -generalized derivative of V in direction F is given by

$$\dot{\bar{V}}_{\mathcal{U}}^F(x) = \max_{q \in \tilde{F}_{\{U\}}(x)} \partial V(x)q = \max_{q \in \tilde{F}_{\{U\}}(x)} 2xq = \begin{cases} -2x & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ 2x & \text{if } x < 0. \end{cases}$$

A.3. Stability of differential inclusions. In this section, we consider the asymptotic stability of solutions of the ODEs with discontinuous right-hand sides and differential inclusions, which contains two parts. The first is stability of Krasovskii solutions and the second is for general differential inclusions.

Let $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ be a set-valued mapping such that F is upper semicontinuous whose values are nonempty, compact, and convex. Consider the differential inclusion

$$(A.4) \quad \dot{\mathbf{X}}(t) \in F(\mathbf{X}(t)).$$

DEFINITION A.6 (See [14], Definition 2.1). The differential inclusion (A.4) is strongly asymptotically stable (in Clarke's sense) if there is no solution exhibiting finite time blow-up and the following properties hold.

(a) Uniform attraction: for any $r > 0, R > 0$, there is $T = T(R, r)$ such that for any solution $\mathbf{X}(\cdot)$ of (A.4) with $|\mathbf{X}(0)| < R$ then $|\mathbf{X}(t)| \leq r$ for all $t \geq T$.

(b) Uniform boundedness: there is a continuous nonincreasing function $m : (0, \infty) \rightarrow (0, \infty)$ such that for any solution $\mathbf{X}(\cdot)$ of (A.4) with $|\mathbf{X}(0)| \leq R$ then $|\mathbf{X}(t)| \leq m(R)$ for all $t \geq 0$.

(c) Lyapunov stability: $\lim_{R \downarrow 0} m(R) = 0$.

DEFINITION A.7 (Classical Lyapunov stability; see [29], Definition 7.1). The differential inclusion $\dot{\mathbf{X}}(t) \in F(\mathbf{X}(t))$ is said to be (strongly) asymptotically stable at $\mathbf{x} = \mathbf{0}$ if every solution is stable at $\mathbf{x} = \mathbf{0}$, [that is, for any $\varepsilon > 0$, there is $\delta > 0$ such that if $|\mathbf{X}(0)| \leq \delta$ then $|\mathbf{X}(t)| < \varepsilon$, $\forall t \geq 0$] and there is $c > 0$ such that if $|\mathbf{X}(0)| \leq c$ then $\lim_{t \rightarrow \infty} |\mathbf{X}(t)| = 0$. Moreover, it is said to be globally asymptotically stable if the constant c can be ∞ .

PROPOSITION A.1 (See [14]). *The strongly asymptotic stability (in Clarke's sense) implies the classical asymptotic stability in the Lyapunov sense.*

The following theorem ([14], Theorem 1.3) provides necessary and sufficient conditions for strongly asymptotic stability of the Karasovskii solutions of the ODEs with discontinuous right-hand sides (see Section A.1 for definition)

$$(A.5) \quad \dot{\mathbf{X}}(t) = f(\mathbf{X}(t)).$$

THEOREM A.1. *Let f be a locally bounded function. Then, Krasovskii solutions of (A.5) are strongly asymptotically stable if and only if there exists a C^∞ -smooth pair of functions (V, \widehat{V}_0) satisfying:*

- (1) $V(\mathbf{x}) > 0$ and $\widehat{V}_0(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$, $V(\mathbf{0}) = 0$;
- (2) the sublevel sets $\{\mathbf{x} \in \mathbb{R}^d : V(\mathbf{x}) \leq l\}$ are bounded for every $l \geq 0$;
- (3) $\limsup_{\mathbf{y} \rightarrow \mathbf{x}} \langle \nabla V(\mathbf{x}), f(\mathbf{y}) \rangle \leq -\widehat{V}_0(\mathbf{x})$, $\forall \mathbf{x} \neq \mathbf{0}$.

REMARK 17. Note that in differential inclusions, the uniqueness of solution is not always guaranteed. Hence, the term “strongly” in definitions of stability means that these definitions hold for all solutions. In contrast, “weak” stability means that there is a solution that is stable. The condition of “weak asymptotic stability” of Krasovskii solutions can be found in [14].

In contrast to Theorem A.1, sufficient conditions for asymptotic stability of general differential inclusions can be found in [2] and references therein. Recently, these sufficient conditions for differential inclusion $\dot{\mathbf{X}}(t) \in F(\mathbf{X}(t))$ are much improved in [29]. We state this result in the following theorem.

THEOREM A.2 ([29], Theorem 7.2). *If there exists a \mathcal{U} -generalized Lyapunov function $V : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\dot{V}_{\mathcal{U}}^F(\mathbf{x}) \leq -\widehat{V}_0(\mathbf{x})$, for some positive definite function \widehat{V}_0 (see Definitions A.4 and A.5), then (A.4) is (strongly) asymptotically stable (in the sense of Lyapunov) at $\mathbf{x} = \mathbf{0}$. Furthermore, if $\{\mathbf{x} \in \mathbb{R}^d : V(\mathbf{x}) \leq l\}$ are compact for all $l > 0$ then (A.4) is (strongly) globally asymptotically stable (in the sense of Lyapunov) at $\mathbf{x} = \mathbf{0}$.*

REMARK 18. Another result on stability of differential inclusions using Lyapunov functional method can be found in [2]. The technique is based on the “set-valued derivative of a regular function V ” with respect to F . However, using the \mathcal{U} -generalized derivative is shown to be much stronger and more effective; see [29]. Moreover, if $U(\cdot)$ satisfies $\partial U(\cdot) = (1, \dots, 1)^\top$, then the $\{U\}$ -generalized derivative is the set-valued derivative.

A.4. Set-valued dynamical systems: Invariant set, limit set, and chain recurrence.

Consider the differential inclusion

$$(A.6) \quad \dot{\mathbf{X}}(t) \in F(\mathbf{X}(t)).$$

We recall some concepts, which are used in this paper; more details can be found in [3, 5, 53] and references therein.

DEFINITION A.8 (See [5], Section 3). Let $\mathbf{X}(\cdot)$ be a solution of (A.6). The limit set of $\mathbf{X}(\cdot)$, denoted by $L(\mathbf{X})$, is defined as $L(\mathbf{X}) = \bigcap_{t \geq 0} \overline{\{\mathbf{X}(s) : s \geq t\}}$.

DEFINITION A.9 (See [5], Definition V). A set $A \subset \mathbb{R}^d$ is said to be invariant if for all $\mathbf{x} \in A$, there exists a solution $\mathbf{X}(\cdot)$ of (A.6) with $\mathbf{X}(0) = \mathbf{x}$ such that $\mathbf{X}(\mathbb{R}) \subset A$.

DEFINITION A.10 (See [5], Definition VI). Let A be a subset of \mathbb{R}^d .

- $\mathbf{x}, \mathbf{y} \in A$ is said to be chain connected in A if for every $\varepsilon > 0$ and $T > 0$, there exist an integer $n \in \mathbb{N}$, and solutions $\mathbf{X}_1(\cdot), \dots, \mathbf{X}_n(\cdot)$ to (A.6), and real numbers $t_1, \dots, t_n > T$ such that:

- (a) $\mathbf{X}_i(s) \in A$ for all $0 \leq s \leq t_i, i = 1, \dots, n$;
- (b) $|\mathbf{X}_i(t_i) - \mathbf{X}_{i+1}(0)| \leq \varepsilon$ for all $i = 1, \dots, n-1$;
- (c) $|\mathbf{X}_1(0) - \mathbf{x}| \leq \varepsilon$ and $|\mathbf{X}_n(t_n) - \mathbf{y}| \leq \varepsilon$.

- A is said to be “internally chain transitive” of (A.6) if A is compact and \mathbf{x}, \mathbf{y} are chain-connected in A for all $\mathbf{x}, \mathbf{y} \in A$.

DEFINITION A.11 (See [5, 37, 53]). θ is said to be a “chain-recurrent point” of (A.6) if for any $\varepsilon > 0$ and $T > 0$, there exist an integer $n \in \mathbb{N}$, and solutions $\mathbf{X}_1(\cdot), \dots, \mathbf{X}_n(\cdot)$ to (A.6) and real numbers $t_1, \dots, t_n > T$ such that

$$|\mathbf{X}_1(0) - \theta| \leq \varepsilon, \quad |\mathbf{X}_i(t_i) - \mathbf{X}_{i+1}(0)| \leq \varepsilon \quad \forall i = 1, \dots, n-1, \quad |\mathbf{X}_n(t_n) - \theta| \leq \varepsilon.$$

Moreover, we say that θ is a “chain-recurrent point” in A of (A.6), if we assume further that $\mathbf{X}_i(s) \in A$ for all $0 \leq s \leq t_i, i = 1, \dots, n$.

The following lemma (see [5], Lemma 3.5) shows the relationship between invariant set and internally chain transitive set.

LEMMA A.5. *An internally chain transitive set is invariant.*

A.5. Set-valued analysis: Continuity and T -differentiability. This section reviews definitions and results of set-valued analysis in [1, 33, 44, 48] and references therein. Recall that $B = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| < 1\}$ and \bar{B} is its closure.

DEFINITION A.12 (See [1], Chapter 1, Section 1, or [48]).

- A set-valued mapping F is said to be lower semicontinuous at $\bar{\mathbf{x}}$ if for every open set U with $F(\bar{\mathbf{x}}) \cap U \neq \emptyset$, there is an open set V such that $\bar{\mathbf{x}} \in V$ and $F(\mathbf{x}) \cap U \neq \emptyset$ for every $\mathbf{x} \in V$.
- F is said to be continuous if it is both lower semicontinuous and upper semicontinuous (see Definition A.2).

LEMMA A.6 (Criteria on continuity, see [33], Chapter 2.2). *If a set-valued mapping $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ has convex and compact values, then F is continuous if and only if for each $\mathbf{p} \in \mathbb{R}^d$, $\sigma(\mathbf{p}, F(\mathbf{x}))$ is continuous (in \mathbf{x}), where $\sigma(\mathbf{p}, A) := \sup\{\mathbf{p}^\top \mathbf{a} : \mathbf{a} \in A\}$.*

DEFINITION A.13 (See [44], Definition 2.1). A set-valued mapping $T : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ is positively homogeneous if $T(\mathbf{0})$ is a cone, and $T(k\mathbf{x}) = kT(\mathbf{x})$ for all $k > 0$, $\mathbf{x} \in \mathbb{R}^d$.

DEFINITION A.14 (See [44], Definition 4.1). Let $T : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ be a positively homogeneous set-valued mapping. We say $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ is outer T -differentiable at \mathbf{x}^* if for any $\delta > 0$, there exists a neighborhood V of \mathbf{x}^* such that

$$(A.7) \quad F(\mathbf{x}) \subset F(\mathbf{x}^*) + T(\mathbf{x} - \mathbf{x}^*) + \delta|\mathbf{x} - \mathbf{x}^*|B \quad \text{for all } \mathbf{x} \in V.$$

The relationship between T -differentiability and others differentiability, and the analysis as well as computation examples of T -differentiability can be found in [44].

A.6. Stochastic differential inclusions. Given a set-valued mapping $F : \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$ taking nonempty values, there exists an $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $f(\mathbf{x}) \in F(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$, such a function f is called a selector of F . For an L^2 -continuous (continuous in mean) \mathcal{F}_t -nonanticipative stochastic process $(\mathbf{X}(t))_{0 \leq t \leq T}$ and set-valued mapping $F_1 : [0, T] \times \mathbb{R}^d \rightarrow 2^{\mathbb{R}^d}$, $F_2 : [0, T] \times \mathbb{R}^d \rightarrow 2^{\mathbb{R}^{d \times m}}$ taking closed (subset) values, we denote $(F_1 \circ \mathbf{X})(t)(\omega) := F_1(t, \mathbf{X}(t)(\omega))$, $(F_2 \circ \mathbf{X})(t)(\omega) = F_2(t, \mathbf{X}(t)(\omega))$ and denote by $S(F_1 \circ \mathbf{X})$, $S(F_2 \circ \mathbf{X})$ the family of all \mathcal{F}_t -nonanticipative selectors of $F_1 \circ \mathbf{X}$ and $F_2 \circ \mathbf{X}$, respectively. Let $(\mathbf{W}(t))_{0 \leq t \leq T}$ be an m -dimensional \mathcal{F}_t -Brownian motion and define the following sets

$$\begin{aligned} \int_s^t (F_1 \circ \mathbf{X})(r) dr &:= \left\{ \int_0^T \mathbf{1}_{[s,t]}(r) f(r) dr : f \in S(F_1 \circ \mathbf{X}) \right\}, \\ \int_s^t (F_2 \circ \mathbf{X})(r) dr &:= \left\{ \int_0^T \mathbf{1}_{[s,t]}(r) g(r) d\mathbf{W}(r) : g \in S(F_2 \circ \mathbf{X}) \right\}. \end{aligned}$$

Consider the stochastic differential inclusion

$$(A.8) \quad d\mathbf{X}(t) \in F_1(t, \mathbf{X}(t)) dt + F_2(t, \mathbf{X}(t)) d\mathbf{W}(t).$$

DEFINITION A.15 (See [31, 33]). We define the (stochastic) weak solution to (A.8) as a system consisting of a complete filtered probability space $\{\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P}\}$, a continuous \mathcal{F}_t -adapted process $(\mathbf{X}(t))_{0 \leq t \leq T}$, and an \mathcal{F}_t -Brownian motion $\mathbf{W}(t)$ satisfying

$$\mathbf{X}(t) - \mathbf{X}(s) \in \int_s^t F_1(r, \mathbf{X}(r)) dr + \int_s^t F_2(r, \mathbf{X}(r)) d\mathbf{W}(r), \quad \forall 0 \leq s < t \leq T, \text{ w.p.1.}$$

Denote by $\mathcal{X}_\mu(F_1, F_2)$ a set of all weak solutions to (A.8) with an initial distribution μ . It is called a (stochastically) strong solution (solution for short) if the complete filtered probability space and the Brownian motion have been given.

LEMMA A.7 (See [31]). Assume that F_1, F_2 are measurable and bounded and have convex values, where F_2 has convex value in the sense of that $\{g \cdot g^\top : g \in F_2(t, \mathbf{x})\}$ is convex for each $(t, \mathbf{x}) \in [0, T] \times \mathbb{R}^d$ and $F_1(t, \cdot), F_2(t, \cdot)$ are continuous for fixed $t \in [0, T]$. Then, for any initial distribution μ , the set $\mathcal{X}_\mu(F_1, F_2)$ is nonempty.

When convexity is absent, the above results were studied in [32]. For more details on stochastic differential inclusions, the reader is referred to [31–33] and references therein.

A.7. Proof of Proposition 3.1.

PROOF. Without loss of generality and for notational simplicity, we assume that $\mathbf{x}^* = \mathbf{0}$ and verify the tightness for sequence of $\frac{\mathbf{X}_n}{\sqrt{a_n}}$. To prove this tightness, it suffices to show that for each small $\kappa > 0$, there are finite constants M_κ and C_κ such that

$$(A.9) \quad \mathbb{P}\left(\frac{\mathbf{X}_n}{\sqrt{a_n}} \geq C_\kappa\right) \leq \kappa, \quad \text{for } n \geq M_\kappa.$$

Let $\varepsilon > 0$ be small. Because $\mathbf{X}_n \rightarrow \mathbf{x}^* = \mathbf{0}$ w.p.1, for any given small $\nu > 0$, there exists an $N_{\nu, \varepsilon}$ such that $|\mathbf{X}_n| \leq \varepsilon$ for $n \geq N_{\nu, \varepsilon}$ with probability $\geq 1 - \nu$. By modifying the processes on a set of probability at most ν , one can assume that $|\mathbf{X}_n| \leq \varepsilon$ for $n \geq N_{\nu, \varepsilon}$ and that all the assumptions continue to hold. Denote the modified sequence by $\{\mathbf{X}_n^\nu\}$ and if we can show $\{\frac{\mathbf{X}_n^\nu}{\sqrt{a_n}}\}$ is tight for each $\varepsilon > 0$, $\nu > 0$, then the original sequence is tight. Hence, for the purposes of the tightness proof and by shifting the time origin if needed, it can be supposed without loss of generality that $|\mathbf{X}_n| \leq \varepsilon$ for all n for the original process, where $\varepsilon > 0$ is arbitrarily small.

Next, denote by \mathbb{E}_n the conditional expectation on the past information up to time n (i.e., the σ -algebra generated by $\{\xi_j : j < n\}$). We have that

$$(A.10) \quad \begin{aligned} & \mathbb{E}_n(V(\mathbf{X}_{n+1}) - V(\mathbf{X}_n)) \\ &= a_n \mathbb{E}_n(V_{\mathbf{X}}(\mathbf{X}_n)(\bar{\mathbf{h}}(\mathbf{X}_n) + \mathbf{b}_n(\mathbf{X}_n))) + a_n \mathbb{E}_n(V_{\mathbf{X}}(\mathbf{X}_n)(\mathbf{h}(\mathbf{X}_n, \xi_n) - \bar{\mathbf{h}}(\mathbf{X}_n))) \\ & \quad + O(a_n^2) \mathbb{E}_n |\mathbf{b}_n(\mathbf{X}_n) + \mathbf{h}(\mathbf{X}_n, \xi_n)|^2 \\ & \leq a_n \mathbb{E}_n \max \dot{\bar{V}}^{G+\bar{\mathbf{h}}}(\mathbf{X}_n) + a_n V_{\mathbf{X}}(\mathbf{X}_n) \mathbb{E}_n(\mathbf{h}(\mathbf{X}_n, \xi_n) - \bar{\mathbf{h}}(\mathbf{X}_n)) \\ & \quad + O(a_n^2)(1 + V(\mathbf{X}_n)) \\ & \leq -\lambda a_n \mathbb{E}_n V(\mathbf{X}_n) + a_n V_{\mathbf{X}}(\mathbf{X}_n) \mathbb{E}_n(\mathbf{h}(\mathbf{X}_n, \xi_n) - \bar{\mathbf{h}}(\mathbf{X}_n)) + O(a_n^2)(1 + V(\mathbf{X}_n)). \end{aligned}$$

Let $V_1(\mathbf{x}; n) := a_n V_{\mathbf{X}}(\mathbf{x}) \mathbb{E}_n(\mathbf{h}(\mathbf{x}, \xi_n) - \bar{\mathbf{h}}(\mathbf{x}))$, and define the perturbed Lyapunov function $\tilde{V}(\mathbf{x}; n) := V(\mathbf{x}) + V_1(\mathbf{x}; n)$. In fact, the idea of perturbed Lyapunov functional method is that the perturbations added are small in terms of order of magnitude, and they lead to desired cancellation of the un-wanted terms in (A.10). Thus, by using the usual computation in the perturbed Lyapunov functional method (see, e.g., [37], Theorem 10.4.2, page 345–346), we can obtain from (A.10) that

$$\mathbb{E}_n V(\mathbf{X}_{n+1}) - V(\mathbf{X}_n) \leq -\lambda_1 a_n V(\mathbf{X}_n) + O(a_n^2),$$

where $0 < \lambda_1 < \lambda$. By taking ε small enough, it can be supposed that λ_1 is arbitrarily close to λ . Thus, there is a real number K_1 such that for all $n \geq 0$

$$(A.11) \quad \mathbb{E} V(\mathbf{X}_{n+1}) \leq \prod_{i=1}^n (1 - \lambda_1 a_i) \mathbb{E} V(\mathbf{X}_0) + K_1 \sum_{i=0}^n \prod_{j=i+1}^n (1 - \lambda_1 a_j) a_i^2.$$

Therefore, it is readily seen that to obtain (A.9), it suffices to prove that the right side of (A.11) is of the order of a_n . However, this fact can be easily proved by approximating this quantity by an exponential approximation. The detail of this argument can be found in [37], Section 10, page 342–343, and is thus omitted here. \square

Funding. This research was supported in part by the National Science Foundation Grant DMS-2204240.

REFERENCES

- [1] AUBIN, J.-P. and CELLINA, A. (1984). *Differential Inclusions. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **264**. Springer, Berlin. MR0755330 <https://doi.org/10.1007/978-3-642-69512-4>
- [2] BACCIOTTI, A. and CERAGIOLI, F. (1999). Stability and stabilization of discontinuous systems and nonsmooth Lyapunov functions. *ESAIM Control Optim. Calc. Var.* **4** 361–376. MR1693045 <https://doi.org/10.1051/cocv:1999113>
- [3] BENAÏM, M. (1996). A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.* **34** 437–472. MR1377706 <https://doi.org/10.1137/S0363012993253534>
- [4] BENAÏM, M. and FAURE, M. (2012). Stochastic approximation, cooperative dynamics and supermodular games. *Ann. Appl. Probab.* **22** 2133–2164. MR3025692 <https://doi.org/10.1214/11-AAP816>
- [5] BENAÏM, M., HOFBAUER, J. and SORIN, S. (2005). Stochastic approximations and differential inclusions. *SIAM J. Control Optim.* **44** 328–348. MR2177159 <https://doi.org/10.1137/S0363012904439301>
- [6] BENAÏM, M., HOFBAUER, J. and SORIN, S. (2012). Perturbations of set-valued dynamical systems, with applications to game theory. *Dyn. Games Appl.* **2** 195–205. MR2922840 <https://doi.org/10.1007/s13235-012-0040-0>
- [7] BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P. (1990). *Stochastic Approximations and Adaptive Algorithms*. Springer, New York.
- [8] BIANCHI, P., HACHEM, W. and SCHECHTMAN, S. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. Available at [arXiv:2005.08513](https://arxiv.org/abs/2005.08513).
- [9] BOLTE, J. and PAUWELS, E. (2021). Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Math. Program.* **188** 19–51. MR4276581 <https://doi.org/10.1007/s10107-020-01501-5>
- [10] BURKE, J., CURTIS, F., LEWIS, A., OVERTON, M. and SIMOES, L. (2020). Gradient sampling methods for nonsmooth optimization. In *Numerical Nonsmooth Optimization* (A. Bagirov, M. Gaudioso, N. Karmitsa, M. Mäkelä and S. Taheri, eds.) Springer, Cham.
- [11] CHEN, X. (2012). Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.* **134** 71–99. MR2947553 <https://doi.org/10.1007/s10107-012-0569-0>
- [12] CHONG, E. K. P. and ZAK, S. H. (2014). *An Introduction to Optimization*. Wiley, New York.
- [13] CLARKE, F. H. (1983). *Optimization and Nonsmooth Analysis. Canadian Mathematical Society Series of Monographs and Advanced Texts*. Wiley, New York. MR0709590
- [14] CLARKE, F. H., LEDYAEV, Y. S. and STERN, R. J. (1998). Asymptotic stability and smooth Lyapunov functions. *J. Differential Equations* **149** 69–114. MR1643670 <https://doi.org/10.1006/jdeq.1998.3476>
- [15] CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines*. Cambridge Univ. Press, Cambridge.
- [16] DAVIS, D., DRUSVYATSKIY, D., KAKADE, S. and LEE, J. D. (2020). Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* **20** 119–154. MR4056927 <https://doi.org/10.1007/s10208-018-09409-5>
- [17] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, New York. MR0838085 <https://doi.org/10.1002/9780470316658>
- [18] EWEDA, E. and MACCHI, O. (1983). Quadratic mean and almost-sure convergence of unbounded stochastic approximation algorithms with correlated observations. *Ann. Inst. Henri Poincaré Sect. B (N. S.)* **19** 235–255. MR0725558
- [19] FRIKHA, N. (2016). Multi-level stochastic approximation algorithms. *Ann. Appl. Probab.* **26** 933–985. MR3476630 <https://doi.org/10.1214/15-AAP1109>
- [20] GIRAUD, C. (2015). *Introduction to High-Dimensional Statistics. Monographs on Statistics and Applied Probability* **139**. CRC Press, Boca Raton, FL. MR3307991
- [21] GRIMMER, B. (2019). Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM J. Optim.* **29** 1350–1365. MR3949713 <https://doi.org/10.1137/18M117306X>
- [22] HÁJEK, O. (1979). Discontinuous differential equations. I, II. *J. Differential Equations* **32** 149–170, 171–185. MR0534546 [https://doi.org/10.1016/0022-0396\(79\)90056-1](https://doi.org/10.1016/0022-0396(79)90056-1)
- [23] HAJINEZHAD, D. and HONG, M. (2019). Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. *Math. Program.* **176** 207–245. MR3960809 <https://doi.org/10.1007/s10107-019-01365-4>
- [24] HELOU NETO, E. S. and DE PIERRO, Á. R. (2009). Incremental subgradients for constrained convex optimization: A unified framework and new methods. *SIAM J. Optim.* **20** 1547–1572. MR2587740 <https://doi.org/10.1137/070711712>

- [25] HIGHAM, D. J., MAO, X., ROJ, M., SONG, Q. and YIN, G. (2013). Mean exit times and the multilevel Monte Carlo method. *SIAM/ASA J. Uncertain. Quantificat.* **1** 2–18. MR3283875 <https://doi.org/10.1137/120883803>
- [26] HILL, R. W. and HOLLAND, P. W. (1977). Two robust alternatives to least squares regression. *J. Amer. Statist. Assoc.* **72** 828–833.
- [27] HURLEY, M. (1995). Chain recurrence, semiflows, and gradients. *J. Dynam. Differential Equations* **7** 437–456. MR1348735 <https://doi.org/10.1007/BF02219371>
- [28] JASSO-FUENTES, H. and YIN, G. (2013). *Advanced Criteria for Controlled Markov-Modulated Diffusions in an Infinite Horizon: Overtaking, Bias, and Blackwell Optimality*. Science Press, Beijing, China.
- [29] KAMALAPURKAR, R., DIXON, W. E. and TEEL, A. R. (2020). On reduction of differential inclusions and Lyapunov stability. *ESAIM Control Optim. Calc. Var.* **26** 24. MR4071314 <https://doi.org/10.1051/cocv/2019074>
- [30] KHAMARU, K. and WAINWRIGHT, M. J. (2019). Convergence guarantees for a class of non-convex and non-smooth optimization problems. *J. Mach. Learn. Res.* **20** 154. MR4030168
- [31] KISIELEWICZ, M. (2001). Weak compactness of solution sets to stochastic differential inclusions with convex right-hand sides. *Topol. Methods Nonlinear Anal.* **18** 149–169.
- [32] KISIELEWICZ, M. (2005). Weak compactness of solution sets to stochastic differential inclusions with non-convex right-hand sides. *Stoch. Anal. Appl.* **23** 871–901. MR2158883 <https://doi.org/10.1080/SAP-200044414>
- [33] KISIELEWICZ, M. (2013). *Stochastic Differential Inclusions and Applications*. Springer Optimization and Its Applications **80**. Springer, New York. MR3097668 <https://doi.org/10.1007/978-1-4614-6756-4>
- [34] KIWIEL, K. C. (2003). Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM J. Optim.* **14** 807–840. MR2085944 <https://doi.org/10.1137/S1052623400376366>
- [35] KUSHNER, H. J. (1983). An averaging method for stochastic approximations with discontinuous dynamics, constraints, and state-dependent noise. In *Recent Advances in Stochastics* (M. H. Rizzi, J. S. Rustagi and D. Siegmund, eds.) Academic Press, New York.
- [36] KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Applied Mathematical Sciences **26**. Springer, New York-Berlin. MR0499560
- [37] KUSHNER, H. J. and YIN, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. *Applications of Mathematics* **35**. Springer, New York. MR1993642
- [38] LANGFORD, J., LI, L. and ZHANG, T. (2009). Sparse online learning via truncated gradient. *Adv. Neural Inf. Process. Syst.* **10** 905–912.
- [39] LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **AC-22** 551–575. MR0465458 <https://doi.org/10.1109/tac.1977.1101561>
- [40] MÉTIVIER, M. and PRIOURET, P. (1987). Théorèmes de convergence presque sure pour une classe d’algorithmes stochastiques à pas décroissant. *Probab. Theory Related Fields* **74** 403–428. MR0873887 <https://doi.org/10.1007/BF00699098>
- [41] NEDIĆ, A. and LEE, S. (2014). On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM J. Optim.* **24** 84–107. MR3150179 <https://doi.org/10.1137/120894464>
- [42] NITANDA, A. (2014). Stochastic proximal gradient descent with acceleration techniques. *Adv. Neural Inf. Process. Syst.* 1574–1582.
- [43] PADEN, B. E. and SASTRY, S. S. (1987). A calculus for computing Filippov’s differential inclusion with application to the variable structure control of robot manipulators. *IEEE Trans. Circuits Syst.* **34** 73–82. MR0871547 <https://doi.org/10.1109/TCS.1987.1086038>
- [44] PANG, C. H. J. (2011). Generalized differentiation with positively homogeneous maps: Applications in set-valued analysis and metric regularity. *Math. Oper. Res.* **36** 377–397. MR2832397 <https://doi.org/10.1287/moor.1110.0497>
- [45] PERKINS, S. and LESLIE, D. S. (2012). Asynchronous stochastic approximation with differential inclusions. *Stoch. Syst.* **2** 409–446. MR3354772 <https://doi.org/10.1214/11-SSY056>
- [46] RAM, S. S., NEDIĆ, A. and VEERAVALLI, V. V. (2009). Incremental stochastic subgradient algorithms for convex optimization. *SIAM J. Optim.* **20** 691–717. MR2515792 <https://doi.org/10.1137/080726380>
- [47] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 <https://doi.org/10.1214/aoms/1177729586>
- [48] ROCKAFELLAR, R. T. and WETS, R. JB. (1997). *Variational Analysis*. Springer, Berlin.
- [49] SHALEV-SHWARTZ, S., SINGER, Y., SREBRO, N. and COTTER, A. (2011). Pegasos: Primal estimated sub-gradient solver for SVM. *Math. Program.* **127** 3–30. MR2776708 <https://doi.org/10.1007/s10107-010-0420-4>
- [50] SHALEV-SHWARTZ, S. and SREBRO, N. (2018). SVM optimization: Inverse dependence on training set size. In *Proceedings of the 25th International Conference on Machine Learning* 928–935.

- [51] SHALEV-SHWARTZ, S. and TEWARI, A. (2011). Stochastic methods for ℓ_1 -regularized loss minimization. *J. Mach. Learn. Res.* **12** 1865–1892. [MR2819020](#)
- [52] SHAMIR, O. and ZHANG, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *Proc. Mach. Learn. Res.* **28** 71–79.
- [53] TADIĆ, V. B. and DOUCET, A. (2017). Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.* **27** 3255–3304. [MR3737925](#) <https://doi.org/10.1214/16-AAP1272>
- [54] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- [55] WANG, H., LI, G. and JIANG, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.* **25** 347–355. [MR2380753](#) <https://doi.org/10.1198/073500106000000251>
- [56] WIDROW, B. and STEARNS, S. D. (1985). *Adaptive Signal Processing*. Prentice-Hall, Upper Saddle River, NJ.
- [57] YIN, G. (1991). On extensions of Polyak’s averaging approach to stochastic approximation. *Stoch. Stoch. Rep.* **36** 245–264. [MR1128497](#) <https://doi.org/10.1080/17442509108833721>
- [58] YIN, G., KRISHNAMURTHY, V. and ION, C. (2003). Iterate-averaging sign algorithms for adaptive filtering with applications to blind multiuser detection. *IEEE Trans. Inf. Theory* **49** 657–671. [MR1967190](#) <https://doi.org/10.1109/TIT.2002.808100>