Delft University of Technology

Convergence of sequences
A survey

Franci, Barbara; Grammatico, Sergio

DOI
10.1016/j.arcontrol.2022.01.003
Publication date
2022
Document Version
Final published version
Published in
Annual Reviews in Control

Citation (APA)
Franci, B., & Grammatico, S. (2022). Convergence of sequences: A survey. *Annual Reviews in Control*, *53*, 161-186. https://doi.org/10.1016/j.arcontrol.2022.01.003

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Annual Reviews in Control 53 (2022) 161–186

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Review article

Convergence of sequences: A survey☆

Barbara Franci [a,*], Sergio Grammatico [b]

[a] Department of Data Science and Knowledge Engineering, Maastricht University, Maastricht, The Netherlands
[b] Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Keywords:
Convergence

ABSTRACT

Convergent sequences of real numbers play a fundamental role in many different problems in system theory, e.g., in Lyapunov stability analysis, as well as in optimization theory and computational game theory. In this survey, we provide an overview of the literature on convergence theorems and their connection with Féjer monotonicity in the deterministic and stochastic settings, and we show how to exploit these results.

1. Introduction

Why Are Convergence Theorems Necessary?
The answer to this "naive" question is not simple.
cit. Boris T. Polyak, 1987 (Polyak, 1987, Section 1.6.2).

While the answer may have become clearer through the years, since many problems in applied mathematics rely on convergence theorems, it is still not simple. Besides the theoretical investigation, in fact, one fundamental aspect is how convergence theorems can be of practical use, i.e., if the assumptions are plausible for a variety of applications, for instance, in systems theory. Moreover, convergence theorems may also give qualitative information, e.g., if convergence is guaranteed for any initial point and in what sense (strongly, weakly, almost surely, in probability), which affects the range of application. The aim of this paper is to collect these results toward a complete overview, thus to be able to find the one that most suits the application at hand. In fact, many convergence results find their use in theoretical applications, such as Lyapunov stability analysis (Benaim, 1996; Benaïm, 1999; Khalil & Grizzle, 2002; Polyak, 1987), variational analysis (Iusem, Jofré, Oliveira, & Thompson, 2017, 2019; Malitsky, 2015, 2020; Yousefian, Nedić, & Shanbhag, 2014, 2017) and game equilibrium seeking (Facchinei & Pang, 2007; Franci & Grammatico, 2020a; Franci, Staudigl, & Grammatico, 2020; Koshal, Nedic, & Shanbhag, 2013), in automatic control, such as model predictive control (Lee & Nedić, 2015) and network control problems (Shi, Johansson, & Johansson, 2013), as well as in other engineering areas, e.g., training and learning in generative adversarial networks (Bot, Sedlmayer and Vuong, 2020; Franci & Grammatico, 2020b, 2021b), vehicle flow control in traffic networks (Duvocelle, Meier, Staudigl, & Vuong, 2019) and in modeling the prosumer behavior in smart power grids (Franci & Grammatico, 2020a; Franci et al., 2020; Kannan, Shanbhag, & Kim, 2013; Yi & Pavel, 2019).

1.1. Lyapunov decrease and Féjer monotonicity

In the mathematical literature, many convergence results hold for sequences of numbers while in system and control theory, the state and decision variables are usually *vectors* of real numbers. It is therefore important to understand the deep connection between the two theories. The bridging idea is to associate a real number to the state vector, i.e., via a function, and then prove convergence exploiting the properties of such a function. The most common example of this approach is that of Lyapunov theory where a suitable Lyapunov function is shown to be decreasing along the evolution of the state variable, thus obtaining convergence of the state vector to a target set (Benaim, 1996; Khalil & Grizzle, 2002; Polyak, 1987). An alternative approach is to consider the distance from a target set and show that such a distance vanishes eventually via a suitable technical result on the convergence of the distance-valued sequence of real numbers.

In this work, we focus mostly on the latter methodology. To explain our choice, let us note that solving an optimization problem consist of *designing a sequence* of vectors that converge to the solution, the minimum of a given cost function. Similarly, in algorithmic game theory, one usually aims at constructing a *sequence* that converge to an equilibrium, e.g., a Nash equilibrium, the optimum for each player given the actions of the other players. The key point here is that, in general, the target set is not known a priori, yet the distance of the

☆ This work was partially supported by NWO under research projects OMEGA (613.001.702) and P2P-TALES (647.003.003), and by the ERC under research project COSMOS (802348).

* Corresponding author.
E-mail addresses: b.franci@maastrichtuniversity.nl (B. Franci), s.grammatico@tudelft.nl (S. Grammatico).

https://doi.org/10.1016/j.arcontrol.2022.01.003
Received 4 November 2021; Received in revised form 20 January 2022; Accepted 21 January 2022
Available online 21 February 2022
1367-5788/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Convergence results for Féjer monotone sequences, deterministic sequences of real numbers and with variable metric (separated by the horizontal lines, respectively). For the applications, MI stands for Monotone Inclusion, VI for variational inequalities, NE for Nash Equilibrium problems, LYAP for Lyapunov analysis and NC for nonconvex optimization.

Result	Reference	Application	Reference
Proposition 3.1	Bauschke et al. (2011, Proposition 5.4)		
Theorem 3.2	Combettes (2001b, Theorem 3.8)		
Lemma 3.3	Opial et al. (1967) (Opial)	MI - Theorem 6.1	Malitsky and Tam (2020, Theorem 2.5)
		VI - Theorem 6.4	Malitsky (2020, Theorem 1)
Lemma 3.4	Combettes (2001b, Lemma 3.1)	NC - Theorem 6.9	Di Lorenzo and Scutari (2016, Theorem 3)
Corollary 3.5	Scutari and Sun (2019, Lemma 9)		
Lemma 3.6	Bauschke et al. (2011, Lemma 5.31)	VI - Theorem 6.4	Malitsky (2020, Theorem 1)
Corollary 3.7	Malitsky (2015, Lemma 2.8)	VI - Theorem 6.5	Malitsky (2015, Theorem 3.2)
		LYAP - Theorem 6.8	Polyak (1987, Theorem 1.4.1)
Corollary 3.8	Polyak (1987, Lemma 2.2.2)		
Lemma 3.9	Polyak (1987, Lemma 2.2.3)	NE - Theorem 6.7	Kannan and Shanbhag (2012, Theorem 2.4)
Lemma 3.10			
Lemma 3.11	Xu (2003, Lemma 2.1)		
Lemma 3.12	Extension of Xu (2002, Lemma 2.5)	NE - Theorem 6.6	Duvocelle et al. (2019, Theorem 3.1)
Corollary 3.13	Lei, Shanbhag and Chen (2020, Proposition 3)		
Corollary 3.14	Qin, Shang, and Su (2008, Lemma 1.1)		
Corollary 3.15	Xu (1998, Lemma 3)	MI - Theorem 6.3	Dadashi and Postolache (2019, Theorem 3.1)
Proposition 3.16	Alber, Iusem, and Solodov (1998, Proposition 2)		
Lemma 3.17	He and Yang (2013, Lemma 7)		
Lemma 3.18	Maingé (2008, Lemma 2.2)		
Lemma 3.19	Malitsky and Tam (2018, Lemma 2.7)	MI - Theorem 6.2	Malitsky and Tam (2020, Theorem 2.9)
Proposition 5.1	Combettes and Vũ (2013, Proposition 3.2)	MI - Theorem 8.1	Vũ (2013, Theorem 3.1)
Theorem 5.2	Combettes and Vũ (2013, Theorem 3.3)	MI - Theorem 8.1	Vũ (2013, Theorem 3.1)
Corollary 5.3	Combettes and Vũ (2013, Proposition 4.1)		

constructed sequence from such set can be analyzed anyways. On the contrary, in Lyapunov stability analysis, the target set is usually known a priori.

By exploiting the relation between the iterations and a suitable distance-like function, we show in this paper that convergence theorems represent a key ingredient for a wide variety of system-theoretic problems in fixed-point theory, game theory and optimization (Bauschke, Combettes, et al., 2011; Combettes, 2001b; Eremin & Popov, 2009; Facchinei & Pang, 2007; Polyak, 1987). In many cases, the study of iterative algorithms allows for a systematic analysis that follows from the concept of Féjer monotone sequence. The basic idea behind Féjer monotonicity is that at each step, each iterate is closer to the target set than the previous one. In a sense, the distance used for Féjer sequences can be seen as a specific class of Lyapunov function and Féjer monotonicity shows that it is decreasing along the iterates. The concept was first introduced in 1922 (Fejér, 1922), but the term Féjer monotone sequence was first used thirty years later in 1954 (Motzkin & Schoenberg, 1954) and a huge part of the studies on its properties was made in the 60s (Eremin, 1968a, 1968b, 1969; Eremin & Popov, 2009) and still continues (Combettes, 2001a, 2001b; Combettes & Pesquet, 2015; Combettes & Vũ, 2013; Kohlenbach, Leuştean, & Nicolae, 2018).

Unfortunately, Féjer monotonicity is hard to obtain, therefore the concept is typically relaxed to a quasi-Féjer property, where a vanishing error must be considered. Such an error term in the distance inequality is common in many equilibrium problems (Bauschke et al., 2011; Duflo, 2013; Duvocelle et al., 2019; Franci & Grammatico, 2020a; Iusem et al., 2017; Kannan et al., 2013; Malitsky & Tam, 2020; Polyak, 1987; Van Nguyen, 2017), especially in the stochastic case where the concept of quasi-Féjer monotone sequence was first introduced (Ermol'Ev, 1969; Ermoliev & Wets, 1988). However, these properties are not necessarily enough to ensure convergence, hence, (quasi) Féjer monotonicity is often used in combination with convergence results on sequences of real numbers. These technical results have been used in many theoretical and computational applications that range from stochastic Nash equilibrium seeking (Franci & Grammatico, 2020a; Franci et al., 2020; Koshal et al., 2013) to machine learning (Bot, Sedlmayer et al., 2020; Duvocelle et al., 2019; Franci & Grammatico, 2020b).

1.2. What this survey is about

In this survey, we present a number of convergence theorems for sequences of real (random) numbers. We show how they can be used in combination with (quasi) Féjer monotone sequences or Lyapunov functions to obtain convergence of an iterative algorithm, essentially a discrete-time dynamical system, to a desired solution. Moreover, we present some applications to show how they can be adopted in a variety of settings. Specifically, we present convergence results for both deterministic and stochastic sequences of real numbers and we also include some results on Féjer monotone sequences and with variable metric. We show that these results help proving not only convergence of an iterative algorithm but also the Law of Large Numbers, with applications in model predictive control (Lee & Nedić, 2015) and opinion dynamics (Shi et al., 2013) among others.

We report in Tables 1 and 2 the results for deterministic and stochastic sequences respectively, with the corresponding bibliographic source and application.

The paper is organized as follows. In the next section, we recall some preliminary notions on the concept of "convergence" and of random variables. Section 3 is devoted to deterministic convergence results while the stochastic case is discussed in Section 4. An extension with variable metric is considered in Section 5. Sections 6–8 propose applications of the convergence lemmas for deterministic, stochastic, and variable metric sequences, respectively.

1.3. What this survey is not about

This is not a survey on solution algorithms for optimization problems and variational inequalities. Some relevant references on iterative methods include Bauschke et al. (2011), Combettes and Pesquet (2021), Doob (1953), Facchinei and Pang (2007), Polyak (1987), Rockafellar (1970) and the references therein.

We also remark that, despite the notion of Féjer sequence is used throughout the paper, this is not a survey on the properties of Féjer monotone sequences. The interested reader may refer to Bauschke et al. (2011), Berg, Engel, Pazderski, and Stolle (1995), Combettes (2001a, 2001b), Combettes and Pesquet (2015), Combettes and Vũ (2013) and Kohlenbach et al. (2018).

Table 2
Convergence results for stochastic sequences of real random variables and stochastic Féjer monotone sequences (separated by double horizontal lines, respectively). For the applications, VI stands for variational inequalities, NE for Nash Equilibrium problems, MPC for Model Predictive Control, LLN for Law of Large Numbers and NET for Network control problems.

Result	Reference	Application	Reference
Lemma 4.1	Robbins and Siegmund (1971) (Robbins–Siegmund)	VI - Theorem 7.1	Bot, Mertikopoulos, Staudigl and Vuong (2020, Theorem 4.5)
		VI - Theorem 7.2	Iusem et al. (2017, Theorem 3.18)
		NE - Proposition 7.3	Koshal et al. (2013, Proposition 3)
		MPC - Proposition 7.4	Lee and Nedić (2015, Proposition 1)
Corollary 4.2	Polyak (1987, Lemma 2.2.9) (Gladyshev)		
Corollary 4.3	Poggio, Voinea, and Rosasco (2011, Theorem B.2)		
Corollary 4.4	Duflo (2013, Corollary 1.3.13)	LLN - Theorem 7.5	Duflo (2013, Theorem 1.3.15)
Lemma 4.6	Combettes and Pesquet (2019, Lemma 2.1)		
Lemma 4.7	Polyak (1987, Lemma 2.2.10)	VI - Theorem 7.6	
		NET - Theorem 7.7	Shi et al. (2013, Theorem 5)
Proposition 4.5	Combettes and Pesquet (2015, Proposition 2.3)		
Proposition 5.4	Vu (2016, Proposition 2.4)		

2. Notation and preliminaries

We use the nomenclature and notation from Bauschke et al. (2011) and Rockafellar (1970).

$\mathbb{N}$ indicates the set of natural numbers and $\mathbb{R}$ ($\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$) is the set of (extended) real numbers. $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ denotes the standard inner product and $\|\cdot\|$ is the associated Euclidean norm. $\mathbb{B} = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$ represents the unit ball. Let $d_{\mathcal{X}}(x) = \min_{y \in \mathcal{X}} \|x - y\|$ be the distance between x and the set $\mathcal{X}$.

We indicate that a matrix A is positive definite, i.e., $x^\top A x > 0$, with $A \succ 0$. Given a symmetric $W \succ 0$, the W-induced inner product is $\langle x, y \rangle_W = \langle Wx, y \rangle$ and the associated norm is defined as $\|x\|_W = \sqrt{\langle Wx, x \rangle}$. Id is the identity operator. Given a continuous linear operator $T : \mathbb{R}^n \to \mathbb{R}^n$, the *adjoint* of T is the unique continuous linear operator T^* such that $\forall x, y \in \mathbb{R}^n$ $\langle Tx, y \rangle = \langle x, T^* y \rangle$. Let $S(\mathbb{R}^n)$ be the set of self-adjoint bounded linear operators of $\mathbb{R}^n$ and let the Loewner partial order be defined for all $T_1, T_2 \in S(\mathbb{R}^n)$ as $T_1 \succeq T_2 \Leftrightarrow \forall x \in \mathbb{R}^n$ $\langle T_1 x | x \rangle \geq \langle T_2 x | x \rangle$. Let $\beta \geq 0$ and $\mathcal{P}_\beta = \{L \in S(\mathbb{R}^n) | L \succeq \beta \mathrm{Id}\}$. Positive semidefinite matrices belongs to $\mathcal{P}_\beta$.

Unless otherwise mentioned, we use v, u and w for (real or random) numbers while we use x, y, z to indicate vectors (of real numbers or random variables), i.e., $v, u, w \in \mathbb{R}$ and $x, y, z \in \mathbb{R}^n$, respectively. Capital letters indicate operators or matrices. Letters from the Greek alphabet are also used for real numbers but they mostly represent errors (ε), step size sequences (α) or coefficients (δ, γ); ξ often indicates random quantities. Since it may be dependent on the context, when necessary, the meaning is introduced along with the symbol. In general, calligraphic capital letters indicate sets, $\mathcal{C}$ indicates a convex set and $\mathcal{S}$ a target or solution set. Throughout the survey, we suppose that the sequence $(x^k)_{k \in \mathbb{N}}$ belongs to a set $\mathcal{X} \subseteq \mathbb{R}^n$. Further assumptions will be made when necessary.

Given a vector $x \in \mathbb{R}^n$, we indicate the maximum entry as $x_{\max} = \max_{1 \leq i \leq N} \{x_i\}$ and, analogously, the minimum entry as $x_{\min} = \min_{1 \leq i \leq N} \{x_i\}$. Most often, the superscript *, e.g., x^*, indicates a solution of the problem, while the bar, i.e., $\bar{x}$, indicates an accumulation point of an iterative process.

With reference to the application sections, we use Standing Assumptions to state technical conditions that implicitly hold throughout the paper, while Assumptions are postulated only when explicitly used.

More notation and definitions related to monotone operator theory, functional to the application sections, are postponed to Appendix.

2.1. Convergence notions

Let us first recall some definitions related to the notion of convergence itself.

Definition 2.1. A sequence $(x^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^n$ is said to converge weakly to a point $\bar{x} \in \mathcal{X}$ if, for all $y \in \mathcal{X}$,

$$\langle x^k, y \rangle \to \langle \bar{x}, y \rangle \text{ as } k \to \infty.$$

A sequence $(x^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^n$ is said to converge strongly to a point $\bar{x} \in \mathcal{X}$ if

$$\lim_{k \to \infty} \|x^k - \bar{x}\| = 0.$$

In general, strong convergence implies weak convergence. In finite dimension, the two notions are equivalent (Bauschke et al., 2011, Lemma 2.51), hence, in this paper, we generally talk about convergence.

Let us also introduce the concept of linear convergence. While the study of rates of convergence is not the focus of this survey, we include a result in this direction for the sake of completeness (Lemma 3.19 in Section 3.2).

Definition 2.2. A sequence $(x^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^n$ is said to converge Q-linearly to a point $\bar{x}$ if there exists $c > 0$ such that

$$\lim_{k \to \infty} \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} = c.$$

A sequence $(x^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^n$ is said to converge R-linearly to a point $\bar{x}$ if there exists a sequence $(\varepsilon_k)_{k \in \mathbb{N}}$ such that

$$\|x^k - \bar{x}\| \leq \varepsilon^k \quad \text{for all } k \in \mathbb{N}$$

and the sequence $(\varepsilon^k)_{k \in \mathbb{N}}$ converges Q-linearly to zero.

Given the definitions of convergence, let us define the concept of cluster point.

Definition 2.3. A point $\bar{x} \in \mathbb{R}^n$ is said to be a cluster point (or limit point or accumulation point) of a sequence $(x^k)_{k \in \mathbb{N}}$ if, for every $\epsilon > 0$ and for $\bar{k} \in \mathbb{N}$ there exists $k \geq \bar{k}$ such that $x^k \in \{\bar{x}\} + \epsilon \mathbb{B}$. In other words, there is at least one $\bar{k} \in \mathbb{N}$ such that x^k lies in a neighborhood of $\bar{x}$ for all $k \geq \bar{k}$.

The set of all cluster points is called limit set.

If a sequence $(x^k)_{k \in \mathbb{N}}$ in $\mathbb{R}^n$ has a subsequence that converges to a point $\bar{x} \in \mathbb{R}^n$, then $\bar{x}$ is called a sequential cluster point of $(x^k)_{k \in \mathbb{N}}$.

Example 2.1 (*A Cluster Point is Also a Sequential Cluster Point*)**.** Consider the sequence $(x^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^2$ defined as $x^k = \frac{1}{k}(cos(k\frac{\pi}{2}), sin(k\frac{\pi}{2}))$. The sequence converges to $\bar{x} = (0, 0)$ as $k \to \infty$, which is a cluster point and a sequential cluster point, as shown in Fig. 1. The limit set is the singleton $\{(0, 0)\}$.

Example 2.2 (*A Sequential Cluster Point is Not Necessarily a Cluster Point*)**.** Consider the sequence $(x^k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^2$ defined as $x^k = \frac{k+3}{k}\left(cos\left(k\frac{\pi}{2}\right), sin\left(k\frac{\pi}{2}\right)\right)$. The sequence does not converge but it has many sequential cluster points (see Fig. 2). For instance, consider $\bar{x} = (1, 0) = (cos(2\pi), sen(2\pi))$. Then, the subsequence (x^{k_n}) with $k_n = 4n$, $n \in \mathbb{N}$ converges to $\bar{x}$ which in turn is a sequential cluster point. However, the limit set is given by the circumference $\{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$, in red in Fig. 2.

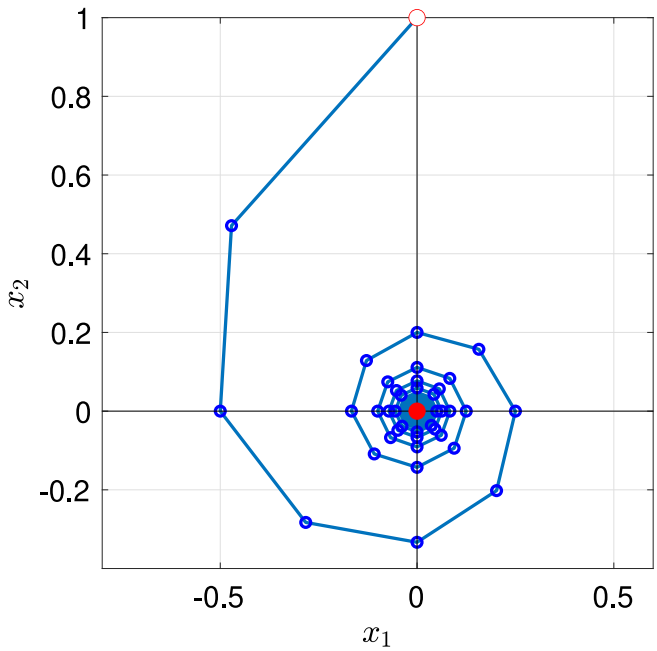

Fig. 1. Sequence converging to a cluster point which is also a sequential cluster point (Example 2.1). The empty dot represents the initial point and the red dot is the cluster point. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

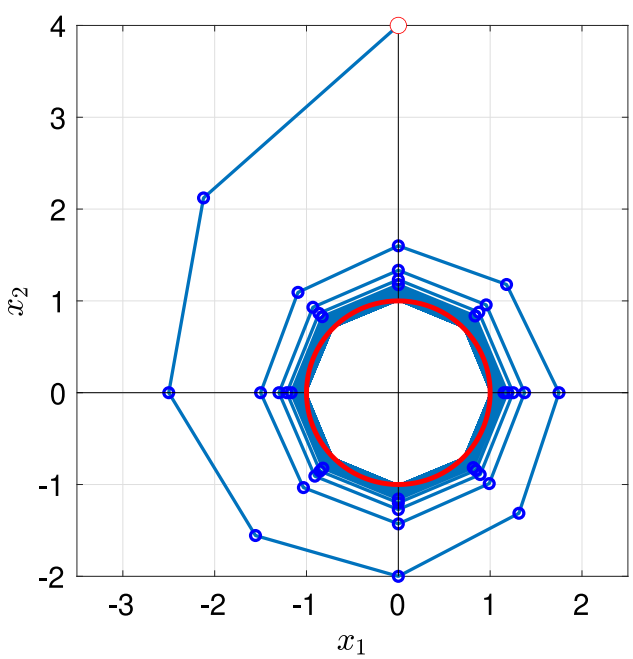

Fig. 2. Sequence converging to a set of sequential cluster points where none of them is a cluster point (Example 2.2). The empty dot represents the initial point and the red circle is the limit set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Example 2.3 (*ω-Limit Set*)**.** The concept of limit set reminds that of ω-limit set (Benaim, 1996; Benaïm, 1999). Given a continuous function $f : \mathbb{R} \to \mathbb{R}$, the ω-limit set is the set of cluster points of the forward orbit of the iterated function f at $x \in \mathbb{R}$, namely,

$$\omega(x, f) = \bigcap_{n\in\mathbb{N}} \{f^k(x) : k > n\}.$$

In particular, given a dynamical system with flow $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, y is a ω-limit point of x if there exist $(t^k)_{k\in\mathbb{N}} \subseteq \mathbb{R}$ such that $\lim_{k\to\infty} t^k = \infty$ and $\lim_{k\to\infty} \phi(t^k, x) = y$.

Let us conclude this section with some preliminary results related to the convergence properties of a given sequence. We consider these results common knowledge and we refer to them throughout the paper, even without a specific reference.

Lemma 2.1 (Bauschke et al., 2011, Lemma 2.45)**.** *Let $(x^k)_{k\in\mathbb{N}}$ be a bounded sequence in $\mathbb{R}^n$. Then, $(x^k)_{k\in\mathbb{N}}$ possesses a convergent subsequence.*

Lemma 2.2 (Bauschke et al., 2011, Lemma 2.46)**.** *Let $(x^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Then, $(x^k)_{k\in\mathbb{N}}$ converges if and only if it is bounded and possesses at most one sequential cluster point.*

Lemma 2.3 (Bauschke et al., 2011, Lemma 2.47)**.** *Let $(x^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$ and let $\mathcal{X}$ be a nonempty subset of $\mathbb{R}^n$. Suppose that, for every $x \in \mathcal{X}$, $(\|x^k - x\|)_{k\in\mathbb{N}}$ converges and that every sequential cluster point of $(x^k)_{k\in\mathbb{N}}$ belongs to $\mathcal{X}$. Then, $(x^k)_{k\in\mathbb{N}}$ converges to a point in $\mathcal{X}$.*

Example 2.4 (*Assumptions of Lemmas 2.1–2.3*)**.** Consider the sequence defined by $x^k = (-1)^k$ for all $k \in \mathbb{N}$. Both 1 and -1 are sequential cluster points but not cluster points. The sequence does not converge. Let us use $(x^k)_{k\in\mathbb{N}}$ to verify Lemmas 2.1–2.3.

The sequence is bounded in $[-1, 1]$ and it has (at least) two convergent subsequences: $x^{k_n} = -1$ and $x^{k_m} = 1$, $n, m \in \mathbb{N}$. Hence, Lemma 2.1 holds. However, the sequence is not convergent. In fact, contrary to Lemma 2.2, it has two sequential cluster points. Concerning Lemma 2.3, we note that the sequence $(\|x^k - \bar{x}\|)_{k\in\mathbb{N}}$ does not converge for any $\bar{x} \in [-1, 1] \setminus \{0\}$. On the other hand, it converges for $\bar{x} = 0$ which is not a cluster point of the sequence $(x^k)_{k\in\mathbb{N}}$.

2.2. Probability theory

Concerning the stochastic case, we focus on almost sure convergence. Let us first introduce the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, $\mathcal{F}$ is the event space, and $\mathbb{P}$ is the probability function defined on the event space. The symbol $\mathbb{E}$ is used for the associated expected values.

Let us recall that a random variable is a function from the sample space to a measurable space (in our case, the set of real numbers), i.e., $x : \Omega \to \mathbb{R}^n$. However, for brevity, we omit the dependency on the sample space and write x instead of $x(\omega)$, $\omega \in \Omega$, when it is clear from the context.

Definition 2.4. A sequence $(x^k)_{k\in\mathbb{N}}$ of random variables converges almost surely (a.s.) toward $\bar{x} \in \mathcal{X}$ if

$$\mathbb{P}\left[\lim_{n\to\infty} x^k = \bar{x}\right] = 1.$$

From now on, results involving random variables are supposed to hold almost surely, even if it is not explicitly mentioned.

Example 2.5 (*a.s. Convergence*)**.** Let $\Omega = [0, 1]$ be a continuous sample space with the uniform probability distribution. For each element of the sample space $\omega \in \Omega$, let us define the sequence of random variables $v^k(\omega) := \omega + (\omega)^k$ and the random variable $\bar{v}(\omega) := \omega$. Then, we have that $\lim_{k\to\infty} |v^k(\omega) - \bar{v}(\omega)| = 0$ for all $\omega \in [0, 1)$. Instead, for $\omega = 1$, we have $v^k(\omega) = 2$ for all $k \in \mathbb{N}$, thus, $v^k(\omega)$ does not converge to $\bar{v}(\omega)$. However, since $\mathbb{P}[0 \le \omega < 1] = 1$, $v^k(\omega)$ converges a.s. to $\bar{v}(\omega)$ as $k \to \infty$. In Fig. 3, we show how the distance from the limit point moves toward zero with high probability as the number of iterations increases, i.e., $\mathbb{P}[\lim_{k\to\infty} |v^k(\omega) - \bar{v}(\omega)| = 0] = 1$.

Let us recall some probabilistic and stochastic definitions that will be useful later on, starting from the definition of filtration.

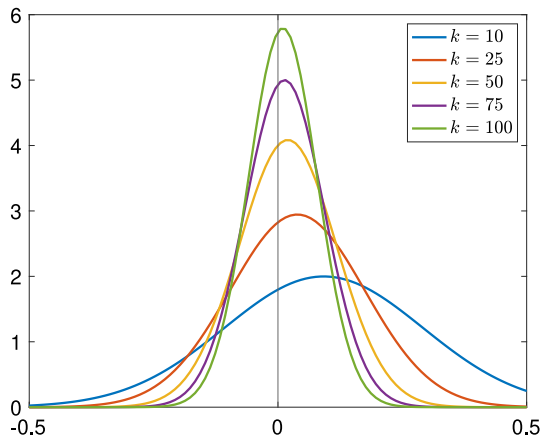

Fig. 3. Graphical representation of a.s. convergence. By increasing the number of iterations, the mass of the probability distribution concentrates on the limit point (Example 2.5).

Definition 2.5. Let $(\xi^k)_{k\in\mathbb{N}} \subseteq \Omega$ be a sequence of random variables and let $\mathcal{F}_k$, $k \in \mathbb{N}$ be the σ-algebra of Ω, generated by the events prior to k, that is, $\mathcal{F}_0 = \sigma\left(X_0\right)$ and $\mathcal{F}_k = \sigma\left(X_0, \xi_1, \xi_2, \dots, \xi_k\right)$ for all $k \geq 1$. Then $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ is called a filtration, if $\mathcal{F}_k \subseteq \mathcal{F}_\ell \subseteq \Omega$ for all $k \leq \ell$.

In words, a filtration is a family of σ-algebras non-decreasingly ordered that collects the history of ξ^k. Given a filtration, a subsequent important concept is that of martingale (Doob, 1953, Chapter 7), (Chung, Williams, & Williams, 1990, Section 1.9), (Kushner & Yin, 2003, Section 4.1).

Definition 2.6. A sequence of random variables $(v^k)_{k\in\mathbb{N}}$ is said to be a martingale adapted to $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ if it is integrable and for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[v^{k+1}|\mathcal{F}_k\right] = v^k.$$

It is a supermartingale if for all $k \in \mathbb{N}$

$$\mathbb{E}\left[v^{k+1}|\mathcal{F}_k\right] \leq v^k,$$

and a submartingale if for all $k \in \mathbb{N}$

$$\mathbb{E}\left[v^{k+1}|\mathcal{F}_k\right] \geq v^k.$$

These notions are the stochastic generalization of the notion of monotone (decreasing or increasing) sequences. Moreover, we note that every martingale is a submartingale and a supermartingale, while every sequence which is *both* a submartingale and a supermartingale is also a martingale.

Example 2.6 (*Martingales*)**.** Let $(x^k)_{k\in\mathbb{N}}$ be the sequence generated by the fortune of a gambler after k tosses of a fair coin. The gambler wins 1 if the coin comes up head (with probability $p = \frac{1}{2}$) and loses 1 otherwise. The expected fortune after the next toss is equal to the present fortune, i.e., $\mathbb{E}[x^{k+1}] = x^k$, hence the sequence is a martingale.

Let us now consider the toss of a biased coin, with head coming up with probability $p \neq \frac{1}{2}$. If $p > \frac{1}{2}$, on average the gambler wins money, i.e., $\mathbb{E}[x^{k+1}] \geq x^k$ and the sequence is a supermartingale. On the other hand, if $p < \frac{1}{2}$ the gambler loses and the sequence is a submartingale. See Borkar (1995), Doob (1953), Duflo (2013), Polyak (1987) and Stroock (2010) for other examples.

We conclude this section with the following result, due to Borkar (1995, Theorem 3.3.1), Doob (1953, Theorem 7.4.1) and Polyak (1987, Lemma 2.2.7), .

Theorem 2.4 (*Martingale Convergence Theorem*)**.** *Suppose* $(x^k)_{k\in\mathbb{N}}$ *is a nonnegative super-martingale which satisfies* $\sup_{k\in\mathbb{N}} \mathbb{E}[|x^k|] < \infty$. *Then, almost surely, there exists* $\bar{x} \geq 0$ *such that* $\lim_{k\to\infty} x^k = \bar{x}$ *and* $\mathbb{E}[|\bar{x}|] < \infty$.

2.3. Properties of sequences

In this section, we collect some properties of sequences of numbers that are used later on as assumptions to claim convergence. We also include some examples and establish some relations between such properties.

The most used assumption involves the summability of a sequence. A sequence $(\delta^k)_{k\in\mathbb{N}}$ of real numbers is *summable* if $\sum_{k=1}^{\infty} \delta^k < \infty$, i.e., if the series converges; it is *square summable* if $\sum_{k=1}^{\infty} (\delta^k)^2 < \infty$. The concept can be generalized to higher powers, but they are not relevant for the sequences we consider here. Moreover, one can consider *absolute summability* of a sequence, i.e., $\sum_{k=1}^{\infty} |\delta^k| < \infty$, which implies convergence of the series, as absolute summability is stronger than summability; clearly, the two concepts are equivalent for nonnegative sequences.

Example 2.7 (*Summable Sequence*)**.** A simple example of summable sequence is $\left(\frac{1}{k^2}\right)_{k\in\mathbb{N}}$. Thus, a square summable is $\left(\frac{1}{k}\right)_{k\in\mathbb{N}}$. In fact, $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$ and $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$.

In light of Example 2.7, let us note that a square summable sequence is not necessarily summable. However, the vice versa holds, i.e., a summable sequence is also square summable.

One of the reasons why the summability condition is used in the convergence lemmas is its relation with the convergence of an infinite product. In fact, given a positive sequence $(\delta^k)_{k\in\mathbb{N}}$, it holds that $\sum_{k=1}^{\infty} \delta^k$ converges if and only if $\prod_{k=1}^{\infty}(1-\delta^k)$ converges. Similarly, the condition for a sequence being *not* summable, i.e., $\sum_{k=1}^{\infty} \delta^k = \infty$, is equivalent to $\prod_{k=1}^{\infty}(1-\delta^k) = 0$ when $\delta^k \in (0,1)$. A proof of these statements can be found in Knopp (1990, Chapter VII).

Example 2.8 (*Summable Sequence Continues*)**.** For the sequence $\left(\frac{1}{k^2}\right)_{k\in\mathbb{N}}$ we have that $\prod_{k=1}^{n}(1-\frac{1}{k^2}) = \frac{n+1}{2n} \to \frac{1}{2}$ as $n \to \infty$. Instead, if one considers the sequence $\left(\frac{1}{k}\right)_{k\in\mathbb{N}}$ it holds that $\prod_{k=1}^{n}(1-\frac{1}{k}) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4}\cdots\frac{n-1}{n} = \frac{1}{n} \to 0$ as $n \to \infty$.

In some cases, it is assumed that $\sum_{k=1}^{\infty}(1-\delta^k) = \infty$ with $\delta^k \in [0,1)$. This condition is not very restrictive in terms of (non) summability of the sequence. In fact, both summable and nonsummable sequences can be found that satisfy the equation above.

Example 2.9 (*Summable Sequence Continues*)**.** While all summable sequences satisfy $\sum_{k=1}^{\infty}(1-\delta^k) = \infty$, a nonsummable sequence is provided by our usual example, $\left(\frac{1}{k}\right)_{k\in\mathbb{N}}$. To see this, consider that $\sum_{k=1}^{\infty}(1-\delta^k) = \infty$ is equivalent to $\prod_{k=1}^{\infty}\delta^k = 0$. Then, we have $\prod_{1=k}^{n}\frac{1}{k} = \frac{1}{n!} \to 0$ as $n \to \infty$ (! being the factorial operator).

Remark 2.1. Without anticipating the convergence results, let us note that the equivalences between infinite sums and infinite products are interesting for proving convergence lemmas because of the following fact. The sequences involved must often satisfy an inequality of the form $v^{k+1} \leq (1-\delta^k)v^k + \epsilon_1^k$, for some positive ϵ_1^k. By iterating such inequality, one obtains $v^{k+1} \leq v^0 \prod_{j=1}^{k}(1-\delta^j) + \epsilon_2^k$, where the coefficient multiplying v^0, $\prod_{j=1}^{k}(1-\delta^j)$, depends on the summability of the sequence $(\delta^k)_{k\in\mathbb{N}}$.

Sometimes we may just use the fact that a sequence is vanishing, in the sense that $\lim_{k\to\infty}\delta^k = 0$, which is implied by summability while the vice versa does not necessarily hold.

To conclude this section, let us remark that the summability of a sequence gives an idea of how fast the sequence is vanishing while such information is lost when just considering the limit.

2.4. Distance from a target set

The basic idea for proving convergence of a sequence is that the distance from the solution should vanish or at least decrease at each iteration. This is particularly important when we consider vectors, i.e., when convergence results for sequences of real numbers cannot be applied directly.

The most used concept in this direction is that of Féjer monotone sequence. The term was coined in Motzkin and Schoenberg (1954) but the concept was first proposed by Féjer in Fejér (1922). These processes have been widely studied in the literature (Bauschke, Borwein, & Combettes, 2003; Combettes, 2001a, 2001b; Combettes & Pesquet, 2015; Eremin, 1969) since they can be applied in solving classical problems as systems of equations or inequalities, operator equations with a priori information, equilibrium problems, and many others (Bot, Mertikopoulos et al., 2020; Combettes, 2001a, 2001b; Eremin & Popov, 2009; Iusem et al., 2017). The key point is that one can take the target set to be the solution set of the problem of interest (even if it is unknown). Then, since the distance from the target decreases, the sequence will eventually reach (a point close to) the solution.

Definition 2.7. A sequence $(x^k)_{k\in\mathbb{N}}$ is Féjer monotone with respect to a target set $\mathcal{S} \neq \emptyset$ if for every $\bar{x} \in \mathcal{S}$, it holds that for all $k \in \mathbb{N}$

$$\|x^{k+1} - \bar{x}\| \leq \|x^k - \bar{x}\|.$$

In words, Definition 2.7 states that the distance between the iterates and any point $\bar{x} \in \mathcal{S}$ does not increase.

Example 2.10 (*Féjer Monotone Sequence of Numbers*)**.** Let us consider the sequence $v^k = \frac{(-1)^k}{k}$. Though the sequence is oscillating, it is convergent to $\bar{v} = 0$ and it is Féjer monotone with respect to $\mathcal{S} = \{0\}$.

Example 2.11 (*Féjer Monotone Sequence of Vectors*)**.** An example of a Féjer monotone sequence is the one generated by the projection (Definition A.1) onto a nonempty, closed and convex set $\mathcal{C}$, which acts as the target set, Berg et al. (1995), Combettes (2001a, 2001b), Eremin and Popov (2009) and Gubin, Polyak, and Raik (1967), i.e.,

$$x^{k+1} = \mathrm{proj}_{\mathcal{C}}(x^k).$$

The claim follows immediately from the fact that the projection operator is firmly nonexpansive (Bauschke et al., 2011, Proposition 4.16), hence nonexpansive (Definition A.3). In fact, any sequence generated by an iteration of the form $x^{k+1} = T(x^k)$ where T is a nonexpansive operator is Féjer monotone (Combettes, 2001a, Equation (2)).

We remark that the diminishing distance from a target point does not necessarily imply convergence to such point. Specifically, we note that a Féjer monotone sequence $(x^k)_{k\in\mathbb{N}}$ with respect to a nonempty set $\mathcal{S}$ may not converge even if the limit set is not empty.

Example 2.12 (*Non-convergent Féjer Monotone Sequence*)**.** The sequence defined as $x^k = (-1)^k x_0$ for all $k \in \mathbb{N}$ is Féjer monotone with respect to $\mathcal{S} = \{0\}$ but it does not converge for any $x_0 \notin \mathcal{S}$ (Combettes, 2001b, page 9), (Combettes, 2001a, page 1).

The notion of Féjer monotonicity can be extended in various directions (Berg et al., 1995; Combettes, 2001b; Combettes & Pesquet, 2015; Combettes & Vũ, 2013; Eremin, 1969; Lin, Rosasco, Villa, & Zhou, 2018). Here we recall only the concept of quasi-Féjer monotone sequence, first introduced in the stochastic case (Ermol'Ev, 1969; Ermoliev & Wets, 1988) (see also Definition 2.9) and later in several (deterministic) variants (Berg et al., 1995; Combettes, 2001b, 2004; Lin et al., 2018).

Definition 2.8. Let $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. A sequence $(x^k)_{k\in\mathbb{N}} \subseteq \mathbb{R}^n$ is quasi-Féjer monotone with respect to a target set $\mathcal{S} \neq \emptyset$ if for every $\bar{x} \in \mathcal{S}$ there exists a nonnegative sequence $(\varepsilon^k)_{k\in\mathbb{N}}$ such that $\sum_{k=0}^{\infty} \varepsilon^k < \infty$ and it holds that

$$\phi(\|x^{k+1} - \bar{x}\|) \leq \phi(\|x^k - \bar{x}\|) + \varepsilon^k \text{ for all } k \in \mathbb{N}.$$

Remark 2.2. Definition 2.8 is perhaps the most general definition of quasi-Féjer monotone sequence, as there are no restrictions on the function ϕ. However, besides some general results (see, e.g., Proposition 5.1 and Theorem 5.2), many convergence theorems hold for a given choice of the function, i.e., $\phi = |\cdot|$ or $\phi = |\cdot|^2$. For details, see Section 3.1 or (Alber et al., 1998; Combettes, 2001b; Combettes & Vũ, 2013; Ermol'Ev, 1969; Ermoliev & Wets, 1988).

Next, we give a definition of Féjer monotone sequence in the stochastic case. Stochastic quasi-Féjer monotone sequences were first introduced in Ermol'Ev (1969) and later discussed in Barty, Roy, and Strugarek (2007) and Combettes and Pesquet (2015). The interpretation is that the expected value of the distance from the target set is non-increasing, which reminds the definition of (super)martingale (Combettes & Pesquet, 2015; Ermol'Ev, 1969).

Definition 2.9. Let $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. A sequence $(x^k)_{k\in\mathbb{N}}$ of random variables is stochastic Féjer monotone with respect to a target set $\mathcal{S} \neq \emptyset$ if for every $\bar{x} \in \mathcal{S}$ it holds that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\phi(\|x^{k+1} - \bar{x}\|)|\mathcal{F}_k\right] \leq \phi(\|x^k - \bar{x}\|).$$

It is called stochastic quasi-Féjer monotone relative to a target set $\mathcal{S} \neq \emptyset$ if for every $\bar{x} \in \mathcal{S}$ there exists a nonnegative sequence $(\varepsilon^k)_{k\in\mathbb{N}}$ such that $\sum_{k=0}^{\infty} \varepsilon^k < \infty$ and it holds that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\phi(\|x^{k+1} - \bar{x}\|)|\mathcal{F}_k\right] \leq \phi(\|x^k - \bar{x}\|) + \varepsilon^k.$$

Definitions 2.8 and 2.9 hold true for any norm of choice, yet other metrics can be considered (see Remark 2.3). Moreover, variable metrics have been considered as well (Combettes & Vũ, 2013; Vũ, 2013). We report here the definition for deterministic quasi-Fejer monotone sequences relative to a variable metric but it can be also extended to the stochastic case (Vu, 2016) or to the Bregman distance (Van Nguyen, 2016) (Remark 2.3).

Definition 2.10. Let $\beta \geq 0$ and $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and let $(W_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\beta$. A sequence $(x^k)_{k\in\mathbb{N}}$ of random variables is quasi-Féjer monotone with respect to a target set $\mathcal{S} \neq \emptyset$ and relative to $(W_k)_{k\in\mathbb{N}}$, if, given a nonnegative sequence $(\eta^k)_{k\in\mathbb{N}}$ such that $\sum_{k=0}^{\infty} \eta^k < \infty$, for every $\bar{x} \in \mathcal{S}$ there exists a nonnegative sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ such that $\sum_{k=0}^{\infty} \varepsilon^k < \infty$ and for all $k \in \mathbb{N}$

$$\phi(\|x^{k+1} - \bar{x}\|_{W_{k+1}}) \leq (1 + \eta_k)\, \phi(\|x^k - \bar{x}\|_{W_k}) + \varepsilon^k.$$

There are many results on (stochastic, quasi) Féjer monotone sequences but they lie outside the scope of this survey. For a deeper insight on this topic we refer to Bauschke et al. (2011), Bauschke, Dao, and Moursi (2015), Combettes (2001b), Combettes and Pesquet (2015) and Combettes and Vũ (2013).

Remark 2.3. An important generalization of Féjer monotonicity is that of Bregman monotonicity (Bauschke et al., 2003; Bregman, 1967; Facchinei & Pang, 2007; Van Nguyen, 2016, 2017). The concept has received a rising interest recently in the system and control community (Alacaoglu & Malitsky, 2021; Ananduta & Grammatico, 2021; Benning, Betcke, Ehrhardt, & Schönlieb, 2021; Bravo, Leslie, & Mertikopoulos, 2018; Mertikopoulos et al., 2018). For the sake of completeness, we report here the definition, and later on we recall when some results hold also with the Bregman distance.

Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed convex set and let $f : \mathcal{C} \to \mathbb{R}$ be a strictly convex continuous function which is continuously differentiable on $int\mathcal{C}$. The Bregman distance associated with f is

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \tag{2.1}$$

and it has the following geometric interpretation: $D_f(x, y)$ is the difference between $f(x)$ and the value at x of the linearized approximation of $f(x)$ at y. $D_f(x, y)$ is nonnegative and it is zero if and only if $x = y$.

We note that in general the Bregman distance is not a "real" distance, since it may fail to satisfy, for instance, the triangular inequality.

An example of a Bregman function is $f = \|\cdot\|^2$ whose associated distance is $D_f(x, y) = \|x - y\|^2/2$. Another example is given by $g(x) = \sum_{i=1}^n x_i \log x_i$ with the convention that $0 \log 0 = 0$. The associated distance is $D_g(x, y) = \sum_{i=1}^n (x_i \log \frac{x_i}{y_i} + y_i - x_i)$ (Facchinei & Pang, 2007, Example 12.7.4), i.e., the Kullback–Leibler divergence (Kullback, 1997; Kullback & Leibler, 1951), widely used in machine learning and generative adversarial networks (Goodfellow, 2016; Goodfellow et al., 2014).

A sequence $(x^k)_{k\in\mathbb{N}}$ in C is Bregman monotone with respect to a set $\mathcal{S}$ if the following conditions hold:

(i) $\mathcal{S} \cap C \neq \emptyset$,
(ii) $(x^k)_{k\in\mathbb{N}}$ lies in $int(C)$,
(iii) for every $\bar{x} \in \mathcal{S} \cap C$, $D_f(\bar{x}, x^{k+1}) \leq D_f(\bar{x}, x^k)$ for all $k \in \mathbb{N}$.

3. Convergence of deterministic sequences

In this section, we walk through a number of convergence results for deterministic sequences of real numbers. When possible, we propose first the most general result and then show its consequences. We start with some results on Féjer monotone sequences and then move to general sequences of real numbers.

3.1. Féjer monotone convergent sequences

The first result we present is related to the concept of Féjer monotone sequences and it was originally proposed in Bauschke et al. (2011). Parts of this result are also in Berg et al. (1995, Theorems 2.7 and 2.10) while in Combettes (2001a, Propositions 1–4) a distinction between strong and weak convergence is made. Other properties of Féjer monotone sequences can be found in Alber et al. (1998), Bauschke et al. (2011), Berg et al. (1995), Combettes (2001a, 2001b) and reference therein.

Proposition 3.1 (*Proposition 5.4, Bauschke et al., 2011*). *Let $\mathcal{X}$ be a nonempty subset of $\mathbb{R}^n$ and let $(x^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Suppose that $(x^k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathcal{X}$. Then, the following statements hold:*

(i) $(x^k)_{k\in\mathbb{N}}$ is bounded;
(ii) For every $\bar{x} \in \mathcal{X}, (\|x^k - \bar{x}\|)_{k\in\mathbb{N}}$ converges;
(iii) $(d_{\mathcal{X}}(x^k))_{k\in\mathbb{N}}$ is decreasing and converges;
(iv) $\|x^{k+m} - x^k\| \leq 2d_{\mathcal{X}}(x^k)$ for all $m, k \in \mathbb{N}$;

Proof. The statements follow from the definition of Féjer monotone sequence (Definition 2.7). □

Remark 3.1. A similar result holds also for quasi-Féjer sequences (Combettes, 2001b, Proposition 3.3), (Alber et al., 1998, Proposition 1). However, in such a case it is not possible to prove that the distance from the target set is decreasing as in Proposition 3.1(iii).

Formally, let $\mathcal{X} \subseteq \mathbb{R}^n$ be nonempty closed convex and let $(x^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Suppose that $(x^k)_{k\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathcal{X}$. Then, the following statements hold:

(i) $(x^k)_{k\in\mathbb{N}}$ is bounded;
(ii) For every $\bar{x} \in \mathcal{X}, (\|x^k - \bar{x}\|)_{k\in\mathbb{N}}$ converges.

We note that having convergence of the sequence as in Proposition 3.1(ii) does not necessarily mean that the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a point in $\mathcal{X}$ (see Examples 2.12). On the other hand, the latter result can be obtained under slightly stronger assumptions (see also Examples 2.4).

Theorem 3.2 (*Theorem 3.8, Combettes, 2001b*). *Let $\mathcal{X}$ be a nonempty set and let $(x^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Suppose that $(x^k)_{k\in\mathbb{N}}$ is quasi-Féjer monotone with respect to $\mathcal{X}$. Then, $(x^k)_{k\in\mathbb{N}}$ converges to a point in $\mathcal{X}$ if and only if every sequential cluster point of $(x^k)_{k\in\mathbb{N}}$ belongs to $\mathcal{X}$.*

Proof. Necessity is straightforward. Sufficiency follows from Remark 3.1 (specifically from Combettes, 2001b, Proposition 3.3). □

Remark 3.2. Since Theorem 3.2 holds for quasi-Féjer monotone sequences, it holds also for Féjer monotone ones (Bauschke et al., 2011, Theorem 5.5). In this case, the proof follows by the fact that for every $x \in \mathcal{X}$ the sequence $(\|x^k - x\|)_{k\in\mathbb{N}}$ converges by Proposition 3.1 and that if every sequential cluster point x belongs to $\mathcal{X}$, then the sequence converges to a point in $\mathcal{X}$ by Lemma 2.3. The result in Theorem 3.2 has been obtained many times in the literature, for weak and strong convergence (Alber et al., 1998; Combettes, 2001a, 2001b; Gubin et al., 1967), but it seems to originate in Schaefer (1957).

Remark 3.3. Under suitable conditions, convergence results as in Proposition 3.1 and Theorem 3.2 can be obtained also for Bregman monotone sequences (Bauschke et al., 2003, Proposition 4.1 and Theorem 4.11).

The following result is known as the Opial Lemma (Opial et al., 1967) and it can be found in many works and with different applications (Abbas & Attouch, 2015; Abbas, Attouch, & Svaiter, 2014; Bauschke et al., 2011; Bot, Sedlmayer et al., 2020; Maingé, 2007; Malitsky, 2020; Malitsky & Tam, 2020), since it often relate to convergence of sequences generated by nonexpansive operators (Naraghirad, Shi, & Wong, 2020; Opial et al., 1967; Peypouquet & Sorin, 2010) (see also Example 2.11). We here show a proof which follows from some results in Bauschke et al. (2011) and we report the discrete time formulation (Attouch & Peypouquet, 2019; Boţ & Csetnek, 2016; Csetnek, Malitsky, & Tam, 2019), but it can be found also in continuous time (Attouch & Peypouquet, 2019; Bot & Csetnek, 2016; Csetnek et al., 2019). For a different proof see Opial et al. (1967) and Peypouquet and Sorin (2010).

Lemma 3.3 (*Opial Lemma*). *Let $(x^k)_{k\in\mathbb{N}}$ be a bounded sequence and let $\mathcal{X} \subseteq \mathbb{R}^n$. If*

1. *for all $z \in \mathcal{X}$ $\lim_{k\to\infty} \|x^k - z\|$ exists;*
2. *every sequential cluster point of $(x^k)_{k\in\mathbb{N}}$ is in $\mathcal{X}$ as $k \to \infty$;*

then, $(x^k)_{k\in\mathbb{N}}$ is convergent to a point in $\mathcal{X}$.

Proof. Since the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded, it has at least one sequential cluster point. We show that, under this assumption, there cannot be two. The proof follows by contradiction. Suppose that $\bar{x}$ and $\bar{y}$ are two sequential cluster points, that is, $x_{k_n} \to \bar{x}$ and $x_{k_l} \to \bar{y}$, for $n, l \in \mathbb{N}$. Since $\bar{x}$ and $\bar{y}$ are sequential cluster points, the sequences $(\|x^k - \bar{x}\|)_{k\in\mathbb{N}}$ and $(\|x^k - \bar{y}\|)_{k\in\mathbb{N}}$ converge. Moreover, it holds that, for all $k \in \mathbb{N}$

$$2\langle x^k, \bar{x} - \bar{y}\rangle = \|x^k - \bar{y}\|^2 - \|x^k - \bar{x}\|^2 + \|\bar{x}\|^2 - \|\bar{y}\|^2.$$

Therefore, $\langle x^k, \bar{x} - \bar{y}\rangle$ converges to some point w. Taking the limit along x_{k_n} and x_{k_l} we have

$$w = \langle \bar{x}, \bar{x} - \bar{y}\rangle = \langle \bar{y}, \bar{x} - \bar{y}\rangle.$$

It follows that $\|\bar{x} - \bar{y}\|^2 = 0$ hence $\bar{x} = \bar{y}$. □

Remark 3.4. The Opial Lemma provides a powerful tool to derive convergence of an iterative process. In fact, condition 2. has been already mentioned in many previous results in this survey. Interestingly, similar results can be extended to the Bregman distance (Huang, Jeng, Kuo, & Hong, 2011; Naraghirad et al., 2020; Naraghirad, Wong, & Yao, 2014).

Table 3
Convergence results for deterministic sequences of real numbers divided by their form. In the first line, the most general inequality is presented. NN stands for a sequence of *nonnegative* real numbers, while ✓(✗) indicates if the inequality in the corresponding lemma contains (or not) a term of that column type. C^k is a general "coefficient", whose specific form can be retrieved from the column.

	Seq($k+1$) v^{k+1}	$\leq$	Coeff. C^k	Seq(k) v^k	Negative $-\theta^k$	Noise $+\varepsilon^k$
Lemma 3.4	NN		γ		✓	✓
Lemma 3.6	NN		$(1+\delta^k)$		✓	✓
Corollary 3.7	NN		1		✓	✗
Corollary 3.8	NN		$(1+\delta^k)$		✗	✓
Lemma 3.9	Real		γ^k		✗	✓
Lemma 3.10	NN		$(1-\delta^k)$		✓	✓
Lemma 3.11	NN		$(1-\delta^k)$		✗	$\delta^k\beta^k$
Lemma 3.12	NN		$(1-\delta^k)$		✗	$\delta^k\beta^k+\varepsilon^k$
Corollary 3.13	NN		$(1-\delta^k)$		✗	$\delta^k(\beta^k+\eta^k)$
Corollary 3.14	NN		$(1-\delta^k)$		✗	✓
Corollary 3.15	NN		$(1-\delta^k)$		✗	$\eta^k+\varepsilon^k$
Proposition 3.16	NN		1		✗	$a\beta^k$
Lemma 3.17	NN		$(1-\delta^k)$		✗	$\delta^k\gamma^k+\beta^k$
			1		✓	✓
Lemma 3.18	Real		$(1+\delta^k)$		$\delta^k v^{k-1}$	✓
Lemma 3.19	NN		$1/\gamma$		β^{k+1}/γ	β^k/γ

3.2. Convergent sequences of real numbers

We now introduce a number of results on sequences of real numbers. We note that even if the following results are for general sequences of real numbers, their importance for system theory lies on the fact that they can be paired with (quasi) Féjer monotonicity (see Remark 3.5). In Table 3, we summarize the results presented in this section, with emphasis on the auxiliary sequences that may affect convergence.

Let us note that, in the first line of Table 3, C^k is a coefficient which, depending on the form, represents the level of expansion or contraction, ε^k can be seen as an additive noise and θ^k is a "negative term", because of the minus sign, which decreases the value of the sequence v^k. For a graphical interpretation of the effects of those sequences, we also refer to Fig. 4 later on, which is specifically related to Lemma 3.6.

The first lemma that we report is widely used and it has a number of consequences that are widely used as well. We do not include the proof since it is very similar to the proof of the forthcoming Lemma 3.10.

Lemma 3.4 (*Lemma 3.1, Combettes, 2001b*)**.** *Let $\gamma \in (0,1]$ and let $(v^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ be nonnegative sequences such that $\sum_{k=0}^{\infty}\varepsilon^k<\infty$ and*

$$v^{k+1} \leq \gamma v^k - \theta^k + \varepsilon^k \text{ for all } k\in\mathbb{N}. \tag{3.1}$$

Then, the following statements hold:

(i) $(v^k)_{k\in\mathbb{N}}$ is bounded;
(ii) $(v^k)_{k\in\mathbb{N}}$ converges;
(iii) $\sum_{k=0}^{\infty}\theta^k<\infty$;
(iv) If $\gamma\neq 1$, then $\sum_{k=0}^{\infty}v^k<\infty$.

Remark 3.5. If $v^k=\|x^k-\bar{x}\|$, for some sequence $(x^k)_{k\in\mathbb{N}}$ and a given $\bar{x}\in\mathcal{X}$, having that $(v^k)_{k\in\mathbb{N}}$ satisfies the inequality (3.1) implies that $(x^k)_{k\in\mathbb{N}}$ is a quasi-Féjer monotone sequence relative to the set $\mathcal{X}$.

We also note that the function $V(x^k)=\|x^k-\bar{x}\|=v^k$ can also be seen as a decreasing Lyapunov function associated to the sequence that tends toward zero when $\varepsilon^k=0$ for all $k\in\mathbb{N}$ (Polyak, 1987, Section 2.2).

Remark 3.6. The case where $(v^k)_{k\in\mathbb{N}}$ can be a negative sequence and with $\gamma=1$ is addressed in Bertsekas and Tsitsiklis (2000, Lemma 1). There, it is also pointed out that if v^k is negative and satisfies (3.1), it may diverge to $-\infty$ when $(\theta^k)_{k\in\mathbb{N}}$ is not summable.

For a specific choice of the noise term instead, the following result can be proven (Kannan & Shanbhag, 2012, Lemma 3.3). Suppose

$$v^{k+1}\leq\gamma v^k+\eta^k\beta, \text{ for all } k\in\mathbb{N}$$

where $\gamma\in(0,1)$, $(\eta^k)_{k\in\mathbb{N}}$ is a decreasing positive sequence such that $\sum_{k=0}^{\infty}(\eta^k)^2<\infty$, and let $0\leq v^k\leq\bar{v}<\infty$ for all $k\in\mathbb{N}$. Then, $\sum_{k=1}^{\infty}\eta^k v^k<\infty$.

With the same arguments as for Lemma 3.4, the following corollary can be proven. Interestingly, this result concerns the finite sum of the sequence.

Corollary 3.5 (*Lemma 9, Scutari & Sun, 2019*)**.** *Let $(v^k)_{k\in\mathbb{N}}$ be a real sequence and let $(\theta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ be nonnegative sequences such that $\sum_{k=0}^{\infty}\varepsilon^k<\infty$ and such that*

$$\sum_{n=0}^{N-1}v^{k+N+n}\leq\sum_{n=0}^{N-1}v^{k+n}-\sum_{n=0}^{N-1}\theta^{k+n}+\sum_{n=0}^{N-1}\varepsilon^{k+n}.$$

for $N\in\mathbb{N}$. Then, either $\sum_{n=0}^{N-1}v^{k+n}\to-\infty$, or $\sum_{n=0}^{N-1}v^{k+n}$ converges to a finite value and $\sum_{k=0}^{\infty}\theta^k<\infty$.

Proof. It suffices to set $v_1^k=\sum_{n=0}^{N-1}v^{k+n}$, $\theta_1^k=\sum_{n=0}^{N-1}\theta^{k+n}$ and $\varepsilon_1^k=\sum_{n=0}^{N-1}\varepsilon^{k+n}$ and then apply Lemma 3.4. □

The next lemma is a consequence and a generalization of Lemma 3.4. It has its stochastic counterpart in the well know Robbins–Siegmund Lemma (Lemma 4.1) (Robbins & Siegmund, 1971). It is taken from Bauschke et al. (2011) yet here we provide a different proof. For a graphical interpretation, we refer to Fig. 4.

Lemma 3.6 (*Lemma 5.31, Bauschke et al., 2011*)**.** *Let $(v^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$, $(\varepsilon^k)_{k\in\mathbb{N}}$ and $(\delta^k)_{k\in\mathbb{N}}$ be nonnegative sequences such that $\sum_{k=0}^{\infty}\varepsilon^k<\infty$ and $\sum_{k=0}^{\infty}\delta^k<\infty$ and*

$$v^{k+1}\leq(1+\delta^k)v^k-\theta^k+\varepsilon^k, \text{ for all } k\in\mathbb{N}. \tag{3.2}$$

Then, $\sum_{k=0}^{\infty}\theta^k<\infty$ and $(v^k)_{k\in\mathbb{N}}$ is bounded and converges to a nonnegative variable.

Proof. Define $\beta^k=\prod_{i=1}^{k}(1+\delta^i)$ and note that β^k converges to some $\bar{\beta}$ since $(\delta^k)_{k\in\mathbb{N}}$ is summable. Moreover, it holds that

$$1+\delta^k=\frac{\beta^k}{\beta^{k-1}}$$

and, for all $k\in\mathbb{N}$

$$v^{k+1}\leq\frac{\beta^k}{\beta^{k-1}}v^k+\varepsilon^k-\theta^k.$$

Since $\beta^k>0$ for all $k\in\mathbb{N}$, we have

$$\frac{v^{k+1}}{\beta^k}\leq\frac{v^k}{\beta^{k-1}}+\frac{\varepsilon^k}{\beta^k}-\frac{\theta^k}{\beta^k}.$$

Now, let

$$\tilde{v}^k=\frac{v^k}{\beta^{k-1}},\quad\tilde{\varepsilon}^k=\frac{\varepsilon^k}{\beta^k},\quad\tilde{\theta}^k=\frac{\theta^k}{\beta^k}$$

and rewrite the inequality as

$$\tilde{v}^{k+1}\leq\tilde{v}^k+\tilde{\varepsilon}^k-\tilde{\theta}^k.$$

Note that $\tilde{v}^k$, $\tilde{\varepsilon}^k$ and $\tilde{\theta}^k$ are nonnegative and $\sum_{k=1}^{\infty}\tilde{\varepsilon}^k\leq\sum_{k=1}^{\infty}\varepsilon^k<\infty$, hence we can apply Lemma 3.4. It follows that $\tilde{v}^k$ is bounded by $\bar{\alpha}$ and convergent to some $\bar{v}$ and that $\sum_{k=1}^{\infty}\tilde{\theta}^k<\infty$. Therefore v^k is convergent, i.e.,

$$v^k=\frac{v^k}{\beta^{k-1}}\beta^{k-1}=v_1^k\beta_{k_1}\to\bar{\alpha}\bar{\beta} \text{ as } k\to\infty,$$

and bounded

$$\frac{v^k}{\beta^{k-1}}<A\Rightarrow v^k<A\beta^{k-1}\to A\beta_\infty \text{ as } k\to\infty.$$

Since $\theta^k=\tilde{\theta}^k\beta^k<\tilde{\theta}^k\beta^\infty$ for all $k\in\mathbb{N}$, we conclude that $(\theta^k)_{k\in\mathbb{N}}$ is summable. □

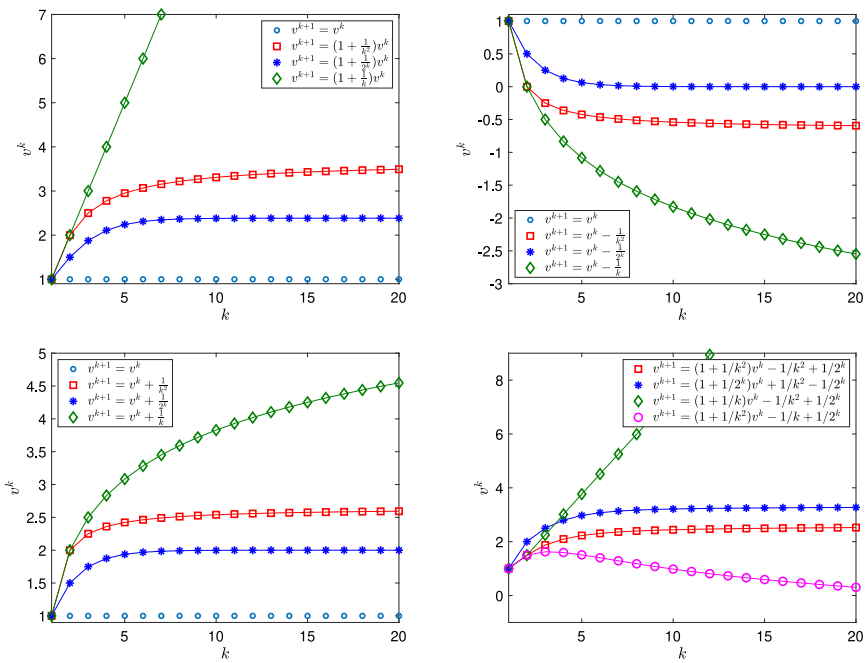

Fig. 4. Influence of the auxiliary sequences of Lemma 3.6 in the behavior of the convergent sequence $(v^k)_{k\in\mathbb{N}}$. If the sequences are non-summable (green lines in the first three plots, green and magenta in the last), then convergence does not necessarily hold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We note that there is a slight difference between Lemmas 3.4 and 3.6. Specifically, in the former, the sequence converges if the coefficient γ is in the interval $(0, 1]$ while in Lemma 3.6 the coefficient can be taken larger than 1 and time varying.

The following results are immediate consequences of Lemmas 3.4 and 3.6. Let us start with removing the noise term and noticing that summability of a sequence implies that its limit goes to zero (Section 2.3).

Corollary 3.7 (*Lemma 2.8, Malitsky, 2015*). *Let $(v^k)_{k\in\mathbb{N}}$ and $(\theta^k)_{k\in\mathbb{N}}$ be nonnegative sequences such that*

$$v^{k+1} \le v^k - \theta^k, \text{ for all } k \in \mathbb{N}.$$

Then, $(v^k)_{k\in\mathbb{N}}$ is bounded and $\lim_{k\to\infty} \theta^k = 0$.

Proof. It follows from Lemmas 3.4 and 3.6 by taking δ^k and ε^k equal to 0. □

This result from Polyak (1987) instead can be obtained as a consequence of Lemma 3.6 by removing the negative term.

Corollary 3.8 (*Lemma 2.2.2, Polyak, 1987*). *Let $(v^k)_{k\in\mathbb{N}}$, $(\varepsilon^k)_{k\in\mathbb{N}}$ and $(\delta^k)_{k\in\mathbb{N}}$ be nonnegative sequences such that*

$$v^{k+1} \le \left(1+\delta^k\right) v^k + \varepsilon^k, \text{ for all } k \in \mathbb{N}$$

and $\sum_{k=0}^{\infty} \delta^k < \infty$, $\sum_{k=0}^{\infty} \varepsilon^k < \infty$. Then v^k converges to some $\bar{v} \ge 0$.

Proof. It follows from Lemma 3.6 by taking $\theta^k = 0$ for all $k \in \mathbb{N}$. See Polyak (1987) for a different proof. □

Concerning the coefficient sequence, other options can be considered. In the next result, the coefficient should be strictly smaller than 1, compared to Lemma 3.6, but need not be constant as in Lemma 3.4.

Lemma 3.9 (*Lemma 2.2.3, Polyak, 1987*). *Let $(v^k)_{k\in\mathbb{N}}$ be a sequence of real numbers such that*

$$v^{k+1} \le \gamma_k v^k + \varepsilon^k, \text{ for all } k \in \mathbb{N},$$

where $(\varepsilon^k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ are nonnegative sequences such that

1. $0 \le \gamma_k < 1$
2. $\sum_{k=0}^{\infty}(1-\gamma_k) = \infty$,
3. $\lim_{k\to\infty} \frac{\varepsilon^k}{1-\gamma_k} = 0$.

Then, $\lim_{k\to\infty} v^k = \bar{v} \le 0$. Moreover, if $v^k > 0$ then $\lim_{k\to\infty} v^k = 0$.

Proof. By definition, given $\varepsilon > 0$ there exists $k_0 \in \mathbb{N}$ such that

$$\frac{\varepsilon^k}{(1-\gamma_k)} \le \varepsilon \quad \text{for all } k \ge k_0.$$

Then

$$\begin{aligned} v^{k+1} &\le \gamma_k v^k + \varepsilon^k \\ &\le \gamma_k v^k + (1-\gamma_k)\varepsilon \\ &\le \gamma_k \gamma_{k-1} v^{k-1} + \left[\gamma_k(1-\gamma_{k-1}) + (1-\gamma_k)\right]\varepsilon \\ &\le \prod_{i=k_0}^{k} \gamma_i v^{k_0} + \varepsilon\left(1 - \prod_{j=k_0}^{k} \gamma_i\right). \end{aligned}$$

Note that $\sum_{k=0}^{\infty}\left(1-\gamma_k\right) = \infty$ implies that $\prod_{k=0}^{\infty} \gamma_k = 0$ therefore taking the lim sup as $k \to \infty$ leads to $\limsup v^k \le \varepsilon$, which proves the claim. □

Remark 3.7. In Kannan and Shanbhag (2012, Lemma 2.1), the result in Lemma 3.9 is proven also for a different condition than 1., i.e.,

1.* there exists $\bar{k} \in \mathbb{N}$ such that $0 < \gamma_k < 1$ for all $k \ge \bar{k}$ and $q_k < \infty$ for all $k \le \bar{k}$.

The proof follows considering the shifted process starting from $\bar{k}$ and using Lemma 3.9 on the resulting sequence.

Many of the previous results have the coefficient $(1+\delta^k)$, therefore, we now consider what happens if we change it to $(1-\delta^k)$ (see also Fig. 5 for a graphical interpretation). This might be a special case of Lemma 3.4 but, in some cases, it allows to study convergence to zero (see Remark 3.8), which relates to the standard Lyapunov based approach for stability analysis. In fact, we have already had a glimpse of the effect of a coefficient smaller than 1 in Lemma 3.4(iv) and Lemma 3.9 and its connection with Lyapunov analysis (Remark 3.5).

The first result of this type extends the previous lemmas to this case. This result is new as we provide a proof that does not follow

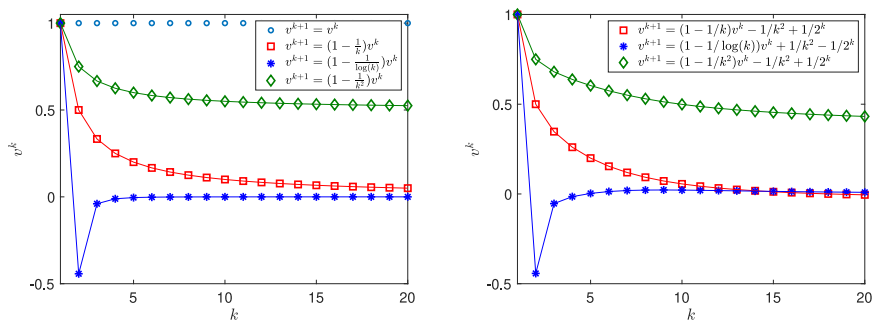

Fig. 5. Influence of the coefficient $(1-\delta^k)$ in the behavior of the convergence of a sequence $v^{k+1} = (1-\delta^k)v^k - \theta^k + \varepsilon^k$. We note that if the sequence δ^k is summable (green lines in the plots), the convergence to zero is not guaranteed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from previous results. Let us note that in the following lemma, we do not assume any condition on the coefficient sequence δ^k besides being smaller that 1.

Lemma 3.10. *Let $(v^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$, $(\varepsilon^k)_{k\in\mathbb{N}}$ and $(\delta^k)_{k\in\mathbb{N}}$ be nonnegative sequences such that $\sum_{k=0}^{\infty}\varepsilon^k < \infty$, $\delta^k \in [0,1)$ for all $k \in \mathbb{N}$ and*

$$v^{k+1} \le (1-\delta^k)v^k + \varepsilon^k - \theta^k \text{ for all } k \in \mathbb{N}.$$

Then, $(v^k)_{k\in\mathbb{N}}$ is bounded and converges to some $\bar{v} \ge 0$ and $\sum_{k=0}^{\infty}\theta^k < \infty$.

Proof. To prove that $(v^k)_{k\in\mathbb{N}}$ is bounded, let $\varepsilon = \sum_{k=0}^{\infty}\varepsilon^k$. Then,

$$0 \le v^{k+1} \le \prod_{i=0}^{k}(1-\delta^i)v^0 + \sum_{i=0}^{k}\prod_{j=0}^{k-i}(1-\delta_j)\varepsilon_i \le v^0 + \varepsilon.$$

Therefore $v^k \in [0, v^0+\varepsilon]$ and the first claim is proven. Now we prove convergence. Let $\bar{v} = \liminf_{k\to\infty} v^k \in [0, v^0+\varepsilon]$. Then there exists a subsequence v^{k_n} such that $\lim_{k\to\infty} v^{k_n} = \bar{v}$. Then, for every $\eta > 0$ there exists $n_0 \in \mathbb{N}$ such that $v^{k_{n_0}} \le \bar{v} + \eta/2$. Since $\sum_{k=0}^{\infty}\varepsilon^k < \infty$, there exists n_1 such that $\sum_{m>n_1}\varepsilon^m \le \eta/2$. Set $n = \max\{n_0, n_1\}$, then, iterating, for every $k \ge k_n$

$$v^k \le v^{k_n} + \sum_{m\ge n}\varepsilon^m \le \frac{\eta}{2} + \bar{v} + \frac{\eta}{2} = \bar{v} + \eta.$$

Hence, $\limsup_{k\to\infty} v^k \le \liminf_{k\to\infty} v^k + \eta$ and, since η can be arbitrarily small, $(v^k)_{k\in\mathbb{N}}$ converges to $\bar{v}$. Lastly, we show that $(\theta^k)_{k\in\mathbb{N}}$ is summable. Since

$$\theta^k \le (1-\delta^k)v^0 - v^{k+1} + \varepsilon^k,$$

we can do a telescopic sum to obtain

$$\sum_{k=1}^{\infty}\theta^k \le (1-\delta_0)v^0 - v^{k+1} + \sum_{k=1}^{\infty}\varepsilon^k \le v^0 + \varepsilon. \quad \square$$

The following lemmas are taken from various works (Qin et al., 2008; Xu, 1998, 2002, 2003) and they are quite similar to each other. We here establish the relations and difference between them. Let us remark that in the following results, the sequence δ^k in the coefficient is not summable, i.e., from now on $\sum_{k=1}^{\infty}\delta^k = \infty$. The advantage of this choice is that convergence to zero can be obtained, as shown in Fig. 5. Hence, the result in Lemma 3.10 can be improved.

The first result considers real (not only positive) noise sequences ε^k and it also provides two alternative conditions on the auxiliary sequences.

Lemma 3.11 (*Lemma 2.1, Xu, 2003*). *Let $(v^k)_{k\in\mathbb{N}}$ be a sequence of nonnegative real numbers such that*

$$v^{k+1} \le \left(1-\delta^k\right)v^k + \delta^k\beta^k, \text{ for all } k \in \mathbb{N}$$

where $(\delta^k)_{k\in\mathbb{N}}$ and $(\beta^k)_{k\in\mathbb{N}}$ are sequences of real numbers such that:

1. *$\delta^k \in [0,1]$ and $\sum_{k=0}^{\infty}\delta_k = \infty$, or equivalently, $\prod_{k=0}^{\infty}\left(1-\delta^k\right) = 0$,*

2a. *$\limsup_{k\to\infty}\beta^k \le 0$,*

2b. *$\sum_{k=0}^{\infty}\delta^k\beta^k < \infty$.*

Then, $\lim_{k\to\infty} v^k = 0$.

Proof. If 1. and 2*a*. hold, then the result can be proven with the same arguments as the proof of Lemma 3.9 by setting $\varepsilon^k = \delta^k\beta^k$. On the other hand, if 1. and 2*b*. hold, we have for all $k > m$

$$\begin{aligned} v^{k+1} &\le (1-\delta^k)v^k + \delta^k\beta^k \\ &\le \prod_{i=m}^{k}(1-\delta^i)v^m + \sum_{i=m}^{k}\delta^i\beta^i. \end{aligned}$$

Taking the limit for $k \to \infty$ and $m \to \infty$ we have $\limsup v^k \le 0$. $\square$

We note that if we set $\gamma_k = 1-\delta^k$ and $\varepsilon^k = \delta^k\beta^k$, we obtain the same statement as Lemma 3.9. Moreover, condition 2*b*. provides an alternative assumption, similar to most results in the literature.

Remark 3.8. Convergence to zero is of particular interest in combination with a Féjer-like property. Specifically, if $v^k = \|x^k - \bar{x}\|$ for some sequence $(x^k)_{k\in\mathbb{N}}$ and $\bar{x} \in \mathcal{X}$, Lemma 3.11 states that $\lim_{k\in\mathbb{N}}\|x^k - \bar{x}\| = 0$, i.e., $\lim_{k\to\infty} x^k = \bar{x}$. We note, however, that the sequence is not quasi-Féjer monotone because of the term δ^k which cannot be 0 (contrary to Remark 3.5). We refer to Section 6 for more details.

Assumption 2*a*. in Lemma 3.11 is used in a previous paper by the same authors (Xu, 2002, Lemma 2.5). Since also Assumption 2*b*. in Lemma 3.11 can be used to prove the following result, let us extend (Xu, 2002, Lemma 2.5) next.

Lemma 3.12 (*Extension of Lemma 2.5, Xu, 2002*). *Let $(v^k)_{k\in\mathbb{N}}$ be a sequence of nonnegative real numbers satisfying*

$$v^{k+1} \le \left(1-\delta^k\right)v^k + \delta^k\beta^k + \varepsilon^k, \text{ for all } k \in \mathbb{N}$$

where $(\delta^k)_{k\in\mathbb{N}}$, $(\beta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ satisfy the following conditions:

1. *$\delta^k \in [0,1]$, $\sum_{k=0}^{\infty}\delta^k = \infty$, or equivalently, $\prod_{k=1}^{\infty}(1-\delta^k) = 0$,*

2a. *$\limsup_{k\to\infty}\beta^k \le 0$,*

2b. *$\sum_{k=0}^{\infty}\delta^k\beta^k < \infty$,*

3. *$\varepsilon^k \ge 0$ and $\sum_{k=0}^{\infty}\varepsilon^k < \infty$.*

Then, $\lim_{k\to\infty} v^k = 0$.

Proof. The proof is similar to the proof of Lemma 3.11. Since ε^k is summable, $\sum_{k=k_0}^{\infty}\varepsilon^k < \varepsilon$ for some $k_0 \in \mathbb{N}$ and $\varepsilon > 0$ arbitrarily small, and

$$\begin{aligned} v^{k+1} &\le \prod_{i=k_0}^{k}(1-\delta^i)v^{k_0} + \varepsilon\left(1 - \prod_{i=k_0}^{k}(1-\delta^i)\right) + \sum_{i=k_0}^{k}\varepsilon^i \quad \square \\ &\le 2\varepsilon. \end{aligned}$$

A particular case of Lemma 3.12 is proposed in Lei, Shanbhag and Chen (2020) as a consequence of Chen (2006, Theorem 3.3.1). Let us note that the assumptions in the following result imply those in Lemma 3.12 which is, in turn, more general.

Corollary 3.13 (*Proposition 3, Lei, Shanbhag and Chen, 2020*). *Let* $(v^k)_{k\in\mathbb{N}}$ *be a nonnegative sequence such that*

$$v^{k+1} \leq (1-\delta_k)v^k + \delta^k(\beta^k + \eta^k)$$

where $(\delta^k)_{k\in\mathbb{N}}$, $(\beta^k)_{k\in\mathbb{N}}$, $(\eta^k)_{k\in\mathbb{N}}$ *are nonnegative sequences such that*

1. $\sum_{k=1}^{\infty} \delta^k = \infty$ *and* $\lim_{k\to\infty} \delta^k = 0$
2. $\lim_{k\to\infty} \beta^k = 0$
3. $\sum_{k=1}^{\infty} \delta^k \eta^k < \infty$

Then, $\lim_{k\to\infty} v^k = 0$.

Proof. The sequence satisfies the assumptions of Lemma 3.12, hence the result holds. □

A very recent result of this type is a consequence of both Lemmas 3.9 and 3.11.

Corollary 3.14 (*Lemma 1.1, Qin et al., 2008*). *Assume that* $(v^k)_{k\in\mathbb{N}}$ *is a sequence of nonnegative real numbers such that*

$$v^{k+1} \leq \left(1-\delta^k\right) v^k + \varepsilon^k, \text{ for all } k \in \mathbb{N}$$

where $(\delta^k)_{k\in\mathbb{N}}$ *and* $(\varepsilon^k)_{k\in\mathbb{N}}$ *are sequences such that*

1. $\delta^k \in (0,1)$ *and* $\sum_{k=1}^{\infty} \delta^k = \infty$,
2a. $\limsup_{k\to\infty} \frac{\varepsilon^k}{\delta^k} \leq 0$,
2b. $\sum_{k=1}^{\infty} |\varepsilon^k| < \infty$.

Then, $\lim_{k\to\infty} v^k = 0$.

Proof. Suppose that 2*a*. holds. Then, the proof follows from Lemma 3.9 by setting $\gamma_k = (1-\delta^k)$. When 2*b*. holds instead, the proof follows applying Lemma 3.11 by defining $\varepsilon^k = \delta^k \frac{\varepsilon^k}{\delta^k} = \delta^k \beta^k$. □

A consequence of Corollary 3.14 is the following result that presents a slightly different notation.

Corollary 3.15 (*Lemma 3, Xu, 1998*). *Let* $(v^k)_{k\in\mathbb{N}}$, $(\eta^k)_{k\in\mathbb{N}}$, $(\varepsilon_k)_{k\in\mathbb{N}}$ *and* $(\delta^k)_{k\in\mathbb{N}}$ *be nonnegative real sequences such that*

$$v^{k+1} \leq (1-\delta^k)v^k + \eta^k + \varepsilon^k \text{ for all } k \in \mathbb{N}$$

and such that

1. $\delta^k \in [0,1]$ *and* $\sum_{k=0}^{\infty} \delta^k = \infty$,
2. $\varepsilon^k = o(\delta^k)$,
3. $\sum_{k=0}^{\infty} \eta^k < \infty$.

Then, $\lim_{k\to\infty} v^k = 0$.

Proof. We note that $\varepsilon^k = o(\delta^k)$ is equivalent to $\lim_{k\to\infty} \varepsilon^k/\delta^k = 0$. Then, the result follows by applying Corollary 3.14 and Lemma 3.12. □

Remark 3.9. A result similar to the last corollaries where the boundedness of the sequence is shown, is presented also in Cholamjiak, Cholamjiak, and Suantai (2018, Lemma 2.5) and reads as follows.

Let $(v^k)_{k\in\mathbb{N}}$ and $(\eta^k)_{k\in\mathbb{N}}$ be sequences of nonnegative real numbers such that $\sum_{k=1}^{\infty} \eta^k < \infty$ and such that

$$v^{k+1} \leq (1-\delta^k)v^k + \eta^k + \varepsilon^k, \text{ for all } k \in \mathbb{N}$$

where $(\delta^k)_{k\in\mathbb{N}} \subseteq (0,1)$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ is a sequence of real numbers. Then, the following results hold:

(i) If $\varepsilon^k \leq \delta^k M$ for some $M \geq 0$, then (v^k) is a bounded sequence.

(ii) If $\sum_{k=1}^{\infty} \delta^k = \infty$ and $\limsup_{n\to\infty} \frac{\varepsilon^k}{\delta^k} \leq 0$, then $\lim_{n\to\infty} v^k = 0$.

We note that (ii) is a consequence of Lemma 3.12 or Corollaries 3.14 and 3.15.

We now consider three results whose conditions for convergence are more involved than the results proposed until now (Alber et al., 1998; Cholamjiak et al., 2018; He & Yang, 2013; Maingé, 2008). The first one is proposed in Alber et al. (1998). It allows for non-summable additive noise but requires a condition that couple the sequences involved.

Proposition 3.16 (*Proposition 2, Alber et al., 1998*). *Let* $(v^k)_{k\in\mathbb{N}}$ *and* $(\beta^k)_{k\in\mathbb{N}}$ *be two nonnegative sequences such that* $\sum_{k=0}^{\infty} \beta_k = \infty$ *and* $\sum_{k=0}^{\infty} \beta_k v^k < \infty$. *Then:*

(i) there exists a subsequence (v^{k_n}), $n \in \mathbb{N}$ *such that* $\lim_{k\to\infty} v^{k_n} = 0$.
(ii) Moreover, if there exists $a > 0$ *such that*

$$v^{k+1} \leq v^k + a\beta^k, \text{ for all } k \in \mathbb{N}$$

then $\lim_{k\to\infty} v^k = 0$.

Proof. Both claims can be proven by contradiction. See Alber et al. (1998) for more details. □

In the next result, the sequence should satisfy two interdependent inequalities (Cholamjiak et al., 2018; He & Yang, 2013).

Lemma 3.17 (*Lemma 7, He & Yang, 2013*). *Let* $(v^k)_{k\in\mathbb{N}}$ *and* $(\eta^k)_{k\in\mathbb{N}}$ *be nonnegative sequences of real numbers, let* $(\delta^k)_{k\in\mathbb{N}} \subseteq (0,1)$ *and let* $(\gamma^k)_{k\in\mathbb{N}}$, $(\varepsilon^k)_{k\in\mathbb{N}}$, *and* $(\beta^k)_{k\in\mathbb{N}}$ *be three sequences of real numbers such that*

$$\begin{aligned} v^{k+1} &\leq (1-\delta^k)v^k + \delta^k\gamma^k + \beta^k, \text{ and,} \\ v^{k+1} &\leq v^k - \eta^k + \varepsilon^k, \text{ for all } k \in \mathbb{N} \end{aligned} \tag{3.3}$$

and such that

1. $\sum_{k=0}^{\infty} \delta^k = \infty$
2. $\lim_{k\to\infty} \varepsilon^k = 0$
3. $\lim_{k\to\infty} \eta^{k_n} = 0$ *implies that* $\limsup_{k\to\infty} \gamma^{k_n} \leq 0$ *for any subsequence* $(k_n) \subset (k)$
4. $\limsup_{n\to\infty} \frac{\beta^k}{\delta^k} \leq 0$.

Then, $\lim_{n\to\infty} v^k = 0$.

Proof. The proof is divided in two cases.

Case 1: $(v^k)_{k\in\mathbb{N}}$ is eventually decreasing and the result follows from Lemma 3.12.

Case 2: $(v^k)_{k\in\mathbb{N}}$ is not eventually decreasing. Then, there exists k_0 such that $v^{k_0} \leq v^{k_0+1}$. Let $J_k = \{k_0 \leq n \leq k : v_n \leq v_{n+1}\}$, $k > k_0$ and let $\tau(k) = \max J_k$. Then, $\tau(k) \to \infty$ as $k \to \infty$ by definition. It follows that $v^k \leq v^{\tau(k)+1}$. Then, using (3.3) and the assumptions it follows that

$$v^{\tau(k)} \leq \gamma^{\tau(k)} + \frac{\beta^{\tau(k)}}{\delta^{\tau(k)}},$$

and $\limsup_{k\to\infty} v^{\tau(k)} \leq 0$. Hence, $\lim_{k\to\infty} v^k = 0$. For more details we refer to He and Yang (2013). □

Remark 3.10. By removing β^k in Eq. (3.3), the result holds as a particular case of Lemma 3.17 and can be proven similarly, using Lemma 3.11 instead of Lemma 3.12.

The next result, instead, uses the sequence at two steps backwards.

Lemma 3.18 (*Lemma 2.2, Maingé, 2008*). *Let* $(v^k)_{k\in\mathbb{N}}$ *and* $(\varepsilon^k)_{k\in\mathbb{N}}$ *be nonnegative sequences such that:*

$$v^{k+1} - v^k \leqslant \delta^k(v^k - v^{k-1}) + \varepsilon^k, \text{ for all } k \in \mathbb{N} \tag{3.4}$$

and such that

1. $\sum_{k=1}^{\infty} \varepsilon^k < \infty$

Table 4
Convergence results for stochastic sequences of real numbers divided by their form. In the first line, the most general inequality is presented. NN stands for a sequence of *nonnegative* real numbers, while ✓(✗) indicates if the inequality in the corresponding lemma contains (or not) a term of that column type. C^k is a general "coefficient", whose specific form can be retrieved from the column.

	Seq$(k+1)$ $\mathbb{E}[v^{k+1}]$	$\leq$	Coeff. Seq(k) $C^k v^k$	Negative $-\theta^k$	Noise $+\varepsilon^k$
Lemma 4.1	NN		$(1+\delta^k)$	✓	✓
Corollary 4.2	NN		$(1+\delta^k)$	✗	✓
Corollary 4.3	NN		1	✓	✗
Corollary 4.4	NN		1	✓	✓
Lemma 4.6	NN		γ^k	✓	✓
Lemma 4.7	NN		$(1-\delta^k)$	✗	✓

2. $(\delta^k) \subset [0,\delta]$, *where* $\delta \in [0,1)$.

Then $(v^k)_{k\in\mathbb{N}}$ *converges and* $\sum_{k=1}^{\infty}[v^{k+1}-v^k]_+ < \infty$, *where* $[t]_+ := \max\{t,0\}$ *(for any* $t \in \mathbb{R}$*).*

Proof. Let $u^k = v^k - v^{k-1}$. Then, $[u^{k+1}]_+ \leq \delta[u^k]_+ + \varepsilon^k$ and $([u^k]_+)_{k\in\mathbb{N}}$ is bounded. It follows that the sequence $(w^k = v^k - \sum_{j=1}^{k}[u_j]_+)_{k\in\mathbb{N}}$ is bounded and non increasing, hence, convergent. Hence, $(v^k)_{k\in\mathbb{N}}$ is convergent. □

Remark 3.11. The result can be extended to the case with a negative term, namely, Eq. (3.4) becomes

$$v^{k+1} - v^k \leqslant \delta^k(v^k - v^{k-1}) + \varepsilon^k - \theta^k, \text{ for all } k \in \mathbb{N}$$

where θ^k is a nonnegative sequence. The conclusions are the same as Lemma 3.18 and, moreover, it holds that $\sum_{k=1}^{\infty}\theta^k < \infty$ (Boţ & Csetnek, 2016, Lemma 2).

Remark 3.12. The inequality in Eq. (3.4) can be rewritten as

$$v^{k+1} \leq (1+\delta^k)v^k - \delta^k v^{k-1} + \varepsilon^k, \text{ for all } k \in \mathbb{N}$$

which is similar to the form of the results presented until now. However, we note that in Lemma 3.18, the sequence v^k need not be nonnegative.

We conclude this section with the following result on the convergence rate which guarantees convergence to zero. However, the study of the convergence rates lays outside the scopes of this survey. For similar results, we refer to Lei and Shanbhag (2018, Lemma 3), Malitsky (2015, Lemma 2.9) and, more generally, to Polyak (1987).

Lemma 3.19 (*Lemma 2.7, Malitsky & Tam, 2018*). *Let* $(v^k)_{k\in\mathbb{N}}$ *and* $(\beta^k)_{k\in\mathbb{N}}$ *be two nonnegative sequences of real numbers. Suppose there exist constants* $\gamma > 1$ *and* $\delta > 0$ *such that*

$$\gamma v^{k+1} + \beta^{k+1} \leq v^k + \beta^k \quad \text{and} \quad \delta\beta^k \leq v^k \text{ for all } k \in \mathbb{N}$$

Then $(v^k)_{k\in\mathbb{N}}$ *and* $(\beta^k)_{k\in\mathbb{N}}$ *converge to zero with R-linear rate.*

4. Convergence of stochastic sequences

In this section, we report the convergence results available for sequences of random variables, summarized in Table 4. We recall that the probability space is $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, $\mathcal{F}$ is the event space, and $\mathbb{P}$ is the probability function defined on the event space. The symbol $\mathbb{E}$ indicates the associated expected values. We also recall that $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ is a filtration.

4.1. Convergent sequences of random variables

Firstly, we recall some results on convergent random sequences. We start with a result by Robbins and Siegmund, first appeared in Robbins and Siegmund (1971), which is the most used in the stochastic literature. In Fig. 6, we provide a graphical interpretation.

Lemma 4.1 (*Robbins–Siegmund Lemma*). *Let* $(v^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$, $(\varepsilon^k)_{k\in\mathbb{N}}$ *and* $(\delta^k)_{k\in\mathbb{N}}$ *be nonnegative sequences such that* $\sum_{k=0}^{\infty}\varepsilon^k < \infty$, $\sum_{k=0}^{\infty}\delta^k < \infty$ *and*

$$\mathbb{E}[v^{k+1}|\mathcal{F}_k] \leq (1+\delta^k)v^k + \varepsilon^k - \theta^k \text{ a.s., for all } k \in \mathbb{N} \tag{4.1}$$

Then, $\sum_{k=0}^{\infty}\theta^k < \infty$ *and* $(v^k)_{k\in\mathbb{N}}$ *converges a.s. to a nonnegative random variable.*

Proof. The proof follows by rewriting the sequence as in Lemma 3.6. Then, it is possible to show that the sequence

$$y^n = \tilde{v}^k - \sum_{k=0}^{n-1}(\tilde{\varepsilon}^k - \tilde{\theta}^k)$$

is a supermartingale. The claim then follows by the Martingale Convergence Theorem (Theorem 2.4). See Robbins and Siegmund (1971) for technical details. □

Remark 4.1. Besides the convergence of the sequence $(v^k)_{k\in\mathbb{N}}$, it is of particular interest also the fact that the sequence $(\theta^k)_{k\in\mathbb{N}}$ is summable. Specifically, this result can be used to obtain more information once related with a (quasi-)Féjer property. In the stochastic case, this term is particularly useful, compared to the deterministic case, because there are not as many results on stochastic Féjer monotone sequences and the techniques available for the deterministic case cannot be used here. We refer to the application sections to see how the negative term is exploited.

The following results are consequences of Robbins–Siegmund Lemma. The first one is attributed to Gladyshev (Ljung, Pflug, & Walk, 2012; Polyak, 1987). In fact, it came implicitly in a work by Gladyshev (Gladyshev, 1965) in which the author provides a proof of the convergence of Robbins–Monro algorithm (Ljung et al., 2012; Robbins & Monro, 1951). Even if it was published prior than the result by Robbins–Siegmund, it is a particular case of Lemma 4.1 (Robbins & Siegmund, 1971, Application 2).

Corollary 4.2 (*Gladyshev Gladyshev, 1965, Lemma 2.2.9, Polyak, 1987*). *Let* $(v^k)_{k\in\mathbb{N}}$ *be a nonnegative sequence of random variables. Let* $\mathbb{E}[v^0] < \infty$, *and let* $(\delta^k)_{k\in\mathbb{N}}$ *and* $(\varepsilon^k)_{k\in\mathbb{N}}$ *be such that* $\sum_{k=0}^{\infty}\delta^k < \infty$ *and* $\sum_{k=0}^{\infty}\varepsilon^k < \infty$

$$\mathbb{E}[v^{k+1}|\mathcal{F}_k] \leq (1+\delta^k)v^k + \varepsilon^k \text{ for all } k \in \mathbb{N} \tag{4.2}$$

Then $\lim_{k\to\infty} v^k = \bar{v} \geq 0$ *a.s. where* $\bar{v}$ *is a random variable.*

Proof. It follows from Lemma 4.1 letting $\theta_k = 0$. Different proofs can be found in Borkar (1995, Theorem 3.3.6), Polyak (1987, Lemma 2.2.9) or Robbins and Siegmund (1971, Application 2). □

Remark 4.2. In Barty et al. (2007, Lemma 2.2), it is shown that if Eq. (4.2) holds, then the sequence $(v^k)_{k\in\mathbb{N}}$ is bounded.

Similarly to Lemmas 3.4 and 3.6 in the deterministic case, many results can be obtained removing or changing the sequences in (4.1). In fact, the next corollary is straightforward from Lemma 4.1.

Corollary 4.3 (*Theorem B.2, Poggio et al., 2011*). *Let* $(v^k)_{k\in\mathbb{N}}$ *and* $(\theta^k)_{k\in\mathbb{N}}$ *be positive sequences adapted to* $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ *and let*

$$\mathbb{E}[v^{k+1}|\mathcal{F}_k] \leq v^k - \theta^k \text{ for all } k \in \mathbb{N}.$$

Then, $(v^k)_{k\in\mathbb{N}}$ *converges a.s. to a finite random variable* $\bar{v}$ *and* $\sum_{k=0}^{\infty}\theta^k < \infty$.

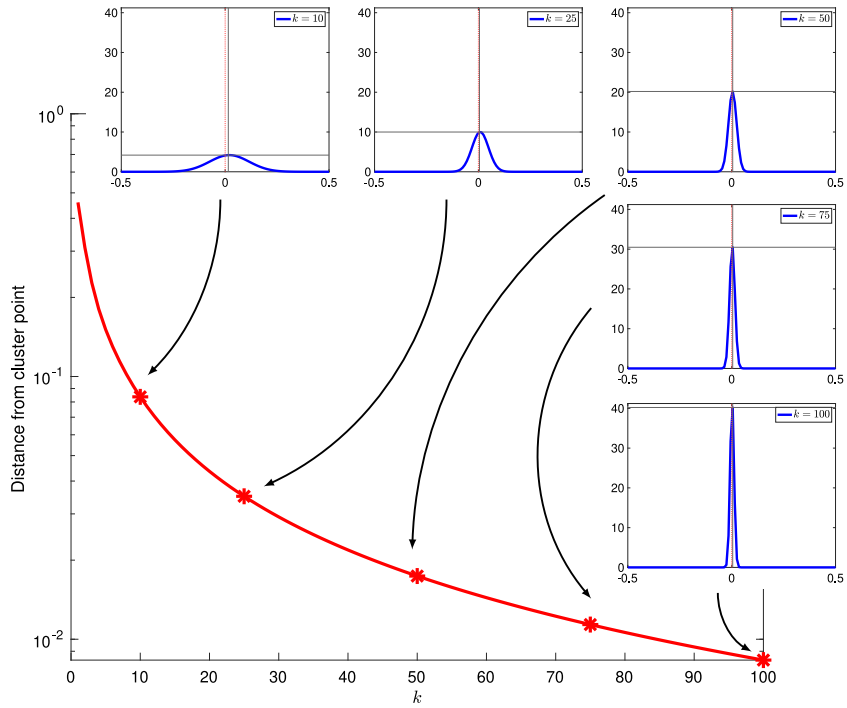

Fig. 6. The big plot shows how the average distance from a cluster point of the sequence $\mathbb{E}[v^{k+1}|\mathcal{F}_k] = (1+\delta^k)v^k + \varepsilon^k - \theta^k$, generated by Robbins–Siegmund Lemma (Lemma 4.1), goes to zero as the number of iterations increases. The small plots show how the distribution of the distance from a cluster point varies with the iteration toward a probability distribution centered at 0, i.e., the sequence converges a.s.

Proof. It follows from Robbins–Siegmund Lemma by taking $\varepsilon^k = 0$. For a different proof, see Poggio et al. (2011). □

Interestingly, we note that besides convergence of the sequence, there is additional information to be derived from Robbins–Siegmund Lemma. For instance, the next corollary is used in Duflo (2013) to prove the Law of Large Numbers for martingales (Duflo, 2013, Theorem 1.3.15) (see also Section 7).

Corollary 4.4 (*Corollary 1.3.13, Duflo, 2013*). *Let $(v^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ be positive sequences adapted to $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ and let $(a_k)_{k\in\mathbb{N}}$ be a strictly positive, increasing sequence adapted to $\mathcal{F}$ such that*

$$\mathbb{E}[v^{k+1}|\mathcal{F}_k] \le v^k + \varepsilon^k - \theta^k \text{ for all } k \in \mathbb{N}.$$

Then, if $\sum_{k=1}^{\infty} a_k^{-1}\varepsilon^k < \infty$ the following hold a.s.:

(i) $\sum_{k=1}^{\infty} a_k^{-1}(v^{k+1} - v^k)$ converges and $\sum_{k=1}^{\infty} a_k\theta^k < \infty$;
(ii) $(v^k)_{k\in\mathbb{N}}$ converges if $(a_k)_{k\in\mathbb{N}}$ is convergent;
(iii) $\lim_{k\to\infty} a_k^{-1}v^k = 0$ and $\lim_{k\to\infty} a_k^{-1}v^{k+1} = 0$ if $(a_k)_{k\in\mathbb{N}}$ is divergent.

Proof. Let

$$u^k = \sum_{i=1}^{k} v^i\left(a_{i-1}^{-1} - a_i^{-1}\right) + v^k a_k^{-1}.$$

Then we can apply Robbins–Siegmund Lemma to the inequality

$$\mathbb{E}\left[u_{k+1}|\mathcal{F}_k\right] \le u^k + a_k^{-1}(\varepsilon^k - \theta^k)$$

and conclude the proof. For technical details, we refer to Duflo (2013). □

The following proposition explicitly connects stochastic quasi-Féjer monotone sequences to Robbins–Siegmund Lemma.

Proposition 4.5 (*Proposition 2.3, Combettes & Pesquet, 2015*). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be nonempty and closed, let $\phi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a strictly increasing function such that $\lim_{t\to\infty}\phi(t) = \infty$, and let $(x^k)_{k\in\mathbb{N}}$ be a sequence of random variables. Let $C(x^k)$ be the set of sequential cluster points of $(x^k)_{k\in\mathbb{N}}$. Suppose that, for every $\bar{x} \in \mathcal{X}$, there exist $(\delta^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$, and $(\varepsilon^k)_{k\in\mathbb{N}}$ positive sequences such that $\sum_{k=0}^{\infty}\delta^k < \infty$, $\sum_{k=0}^{\infty}\varepsilon^k < \infty$ and*

$$\mathbb{E}(\phi(\|x^{k+1} - \bar{x}\|)|\mathcal{F}_n) + \theta^k(\bar{x})$$
$$\le (1+\delta^k(\bar{x}))\phi(\|x^k - \bar{x}\|) + \varepsilon^k(\bar{x}) \text{ a.s. for all } k \in \mathbb{N}$$

Then the following hold:

(i) $\sum_{k=1}^{\infty}\theta^k < \infty$ a.s.;
(ii) $(x^k)_{k\in\mathbb{N}}$ is bounded a.s.;
(iii) $(\|x^k - \bar{x}\|)_{k\in\mathbb{N}}$ converges a.s.;
(iv) Let $C(x^k) \subset \mathcal{X}$ a.s., then $(x^k)_{k\in\mathbb{N}}$ converges a.s.

Proof. (i) It follows from Lemma 4.1.

(ii) Let $v^k = \|x^k - \bar{x}\|$. Then, $\lim_{k\to\infty}\phi(v^k) = \bar{v} \in \mathbb{R}_{\ge 0}$ by Lemma 4.1. Since $\lim_{t\to\infty}\phi(t) = \infty$, v^k is bounded and, therefore, also x^k is bounded.

(iii) It follows from (ii), for more details see Combettes and Pesquet (2015, Proposition 2.3).

(iv) Let $\bar{x}, \bar{y} \in C(x^k)$. Then, there exist two subsequences (x_{k_n}) and (x_{k_m}) such that $x_{k_n} \to \bar{x}$ and $x_{k_m} \to \bar{y}$ as $k \to \infty$. By (iii) the sequences $(\|x^k - \bar{x}\|)_{k\in\mathbb{N}}$ and $(\|x^k - \bar{y}\|)_{k\in\mathbb{N}}$ converge and it holds that $\langle x^k, \bar{x} - \bar{y}\rangle \to \rho$ for some $\rho \in \mathbb{R}^n$. Then, $\langle \bar{x}, \bar{x} - \bar{y}\rangle = \rho$, $\langle \bar{y}, \bar{x} - \bar{y}\rangle = \rho$ and

$$0 = \langle \bar{x}, \bar{x} - \bar{y}\rangle - \langle \bar{y}, \bar{x} - \bar{y}\rangle = \|\bar{x} - \bar{y}\|^2.$$

Therefore, $\bar{x} = \bar{y}$ and $\lim_{k\to\infty} x^k = \bar{x}$. □

Remark 4.3. A specific case of Proposition 4.5 was also presented in Barty et al. (2007, Lemma 2.3) without the negative term θ^k, using $\phi = |\cdot|^2$ and setting $\delta^k = 0$.

More generally, an analogous result holds with $\phi = |\cdot|^p$, $p > 0$ (Combettes & Pesquet, 2019, Lemma 2.2).

Analogously to the deterministic case, also for sequences of random variables, we can find results for sequences with a coefficient strictly smaller than 1. This is the case of the following results.

Lemma 4.6 (*Lemma 2.1, Combettes & Pesquet, 2019*). *Let $(v^k)_{k\in\mathbb{N}}$, $(\theta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ be sequences of nonnegative random variables and suppose that there exists a nonnegative sequence $(\gamma^k)_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\gamma^k < 1$ and*

$$\mathbb{E}[v^{k+1} \mid \mathcal{F}_k] \le \gamma^k v^k + \varepsilon^k - \theta^k \text{ for all } k \in \mathbb{N}.$$

Moreover, let $\mathbb{E}[v^0] < \infty$ and $\sum_{k=1}^{\infty}\mathbb{E}[\varepsilon^k] < \infty$. Then $\sum_{k=1}^{\infty}\mathbb{E}[v^k] < \infty$ and $\sum_{k=1}^{\infty}\mathbb{E}[\theta^k] < \infty$.

Proof. The proof follows with arguments similar to Lemmas 3.12 and 3.10 but it can be proven also as a consequence of Lemma 4.1. For technical details we refer to Combettes and Pesquet (2019). □

We conclude this section with a lemma that is quite popular in the literature (Koshal et al., 2013; Polyak, 1987; Yousefian et al., 2017) and cited along with Robbins–Siegmund Lemma. It is the stochastic counterpart of Lemma 3.9 even if it has a slightly different notation.

Lemma 4.7 (*Lemma 2.2.10, Polyak, 1987*). *Let $(v^k)_{k\in\mathbb{N}}$ be a sequence of nonnegative random variables such that $\mathbb{E}[v^0] < \infty$ and let $(\delta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ be deterministic nonnegative sequences such that $0 \le \delta^k \le 1$ for all $k \in \mathbb{N}$, $\sum_{k=0}^{\infty}\delta^k = \infty$, $\sum_{k=0}^{\infty}\varepsilon^k < \infty$, $\lim_{k\to\infty}\frac{\varepsilon^k}{\delta^k} = 0$ and*

$$\mathbb{E}[v^{k+1}|\mathcal{F}_k] \le (1-\delta^k)v^k + \varepsilon^k \text{ a.s., for all } k \in \mathbb{N}.$$

Then, $\lim_{k\to\infty} v^k = 0$ a.s.

Proof. The proof follows by applying Lemma 3.9 to

$$\mathbb{E}[v^{k+1}] \le (1-\delta^k)\mathbb{E}[v^k] + \varepsilon^k \tag{4.3}$$

and showing that

$$u^k = v^k - \sum_{i=k}^{\infty}\varepsilon_i$$

is a supermartingale. Then, the claim follows by the Martingale Convergence Theorem (Theorem 2.4). See Polyak (1987) for technical details. □

Remark 4.4. To retrieve the same form of Lemma 3.9, one can refer to Lei, Shanbhag, Pang and Sen (2020, Lemma 1(a)) where the following statement is provided.

Let $(v_k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ be sequences of nonnegative random variables such that

$$\mathbb{E}[v^{k+1} \mid \mathcal{F}_k] \le \gamma^k v^k + \varepsilon^k \text{ for all } k \in \mathbb{N},$$

$\mathbb{E}[v_0] < \infty$, $0 \le \gamma^k < 1$, $\sum_{k=0}^{\infty}(1-\gamma^k) = \infty$, $\sum_{k=0}^{\infty}\varepsilon^k < \infty$, and $\lim_{k\to\infty}\frac{\varepsilon^k}{1-\gamma^k} = 0$. Then, $\lim_{k\to\infty} v_k = 0$ a.s.

Remark 4.5. Convergence to zero in Lemma 4.7 can also be derived from Lemma 4.1 by exploiting the properties of the negative term in Eq. (4.3). In fact, from Lemma 4.1 and Eq. (4.3) we have that $\sum_{k=1}^{\infty}\delta^k v^k < \infty$ and since δ^k is not summable, it must be $\lim_{k\to\infty} v^k = 0$.

5. Convergence with variable metric

Let us consider in this section the more general setting with variable metric, i.e., cases in which the metric is allowed to change at each iteration. Applications of these results involve theoretical problems as monotone inclusions (Combettes & Vũ, 2014; Vũ, 2013), as well as inverse problems (Combettes & Vũ, 2013), convex feasibility problems (Combettes & Vũ, 2013; Van Nguyen, 2016) and constrained convex minimization (Cui, Tang, & Zhu, 2019). All the results in this section concern Féjer properties and we consider mostly the deterministic case. The first result that we propose is an extension of Proposition 3.1.

Proposition 5.1 (*Proposition 3.2, Combettes & Vũ, 2013*). *Let $\beta > 0$ and let $(W_k)_{k\in\mathbb{N}}$ be in $\mathcal{P}_\beta$. Let $\phi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be strictly increasing and such that $\lim_t \phi(t) = \infty$. Let $\mathcal{S} \subseteq \mathbb{R}^n$ be nonempty, and $(x^k)_{k\in\mathbb{N}}$ be a quasi-Féjer monotone sequence in $\mathbb{R}^n$ with respect to $\mathcal{S}$ relative to $(W_k)_{k\in\mathbb{N}}$. Then the following hold:*

(i) $(x^k)_{k\in\mathbb{N}}$ is bounded;
(ii) Let $\bar{x} \in \mathcal{S}$. Then $(\|x^k - \bar{x}\|_{W_k})_{k\in\mathbb{N}}$ converges.

Proof. (i) follows by the fact that $(W_k)_{k\in\mathbb{N}}$ is in $\mathcal{P}_\beta$. (ii) follows from Corollary 3.8 and by showing that there cannot be two cluster points. □

Remark 5.1. A similar result holds also for quasi-Bregman monotone sequences and it can be proven analogously by applying Corollary 3.8 (Van Nguyen, 2016, 2017).

Analogously to Section 3, under stronger assumptions, we can obtain stronger convergence results. In fact, the next result is a generalization of Theorem 3.2.

Theorem 5.2 (*Theorem 3.3, Combettes & Vũ, 2013*). *Let $\beta > 0$ and let $(W_k)_{k\in\mathbb{N}}$ and W be operators in $\mathcal{P}_\beta$ such that $W_k \to W$ pointwise. Let $\phi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be strictly increasing and such that $\lim_t \phi(t) = \infty$. Let $\mathcal{S} \subseteq \mathbb{R}^n$ be nonempty and let $(x^k)_{k\in\mathbb{N}}$ be a quasi-Féjer monotone sequence in $\mathbb{R}^n$ with respect to $\mathcal{S}$ and relative to $(W_k)_{k\in\mathbb{N}}$. Then, $(x^k)_{k\in\mathbb{N}}$ converges to a point in $\mathcal{S}$ if and only if every sequential cluster point of $(x^k)_{k\in\mathbb{N}}$ is in $\mathcal{S}$.*

Proof. Necessity is straightforward while sufficiency follows by Proposition 5.1 and Lemma 2.2. □

Let us conclude this section with a result that is particularly interesting for the conditions on the sequence that induce the metric. The result is a particular case of Proposition 5.1.

Corollary 5.3 (*Proposition 4.1, Combettes & Vũ, 2013*). *Let $\beta > 0$. Let $(\eta^k)_{k\in\mathbb{N}}$ be a nonnegative sequence such that $\sum_{k=1}^{\infty}\eta^k < \infty$, and let $(W_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\beta$ such that*

$$\begin{aligned} &\mu = \sup_{k\in\mathbb{N}}\|W_k\| < \infty \quad \text{and} \\ &(1+\eta^k)W_k \succeq W_{k+1} \text{ for all } k \in \mathbb{N}. \end{aligned} \tag{5.1}$$

Let $\mathcal{S} \subseteq \mathbb{R}^n$ be nonempty, closed and convex and let $(x^k)_{k\in\mathbb{N}}$ be a quasi-Fejér monotone sequence with respect to $\mathcal{S}$ relative to $(W_k)_{k\in\mathbb{N}}$ with $\phi = \|\cdot\|^2$, i.e.,

$$\|x^{k+1} - \bar{x}\|^2_{W_{k+1}} \le (1+\eta_k)\|x^k - \bar{x}\|^2_{W_k} + \varepsilon^k \quad \text{for all } k \in \mathbb{N}.$$

Then, for every $\bar{x} \in \mathcal{S}$, the sequence $(\|x^k - \bar{x}\|_{W_k})_{k\in\mathbb{N}}$ converges.

Proof. It follows from Corollary 3.8 and Proposition 5.1. For technical details we refer to Combettes and Vũ (2013). □

The condition in (5.1) is not hard to check on the problem data and it can be helpful for application purposes.

We conclude this section with an adaptation of Proposition 4.5 to the variable metric setup, i.e., an extension of Robbins–Siegmund Lemma (Lemma 4.1) to variable metric stochastic quasi-Fejer monotone sequences.

Proposition 5.4 (*Proposition 2.4, Vu, 2016*). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a non-empty closed set and let $\phi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$. Let $\beta > 0$, let W and $(W_k)_{k\in\mathbb{N}}$ be operators in $\mathcal{P}_\beta$ such that $W_k \to W$ pointwise. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence of random vectors. Suppose that, for every $\bar{x} \in \mathcal{X}$, there exist $(\delta^k)_{k\in\mathbb{N}}$ and $(\varepsilon^k)_{k\in\mathbb{N}}$ nonnegative sequences such that $\sum_{k=0}^{\infty}\delta^k < \infty$ and $\sum_{k=0}^{\infty}\varepsilon^k < \infty$ and such that,*

$$\begin{aligned} &\mathbb{E}[\phi(\|x_{k+1} - \bar{x}\|_{W_{k+1}}) \mid \mathcal{F}_k] \\ &\qquad \le (1+\delta_k(\bar{x}))\phi(\|x_k - \bar{x}\|_{W_k}) + \varepsilon_k(\bar{x}) \text{ a.s., for all } k \in \mathbb{N} \end{aligned}$$

Suppose that ϕ is strictly increasing and $\lim_{t\to\infty}\phi(t) = +\infty$. Then, the following hold.

(i) $(\|x_k - \bar{x}\|_{W_k})_{k\in\mathbb{N}}$ *is bounded and converges a.s.;*

(ii) $(x_k)_{k\in\mathbb{N}}$ *converges a.s. to random vector in $\mathcal{X}$ if and only if every cluster point is in $\mathcal{X}$ a.s.*

Proof. (i) follows from Lemma 4.1 and by the fact the ϕ is strictly increasing.

(ii) Necessity is straightforward while sufficiency follows from the properties of a cluster point and Lemma 2.2. For more details we refer to Vu (2016). □

6. Applications of convergent deterministic sequences

Since variational inequalities are the mathematical foundations of optimization-related problems, such as Nash equilibrium seeking (Facchinei & Pang, 2007; Franci & Grammatico, 2020a; Yi & Pavel, 2019), convex optimization (Bauschke et al., 2011; Facchinei & Pang, 2007; Jofré & Thompson, 2019) and machine learning (Franci & Grammatico, 2020b; Gidel, Berard, Vignoud, Vincent, & Lacoste-Julien, 2018), many works in the literature rely on the results presented in the previous sections to prove convergence of a given algorithm to a solution of a variational equilibrium problem. Specifically, they are applied to prove that a given algorithm converges to the solution of a variational inequality or to a zero of the sum of (monotone) operators. Thus, let us first describe the variational problem, starting by the definition of variational inequality (Bauschke et al., 2011; Facchinei & Pang, 2007).

Definition 6.1. Given a set $\mathcal{X} \subseteq \mathbb{R}^n$ and a mapping $F : \mathcal{X} \to \mathbb{R}^n$, a variational inequality, denoted $\mathrm{VI}(\mathcal{X}, F)$, is the problem

$$\text{find } x^* \in \mathcal{X} \text{ such that } \langle F(x^*), y - x^*\rangle \geq 0, \text{ for all } y \in \mathcal{X} \tag{6.1}$$

The set of solutions to this problem is denoted by $\mathrm{SOL}(\mathcal{X}, F)$.

The geometric interpretation of (6.1) is that a point $x^* \in \mathcal{X}$ is a solution of $\mathrm{VI}(\mathcal{X}, F)$ if and only if $F(x^*)$ forms an acute angle with every vector of the form $y - x^*$ for all $y \in \mathcal{X}$. In other words, (6.1) also says that a vector $x \in \mathcal{X}$ solves $\mathrm{VI}(\mathcal{X}, F)$ if and only if $-F(x^*)$ is a vector in the normal cone of $\mathcal{X}$ at x^* (see Appendix for the definition), i.e.,

$$0 \in F(x^*) + \mathrm{N}_{\mathcal{X}}(x^*). \tag{6.2}$$

Sometimes, instead of problem (6.1), a more general definition is proposed:

$$\text{find } x^* \in \mathcal{X} \text{ s.t. } \langle F(x^*), y - x^*\rangle + g(y) - g(x^*) \geq 0, \quad \text{for all } y \in \mathcal{X} \tag{6.3}$$

where g is a proper lower semi-continuous and convex function. Examples for the function g are indicator functions to enforce the set constraints, or penalty functions that promote sparsity, or other desirable structure.

Similarly to (6.1) and (6.2), problem (6.3) can be rewritten as

$$\text{find } x^* \in \mathcal{X} \text{ s.t. } 0 \in (F + \partial g)(x^*), \tag{6.4}$$

where ∂g is the subdifferential of g (definition in Appendix). In fact, if in (6.3) we take g as the indicator function, i.e., $g(x) = \iota_{\mathcal{X}}(x)$, we obtain the standard variation inequality (6.1), and instead of (6.4) we obtain the inclusion in (6.2) since $\partial g = \partial \iota_{\mathcal{X}} = \mathrm{N}_{\mathcal{X}}$ (Combettes & Pesquet, 2021, Equation (14)).

Problems of the form (6.2) and (6.4) are usually called (monotone) inclusion problems, which aim, in the general form, at finding $x^* \in \mathcal{X}$ such that $0 \in T(x^*)$ with $T : \mathbb{R}^n \to \mathbb{R}^n$. Moreover, in many cases it is possible to write a mapping as the summation of two (monotone) operators through an operator splitting technique (Bauschke et al., 2011; Ryu & Boyd, 2016). In this case, the problem of finding a zero of a monotone operator $T = A + B$ can be rewritten as

$$\text{find } x^* \in \mathcal{X} \text{ such that } 0 \in (A + B)(x^*). \tag{6.5}$$

Usually, $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and $B : \mathbb{R}^n \to \mathbb{R}^n$ are a set valued and a single valued monotone operator, respectively. Inclusions as the above arise systematically in convex optimization (Bot, Sedlmayer et al., 2020; Csetnek et al., 2019; Malitsky, 2015; Malitsky & Tam, 2020) and generalized Nash equilibrium problems in convex-monotone games (Franci & Grammatico, 2020a; Franci et al., 2020; Gadjov & Pavel, 2019, 2020; Pavel, 2019; Yi & Pavel, 2019).

Example 6.1 (*Inclusion Problem*)**.** Consider the minimization problem

$$\min_{x\in\mathcal{X}} f(x) + g(x) \tag{6.6}$$

where $g : \mathcal{X} \to \bar{\mathbb{R}}$ is proper, lower semicontinuous and convex and $f : \mathcal{X} \to \mathbb{R}$ is convex with Lipschitz continuous gradient. The solutions of the minimization problem in (6.6) are the points $x \in \mathcal{X}$ such that

$$0 \in (\nabla f + \partial g)(x). \tag{6.7}$$

where ∂g denotes the subdifferential of g and ∇f is the gradient of f. Eq. (6.7) is a monotone inclusion and it is equivalent to the generalized VI in (6.3) with $F = \nabla f$.

We are now ready to present some algorithms where the lemmas of Section 3 are used. The algorithms often rely on the monotonicity properties of the operators involved (see Definitions A.2 and A.3) and, unless otherwise mentioned, we suppose the following assumption to hold.

Standing Assumption 6.1. *The solution set of $\mathrm{VI}(\mathcal{X}, F)$ is not empty, i.e., $\mathrm{SOL}(\mathcal{X}, F) \neq \emptyset$, and $x_0 \in \mathcal{X}$, i.e., the sequence starts in the set $\mathcal{X}$ which is closed and convex.*

For every algorithm, we also propose a sketch of the convergence proof to show how the lemmas are used. A schematic representation of the necessary steps is provided in Fig. 7. The main idea to prove convergence of an algorithm is to obtain a (quasi) Féjer inequality with respect to the solution set and then apply one of the lemmas to the sequence $v^k = \|x^k - x^*\|^2$ where $x^* \in \mathrm{SOL}(\mathcal{X}, F)$ (see also Remark 3.8). Analogously, one can show that a suitable Lyapunov function asymptotically goes to zero.

We list the application depending on the type of problem but we name them after the convergence result that is used. We start with monotone inclusions, then move to VIs and Nash equilibrium problems, and finally consider an example of Lyapunov decrease.

6.1. Applications to monotone inclusions

Application of Lemmas 3.3 and 3.19. Lemma 3.3 is used in Malitsky and Tam (2018, 2020) to prove convergence in the inclusion problem:

find $x \in \mathcal{X}$ such that $0 \in (A + B)(x)$

where $A : \mathcal{X} \rightrightarrows \mathcal{X}$ and $B : \mathcal{X} \to \mathcal{X}$ are monotone operators. The sequence $(x^k)_{k\in\mathbb{N}}$, generated by the algorithm, is defined according to

$$x^{k+1} = J_{\alpha_k A}\left(x^k - \alpha_k B(x^k) - \alpha_{k-1}(B(x^k) - B(x^{k-1}))\right) \tag{6.8}$$

where $J_{\alpha_k A} = (I + \alpha_k A)^{-1}$ is the resolvent of A (Definition A.1). The algorithm is named *forward-reflected-backward splitting* and it is proven to converge to a zero of $A + B$.

Theorem 6.1 (*Theorem 2.5, Malitsky & Tam, 2020*)**.** *Let $A : \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone and $B : \mathcal{H} \to \mathcal{H}$ be monotone and ℓ-Lipschitz continuous. Let $\epsilon > 0$ and suppose $\alpha_k \in \left[\epsilon, \frac{1-2\epsilon}{2\ell}\right]$ for all $k \in \mathbb{N}$. Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (6.8) converges to a point $x^* \in \mathcal{X}$ such that $0 \in (A + B)(x^*)$.*

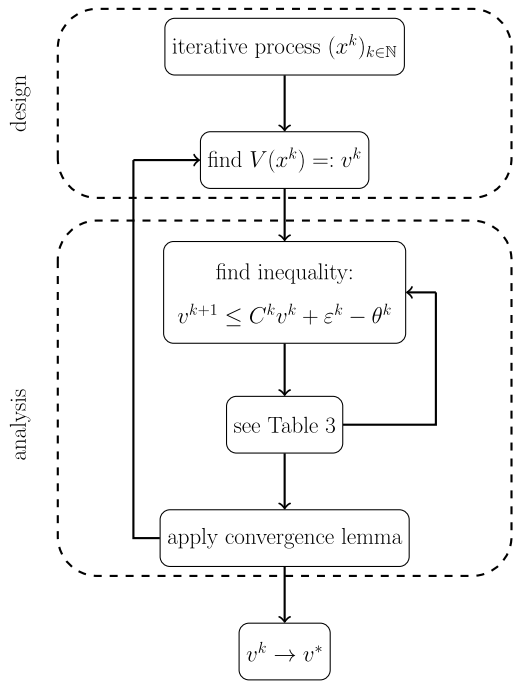

Fig. 7. Schematic representation of how the convergence lemmas for sequences can be used. Given the iterative process, a suitable nonnegative function (Lyapunov or distance-like) should be designed. Then, exploiting the properties of the application at hand, an inequality involving the iterates at times $k+1$ and k can be retrieved. Hence, one should check if the inequality corresponds to a known result (Table 3 for sequences of real numbers) and use the corresponding result to prove convergence. The whole process may take repeated steps to find a suitable function and/or inequality. The same reasoning applies to the stochastic case, in which one should have an expected valued inequality (with $\mathbb{E}[v^{k+1}]$) and refer to Table 4 for a convergence result on stochastic sequences. See also Fig. 8 for an example.

Proof. Let $x^* \in (A+B)^{-1}(0)$. It is possible to show, by using monotonicity and some norm properties, that the following inequality holds:

$$\begin{aligned}&\|x^{k+1}-x^*\|^2 + 2\alpha_k\langle B(x^{k+1}) - B(x^k), x^* - x^{k+1}\rangle + \\ &+ \left(\frac{1}{2}+\epsilon\right)\|x^{k+1}-x^k\|^2 \\ &\leq \|x^k - x^*\|^2 + 2\alpha_{k-1}\langle B(x^k) - B(x^{k-1}), x^* - x^k\rangle \\ &+ \frac{1}{2}\|x^k - x^{k-1}\|^2.\end{aligned} \tag{6.9}$$

Then, by doing a telescopic sum, using Lipschitz continuity and the properties of the parameters involved, the inequality in (6.9) can be rewritten as

$$\begin{aligned}&\frac{1}{2}\|x_{k+1}-x^*\|^2 + \varepsilon\sum_{i=0}^{k}\|x_{i+1}-x_i\|^2 \\ &\quad\leq \|x_0 - x^*\|^2 + 2\lambda_{-1}\langle B(x_0) - B(x_{-1}), x^* - x_0\rangle + \frac{1}{2}\|x_0 - x_{-1}\|^2\end{aligned}$$

from which we deduce that $(x^k)_{k\in\mathbb{N}}$ is bounded and that $\lim_{k\to\infty}\|x^k - x^{k+1}\| = 0$. Now, let $\bar{x}$ be a cluster point of $(x^k)_{k\in\mathbb{N}}$. From the definition of the algorithm in (6.8) and the properties of $A+B$, it follows that $0 \in (A+B)(\bar{x})$. Using again (6.9) and Lipschitz continuity it can be proven that $\lim_{k\to\infty}\|x^k - \bar{x}\|^2$ exists. Then, by Lemma 3.3, the sequence is convergent. □

The authors propose in the same paper also a variant of the algorithm with line search and a second one with inertia, but the convergence proof does not change its essence; in the first case, the authors use locally Lipschitz continuity (Malitsky & Tam, 2020, Theorem 3.4), while in the second they exploit the $1/\ell$-cocoercivity of the operator B (Malitsky & Tam, 2020, Theorem 4.3). Moreover, under the assumption of strong monotonicity of the operator A, they also prove convergence with linear rate, using Lemma 3.19.

Theorem 6.2 (*Theorem 2.9, Malitsky & Tam, 2020*)**.** *Let $A : \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone and μ-strongly monotone and $B : \mathcal{H} \to \mathcal{H}$ be monotone and ℓ-Lipschitz continuous. Suppose $\alpha \in \left(0, \frac{1}{2\ell}\right)$. Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (6.8) converges R-linearly to the unique point $\bar{x} \in \mathcal{X}$ such that $0 \in (A+B)(\bar{x})$.*

Proof. Similarly to the proof of Theorem 6.1 but using strong monotonicity, one obtains the inequality

$$\begin{aligned}&(1+2\mu\alpha)\|x^{k+1}-x^*\|^2 + 2\alpha\langle B(x^{k+1}) - B(x^k), x^* - x^{k+1}\rangle \\ &\qquad + (1-\alpha\ell)\|x^{k+1}-x^k\|^2 \\ &\quad\leq \|x^k - x^*\|^2 + 2\alpha\langle B(x^k) - B(x^{k-1}), x^* - x^k\rangle \\ &\qquad + \frac{1}{2}\|x^k - x^{k-1}\|^2.\end{aligned} \tag{6.10}$$

Setting $\gamma = (1+2\mu\alpha) > 1$, $v_k := \frac{1}{2}\|x^k - x^*\|^2$ and $\beta_k := \frac{1}{2}\|x^k - x^*\|^2 + 2\alpha\langle B(x^k) - B\left(x^{k-1}\right), x^* - x^k\rangle + \frac{1}{2}\|x^k - x^{k-1}\|^2$, one can apply Lemma 3.19 to conclude that the sequence $(x^k)_{k\in\mathbb{N}}$ converges to the unique solution $\bar{x}$ and with a linear rate. □

Application of Corollary 3.15. As an application of Corollary 3.15, let us consider the *inertial forward–backward* algorithm proposed in Dadashi and Postolache (2019) for approximating a zero of an inclusion problem $x \in (A+B)^{-1}(0)$:

$$\begin{cases} y^k = J_{\alpha_k A}\left(x^k - \alpha_k B x^k\right) \\ x^{k+1} = \nu_k x^k + \beta_k y^k + \gamma_k e^k \end{cases} \tag{6.11}$$

where $J_{\alpha_k A}$ is the resolvent of A (Definition A.1) and e^k is an error vector. By using Corollary 3.15 the authors prove the following result.

Theorem 6.3 (*Theorem 3.1, Dadashi & Postolache, 2019*)**.** *Let B be α-cocoercive and let A be maximally monotone. Let $\nu_k, \beta_k, \gamma_k \in (0,1)$ be such that $\nu_k + \beta_k + \gamma_k = 1$ and*

1. *$\lim_{k\to\infty}\gamma_k = 0$, and $\sum_{k=1}^{\infty}\gamma_k = \infty$,*
2. *$\lim_{k\to\infty} e^k = 0$,*
3. *$0 < a \leq \nu_k \leq b < 1$ and $0 < c \leq \beta_k \leq d < 1$,*
4. *$0 < c \leq \alpha_k < 2\alpha$ and $\lim_{k\to\infty}\left(\alpha_k - \alpha_{k+1}\right) = 0$.*

Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (6.11) converges to the point $x^* \in (A+B)^{-1}(0)$, where $x^* = \mathrm{proj}_{(A+B)^{-1}(0)}(0)$.

Proof. Using the nonexpansiveness of the resolvent of a maximally monotone operator (Bauschke et al., 2011, Corollary 23.9) and the cocoercivity of the mapping B, one can prove that the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded. Then, using some properties of the resolvent (Dadashi & Postolache, 2019, Lemma 2.6) and of the convex combination of bounded sequences (Dadashi & Postolache, 2019, Lemma 2.8) and using the monotonicity of A, the following inequality hold:

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \delta^k,$$

where δ^k is a quantity depending on the error e_k and on x^* and such that the assumption of Corollary 3.15 are satisfied. Therefore, convergence holds. □

6.2. Applications to variational inequalities

Application of Lemmas 3.3 and 3.6. The authors in Malitsky (2020) consider the general variational inequality problem in (6.3) where $g : \mathcal{X} \to \bar{\mathbb{R}}$ is a proper convex lower semicontinuous function and $F : \mathrm{dom}\, g \to \mathcal{X}$ is monotone. They propose the *Golden Ratio Algorithm* (GRAAL) whose iterations are given by

$$\begin{aligned}\tilde{x}^k &= \tfrac{(\varphi-1)x^k + \bar{x}^{k-1}}{\varphi} \\ x^{k+1} &= \mathrm{prox}_{\alpha g}(\tilde{x}^k - \alpha F(x^k))\end{aligned} \tag{6.12}$$

where $\varphi = \frac{\sqrt{5}+1}{2}$ is the golden ratio, i.e., $\varphi^2 = 1 + \varphi$. To prove convergence, they use Lemmas 3.3 and 3.6.

Theorem 6.4 (*Theorem 1, Malitsky, 2020*). *Let F be ℓ-Lipchitz continuous and monotone that g be lower semicontinuous and let $\alpha \in \left(0, \frac{\varphi}{2\ell}\right]$. Then the sequences $(x^k)_{k\in\mathbb{N}}$ and $(\tilde{x}^k)_{k\in\mathbb{N}}$, generated by (6.12), converge to a solution of the VI in (6.3).*

Proof. Using the fact that F is Lipschitz continuous and monotone and that the proximal operator is firmly nonexpansive, it holds that

$$\begin{aligned}&(1+\varphi)\|\tilde{x}^{k+1} - x^*\|^2 + \frac{\varphi}{2}\|x^{k+1} - x^k\|^2 \\ &\le (1+\varphi)\|\tilde{x}^k - x^*\|^2 + \frac{\varphi}{2}\|x^k - x^{k-1}\|^2 - \varphi\|x^k - \tilde{x}^k\|^2.\end{aligned} \tag{6.13}$$

Then, $(\tilde{x}^k)_{k\in\mathbb{N}}$ is bounded and $\lim_{k\to\infty}\|x^k - \tilde{x}^k\| = 0$ by Lemma 3.6. Hence, $(x^k)_{k\in\mathbb{N}}$ has at least one cluster point. Then, using the properties of g, all cluster points of $\tilde{x}^k$ are solutions of $\mathrm{VI}(\mathcal{X}, F)$. Since the sequence on the right hand side is non increasing, it is also convergent to a point in the solution set $\mathrm{SOL}(\mathcal{X}, F)$. Therefore, using the fact that $\lim_{k\to\infty}\|x^k - \tilde{x}^k\| = 0$ and the definition of $\tilde{x}^k$ in (6.12), Lemma 3.3 can be applied to conclude that $(x^k)_{k\in\mathbb{N}}$ converges to a solution of (6.3). □

Remark 6.1. Interestingly, in a preliminary version of the paper (Malitsky, 2018), the authors use Theorem 3.2 to prove convergence. In fact, given Eq. (6.13) and using the properties of the mapping g, they obtain that $(x^k)_{k\in\mathbb{N}}$ has a cluster point and they can directly apply Theorem 3.2 to conclude convergence.

In Malitsky (2018), the authors prove convergence of the *explicit GRAAL*, a variation of algorithm (6.12) with an adaptive step size rule. In this case, they only use locally Lipschitz continuity and conclude convergence via Lemma 3.3 (Malitsky, 2020, Theorem 2).

The algorithm has been recently extended to the stochastic case and for stochastic generalized Nash equilibrium problems (Franci & Grammatico, 2021a) and generative adversarial networks (Franci & Grammatico, 2021b) with a proof that relies on Lemma 4.1 on the same line of Section 7.1.

Application of Corollary 3.7. Corollary 3.7 is used in Malitsky (2015) to prove convergence of the *projected reflected gradient method* for variational inequalities as in (6.1). In details, the algorithm reads as

$$x^{k+1} = \mathrm{proj}_{\mathcal{X}}(x^k - \alpha F(2x^k - x^{k-1})) \tag{6.14}$$

and they show that the following result holds.

Theorem 6.5 (*Theorem 3.2, Malitsky, 2015*). *Let F be monotone and ℓ-Lipschitz continuous and $\alpha \in (0, \frac{\sqrt{2}-1}{\ell})$. Then the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (6.14) converges to a solution of $\mathrm{VI}(\mathcal{X}, F)$ in (6.1).*

Proof. Using the firmly nonexpansiveness of the projection, the fact that the mapping is monotone and ℓ-Lipschitz continuous and the bound on the step sizes, the following inequality holds:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 + \alpha\ell\|x^{k+1} - y_k\|^2 + 2\alpha\langle F(z), x^k - x^*\rangle \\ \le \|x^k - x^*\|^2 + \alpha\ell\|x^k - y^{k-1}\|^2 \\ + 2\alpha\langle F(x^*), x^{k-1} - x^*\rangle \\ - (1 - \alpha\ell(1+\sqrt{2}))\|x^k - x^{k-1}\|^2,\end{aligned}$$

where $x^* \in \mathrm{SOL}(\mathcal{X}, F)$. Now, by letting

$$\begin{aligned}\upsilon^k &= \|x^k - x^*\|^2 + \alpha\ell\|x^k - y^{k-1}\|^2 + 2\alpha\langle F(x^*), x^{k-1} - x^*\rangle \\ \theta^k &= (1 - \alpha\ell(1+\sqrt{2}))\|x^k - x^{k-1}\|^2,\end{aligned}$$

it follows that $\upsilon^{k+1} \le \upsilon^k - \theta^k$ as in Corollary 3.7, from which it is possible to deduce that $(x^k)_{k\in\mathbb{N}}$ is bounded and has at least one cluster point $\bar{x}$ and that $\lim_{k\to\infty}\|x^k - x^{k-1}\| = 0$. By Minty Theorem (Malitsky, 2015, Lemma 2.2) one have that any cluster point $\bar{x}$ is also a solution of the VI. By contradiction, it is possible to prove that $(x^k)_{k\in\mathbb{N}}$ cannot have two cluster points, therefore $\lim_{k\to\infty} x^k = \bar{x} \in \mathrm{SOL}(\mathcal{X}, F)$. □

Since the constant ℓ can be hard to compute, to avoid using ℓ-Lipschitz continuity, in the same paper, the authors also propose a variant of the algorithm in (6.14) that includes a prediction–correction technique to select the step sizes. The convergence result (Malitsky, 2015, Theorem 4.4) is proved similarly to the original result, using Corollary 3.7. Moreover, they also provide an estimation of the convergence rate when the mapping F is strongly monotone (similarly to Theorem 6.2) using a result similar to Lemma 3.19 (Malitsky, 2015, Lemma 2.9). This algorithm has been recently extended to the stochastic case (Cui & Shanbhag, 2016, 2021) and proved similarly, by exploiting Lemma 4.7.

6.3. Applications to Nash equilibrium problems

Application of Lemma 3.12. The fact that Lemma 3.12 guarantees convergence to zero (Remark 3.8) is used in Duvocelle et al. (2019) to compute a Nash equilibrium in traffic networks. In a dynamic traffic assignment problem, travelers participate in a non-cooperative game choosing a departure time and a route. The authors propose a *forward–backward–forward* algorithm (inspired by Tseng (2000)), i.e.,

$$\begin{aligned}z^k &= \mathrm{proj}_{\mathcal{X}}[x^k - \alpha F(x^k)] \\ y^k &= z^k + \alpha(F(x^k) - F(z^k)) \\ x^{k+1} &= (1 - \nu_k - \gamma_k)x^k + \gamma_k z^k,\end{aligned} \tag{6.15}$$

to solve the associated variational problem. The convergence result is stated next and it shows convergence to the solution of the VI associated to the Nash equilibrium problem (Facchinei & Pang, 2007, Proposition 1.4.2).

Theorem 6.6 (*Theorem 3.1, Duvocelle et al., 2019*). *Let F be pseudomonotone and ℓ-Lipschitz continuous. Let $(\nu^k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$ be sequences in $(0,1)$, such that $(\gamma_k)_{k\in\mathbb{N}} \subset (\nu, 1-\nu_k)$ for some $\nu > 0$, and let $\lim_{k\to\infty}\nu_k = 0$ and $\sum_{k=1}^{\infty}\nu_k = \infty$. Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (6.15) converges to $x^* \in \mathrm{SOL}(\mathcal{X}, F)$ where $x^* = \mathrm{argmin}\{\|z\| : z \in \mathrm{SOL}(\mathcal{X}, F)\}$.*

Proof. Using the definition of the algorithm in (6.15) and some preliminary inequalities (Duvocelle et al., 2019, Lemma 4.1), it holds that (Duvocelle et al., 2019, Lemma 4.3)

$$\begin{aligned}&\|x^{k+1} - x^*\|^2 \le (1 - \nu_k)\|x^k - x^*\|^2 + \\ &\nu_k[2\gamma_k\|x^k - y^k\| \cdot \|x^{k+1} - x^*\| + 2\langle x^*, x^* - x^{k+1}\rangle]\end{aligned} \tag{6.16}$$

To apply Lemma 3.12 to the sequence $\upsilon^k = \|x^{k+1} - x^*\|^2$, the authors check the conditions on $\beta^k = 2\gamma_k\|x^k - y^k\| \cdot \|x^{k+1} - x^*\| + 2\langle x^*, x^* - x^{k+1}\rangle$. First, note that since $\mathcal{X}$ is closed and convex, there exists a unique $x^* \in \mathrm{SOL}(\mathcal{X}, F)$ such that $x^* = \mathrm{proj}_{\mathrm{SOL}(\mathcal{X},F)}(0)$. Now, suppose that there exists $k_0 \in \mathbb{N}$ such that $\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2$ for all $k \ge k_0$. Then, $\lim_{k\to\infty}\|x^k - x^*\|^2$ exists. Then, exploiting the properties of the step size and using monotonicity and Lipschitz continuity, it can be proven that $\lim_{k\to\infty}\|x^k - z^k\|^2 = 0$ and that, by the definition of y_k, $\lim_{k\to\infty}\|y^k - x^k\|^2 = 0$. Therefore, $\lim_{k\to\infty}\|x^{k+1} - x^k\|^2 = 0$. Since the sequence is bounded (Duvocelle et al., 2019, Lemma 4.2), there exists a subsequence (x^{k_j}) such that $x^{k_j} \to \bar{y}$ and $\limsup_{k\to\infty}\langle x^*, x^* - x^k\rangle = \langle x^*, x^* - \bar{y}\rangle \le 0$, by the definition of x^*. Therefore, by Duvocelle et al. (2019, Lemma 4.4), also for a subsequence (z^{k_j}) it holds $z^{k_j} \to \bar{y}$. Then, $\lim_{k\to\infty}\langle x^*, x^* - x^{k+1}\rangle = \langle x^*, x^* - \bar{y}\rangle \le 0$ and by Lemma 3.12, $\lim_{k\to\infty}\|x^k - x^*\|^2 = 0$. For more details, we refer to Duvocelle et al. (2019). □

Application of Lemma 3.9. An instance of how Lemma 3.9 can be used to prove convergence is given in Kannan and Shanbhag (2012) where the authors propose a Nash equilibrium seeking algorithm via a Tikhonov regularization. The iterations, for each agent $i \in \mathcal{I} = \{1, \dots, N\}$, read as

$$x_i^{k+1} = \mathrm{proj}_{C_i}(x_i^k - \gamma_i^k(F_i(x^k) + \epsilon_i^k x_i^k)) \tag{6.17}$$

where $\gamma_k = (\gamma_i^k)_{i=1}^N$ and $\epsilon_k = (\epsilon_i^k)_{i=1}^N$ are the step size and regularization sequences, respectively, and C_i is the local feasible set for each player i. Then, the following result holds.

Theorem 6.7 (*Theorem 2.4, Kannan & Shanbhag, 2012*)**.** *Suppose F is monotone and ℓ-Lipschitz continuous over a closed convex set C and let $(\gamma^k)_{k\in\mathbb{N}}$ and $(\epsilon^k)_{k\in\mathbb{N}}$ be such that*

1. $\sum_{k=1}^{\infty} \gamma_j^k \epsilon_j^k = \infty$
2. $\lim_{k\to\infty} \frac{(\gamma_{\max}^k)^2}{\gamma_{\min}^k \epsilon_{\min}^k} = 0$
3. $\sum_{k=1}^{\infty} (\gamma_j^k)^2 < \infty$
4. $\sum_{k=1}^{\infty} (\epsilon_j^k \gamma_j^k)^2 < \infty$
5. $\lim_{k\to\infty} \frac{\epsilon_{\max}^{k-1} - \epsilon_{\min}^k}{\epsilon_{\min}^k (\gamma_{\min}^k)^2} = 0$
6. $\lim_{k\to\infty} \frac{\gamma_{\max}^k \epsilon_{\max}^k - \gamma_{\min}^k \epsilon_{\min}^k}{\gamma_{\min}^k \epsilon_{\min}^k} = 0$
7. $\lim_{k\to\infty} \epsilon_j^k = 0$ *for all* $j = 1, \dots, N$.

Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (6.17) converges to a Nash equilibrium as $k \to \infty$.

Proof. Since the classic Tikhonov relaxation, i.e., the iterative process where y^{k+1} solves $\mathrm{VI}(\mathcal{X}, F^k)$ and $F^k(y) = F(y) + \epsilon^k y$, is convergent (Bauschke et al., 2011; Tikhonov, 1963), the authors first show that (Kannan & Shanbhag, 2012, Proposition 2.3)

$$\|z^{k+1} - y^k\| \le q_k \|z^k - y^{k-1}\| + \frac{q_k M \sqrt{N}(\epsilon_{\max}^{k-1} - \epsilon_{\min}^k)}{\epsilon_{\min}^k},$$

where $q_k^2 = (1 - \gamma_{\min}^k \epsilon_{\min}^k)^2 + (\gamma_{\max}^k)^2 \ell^2 + 2(\gamma_{\max}^k \epsilon_{\max}^k - \gamma_{\min}^k \epsilon_{\min}^k)\ell$. Then, once they have $v^{k+1} \le q^k v^k + \epsilon^k$ with $v^k = \|x^k - y^{k-1}\|$, they prove that there exists a $\bar{k}$ such that $q^k < 1$ for all $k \ge \bar{k}$ (as in Remark 3.7). Thus, Lemma 3.9 can be applied to conclude convergence. □

6.4. Application to Lyapunov decrease

Application of Corollary 3.7. In this application, we show how the convergence results can be used in combination with a Lyapunov function. Let us consider the classic gradient method (Bauschke et al., 2011; Polyak, 1987)

$$x^{k+1} = x^k - \gamma \nabla f(x^k) \tag{6.18}$$

to find the minimum of a function $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$, f^*, and let us consider $V(x) = f(x) - f^*$ as a Lyapunov function candidate for the discrete time system in (6.18).

Theorem 6.8 (*Theorem 1.4.1, Polyak, 1987*)**.** *Let $f(x)$ be differentiable on $\mathbb{R}^n$ and bounded from below, i.e., $f(x) \ge f^* > -\infty$. Let ∇f be ℓ-Lipschitz continuous and let $\gamma \in \left(0, \frac{2}{\ell}\right)$. Then, in method (6.18) the gradient tends to zero, i.e., $\lim_{k\to\infty} \nabla f(x^k) = 0$ and the function $f(x)$ monotonically decreases, i.e., $f(x^{k+1}) \le f(x^k)$.*

Proof. Using differentiability and Lipschitz continuity, we obtain

$$f(x^{k+1}) \le f(x^k) - \gamma\left(1 - \frac{\ell\gamma}{2}\right)\|\nabla f(x^k)\|^2,$$

and, equivalently,

$$V(x^{k+1}) \le V(x^k) - \gamma\left(1 - \frac{\ell\gamma}{2}\right)\|\nabla V(x^k)\|^2$$

(see also Polyak, 1987, Section 2.2). Then, the claim follows by applying Corollary 3.7. □

6.5. Application to nonconvex optimization

Application of Lemma 3.4 and Corollary 3.5. Some of the lemmas introduced in Section 3 find application in nonconvex and nonsmooth problems as well (Di Lorenzo & Scutari, 2016; Facchinei, Lampariello, & Scutari, 2017; Scutari & Sun, 2019). In particular, let us consider the in-network successive convex approximation algorithm (NEXT) proposed in Di Lorenzo and Scutari (2016) for the optimization problem

$$\min_{x\in\mathcal{X}} \quad U(x) = F(x) + G(x) \tag{6.19}$$

where $F(x) = \sum_{i=1}^n f_i(x)$. In this model, n agents cooperate to solve a minimization problem while knowing only their own cost function f_i, $i \in \mathcal{I} = \{1, \dots, n\}$, possibly nonconvex, the common function G and the feasible set $\mathcal{X}$. The agents are connected over a graph $\mathcal{G} = (\mathcal{I}, \mathcal{E}^k)$ whose edge set may be time varying ($k \in \mathbb{N}$) and whose weight matrix is indicated by $W^k = [w_{ij}^k]_{i,j\in\mathcal{I}}$ for all $k \in \mathbb{N}$. The graph is assumed to be strongly connected and W^k to be row and column stochastic. The goal is to find a stationary solution of the problem via a distributed algorithm. Since the agents communicate with their neighbors only, each of them keeps an estimate, $\tilde{x}_i$, of the optimization variable. Moreover, since f_i is possibly nonconvex, it is replaced by a strongly convex approximation $\tilde{f}_i(\cdot, \tilde{x}_i^k)$. With the aim of minimizing $F(x) = f_i(x) + \sum_{j\neq i} f_j(x)$, the part corresponding to the other agents is linearized around the estimate $\tilde{x}_i^k$. Thus, at each iteration, each agent solves the following:

$$\bar{x}_i(\tilde{x}_i^k) = \operatorname*{argmin}_{x_i\in\mathcal{X}} \quad \tilde{f}_i(x_i; \tilde{x}_i^k) + \pi_i(\tilde{x}_i^k)^\top (x_i - \tilde{x}_i^k) + G(x_i)$$

where $\pi_i(\tilde{x}_i^k) = \sum_{j\neq i} \nabla_x f_j(x_i^k)$. However, this quantity requires the knowledge of the functions f_j, for $j \neq i$, therefore it is replaced by a local estimate $\tilde{\pi}_i^k$:

$$\begin{aligned}\hat{x}_i(\tilde{x}_i^k) &= \operatorname*{argmin}_{x_i\in\mathcal{X}} \quad \tilde{f}_i(x_i; \tilde{x}_i^k) + \tilde{\pi}_i^\top (x_i - \tilde{x}_i^k) + G(x_i) \\ &= \operatorname*{argmin}_{x_i\in\mathcal{X}} \quad \tilde{U}(x_i, \tilde{x}_i, \tilde{\pi}_i)\end{aligned}$$

To update $\tilde{\pi}_i^k$ an auxiliary variable y_i^k is introduced, to track the average gradient $\overline{\nabla f}(\tilde{x}) = \sum_{i=1}^n \nabla f_i(\tilde{x}_i)$. Moreover, to force asymptotic agreement, a consensus update of the estimated variables is used. Therefore, the algorithm reads as follows:

$$\begin{aligned}\hat{x}_i^k &= \operatorname*{argmin}_{x_i\in\mathcal{X}} \tilde{U}_i(x_i, \tilde{x}_i^k, \tilde{\pi}_i^k) \\ z_i^k &= x_i^k + \alpha^k(\hat{x}_i^k - \tilde{x}_i^k) \\ \tilde{x}_i^{k+1} &= \sum_{j=1}^n w_{ij}^k z_j^k \\ y_i^{k+1} &= \sum_{j=1}^n w_{ij}^k y_j^k + (\nabla f_i(\tilde{x}_i^{k+1}) - \nabla f_i(\tilde{x}_i^k)) \\ \tilde{\pi}^{k+1} &= n y_i^{k+1} - \nabla f_i(\tilde{x}_i^{k+1})\end{aligned} \tag{6.20}$$

where the first two iterates represent the optimization step while the last three are the consensus updates.

Theorem 6.9 (*Theorem 3, Di Lorenzo & Scutari, 2016*)**.** *Let $(\tilde{x}^k)_{k\in\mathbb{N}}$ be the sequence generated by (6.20), and define $\mathrm{avg}(\tilde{x}^k) = \frac{1}{n}\sum_{i=1}^n \tilde{x}_i^k$. Suppose that each f_i is continuously differentiable and each ∇f_i is Lipschitz continuous with constant L_i, respectively. Suppose that ∇F is bounded, i.e., there exists $L_F > 0$ such that $\sup_{\tilde{x}\in\mathcal{X}} \|\nabla F(\tilde{x})\| \le L_F$ and that G is a convex function with bounded subgradients, i.e., there exists $L_G > 0$*

such that $\sup_{\tilde{x}\in\mathcal{X}} \sup_{g\in\partial G(\tilde{x})} \|g\| \leq L_G$. *Suppose that* U *is coercive on* $\mathcal{X}$, *i.e.,* $\lim_{\tilde{x}\in\mathcal{X},\|\tilde{x}\|\to\infty} U(\tilde{x}) = \infty$ *and let the step size sequence* $(\alpha^k)_{k\in\mathbb{N}} \subseteq (0,1]$ *be such that* $\sum_{k=0}^{\infty} \alpha^k = \infty$ *and* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. *Then, the sequence* $(\mathrm{avg}(\tilde{x}^k))_{k\in\mathbb{N}}$ *is bounded and all its limit points are stationary point of* (6.19).

Proof. Given some preliminary results (Di Lorenzo & Scutari, 2016, Proposition 9, Section C) and introducing the quantities $\hat{x}_{i,\mathrm{av}}^k = \operatorname{argmin} \tilde{U}(x_i, \mathrm{avg}(\tilde{x}^k), \pi_{i,\mathrm{av}}^k)$, $\tilde{\pi}_{i,\mathrm{av}}^k = n y_{i,\mathrm{av}}^k - \nabla f_i(\mathrm{avg}(\tilde{x}^k))$ and $y_{i,\mathrm{av}}^k = \sum_{j\neq i} n_{ij}^k y_{j,\mathrm{av}}^k + (\nabla f_i(\mathrm{avg}(\tilde{x}^{k+1})) - \nabla f_i(\mathrm{avg}(\tilde{x}_i^k)))$, it holds that

$$\begin{aligned} U(\mathrm{avg}(\tilde{x}^{k+1})) \leq U(\mathrm{avg}(\tilde{x}^k)) &- c_1\alpha^k \sum_{i=1}^n \|\bar{x}_i(\mathrm{avg}(\tilde{x}^k)) - \mathrm{avg}(\tilde{x}^k)\|^2 \\ &+ c_2\alpha^k \sum_{i=1}^n \|\hat{x}_{i,\mathrm{av}}^k - \bar{x}_i(\mathrm{avg}(\tilde{x}^k))\| + c_2\alpha^k \sum_{i=1}^n \varepsilon_i^k \\ &+ c_2\alpha^k \sum_{i=1}^n \|\hat{x}_i^k - \hat{x}_{i,\mathrm{av}}^k\| + c_3\alpha_k^2 \end{aligned}$$

where the constants $c_1, c_2, c_3 > 0$ depend on the Lipschitz constants. Then, by Lemma 3.4, since U is coercive, $U(\mathrm{avg}(\tilde{x}^k))$ converges to a finite value and $\lim_{k\to\infty} \sup_{i\in\mathcal{I}} \|\bar{x}_i(\mathrm{avg}(\tilde{x}^k)) - \mathrm{avg}(\tilde{x}^k)\| = 0$. Since $(\mathrm{avg}(\tilde{x}^k))_{k\in\mathbb{N}}$ is bounded, it has at least one cluster point $w \in \mathcal{X}$, hence $\bar{x}_i(w) = w$ for all $i \in \mathcal{I}$. It follows that w is a stationary point (Di Lorenzo & Scutari, 2016, Proposition 5). □

The algorithm in (6.20) has been improved in Scutari and Sun (2019) where Corollary 3.5 is used to prove convergence.

6.6. Other applications

Opial Lemma (Lemma 3.3) is widely used for deterministic problems, in discrete (Boţ & Csetnek, 2016; Csetnek et al., 2019) and continuous time (Bot & Csetnek, 2016; Csetnek et al., 2019). Moreover, another application of Lemma 3.3 can be found in Bot, Sedlmayer et al. (2020) where the authors propose a forward–backward–forward algorithm (Bot, Mertikopoulos et al., 2020; Tseng, 2000) with an application to generative adversarial networks (Goodfellow, 2016; Goodfellow et al., 2014).

Concerning inclusion problems, the interested reader may find an application of Lemma 3.18 in Boţ and Csetnek (2016) while, for a different iterative scheme, Corollary 3.14 is used in Dadashi and Postolache (2019); finally, an application of Lemma 3.17 can be found in Cholamjiak et al. (2018).

Lemma 3.17 is used also for a variational problem in He and Yang (2013), along with Lemma 3.11. Moving to Nash equilibrium problems, Lemma 3.9 is used in Kannan and Shanbhag (2012) and Lei, Shanbhag, Pang et al. (2020) while Lemma 3.12 is used in Lei, Shanbhag and Chen (2020).

7. Applications of convergent stochastic sequences

Similarly to the deterministic case, many applications of the lemmas for random sequences concern the study of convergent algorithms for stochastic variational inequalities. Most of the literature relies on Robbins–Siegmund Lemma and on the monotone and Lipschitz properties of the operator (see Definitions A.2 and A.3 in Appendix A.2).

Before entering the details on how the lemmas are applied, we recall some preliminary notions on stochastic VIs (SVIs). For an extensive overview, we refer to Shanbhag (2013) and reference therein. More precisely, we are interested in solving SVI($\mathcal{X}, \mathbb{F}$), where $\mathbb{F}$ is an expected value function $\mathbb{F}(x) = \mathbb{E}[f(x, \xi(\omega))]$, for some measurable mapping $f : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$. $\xi : \Omega \to \mathbb{R}^d$ is a random variable and $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability space. For brevity, ξ is used to denote $\xi(\omega)$. Analogously to (6.1), we say that $x^* \in \mathcal{X}$ solves the SVI($\mathcal{X}, \mathbb{F}$) if

$$\langle \mathbb{F}(x^*), y - x^* \rangle \geq 0, \text{ for all } y \in \mathcal{X}, \tag{7.1}$$

and analogously to the deterministic case, we can consider the general variational inequality as in (6.3)

$$\text{find } x^* \in \mathcal{X} \text{ s.t. } \langle \mathbb{F}(x^*), y - x^* \rangle + g(y) - g(x^*) \geq 0 \text{ for all } y \in \mathcal{X}$$

or a monotone inclusion as in (6.5), i.e., find $x^* \in \mathcal{X}$ such that $0 \in (\mathbb{F} + \partial g)(x^*)$. We do not consider the case of stochastic functions g.

If the expected value of $f(x, \xi)$ is known, then the stochastic variational inequality can be solved with a standard solution technique for deterministic variational problems. However, the operator $\mathbb{F}(x)$ is usually not directly accessible, due to the computational burden or lack of information on the distribution of the random variable. Therefore, in general the focus is on $\hat{F}(x, \xi)$, an approximation of $\mathbb{F}(x)$, given some realizations ξ of the random variable.

There are two main methodologies available: stochastic approximation (SA) and sample average approximation (SAA). In the first case, $\mathbb{F}(x)$ is approximated by considering only one (or a finite number of) realization, at each iteration, of the random variable ξ (Cui & Shanbhag, 2021; Iusem et al., 2017; Koshal et al., 2013; Kushner & Yin, 2003; Robbins & Monro, 1951). In the second approach, instead, an infinite number of samples is taken at each iteration, then the approximation is given by the average over all the samples. The SAA scheme is mostly used to study existence of a solution (Kleywegt, Shapiro, & Homem-de Mello, 2002; Shapiro, 2003; Shapiro & Xu, 2008), and it is essentially a deterministic problem, therefore, in this work, we focus on the SA scheme. Hence, let us formalize it. If only one sample is available, the expected value mapping is approximated at each iteration as

$$F^{\mathrm{SA}}(x^k, \xi^k) = f(x^k, \xi^k), \tag{7.2}$$

where ξ^k is a realization of the random variable at time k. This approach is computationally cheap, but it requires, in general, stronger assumptions on the monotonicity of the mappings involved. Therefore, sometimes it is used in combination with the so-called variance reduction (VR). In this case, at each iteration, the approximation of $\mathbb{F}(x)$ has the form

$$\begin{aligned} F^{\mathrm{VR}}(x, \xi^k) &= \frac{1}{\mathcal{N}_k} \sum_{i=1}^{\mathcal{N}_k} f(x, \xi_i^k) \\ &= \frac{1}{\mathcal{N}_k} \sum_{i=1}^{\mathcal{N}_k} F^{\mathrm{SA}}(x, \xi_i^k) \text{ for all } x \in \mathcal{X}. \end{aligned} \tag{7.3}$$

The batch size sequence $(\mathcal{N}_k)_{k\in\mathbb{N}}$ determines the number of samples taken at each iteration. The sequence $\xi^k = (\xi_1^k, \dots, \xi_{\mathcal{N}_k}^k)$ is an i.i.d. random sequence. We suppose that $\mathcal{N}_k$ satisfy the following assumption any time the approximation scheme in (7.3) is used.

Standing Assumption 7.1. *The batch size sequence* $(\mathcal{N}_k)_{k\geq 1}$ *is such that, for some* $c, k_0, a > 0$,

$$\mathcal{N}_k \geq c(k + k_0)^{a+1}, \text{ for all } k \in \mathbb{N}. \tag{7.4}$$

It follows from Standing Assumption 7.1 that the batch size sequence is summable and this is fundamental to control the error committed in the approximation (see also Lemma A.2).

From now on, whenever we refer to an approximation without specifying the type, we use the symbol $\hat{F}$, while if it is one of the two schemes we explicitly use F^{SA} or F^{VR}.

Since we study an approximation (independently on the scheme), let us indicate the stochastic error, that is, the distance between the expected value and its approximation, with

$$\epsilon^k = F^{\mathrm{SA}}(x^k, \xi^k) - \mathbb{F}(x^k),$$

where ξ^k is a (vector of) realization of the random variable at iteration $k \in \mathbb{N}$. Sometimes this term is also called martingale difference (Definition 2.6) (Kushner & Yin, 2003; Ljung et al., 2012).

Standard assumptions on the stochastic error ϵ^k are that it has zero mean and bounded variance (Bot, Mertikopoulos et al., 2020; Franci & Grammatico, 2020a; Iusem et al., 2017; Lei & Shanbhag, 2018).

Standing Assumption 7.2. *The stochastic error is such that*

$\mathbb{E}\left[\epsilon^k|\mathcal{F}_k\right] = 0$ *a.s., for all* $k \in \mathbb{N}$.

Moreover, for all $x \in \mathcal{X}$ *and* $p \geq 1$ *let*

$$s_p(x) = \mathbb{E}\left[\|F^{\mathrm{SA}}(x,\xi) - \mathbb{F}(x)\|^p\right]^{\frac{1}{p}}.$$

There exist $p \geq 2$, $\sigma_0 \geq 0$ *and a measurable locally bounded function* $\sigma : \mathrm{SOL}(\mathcal{X},\mathbb{F}) \to \mathbb{R}$ *such that for all* $x \in \mathcal{X}$ *and all* $x^* \in \mathrm{SOL}(\mathcal{X},\mathbb{F})$

$$s_p(x) \leq \sigma(x^*) + \sigma_0\|x - x^*\|. \tag{7.5}$$

In the following, for ease of reading, we use a stronger condition than that in (7.5), namely,

$$\mathbb{E}\left[\|F^{\mathrm{SA}}(x,\xi) - \mathbb{F}(x)\|^p\right]^{\frac{1}{p}} \leq \sigma. \tag{7.6}$$

While Condition (7.5) is known in the literature as variance reduction, the stronger formulation (7.6) is called uniform bounded variance. Assumption (7.5) is more realistic in those cases where the feasible set $\mathcal{X}$ is unbounded, and it is always satisfied when the mapping f is Carathéodory and random Lipschitz continuous (Bot, Mertikopoulos et al., 2020, Example 1). Since in many realistic examples the feasible set is bounded, we use (7.6) as a variance control assumption. We also remark that many of the following results hold also in the more general case given by Assumption 7.2 and using the L_p norm for any $p \geq 2$. We refer to Bot, Mertikopoulos et al. (2020), Iusem et al. (2017) and references therein for a more detailed insight on this general case.

Remark 7.1. When we use the SA scheme with variance reduction, the following relation between the stochastic error and the batch size sequence holds (see Lemma A.2): for all $k \geq 0$, $c > 0$, σ as in (7.6) and $\mathcal{N}_k$ as in (7.4),

$$\mathbb{E}\left[\|\epsilon^k\|^2|\mathcal{F}_k\right] \leq \frac{c\sigma^2}{\mathcal{N}_k} \quad a.s.. \tag{7.7}$$

Essentially, Lemma A.2 says that the second moment of the error decreases with the increasing number of samples of the random variable.

Sometimes more general results hold for the bound in (7.7) (see, e.g., Bot, Mertikopoulos et al., 2020; Franci & Grammatico, 2020a; Iusem et al., 2017) but they lie outside the scopes of the survey.

We are now ready to describe how the lemmas are used. The first applications that we present are all related to Robbins–Siegmund Lemma (Lemma 4.1). We differentiate the applications on how the negative term $-\theta_k$ is exploited (Remark 4.1). Nonetheless, in all of them, the summability of the term is used differently to obtain convergence. For the first application we also provide a scheme (inspired by Fig. 7) of the step that should be taken to use a lemma for sequences of random numbers (Fig. 8). The section ends with an application of Lemma 4.7. As the reader may note, the forthcoming applications rely on the existence of a martingale, associated to the process, that the lemmas prove to be convergent (Benaim, 1996).

7.1. Applications of Robbins–Siegmund Lemma

Application of Lemma 4.1 with residual. In Bot, Mertikopoulos et al. (2020) and Iusem et al. (2017), the residual $(r_\alpha(x))$ is used to prove convergence (see Appendix for a definition and Remark A.1). Specifically, in Bot, Mertikopoulos et al. (2020), the authors formulate a *stochastic forward–backward–forward algorithm*, inspired by Tseng (2000), given by the following updating rule:

$$\begin{aligned} y^k &= \mathrm{proj}_{\mathcal{X}}(x^k - \alpha^k F^{\mathrm{VR}}(x^k,\xi^k)) \\ x^{k+1} &= y^k + \alpha^k(F^{\mathrm{VR}}(x^k,\xi^k) - F^{\mathrm{VR}}(y^k,\eta^k)) \end{aligned} \tag{7.8}$$

where ξ_n and η^k are i.i.d. random variables and F^{VR} is as in (7.3).

Robbins–Siegmund Lemma is used for concluding that the sequence $(x^k)_{k\in\mathbb{N}}$ converges a.s. to a solution of the SVI in (7.1), proving that the residual goes to zero (Remark A.1). A scheme of the proof and of how Lemma 4.1 is used can be found in Fig. 8.

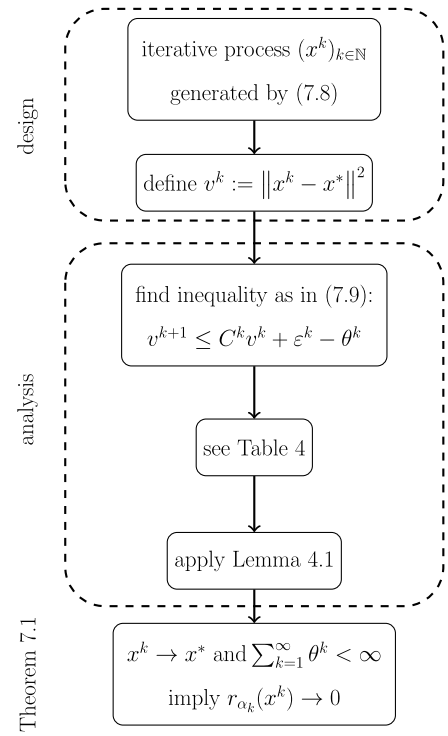

Fig. 8. Schematic representation of Theorem 7.1. First, a distance like function is defined, to obtain a quasi-Fejer inequality. Since the inequality correspond to Robbins–Siegmund Lemma, Lemma 4.1 can be applied. Not only convergence is proved, but also the fact that the negative term is summable contributes to showing that, asymptotically, a solution is reached. For general guidelines, see also Fig. 7.

Theorem 7.1 (*Theorem 1, Bot, Mertikopoulos et al., 2020*)**.** *Let* f *be a Carathéodory map and let* $\mathbb{F}$ *be pseudomonotone and* ℓ*-Lipschitz continuous with* $\ell > 0$. *Let* $0 < \inf_{k\geq 0}\alpha_k \leq \alpha_k \leq \sup_{k\geq 1}\alpha_k < \frac{1}{\sqrt{2}\ell}$. *Then, the sequence* $(x^k)_{k\in\mathbb{N}}$ *generated by* (7.8) *converges a.s. to a limit random variable* $x^* \in \mathrm{SOL}(\mathcal{X},\mathbb{F})$, *and* $\lim_{k\to\infty}\mathbb{E}[r_{\alpha_k}(x^k)^2] = 0$.

Proof. Using monotonicity and Lipschitz continuity of the mapping $\mathbb{F}$ and the definition of the algorithm in (7.8), it is possible to prove a recursion (Bot, Mertikopoulos et al., 2020, Lemma 5) that, taking the expected value (Bot, Mertikopoulos et al., 2020, Proposition 1) and using some bounds on the stochastic error (Bot, Mertikopoulos et al., 2020, Lemma 6) (see also Lemma A.2), reads as

$$\mathbb{E}[\|x^{k+1} - x^*\|^2|\mathcal{F}_k] \leq \|x^k - x^*\|^2 - \frac{\rho_k}{2}r_{\alpha_k}(x^k)^2 + \frac{\kappa_k\sigma^2}{\mathcal{N}_k}, \tag{7.9}$$

where $\rho_k = 1 - 2\ell^2\alpha_k^2$ and κ_k is a constant that depends on the Lipschitz constant and on the step size. To use Lemma 4.1, let $v^k = \|x^k - x^*\|^2$, $\theta_k = \frac{\rho_k}{2}r_{\alpha_k}(x^k)^2$ and $\varepsilon^k = \frac{\kappa_k\sigma_0^2}{\mathcal{N}_k}$. Then the claim follows using the fact that θ^k is summable and therefore the residual tends to zero. □

The use of the residual to prove convergence to the solution of the SVI in (7.1) was previously introduced in Iusem et al. (2017) where the authors propose a *stochastic extragradient method* inspired by Korpelevich (1976). The iterations are given by

$$\begin{aligned} z_i^k &= \mathrm{proj}_{\mathcal{X}}\left[x_i^k - \alpha^k F_i^{\mathrm{VR}}(x^k,\xi_i^k)\right] \\ x_i^{k+1} &= \mathrm{proj}_{\mathcal{X}}\left[x_i^k - \alpha^k F_i^{\mathrm{VR}}(z^k,\eta_i^k)\right], \end{aligned} \tag{7.10}$$

where $(\xi^k)_{k\in\mathbb{N}}$ and $(\eta^k)_{k\in\mathbb{N}}$ are i.i.d. samples of the random variable such that $(\xi^k)_{k\in\mathbb{N}}$ and $(\eta^k)_{k\in\mathbb{N}}$ are independent of each other. They have assumptions on the parameters similar to Bot, Mertikopoulos et al. (2020) and the variance reduction hypothesis. The main result is the asymptotic convergence of the algorithm.

Theorem 7.2 (*Theorem 3.18, Iusem et al., 2017*). *Let f be a Carathéodory map such that $\mathbb{E}[\|f(x,\xi)\|] < \infty$. Let $\mathbb{F}$ be pseudomonotone and ℓ-Lipschitz continuous mapping. Let $0 < \inf_{k\in\mathbb{N}} \alpha_k \le \alpha_k \le \sup_{k\in\mathbb{N}} \alpha_k < \frac{1}{\sqrt{6}\ell}$ for all $k \in \mathbb{N}$. Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (7.10) is bounded, $\lim_{k\to\infty} d(x^k, \mathrm{SOL}(\mathcal{X},\mathbb{F})) = 0$ and $r_{\alpha_k}(x^k)$ converges to 0. In particular, any cluster point of $(x^k)_{k\in\mathbb{N}}$ belongs to $\mathrm{SOL}(\mathcal{X},\mathbb{F})$.*

Proof. Given the properties of the operator $\mathbb{F}$ (Iusem et al., 2017, Lemma 3.11) and of the parameters involved (Iusem et al., 2017, Lemma 3.12), it holds that Iusem et al. (2017, Proposition 3.15)

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 | \mathcal{F}_k\right] \le \left(1 + \frac{C(\sigma^2, x^*)}{\mathcal{N}_k}\right) \|x^k - x^*\|^2 - \frac{\rho_k}{2} r_{\alpha_k}(x^k)^2 + \frac{C(\sigma^2, x^*)}{\mathcal{N}_k}$$

where $\rho_k = (1 - 6\ell^2\alpha_k^2)$, $C(\sigma^2, x^*)$ is a bounded quantity that depends on the solution x^* and on the variance (Iusem et al., 2017, Remark 3.17), and $r_{\alpha_k}(x^k)$ is the residual of x^k. Then the claim follows as in the proof of Theorem 7.1, using Robbins–Siegmund Lemma. □

Application of Lemma 4.1 with strict monotonicity. Robbins–Siegmund Lemma can also be used to prove the convergence of the *partially coordinated iterative proximal point scheme* to a Nash equilibrium (Koshal et al., 2013). The possibility to reach a Nash equilibrium in a game theoretic framework is related to the fact that they can be obtained as the solution of a suitable (S)VI (Facchinei & Pang, 2007, Proposition 1.4.2). The updating rule of the algorithm is given by:

$$x^{k+1} = \mathrm{proj}_{\mathcal{X}}[x^k - \alpha_k(\hat{F}(x^k, \xi^k) + \mu^k(x^k - x^{k-1}))] \tag{7.11}$$

where $\alpha_k \in \mathbb{R}^n$ and $\mu^k \in \mathbb{R}^n$ are the step size and the centering parameters, respectively, and n is the number of agents in the Nash equilibrium problem.

Proposition 7.3 (*Proposition 3, Koshal et al., 2013*). *Let $\mathbb{F} : \mathcal{X} \to \mathbb{R}^n$ be strictly monotone and ℓ-Lipschitz continuous over $\mathcal{X}$. Let the following conditions hold:*

1. $\alpha_{k,\max}\mu_{k,\max} \le (1 + 2\alpha_{k,\max}^2\ell^2)\cdot\ \alpha_{k-1,\min}\mu_{k-1,\min}$ *for all* $k \in \mathbb{N}$;
2. $\lim_{k\to\infty} \frac{\alpha_{k,\max}^2\mu_{k,\min}^2}{\alpha_{k,\min}\mu_{k,\min}} = c$ *with* $c \in \left[0, \frac{1}{2}\right)$;
3. $\sum_{k=0}^{\infty} \alpha_{k,i} = \infty$ *and* $\sum_{k=0}^{\infty} \alpha_{k,i}^2 < \infty$ *for all* $i \le n$;
4. $\sum_{k=0}^{\infty} (\alpha_{k,\max} - \alpha_{k,\min}) < \infty$;
5. $\sum_{k=0}^{\infty} \alpha_{k,\max}^2 \mathbb{E}[\|\epsilon_k\|^2 | \mathcal{F}_k] < \infty$ *a.s.*

Then, the sequence $(x^k)_{k\in\mathbb{N}}$ generated by (7.11) converges a.s. to a solution of $\mathrm{SVI}(\mathcal{X},\mathbb{F})$.

Proof. Using the nonexpansiveness of the projection and some norm properties, one can obtain

$$\begin{aligned}
&\mathbb{E}[\|x^{k+1} - x^*\|^2 | \mathcal{F}_k] + \alpha_{k,\min}\mu_{k,\min}\|x^k - x^*\|^2 \\
&\le (1 + \delta_k)(\|x^k - x^*\|^2 + \alpha_{k-1,\min}\mu_{k-1,\min}\|x^{k-1} - x^*\|^2) \\
&\quad - \alpha_{k,\min}\mu_{k,\min}(1 - d)\|x^k - x^{k-1}\|^2 + \alpha_{k,\max}^2\mathbb{E}[\|\epsilon_k\|^2 | \mathcal{F}_k] \\
&\quad - 2\alpha_{k,\min}(x^k - x^*)^T(\mathbb{F}(x^k) - \mathbb{F}(x^*)),
\end{aligned}$$

where $x^* \in \mathrm{SOL}(\mathcal{X},\mathbb{F})$, δ^k depends on the Lipschitz constant and on the step sizes and $d \in (0,1)$. To apply Lemma 4.1, let $\upsilon^k = \|x^k - x^*\|^2 + \alpha_{k-1,\min}\mu_{k-1,\min}\|x^{k-1} - x^*\|^2$, $\theta^k = \alpha_{k,\min}\mu_{k,\min}(1-d)\|x^k - x^{k-1}\|^2 + 2\alpha_{k,\min}\langle x^k - x^*, \mathbb{F}(x^k) - \mathbb{F}(x^*)\rangle$ and $\varepsilon^k = \alpha_{k,\max}^2\mathbb{E}[\|\epsilon_k\|^2 | \mathcal{F}_k]$. Then, it follows that $(x^k)_{k\in\mathbb{N}}$ is bounded and has a cluster point $\bar{x}$. Since θ^k is summable, $\langle x^k - x^*, \mathbb{F}(x^k) - \mathbb{F}(x^*)\rangle \to 0$ and taking the limit for $k \to \infty$, $\langle \bar{x} - x^*, \mathbb{F}(\bar{x}) - \mathbb{F}(x^*)\rangle = 0$. Since the mapping is strictly monotone (Definition A.2) and the solution set is not empty (Standing Assumption 6.1), there is only one solution x^* (Facchinei & Pang, 2007, Theorem 2.3.3), and we have that $\bar{x} = x^*$. □

7.2. Applications of Robbins–Siegmund Lemma to specific problems

Application of Lemma 4.1 to model predictive control. An interesting application of Robbins–Siegmund Lemma is provided in Lee and Nedić (2015), where the authors propose the *gossip-based random projections* (GRP) algorithm for distributed robust model predictive control (MPC). In their problem, m private facilities aim at finding an optimal control law $u = \mathrm{col}(u(1), \dots, u(T))$ of a dynamic system such that the resulting trajectory $x(t)$, for $t = 1, \dots, T$, remains close to the locally known facilities and the terminal state $x(T)$ is inside some uncertain box with minimum control effort. Formally, the distributed MPC optimization problem is given by

$$\begin{aligned}
\min_u \quad & f(u) = \sum_{i=1}^{m} f_i(u) \\
s.t. \quad & x(t+1) = Ax(t) + Bu(t) \\
& u \in \mathcal{X} \\
& \max_{\ell=1,2,3,4} \{(a_\ell + c_\ell)^\top x(T) - b_\ell\} \le 0
\end{aligned}$$

where $u \in \mathcal{X}$ represent the uncertain input constraint and the last inequality describe the random terminal constraint ($\mathcal{T}$ from now on). We refer to Lee and Nedić (2015) for a specific choice of f, A and B. The algorithm is based on random projections and a gossip communication protocol inspired by Boyd, Ghosh, Prabhakar, and Shah (2006). At each time k, only an agent $I_k \in \mathcal{I}$ and its neighbor $J_k \in \mathcal{I}$ wake up. They draw a sample of one of the linear inequality terminal constraints and they update their estimate while the other agents do nothing. Then, they project their current iterate on the selected constraint $\mathcal{T}$ and on $\mathcal{X}$. The GRP algorithm reads, for $i \in \{I_k, J_k\}$, as

$$\begin{aligned}
w_i^k &= \frac{u_{I_k}^{k-1} + u_{J_k}^{k-1}}{2} \\
u_i^k &= \mathrm{proj}_{\mathcal{X}\cap\mathcal{T}}[w_i^k - \alpha_i^k \nabla f_i(w_i^k)]
\end{aligned} \tag{7.12}$$

where $\{\alpha_i^k\}_{k\in\mathbb{N}}$ is the step size sequence, defined such that $\alpha_i^k = 1/\Gamma_i^k$, where Γ_i^k is the number of updates i has performed until time k.

Let us denote $z_i^k = \mathrm{proj}_{\mathcal{X}}(w_i^k)$ and let $\tilde{w} = \frac{1}{m}\sum_{i=1}^m w_i$ (analogously $\tilde{z}$). Then, convergence of the algorithm is proven as follows.

Proposition 7.4 (*Lee & Nedić, 2015, Proposition 1*). *Let the communication graph be connected and let the set $\mathcal{X}$ be closed and convex. Let the functions $f_i : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable and let their gradients ∇f_i be Lipschitz continuous and bounded over $\mathcal{X}$, i.e., $\|\nabla f_i(u)\| \le G_f$ for all $u \in \mathcal{X}$ and all $i \in \mathcal{I}$. Let $\mathcal{X}^*$ be a nonempty optimal set. Then, the sequences $\{u_i^k\}_{k\in\mathbb{N}}$, $i \in \mathcal{I}$, generated by (7.12) converge to some random point $u^* \in \mathcal{X}^*$ a.s., i.e., $\lim_{k\to\infty} u_i^k = u^*$ a.s. for all $i \in \mathcal{I}$.*

Proof. First, it is possible to show (by using Lemma 4.1) that w_i^k approaches $\mathcal{X}$ (Lee & Nedić, 2015, Lemma 3), and that any two sequences $\{w_i^k\}_{k\in\mathbb{N}}$ and $\{w_j^k\}_{k\in\mathbb{N}}$ have the same limit points a.s. Lee and Nedić (2015, Lemma 4). Then, it holds by Lee and Nedić (2015, Lemma 2) and some properties of the projection, of the norms and of the mappings involved, that

$$\begin{aligned}
&\sum_{i=1}^{m} \mathbb{E}[\|u_i^k - u^*\|^2 \mid \mathcal{F}_{k-1}] \\
&\quad \le \left(1 + \frac{a_4}{k^{\frac{3}{2}} - q}\right) \sum_{j=1}^{m} \|u_j^{k-1} - u^*\|^2 \\
&\quad - \frac{2}{k}\mathbb{E}\left[f(\tilde{z}^k) - f^* \mid \mathcal{F}_{k-1}\right] \\
&\quad + \frac{4G_f}{k} \sum_{i=1}^{m} \mathbb{E}[\|w_i^k - \tilde{w}^k\| \mid \mathcal{F}_{k-1}] + \frac{a_2 m}{k^{\frac{3}{2}} - q}
\end{aligned} \tag{7.13}$$

where q, a_2 and a_4 are constants. Eq. (7.13) satisfies the conditions from Lemma 4.1 (Lee & Nedić, 2015, Lemma 4). Hence, $\{\|u_i^k - u^*\|^2\}_{k\in\mathbb{N}}$ is convergent a.s. for any $i \in \mathcal{I}$ and $u^* \in \mathcal{X}^*$. Moreover, by Lemma 4.1,

$\liminf_{k\to\infty} f(\tilde{z}^k) - f^* = 0$ a.s. and by Lee and Nedić (2015, Lemma 3) $\lim_{k\to\infty} \|w_i^k - z_i^k\| = 0$ for all $i \in \mathcal{I}$ a.s. Hence also the sequences $\{\|w_i^k - u^*\|^2\}_{k\in\mathbb{N}}$ and $\{\|z_i^k - u^*\|^2\}_{k\in\mathbb{N}}$ and their averages $\{\|\tilde{w}^k - u^*\|^2\}_{k\in\mathbb{N}}$ and $\{\|\tilde{z}^k - u^*\|^2\}_{k\in\mathbb{N}}$ are convergent and the sequences $\{\tilde{w}^k\}_{k\in\mathbb{N}}$ and $\{\tilde{z}^k\}_{k\in\mathbb{N}}$ are bounded and have an accumulation point in $\mathcal{X}^*$. Since $\liminf_{k\to\infty} \|w_i^k - \tilde{w}^k\| = 0$ for all $i \in \mathcal{I}$ a.s., it follows that $\lim_{k\to\infty} \|w_i^k - u^*\| = 0$ for all $i \in \mathcal{I}$ a.s. Finally, by Lee and Nedić (2015, Lemma 3), $\lim_{k\to\infty} \|u_i^k - w_i^k\| = 0$ for all $i \in \mathcal{I}$ a.s., which leads to, $\lim_{k\to\infty} u_i^k = u^*$ for all $i \in \mathcal{I}$ a.s. □

Application of Corollary 4.4 to the Law of Large Numbers. Remarkably, the convergence results for sequences can be used also for others scopes beside convergence of an algorithm. This is the case of Corollary 4.4 which is used to prove the Law of Large Numbers. To introduce this application, let us define the notion of increasing process associated to a martingale (Chung et al., 1990; Duflo, 2013).

Let y_k be a martingale such that $\mathbb{E}[y_k^2] < \infty$ for all $k \in \mathbb{N}$. Then, its increasing process is the sequence $(\langle y\rangle_k)_{k\in\mathbb{N}}$ defined by $\langle y\rangle_0 = 0$ and $\langle y\rangle_{k+1} - \langle y\rangle_k = \mathbb{E}[y_{k+1}^2 - y_k^2 \mid \mathcal{F}_k]$ (Duflo, 2013, Proposition 1.3.7). For instance, if $(z_k)_{k\in\mathbb{N}}$ is a sequence of identically distributed random variables with mean μ and variance σ^2 then $y_k = z_1 + \cdots + z_k - k\mu$ satisfies $\mathbb{E}[y_k^2] < \infty$ and is such that $\langle y\rangle_k = k\sigma^2$. Then, the following generalization of the Law of Large Numbers for martingales holds.

Theorem 7.5 (*Theorem 1.3.15, Duflo, 2013*). *Let $(y_k)_{k\in\mathbb{N}}$ be a martingale such that $\mathbb{E}[y_k^2] < \infty$ for all $k \in \mathbb{N}$ and let $(\langle y\rangle_k)_{k\in\mathbb{N}}$ be its increasing process. Then, a.s. $\lim_{k\to\infty} \frac{y_k}{\langle y\rangle_k} = 0$.*

Proof. To apply Corollary 4.4, let $\upsilon_k = y_k^2$, $\epsilon_k = \langle y\rangle_{k+1} - \langle y\rangle_k$ and $a_k = \langle y\rangle_{k+1}(\ln(\langle y\rangle_{k+1}))^{1+\gamma}$. Then, if $\langle y\rangle_{k_0} > 1$, $\sum_{k=k_0}^{\infty} a_k^{-1}\epsilon_k < \infty$, $\lim_{k\to\infty} a_k^{-1}\upsilon_k = 0$ and the claim follows. □

Remark 7.2. We note that, even if it does not involve an algorithm, Theorem 7.5 is in agreement with Figs. 7 and 8. In fact, we have an iterative process $((y^k)_{k\in\mathbb{N}})$, to which we assign a sequence $(\upsilon^k = y_k^2)$ to obtain an inequality as in Corollary 4.4, retrieved from Table 4, and prove the result.

7.3. Applications of Lemma 4.7

Application of Lemma 4.7 to a variational problem. In Yousefian et al. (2014), a *smoothing extragradient scheme* with stochastic approximation, similar to (7.10), is proposed. The iterations read as

$$\begin{aligned} y^{k+1} &= \mathrm{proj}_{\mathcal{X}}[x^k - \gamma_k \hat{F}(x^k + w^k, \eta^k)] \\ x^{k+1} &= \mathrm{proj}_{\mathcal{X}}[x^k - \gamma_k \hat{F}(y^{k+1} + z^k, \xi^k)] \end{aligned} \tag{7.14}$$

where $(\gamma_k)_{k\in\mathbb{N}}$ is the step size sequence, η^k and ξ^k are i.i.d. samples of the random variable and the sequences $(w^k)_{k\in\mathbb{N}}$ and $(z^k)_{k\in\mathbb{N}}$ are also i.i.d. random variables drawn from an uniform distribution on $\left[-\frac{\delta_k}{2}, \frac{\delta_k}{2}\right]$ where δ^k is the smoothing sequence. Let $\mathcal{X}_\delta = \mathcal{X} + C_n(0,\delta)$ where C_n is a n-dimensional cube centered at the origin and δ is an upper bound on δ_k. Then, the following holds.

Theorem 7.6 (*Theorem 2, Yousefian et al., 2014*). *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be closed, convex, and M-bounded. Let $\mathbb{F}$ be strictly monotone over $\mathcal{X}$ and bounded on $\mathcal{X}^\delta$ for some $C > 0$. Suppose the sequence $(\gamma_k)_{k\in\mathbb{N}}$ is such that $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$ and $\sum_{k=1}^{\infty} \gamma^k = \infty$ that the sequence δ_k is diminishing according to Yousefian et al. (2014, Equation 19). Then, the sequence $(x_k)_{k\in\mathbb{N}}$ generated by (7.14) converges to a solution x^* of $\mathrm{SVI}(\mathcal{X},\mathbb{F})$ a.s. as $k \to \infty$.*

Proof. Using the assumptions, from Yousefian et al. (2014, Lemma 4) the following inequality holds:

$$\mathrm{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] \le \left(1 - \frac{\alpha\gamma_k}{M}\right)\|x^k - x^*\|^2 + q\gamma_k^2$$

where q is a constant that depends on the constant C and on the variance of the stochastic error and M is the bound on the set $\mathcal{X}$. Then, convergence follows applying Lemma 4.7 to $\upsilon^k = \|x^k - x^*\|^2$, $\epsilon^k = q\gamma_k^2$ and $\delta^k = \frac{\alpha\gamma_k}{M}$. □

Application of Lemma 4.7 to opinion dynamics. The fact Lemma 4.7 provides convergence to zero is used in Shi et al. (2013) to prove agreement in an opinion dynamics model. Let us consider the spreading of true or false information over a communication network or faults propagations in large scale control systems. In these models, there are n nodes and each of them (node i) activates with a probability $1/n$, then it picks a neighbor j with probability a_{ij}. The probabilities are collected in the interaction matrix $A = [a_{ij}]_{i,j=1}^n$. The dynamics is described as follows, given $\alpha + \beta + \gamma = 1$:

(i) (Attraction) With probability α, node i updates its opinion toward that of its neighbor j,

$$x_i^{k+1} = x_i^k + T_k(x_j^k - x_i^k),$$

where $0 < T_k \le 1$ is the trust level;

(ii) (Neglect) With probability β, node i keeps its own opinion,

$$x_i^{k+1} = x_i^k;$$

(iii) (Repulsion) With probability γ, node i moves away from j, i.e., it updates with a negative coefficient,

$$x_i^{k+1} = x_i^k - S_k(x_j^k - x_i^k)$$

where $S_k > 0$.

The authors propose in Shi et al. (2013) some conditions on the quantities involved under which agreement or disagreement can be obtained with a time-invariant trust level. As a measure of disagreement, let $L^k = \sum_{i=1}^n \|x_i^k - x_{\mathrm{ave}}\|^2$, where $x_{\mathrm{ave}} = \sum_{i=1}^n \frac{x_i^0}{n}$ is the average of the initial values. Then, the following result holds.

Theorem 7.7 (*Shi et al., 2013, Theorem 5*). *Let the communication graph be weakly connected and suppose that the updates are symmetric. Let λ_2^* be the second smallest eigenvalue of $D-(A+A^\top)$ with $D = \mathrm{diag}\left(d_1 \dots d_n\right)$, $d_i = \sum_{j=1}^n \left(a_{ij} + a_{ji}\right)$. Let $T_k \equiv T^* \in [0,1]$ and $S_k \equiv S^* > 0$. Then*

$$D^* = S^*\left(1 + S^*\right)\gamma - T^*\left(1 - T^*\right)\alpha$$

is a critical convergence measure regarding the state convergence of the considered network. Specifically, if $D_* < 0$, then global agreement convergence is achieved, i.e., $\lim_{k\to\infty} \mathbb{E}[L^k] = 0$ a.s.

Proof. Given some preliminary results (Shi et al., 2013, Proposition 3), it holds that

$$\mathbb{E}[L^{k+1} \mid^k] \le \left(1 - \frac{2}{n} D^* \lambda_2^*\right) L^k$$

then $\lim_{k\to\infty} \mathbb{E}[L^k] = 0$ a.s. by Lemma 4.7. □

7.4. Other applications

Other applications of Robbins–Siegmund Lemma (Lemma 4.1) can be found in Alacaoglu and Malitsky (2021), Cui and Shanbhag (2021), Iusem et al. (2017, 2019), Jiang and Xu (2008), Kannan and Shanbhag (2019) and Wang and Bertsekas (2015) for variational problems and monotone inclusions. Concerning Nash equilibrium problems, it is used in Franci and Grammatico (2020a, 2021a) and Koshal et al. (2013). In the specific case of generative adversarial networks, Lemma 4.1 is used in Franci and Grammatico (2020b, 2021b). For an application of this stochastic result to a deterministic problem, we refer to Koshal, Nedić, and Shanbhag (2016). Regarding dynamic systems and Lyapunov analysis, other utilizations of Robbins–Siegmund Lemma are in Benaim (1996, Section 3), Bharath and Borkar (1999) and Ljung et al. (2012, Section I.1).

For other applications of Lemma 4.7 instead, the interested reader may refer to Cui and Shanbhag (2021), Iusem et al. (2019), Kannan and Shanbhag (2019, 2019), Koshal et al. (2013), Lei, Shanbhag, Pang et al. (2020) and Yousefian et al. (2017).

8. Applications of convergent sequences with variable metric

The variable metric framework is not studied as much as the classic setting. We refer to Combettes and Vũ (2013), Vũ (2013) and Cui et al. (2019) for some applications. Next, we illustrate the application to a monotone inclusion problem.

Application of Proposition 5.1 and Theorem 5.2. A study of the *forward–backward–forward* algorithm (Bot, Mertikopoulos et al., 2020; Tseng, 2000) with variable metric is considered in Vũ (2013). There, the authors consider the splitting of a sum of a maximally monotone operator A and a monotone, Lipschitz continuous operator B of the form (6.5) and they suppose that multiple errors (sequences a^k, b^k and c^k) can be made at each iteration. Formally, their proposed algorithm reads as

$$\begin{aligned} y^k &= x^k - \gamma^k W_k(Bx^k + a^k) \\ v^k &= (\mathrm{Id} + \gamma^k W_k A)^{-1} y^k + b^k \\ u^k &= y^k + \gamma^k W_k(Bv^k + c^k) \\ x^{k+1} &= x^k - y^k + u^k \end{aligned} \tag{8.1}$$

where $(W_k)_{k\in\mathbb{N}}$ is the sequence of operators used to induce the metric. Then, the following convergence result holds true.

Theorem 8.1 (*Theorem 3.1, Vũ, 2013*)**.** *Let $\beta, \ell > 0$, let $(\eta^k)_{k\in\mathbb{N}}$ be a nonnegative sequence such that $\sum_{k=1}^{\infty} \eta^k < \infty$ and let $(W_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\beta$ such that*

$$\mu = \sup_{k\in\mathbb{N}} \|W_k\| < +\infty$$

$$(1+\eta^k)W_{k+1} \succeq W_k, \text{ for all } k \in \mathbb{N}.$$

Let $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximally monotone, let $B : \mathbb{R}^n \to \mathbb{R}^n$ be monotone and ℓ-Lipschitz continuous. Let $(a^k)_{k\in\mathbb{N}}, (b^k)_{k\in\mathbb{N}}$ and $(c^k)_{k\in\mathbb{N}}$ be such that $\sum_{k=1}^{\infty} |a^k| < \infty$, $\sum_{k=1}^{\infty} |b^k| < \infty$ and $\sum_{k=1}^{\infty} |c^k| < \infty$. Let $x_0 \in \mathbb{R}^n$, $\varepsilon \in (0, 1/(\ell\mu + 1))$ and let $(\gamma^k)_{k\in\mathbb{N}}$ be a sequence in $[\varepsilon, (1-\varepsilon)/(\ell\mu)]$. Let $x^* \in \mathrm{zer}(A+B)$ and let $(x^k)_{k\in\mathbb{N}}$ be the sequence generated by (8.1). Then, the following hold:

(i) $\sum_{k=1}^{\infty} \|x^k - v^k\|^2 < +\infty$,
(ii) $\lim_{k\to\infty} x^k = x^*$.

Proof. After using some results from Combettes and Vũ (2013) to guarantee that the sequences are well defined and that the monotonicity properties of the operators $W_k A$ and $W_k B$ hold, a quasi-Féjer inequality can be proven, i.e.,

$$\begin{aligned} \|x^{k+1} - x^*\|^2_{W_{k+1}^{-1}} &\le (1+\eta^k)\|x^k - x^*\|^2_{W_k^{-1}} \\ &\quad - \mu^{-1}(1 - \gamma_k^2 \ell^2 \mu^2)\|x^k - v^k\|^2 + \varepsilon^k, \end{aligned} \tag{8.2}$$

where $(\varepsilon^k)_{k\in\mathbb{N}}$ is a summable sequence depending on the error sequences $(a^k)_{k\in\mathbb{N}}$, $(b^k)_{k\in\mathbb{N}}$ and $(c^k)_{k\in\mathbb{N}}$. Then, from Proposition 5.1, it holds that $(\|x^k - x^*\|_{W_k^{-1}})_{k\in\mathbb{N}}$ is bounded and from Lemma 3.4, it follows that $\sum_{k=1}^{\infty} \|x^k - v^k\|^2 < \infty$. Convergence holds as a consequence of Theorem 5.2 by letting $\bar{x}$ be a cluster point. □

9. Conclusion

By trying to answer the question posed by Polyak in 1987, this survey highlights the importance of convergence theorems for mathematical system theory, in a variety of mathematical applications areas.

Thanks to the notions of Lyapunov decrease and (quasi) Féjer monotonicity, results showing the convergence of sequences of (random) real numbers can be exploited in game equilibrium problems, machine learning and optimization, where these results are key to analyze and design iterative processes with guaranteed convergence.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Auxiliary notions

A.1. Preliminaries

In this section, we recall some notions from operator theory (Bauschke et al., 2011; Facchinei & Pang, 2007; Ryu & Boyd, 2016). Let us start with some notation.

$\iota_{\mathcal{X}}$ is the indicator function of the set $\mathcal{X}$, that is, $\iota_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $\iota_{\mathcal{X}}(x) = \infty$ otherwise. The set-valued mapping $\mathrm{N}_{\mathcal{X}} : \mathbb{R}^n \to \mathbb{R}^n$ denotes the normal cone operator of the set $\mathcal{X}$, i.e., $\mathrm{N}_{\mathcal{X}}(x) = \emptyset$ if $x \notin \mathcal{X}$, $\mathrm{N}_{\mathcal{X}}(x) = \{v \in \mathbb{R}^n | \sup_{z\in\mathcal{X}} \langle v, z - x\rangle \le 0\}$ otherwise. Given $\psi : \mathbb{R}^n \to \mathbb{R}$, $\mathrm{dom}(\psi) := \{x \in \mathbb{R}^n | \psi(x) < \infty\}$ is the domain of ψ and its subdifferential is the set-valued mapping $\partial\psi(x) := \{v \in \mathbb{R}^n | \psi(z) \ge \psi(x) + \langle v, z - x\rangle, \ \forall z \in \mathrm{dom}(\psi)\}$.

Let us now provide the definition of projection, proximal operator and resolvent which are used in many algorithms.

Definition A.1. Let $x \in \mathbb{R}^n$ and let $C \subseteq \mathbb{R}^n$ be nonempty closed and convex.

- The projection operator onto C is the operator defined as

  $$\mathrm{proj}_C(x) = \operatorname*{argmin}_{z\in C} \|z - x\|^2.$$

  The point $\mathrm{proj}_C(x)$ is the closest point to x in C. It always exists and it is unique.
- Given a proper lower semicontinuous convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, the proximity operator of f is the operator defined as

  $$\mathrm{prox}_f(x) = \operatorname*{argmin}_{y\in\mathbb{R}^n} \left(f(y) + \frac{1}{2}\|x - y\|^2 \right)$$

- Given $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, the resolvent of A is the operator defined as

  $$J_A = (\mathrm{Id} + A)^{-1}$$

  where Id is the identity function.

The notions discussed until now are related by the following example.

Example A.1 (*Example 23.4, Bauschke et al., 2011*)**.** Let $f = \iota_C$ to be the indicator function of the set $C \subseteq \mathbb{R}^n$, then

$$J_{\mathrm{N}_C} = \left(\mathrm{Id} + \mathrm{N}_C\right)^{-1} = \mathrm{prox}_{\iota_C} = \mathrm{proj}_C$$

where N_C is the normal cone of C.

Given a closed and convex set $C \subseteq \mathbb{R}^n$, $F : C \to \mathbb{R}^n$ and $\alpha > 0$, let $\mathrm{res}_\alpha(x) = \|x - \mathrm{proj}_{\mathcal{X}}(x - \alpha F(x))\|$ be the residual function.

Remark A.1. Let $F : C \to \mathbb{R}^n$ and $C \subseteq \mathbb{R}^n$ be closed and convex. Then, Facchinei and Pang (2007, Proposition 1.5.8)

$$\mathrm{res}(x^*) = 0 \text{ if and only if } x^* \in \mathrm{SOL}(C, F).$$

Moreover, it holds that

$$x^* \in \mathrm{SOL}(C, F) \text{ if and only if } x^* = \mathrm{proj}_C(x^* - \alpha F(x^*)).$$

A.2. Operator theory

The convergence properties of the algorithms proposed for VIs or monotone inclusions are strictly related to the properties of the operators, to its monotonicity in particular. For this reason, here we recall some definition that are useful for the applications.

Definition A.2. Let $\mathrm{gra}(F) = \{(x,u) : u \in F(x)\}$ be the graph of $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^n$. Then, F is said to be

(a) pseudomonotone on $\mathcal{X}$ if $\langle F(y), x-y\rangle \geq 0 \Rightarrow \langle F(x), x-y\rangle \geq 0$ for all $x, y \in \mathcal{X}$;
(b) monotone on $\mathcal{X}$ if $\langle F(x)-F(y), x-y\rangle \geq 0$, for all $x, y \in \mathcal{X}$;
(c) strictly monotone on $\mathcal{X}$ if $\langle F(x)-F(y), x-y\rangle > 0$, for all $x, y \in \mathcal{X}$ and $x \neq y$;
(e) μ-strongly monotone on $\mathcal{X}$ if there exists a constant $\mu > 0$ such that $\langle F(x)-F(y), x-y\rangle \geq \mu\|x-y\|^2$, for all $x, y \in \mathcal{X}$;
(f) maximally monotone (or maximal monotone) if there exists no monotone operator $G : \mathcal{X} \to \mathbb{R}^n$ such that $\mathrm{gra}\, G$ properly contains $\mathrm{gra}\, F$, i.e., $(x,u) \in \mathrm{gra}\, F \Leftrightarrow$ for all $(y,v) \in \mathrm{gra}\, F$ it holds $\langle x-y|u-v\rangle \geq 0$;
(g) uniformly monotone at y if there exist an increasing function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ vanishing only at 0 such that $\langle F(x)-F(y), x-y\rangle \geq \phi(\|x-y\|)$.

The weakest assumption is pseudomonotonicity while strong monotonicity implies all the other notions. Strictly monotone operators are widely used in variational inequalities problems since this is the weaker assumption that guarantees uniqueness of the solution (Facchinei & Pang, 2007, Theorem 2.3.3). It implies monotonicity that, in turn, implies pseudomonotonicity.

Many results are also related to the Lipschitz and cocoercivity constants of the operator.

Definition A.3. A mapping $F : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is said to be

(a) ℓ-Lipschitz continuous with constant $\ell > 0$ if for all $x, y \in \mathcal{X}$ it holds $\|F(x)-F(y)\| \leq \ell\|x-y\|$; if $\ell < 1$, F is called a contraction;
(b) nonexpansive if it is 1-Lipschitz continuous, i.e., $\|F(x)-F(y)\| \leq \|x-y\|$ for all $x, y \in \mathcal{X}$;
(c) firmly nonexpansive if $\|F(x)-F(y)\|^2 + \|(\mathrm{Id}-F)(x)-(\mathrm{Id}-F)(y)\|^2 \leq \|x-y\|^2$, for all $x, y \in \mathcal{X}$
(d) β-cocoercive on $\mathcal{X}$ if there exists a constant $\beta > 0$ such that $\langle F(x)-F(y), x-y\rangle \geq \beta\|F(x)-F(y)\|^2$, for all $x, y \in \mathcal{X}$.

It follows (using Cauchy–Schwartz inequality) that if a map is β-cocoercive, it is also $1/\beta$-Lipschitz continuous. Moreover, cocoercivity implies monotonicity.

Sometimes, in the stochastic case, we mention that the map is Carathéodory.

Definition A.4. A mapping $F : \mathbb{R}^d \times \Xi \to \mathbb{R}^n$ is a Carathéodory map if $x \mapsto F(x,\xi)$ is continuous for almost every $\xi \in \Xi$, and $\xi \mapsto F(x,\xi)$ is measurable for all $x \in \mathbb{R}^d$ where ξ is a random variable with values in Ξ, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A.3. Auxiliary results

Let us recall some results on martingales as this property for L_p norms, known as Burkholder–Davis–Gundy inequality (Kushner & Yin, 2003; Stroock, 2010).

Lemma A.1 (*Burkholder–Davis–Gundy Inequality*). *Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be a filtration and $\{u^k\}_{k\geq 0}$ a vector-valued martingale relative to this filtration. Then, for all $p \in [1,\infty)$, there exists a universal constant $c_p > 0$ such that for every $k \geq 1$*

$$\mathbb{E}\left[\left(\sup_{0\leq i\leq N}\|u_i\|\right)^p\right]^{\frac{1}{p}} \leq c_p \mathbb{E}\left[\left(\sum_{i=1}^{N}\|u_i-u_{i-1}\|^2\right)^{\frac{p}{2}}\right]^{\frac{1}{p}}.$$

We also recall the Minkowski inequality: for given functions $f, g \in L^p(\Xi, \mathcal{F}, \mathbb{P})$, $\mathcal{G} \subseteq \mathcal{F}$ and $p \in [0;\infty]$

$$\mathbb{E}\left[\|f+g\|^p|\mathcal{G}\right]^{\frac{1}{p}} \leq \mathbb{E}\left[\|f\|^p|\mathcal{G}\right]^{\frac{1}{p}} + \mathbb{E}\left[\|g\|^p|\mathcal{G}\right]^{\frac{1}{p}}.$$

When combined with the Burkholder–Davis–Gundy inequality, it leads to the fact that for all $p \geq 2$, there exists a constant $c_p > 0$ such that, for every $k \geq 1$,

$$\mathbb{E}\left[\left(\sup_{0\leq i\leq N}\|u_i\|\right)^p\right]^{\frac{1}{p}} \leq c_p\sqrt{\sum_{k=1}^{N}\mathbb{E}\left(\|u_i-u_{i-1}\|^p\right)^{\frac{2}{p}}}.$$

The following result is presented for uniformly bounded variance but holds also for more general assumptions. For similar results, one can refer to Bot, Mertikopoulos et al. (2020), Franci and Grammatico (2020a) and Iusem et al. (2017).

Lemma A.2. *Let $c > 0$. Let σ be as in Eq. (7.6) and $\mathcal{N}_k$ as in Standing Standing Assumption 7.1. Then, it holds a.s. that*

$$\mathbb{E}\left[\|\epsilon^k\|^2|\mathcal{F}_k\right] \leq \frac{c\sigma^2}{\mathcal{N}_k}, \text{ for all } k \in \mathbb{N}.$$

Proof. We first prove that

$$\mathbb{E}\left[\|\epsilon^k\|^2|\mathcal{F}_k\right]^{\frac{1}{2}} \leq \frac{c_2\sigma}{\sqrt{\mathcal{N}_k}}$$

then the claim follows immediately. To this aim, let us first notice that $F^{\mathrm{VR}}(x,\xi) = \frac{1}{\mathcal{N}}\sum_{k=1}^{\mathcal{N}} F^{\mathrm{SA}}(x,\xi^k)$. Then, let us define the process $\{M_S^S(x)\}_{i=0}^S$ as $M_0(x) = 0$ and for $1 \leq i \leq S$

$$M_i^S(x) = \frac{1}{S}\sum_{j=1}^{i} F^{\mathrm{SA}}(x,\xi_j) - \mathbb{F}(x).$$

Let $\mathcal{F}_i = \sigma(\xi_1, \dots, \xi_i)$. Then $\{M_i^S(x), \mathcal{F}_i\}_{i=1}^S$ is a martingale starting at 0. Let

$$\begin{aligned}\Delta M_{i-1}^S(x) &= M_i^S(x) - M_{i-1}^S(x)\\ &= \frac{1}{S}(F^{\mathrm{SA}}(\boldsymbol{x},\xi_i) - \mathbb{F}(\boldsymbol{x})).\end{aligned}$$

Then, by Eq. (7.6), we have

$$\mathbb{E}\left[\|\Delta M_{i-1}^S\|^2\right]^{\frac{1}{2}} = \frac{1}{S}\mathbb{E}\left[\|F^{\mathrm{SA}}(x,\xi_i) - \mathbb{F}(x)\|^2\right]^{\frac{1}{2}} \leq \frac{\sigma}{S}.$$

By applying Lemma A.1, we have

$$\begin{aligned}\mathbb{E}\left[\|M_S^S(x)\|^2\right]^{\frac{1}{2}} &\leq c_2\sqrt{\sum_{i=1}^{N}\mathbb{E}\left[\|\frac{F^{\mathrm{SA}}(\boldsymbol{x},\xi_i)-\mathbb{F}(x)}{S}\|^2\right]}\\ &\leq c_2\sqrt{\frac{1}{S^2}\sum_{i=1}^{N}\mathbb{E}\left[\|F^{\mathrm{SA}}(\boldsymbol{x},\xi_i)-\mathbb{F}(x)\|^2\right]}\\ &\leq \frac{c_2\sigma}{\sqrt{S}}.\end{aligned}$$

We note that $M_S^S(x^k) = \epsilon^k$, hence by taking the square we conclude that

$$\mathbb{E}\left[\|\epsilon^k\|^2|\mathcal{F}_k\right] \leq \frac{c\sigma^2}{\mathcal{N}_k}. \quad \square$$

References

Abbas, B., & Attouch, H. (2015). Dynamical systems and forward–backward algorithms associated with the sum of a convex subdifferential and a monotone cocoercive operator. *Optimization*, *64*(10), 2223–2252.

Abbas, B., Attouch, H., & Svaiter, B. F. (2014). Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces. *Journal of Optimization Theory and Applications*, *161*(2), 331–360.

Alacaoglu, A., & Malitsky, Y. (2021). Stochastic variance reduction for variational inequality methods. arXiv preprint arXiv:2102.08352.

Alber, Y. I., Iusem, A. N., & Solodov, M. V. (1998). On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming*, *81*(1), 23–35.

Ananduta, W., & Grammatico, S. (2021). Bregman algorithms for a class of mixed-integer generalized Nash equilibrium problems. arXiv preprint arXiv:2105.05687.

Attouch, H., & Peypouquet, J. (2019). Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. *Mathematical Programming*, *174*(1), 391–432.

Barty, K., Roy, J.-S., & Strugarek, C. (2007). Hilbert-valued perturbed subgradient algorithms. *Mathematics of Operations Research*, *32*(3), 551–562.

Bauschke, H. H., Borwein, J. M., & Combettes, P. L. (2003). Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, *42*(2), 596–636.

Bauschke, H. H., Combettes, P. L., et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces, Vol. 408*. Springer.

Bauschke, H. H., Dao, M. N., & Moursi, W. M. (2015). On Fejér monotone sequences and nonexpansive mappings. arXiv preprint arXiv:1507.05585.

Benaim, M. (1996). A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, *34*(2), 437–472.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII* (pp. 1–68). Springer.

Benning, M., Betcke, M. M., Ehrhardt, M. J., & Schönlieb, C.-B. (2021). Choose your path wisely: gradient descent in a Bregman distance framework. *SIAM Journal on Imaging Sciences*, *14*(2), 814–843.

Berg, L., Engel, W., Pazderski, G., & Stolle, H. (1995). Basic properties of Féjer monotone sequences. In *Rostok. Math. Kolloq, Vol. 49* (pp. 57–74). Citeseer.

Bertsekas, D. P., & Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, *10*(3), 627–642.

Bharath, B., & Borkar, V. S. (1999). Stochastic approximation algorithms: Overview and recent trends. *Sadhana*, *24*(4), 425–452.

Borkar, V. S. (1995). *Probability theory: An advanced course*. Springer Science & Business Media.

Boţ, R. I., & Csetnek, E. R. (2016). An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems. *Numerical Algorithms*, *71*(3), 519–540.

Bot, R. I., & Csetnek, E. R. (2016). Second order forward-backward dynamical systems for monotone inclusion problems. *SIAM Journal on Control and Optimization*, *54*(3), 1423–1443.

Bot, R., Mertikopoulos, P., Staudigl, M., & Vuong, P. (2020). Mini-batch forward-backward-forward methods for solving stochastic variational inequalities. *Stochastic Systems*.

Bot, R. I., Sedlmayer, M., & Vuong, P. T. (2020). A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs. arXiv preprint arXiv:2003.07886.

Boyd, S., Ghosh, A., Prabhakar, B., & Shah, D. (2006). Randomized gossip algorithms. *IEEE Transactions on Information Theory*, *52*(6), 2508–2530.

Bravo, M., Leslie, D. S., & Mertikopoulos, P. (2018). Bandit learning in concave N-person games. arXiv preprint arXiv:1810.01925.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 200–217.

Chen, H.-F. (2006). *Stochastic approximation and its applications, Vol. 64*. Springer Science & Business Media.

Cholamjiak, W., Cholamjiak, P., & Suantai, S. (2018). An inertial forward–backward splitting method for solving inclusion problems in Hilbert spaces. *Journal of Fixed Point Theory and Applications*, *20*(1), 1–17.

Chung, K. L., Williams, R. J., & Williams, R. (1990). *Introduction to stochastic integration, Vol. 2*. Springer.

Combettes, P. (2001a). Fejér-monotonicity in convex optimization. In *Encyclopedia of optimization, Vol. 2* (pp. 106–114). Kluwer.

Combettes, P. L. (2001b). Quasi-Fejérian analysis of some optimization algorithms. In *Studies in computational mathematics, Vol. 8* (pp. 115–152). Elsevier.

Combettes, P. L. (2004). Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, *53*(5–6), 475–504.

Combettes, P., & Pesquet, J. (2015). Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, *25*(2), 1221–1248. http://dx.doi.org/10.1137/140971233.

Combettes, P. L., & Pesquet, J.-C. (2019). Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping II: mean-square and linear convergence. *Mathematical Programming*, *174*(1), 433–451.

Combettes, P. L., & Pesquet, J.-C. (2021). Fixed point strategies in data science. *IEEE Transactions on Signal Processing*, *69*, 3878–3905. http://dx.doi.org/10.1109/TSP.2021.3069677.

Combettes, P. L., & Vũ, B. C. (2013). Variable metric quasi-Fejér monotonicity. *Nonlinear Analysis. Theory, Methods & Applications*, *78*, 17–31.

Combettes, P. L., & Vũ, B. C. (2014). Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, *63*(9), 1289–1318.

Csetnek, E. R., Malitsky, Y., & Tam, M. K. (2019). Shadow Douglas–Rachford splitting for monotone inclusions. *Applied Mathematics and Optimization*, *80*(3), 665–678.

Cui, S., & Shanbhag, U. V. (2016). On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems. In *2016 IEEE 55th conference on decision and control (CDC)* (pp. 4510–4515). IEEE.

Cui, S., & Shanbhag, U. V. (2021). On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *Set-Valued and Variational Analysis*, *29*(2), 453–499.

Cui, F., Tang, Y., & Zhu, C. (2019). Convergence analysis of a variable metric forward–backward splitting algorithm with applications. *Journal of Inequalities and Applications*, *2019*(1), 1–27.

Dadashi, V., & Postolache, M. (2019). Forward–backward splitting algorithm for fixed point problems and zeros of the sum of monotone operators. *Arabian Journal of Mathematics*, 1–11.

Di Lorenzo, P., & Scutari, G. (2016). Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, *2*(2), 120–136.

Doob, J. L. (1953). *Stochastic processes, Vol. 101*. New York: Wiley.

Duflo, M. (2013). *Random iterative models, Vol. 34*. Springer Science & Business Media.

Duvocelle, B., Meier, D., Staudigl, M., & Vuong, P. T. (2019). Strong convergence of forward-backward-forward methods for pseudo-monotone variational inequalities with applications to dynamic user equilibrium in traffic networks. arXiv preprint arXiv:1908.07211.

Eremin, I. I. (1968a). Methods of Féjer's approximations in convex programming. *Mathematical Notes of the Academy of Sciences of the USSR*, *3*(2), 139–149.

Eremin, I. I. (1968b). On the speed of convergence in the method of Féjer approximations. *Mathematical Notes of the Academy of Sciences of the USSR*, *4*(1), 522–527.

Eremin, I. (1969). Fejér mappings and convex programming. *Siberian Mathematical Journal*, *10*(5), 762–772.

Eremin, I. I., & Popov, L. D. (2009). Fejér processes in theory and practice: Recent results. *Russian Mathematics*, *53*(1), 36–55.

Ermol'Ev, Y. M. (1969). On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics*, *5*(2), 208–220.

Ermoliev, Y. M., & Wets, R.-B. (1988). *Numerical techniques for stochastic optimization*. Springer-Verlag.

Facchinei, F., Lampariello, L., & Scutari, G. (2017). Feasible methods for nonconvex nonsmooth problems with applications in green communications. *Mathematical Programming*, *164*(1), 55–90.

Facchinei, F., & Pang, J.-S. (2007). *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media.

Fejér, L. (1922). Über die lage der nullstellen von polynomen, die aus minimumforderungen gewisser art entspringen. In *Festschrift David Hilbert zu seinem sechzigsten geburtstag Am 23. Januar 1922* (pp. 41–48). Springer.

Franci, B., & Grammatico, S. (2020a). A distributed forward-backward algorithm for stochastic generalized Nash equilibrium seeking. *IEEE Transactions on Automatic Control*, http://dx.doi.org/10.1109/TAC.2020.3047369.

Franci, B., & Grammatico, S. (2020b). A game–theoretic approach for generative adversarial networks. In *2020 59th IEEE conference on decision and control (CDC)* (pp. 1646–1651). IEEE.

Franci, B., & Grammatico, S. (2021a). Stochastic generalized Nash equilibrium seeking in merely monotone games. *IEEE Transactions on Automatic Control*.

Franci, B., & Grammatico, S. (2021b). Training generative adversarial networks via stochastic Nash games. *IEEE Transactions on Neural Networks and Learning Systems*, 1–10. http://dx.doi.org/10.1109/TNNLS.2021.3105227.

Franci, B., Staudigl, M., & Grammatico, S. (2020). Distributed forward-backward (half) forward algorithms for generalized Nash equilibrium seeking. In *2020 European control conference (ECC)* (pp. 1274–1279). IEEE.

Gadjov, D., & Pavel, L. (2019). Distributed GNE seeking over networks in aggregative games with coupled constraints via forward-backward operator splitting. In *2019 IEEE 58th conference on decision and control (CDC)* (pp. 5020–5025). IEEE.

Gadjov, D., & Pavel, L. (2020). Single-timescale distributed GNE seeking for aggregative games over networks via forward-backward operator splitting. *IEEE Transactions on Automatic Control*.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., & Lacoste-Julien, S. (2018). A variational inequality perspective on generative adversarial networks. arXiv preprint arXiv:1802.10551.

Gladyshev, E. (1965). On stochastic approximation. *Theory of Probability and its Applications*, *10*(2), 275–278.

Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gubin, L., Polyak, B. T., & Raik, E. (1967). The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, *7*(6), 1–24.

He, S., & Yang, C. (2013). Solving the variational inequality problem defined on intersection of finite level sets. In *Abstract and applied analysis, Vol. 2013*. Hindawi.

Huang, Y.-Y., Jeng, J.-C., Kuo, T.-Y., & Hong, C.-C. (2011). Fixed point and weak convergence theorems for point-dependent λ-hybrid mappings in Banach spaces. *Fixed Point Theory and Applications*, *2011*(1), 1–15.

Iusem, A., Jofré, A., Oliveira, R. I., & Thompson, P. (2017). Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, *27*(2), 686–724.

Iusem, A. N., Jofré, A., Oliveira, R. I., & Thompson, P. (2019). Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization*, *29*(1), 175–206.

Jiang, H., & Xu, H. (2008). Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control*, *53*(6), 1462–1475.

Jofré, A., & Thompson, P. (2019). On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Mathematical Programming*, *174*(1), 253–292.

Kannan, A., & Shanbhag, U. V. (2012). Distributed computation of equilibria in monotone Nash games via iterative regularization techniques. *SIAM Journal on Optimization*, *22*(4), 1177–1205.

Kannan, A., & Shanbhag, U. V. (2019). Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, *74*(3), 779–820.

Kannan, A., Shanbhag, U. V., & Kim, H. M. (2013). Addressing supply-side risk in uncertain power markets: stochastic Nash models, scalable algorithms and error analysis. *Optimization Methods & Software*, *28*(5), 1095–1138.

Khalil, H. K., & Grizzle, J. W. (2002). *Nonlinear systems, Vol. 3*. Upper Saddle River, NJ: Prentice hall.

Kleywegt, A. J., Shapiro, A., & Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479–502.

Knopp, K. (1990). *Theory and application of infinite series*. Courier Corporation.

Kohlenbach, U., Leuştean, L., & Nicolae, A. (2018). Quantitative results on Fejér monotone sequences. *Communications in Contemporary Mathematics*, *20*(02), Article 1750015.

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, *12*, 747–756.

Koshal, J., Nedic, A., & Shanbhag, U. V. (2013). Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, *58*(3), 594–609.

Koshal, J., Nedić, A., & Shanbhag, U. V. (2016). Distributed algorithms for aggregative games on graphs. *Operations Research*, *64*(3), 680–704.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kushner, H., & Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications, Vol. 35*. Springer Science & Business Media.

Lee, S., & Nedić, A. (2015). Asynchronous gossip-based random projection algorithms over networks. *IEEE Transactions on Automatic Control*, *61*(4), 953–968.

Lei, J., & Shanbhag, U. V. (2018). Distributed variable sample-size gradient-response and best-response schemes for stochastic Nash equilibrium problems over graphs. arXiv preprint arXiv:1811.11246.

Lei, J., Shanbhag, U. V., & Chen, J. (2020). Distributed computation of Nash equilibria for monotone aggregative games via iterative regularization. In *2020 59th IEEE conference on decision and control (CDC)* (pp. 2285–2290). IEEE.

Lei, J., Shanbhag, U. V., Pang, J.-S., & Sen, S. (2020). On synchronous, asynchronous, and randomized best-response schemes for stochastic Nash games. *Mathematics of Operations Research*, *45*(1), 157–190.

Lin, J., Rosasco, L., Villa, S., & Zhou, D.-X. (2018). Modified Fejér sequences and applications. *Computational Optimization and Applications*, *71*(1), 95–113.

Ljung, L., Pflug, G., & Walk, H. (2012). *Stochastic approximation and optimization of random systems, Vol. 17*. Birkhäuser.

Maingé, P.-E. (2007). Inertial iterative process for fixed points of certain quasi-nonexpansive mappings. *Set-Valued Analysis*, *15*(1), 67–79.

Maingé, P.-E. (2008). Convergence theorems for inertial KM-type algorithms. *Journal of Computational and Applied Mathematics*, *219*(1), 223–236.

Malitsky, Y. (2015). Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, *25*(1), 502–520.

Malitsky, Y. (2018). Golden ratio algorithms for variational inequalities. arXiv preprint arXiv:1803.08832.

Malitsky, Y. (2020). Golden ratio algorithms for variational inequalities. *Mathematical Programming*, *184*, 383–410.

Malitsky, Y., & Tam, M. K. (2018). A forward-backward splitting method for monotone inclusions without cocoercivity. arXiv preprint arXiv:1808.04162.

Malitsky, Y., & Tam, M. K. (2020). A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, *30*(2), 1451–1472.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., & Piliouras, G. (2018). Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. arXiv preprint arXiv:1807.02629.

Motzkin, T. S., & Schoenberg, I. J. (1954). The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, *6*, 393–404.

Naraghirad, E., Shi, L., & Wong, N.-C. (2020). The Bregman–Opial property and bregman generalized hybrid maps of reflexive Banach spaces. *Mathematics*, *8*(6), 1022.

Naraghirad, E., Wong, N.-C., & Yao, J.-C. (2014). Applications of bregman-opial property to bregman nonspreading mappings in Banach spaces. In *Abstract and applied analysis, Vol. 2014*. Hindawi.

Opial, Z., et al. (1967). Weak convergence of the sequence of successive approximations for nonexpansive mappings. *American Mathematical Society. Bulletin*, *73*(4), 591–597.

Pavel, L. (2019). Distributed GNE seeking under partial-decision information over networks via a doubly-augmented operator splitting approach. *IEEE Transactions on Automatic Control*, *65*(4), 1584–1597.

Peypouquet, J., & Sorin, S. (2010). An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems. *Journal of Convex Analysis*, *17*(3-4), 1113–1163.

Poggio, T., Voinea, S., & Rosasco, L. (2011). Online learning, stability, and stochastic gradient descent. arXiv preprint arXiv:1105.4701.

Polyak, B. (1987). *Introduction to optimization*. New York: Optimization Software, Inc..

Qin, X., Shang, M., & Su, Y. (2008). Strong convergence of a general iterative algorithm for equilibrium problems and variational inequality problems. *Mathematical and Computer Modelling*, *48*(7–8), 1033–1046.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Robbins, H., & Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics* (pp. 233–257). Elsevier.

Rockafellar, R. T. (1970). Monotone operators associated with saddle-functions and minimax problems. *Nonlinear Functional Analysis*, *18*(part 1), 397–407.

Ryu, E. K., & Boyd, S. (2016). Primer on monotone operator methods. *Applied and Computational Mathematics*, *15*(1), 3–43.

Schaefer, H. (1957). Über die methode sukzessiver approximationen. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, *59*, 131–140.

Scutari, G., & Sun, Y. (2019). Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, *176*(1), 497–544.

Shanbhag, U. V. (2013). Stochastic variational inequality problems: Applications, analysis, and algorithms. In *Theory driven by influential applications* (pp. 71–107). INFORMS.

Shapiro, A. (2003). Monte Carlo sampling methods. *Handbooks in Operations Research and Management Science*, *10*, 353–425.

Shapiro, A., & Xu, H. (2008). Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. *Optimization*, *57*(3), 395–418.

Shi, G., Johansson, M., & Johansson, K. H. (2013). How agreement and disagreement evolve over random dynamic networks. *IEEE Journal on Selected Areas in Communications*, *31*(6), 1061–1071.

Stroock, D. W. (2010). *Probability theory: An analytic view*. Cambridge University Press.

Tikhonov, A. (1963). On the solution of incorrectly put problems and the regularisation method. In *Outlines joint sympos. Partial differential equations (Novosibirsk, 1963)* (pp. 261–265).

Tseng, P. (2000). A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, *38*(2), 431–446.

Vũ, B. C. (2013). A variable metric extension of the forward–backward–forward algorithm for monotone operators. *Numerical Functional Analysis and Optimization*, *34*(9), 1050–1065.

Van Nguyen, Q. (2016). Variable quasi-Bregman monotone sequences. *Numerical Algorithms*, *73*(4), 1107–1130.

Van Nguyen, Q. (2017). Forward-backward splitting with Bregman distances. *Vietnam Journal of Mathematics*, *45*(3), 519–539.

Vu, B. C. (2016). Almost sure convergence of the forward–backward–forward splitting algorithm. *Optimization Letters*, *10*(4), 781–803.

Wang, M., & Bertsekas, D. P. (2015). Incremental constraint projection methods for variational inequalities. *Mathematical Programming*, *150*(2), 321–363.

Xu, Y. (1998). Ishikawa and Mann iterative processes with errors for nonlinear strongly accretive operator equations. *Journal of Mathematical Analysis and Applications*, *224*(1), 91–101.

Xu, H.-K. (2002). Iterative algorithms for nonlinear operators. *Journal of the London Mathematical Society*, *66*(1), 240–256.

Xu, H.-K. (2003). An iterative approach to quadratic optimization. *Journal of Optimization Theory and Applications*, *116*(3), 659–678.

Yi, P., & Pavel, L. (2019). An operator splitting approach for distributed generalized Nash equilibria computation. *Automatica*, *102*, 111–121.

Yousefian, F., Nedić, A., & Shanbhag, U. V. (2014). Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. In *53rd IEEE conference on decision and control* (pp. 5831–5836). IEEE.

Yousefian, F., Nedić, A., & Shanbhag, U. V. (2017). On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming*, *165*(1), 391–431.