On Smoothing, Regularization and Averaging in Stochastic Approximation Methods for Stochastic Variational Inequalities

Farzad Yousefian∗, Angelia Nedić† and Uday V. Shanbhag‡

November 4, 2014

Abstract

This paper considers stochastic variational inequality (SVI) problems where the mapping is merely monotone and not necessarily Lipschitz continuous. Traditionally, stochastic approximation schemes for SVIs have relied on strong monotonicity and Lipschitzian properties of the underlying map. In the first part of the paper, we weaken these assumptions by presenting a regularized smoothed SA (RSSA) scheme wherein the stepsize, smoothing, and regularization parameters are diminishing sequences updated after every iteration. Under suitable assumptions on the sequences, we show that the algorithm generates iterates that converge to a solution in an almost-sure sense. Additionally, we provide rate estimates that relate iterates to their counterparts derived from a smoothed Tikhonov trajectory associated with a deterministic problem. Motivated by the need to develop non-asymptotic rate statements, we consider a variant of the RSSA scheme, denoted by aRSSA, in which we employ weighted iterate-averaging, rather than the traditional averaging. First, we show that the gap function associated with the sequences by the aRSSA scheme tends to zero both in an almost sure and an expected-value sense. Second, we show that the gap function associated with the averaged sequence diminishes to zero at the optimal rate $O(1/\sqrt{k})$ when smoothing and regularization are suppressed. Third, we develop a window-based variant of this scheme that also displays the optimal rate and note the superiority in the constant factor of the bound when using an increasing set of weights rather than the traditional choice of decreasing weights seen in the classic averaging scheme. We conclude by presenting some preliminary numerical results on a stochastic Nash-Cournot game.

1 Introduction

Given a set $X$ and a mapping $F : X \rightarrow \mathbb{R}^n$, a variational inequality (VI) problem, denoted by $VI(X,F)$, requires a vector $x^* \in X$ such that $F(x^*)^T(x - x^*) \geq 0$ for all $x \in X$. Over the last several decades, variational inequality problems have been applied in capturing a wide range of optimization and equilibrium problems in engineering, economics, game theory, and finance.

∗Industrial & Manufacturing Engineering, Pennsylvania State University, University Park, State College, PA 16802, USA, szy5@psu.edu
†Industrial & Enterprise Systems Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, angelia@illinois.edu
‡Industrial & Manufacturing Engineering, Pennsylvania State University, University Park, State College, PA 16802, USA, udaybag@psu.edu; Nedić and Shanbhag gratefully acknowledge the support of the NSF through awards CMMI 0948905 ARRA (Nedić and Shanbhag), CMMI-0742538 (Nedić) and CMMI-1246887 (Shanbhag). A part of this paper has appeared in [37].
In this paper, we consider a stochastic generalization of this problem where the components of the mapping \( F \) are expectation-valued. More precisely, we are interested in solving \( \text{VI}(X,F) \) where mapping \( F : X \rightarrow \mathbb{R}^n \) represents the expected value of a stochastic mapping \( \Phi : X \times \Omega \rightarrow \mathbb{R}^n \), i.e., \( F(x) = \mathbb{E}[\Phi(x,\xi(\omega))] \) where \( \xi : \Omega \rightarrow \mathbb{R}^d \) is a \( d \)-dimensional random variable and \((\Omega,\mathcal{F},\mathbb{P})\) represents the probability space. Then \( x^* \in X \) solves \( \text{VI}(X,F) \) if

\[
\mathbb{E}[\Phi(x^*, \xi(\omega))]^T (x - x^*) \geq 0, \quad \text{for any } x \in X.
\]

For brevity, throughout this paper, \( \xi \) is used to denote \( \xi(\omega) \).

The stochastic variational inequality (SVI) problem (1) assumes relevance in a range of settings. Such models have immediate utility as they represent the (sufficient) optimality conditions of stochastic convex optimization problems [2, 32] as well as the equilibrium conditions of stochastic convex Nash games [28, 12, 15]. Such models find further applicability when the evaluation of the map is corrupted by errors. While SVIs represent a natural extension of their deterministic counterparts, generally deterministic schemes cannot be applied directly, particularly when the expectation cannot be evaluated efficiently or the underlying distribution \( \mathbb{P} \) is unavailable. Our interest lies in developing schemes that produce asymptotically exact solutions in precisely such regimes. A broad avenue for solving SVI problems is employing Monte-Carlo sampling methods. Of these, sample average approximation methods (SAA) and stochastic approximation methods (SA) are the most well-known approaches. In the context of SAA methods for solving stochastic optimization problems, asymptotic convergence of estimators and exponential rate analysis have been studied comprehensively by Shapiro [31]. Extensions to SVI problems have been provided by Xu in [35], where the exponential convergence rate of the estimators was established under more general assumptions on sampling. More recently, there has also been an effort to develop confidence statements for such problems [20, 21].

### Table 1: A comparison of stochastic approximation methods for solving SVIs (S: Strongly Monotone, *: Approximate solution, a.s.: almost sure, Y: Yes, N: No)

| Ref. | Algorithm | Monotonicity | Lipschitz | Metric | Convergence | Rate |
|------|-----------|--------------|-----------|--------|-------------|------|
| 14   | Mirror-descent (averaging) | Y | N | Gap | mean | \( \mathcal{O} \left( \frac{1}{k} \right) \) |
| 13   | Mirror-Prox (averaging, opt. problems) | Y | Y | Gap* | mean | \( \mathcal{O} \left( \frac{1}{k} \right) \) |
| 11   | Regularized SA | Y | Y | Soln | a.s. |
| 12   | Self-tuned smoothing SA | S | N | Soln* | a.s. |
| 10   | Incremental constraint projection | S | Y | Soln | a.s. |

This paper

| Algorithm | Monotonicity | Lipschitz | Metric | Convergence | Rate |
|-----------|--------------|-----------|--------|-------------|------|
| Regularized smoothing SA (HSSA) | Y | N | Soln | a.s. | |
| Averaging (aSAA) | Y | N | Gap | mean | \( \mathcal{O} \left( \frac{1}{k^2} \right) \) |
| Window-based Averaging (aRSSA\_r) | Y | N | Gap | mean | \( \mathcal{O} \left( \frac{1}{k^2} \right) \) |

A different tack is adopted by stochastic approximation (SA) schemes which were first introduced by Robbins and Monro [29] for stochastic root-finding problems that require an \( x^* \in \mathbb{R}^n \) such that \( \mathbb{E}[g(x,\xi)] = 0 \), where \( \xi : \Omega \rightarrow \mathbb{R}^d \) is a random variable, \( g(\cdot,\xi) : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a continuous map for any realization of \( \xi \) and \( \mathbb{E}[g(x,\xi)] \) is assumed to be monotone. The standard SA scheme is based on the iterative scheme \( x_{k+1} := x_k - \gamma_k g(x_k,\xi_k) \) for all \( k \geq 0 \), where \( \gamma_k > 0 \) denotes the stepsize while \( \xi_k \) represents a realization of a random variable \( \xi \) at the \( k \)-th iteration. The SA schemes have been applied extensively in solving convex stochastic optimization problems [6, 7, 36, 4, 8]. There has been a surge of interest in the solution of SVIs via stochastic approximation schemes. Amongst the earliest work was by Jiang and Xu [11], who considered SVIs with strongly monotone and Lipschitz continuous maps over a closed and convex set and proved that the sequence of solution iterates converge to the unique solution in an almost sure sense. In an extension of that
work, motivated by Tikhonov regularization scheme, a regularized SA method was developed for solving SVIs with a merely monotone but continuous mapping [16]. A comprehensive summary of the various schemes for solving SVIs via SA schemes is provided in Table 1.

The first part of this paper is motivated by the need to weaken Lipschitzian and strong monotonicity requirements. Employing a smoothing technique, our first goal is to weaken the typical conditions for almost sure convergence of SA methods by allowing for a non-Lipschitzian mapping. Our work is motivated by a class of averaged functions first introduced by Steklov [33], employed for the solution of stochastic optimization problems [1, 25, 18, 5]. It is well-known that given a convex function \( f(x) : \mathbb{R}^n \to \mathbb{R} \) and a random variable \( \omega \) with probability distribution \( P(\omega) \), the function \( \hat{f} \) defined by \( \hat{f}(x) \triangleq \int_{\mathbb{R}^n} f(x + \omega)P(\omega)d\omega = \mathbb{E}[f(x + \omega)] \) is a differentiable function. Employing this technique allowed us to address non-smoothness in developing adaptive stepsize SA schemes for stochastic convex optimization problems and Cartesian SVIs in absence or unavailability of a Lipschitz constant [36, 38]. To accommodate merely monotone maps, we introduce a regularization term, inspired by Tikhonov regularization [7] and their iterative counterparts [9, 14, 16]. In the resulting scheme, referred to as a \textit{regularized smoothed stochastic approximation} (RSSA) scheme in which both the smoothing parameter and the regularization parameter are updated after every iteration and are driven to zero in the limit. Note that this is in contrast with our prior work where the smoothing parameter is assumed to be fixed, here the smoothing parameter is updated at each iteration and converges to zero, allowing us to prove convergence to the solution set of the original SVI rather than an approximate problem. The key distinction with prior schemes is that the RSSA scheme can cope with merely monotone SVIs without a Lipschitzian assumption on the map and are equipped with almost sure asymptotic convergence guarantees. Unfortunately, we cannot derive non-asymptotic rate statements without reverting to averaging and this forms the focus of the second part of the paper.

In the second part of this paper, motivated by the need to derive rate statements, we consider counterparts of the RSSA scheme where we examine the averaged sequence. Averaging approaches have proved very useful in constructing subgradient methods (cf. [10]) as well as in developing rate statements (cf. [27, 13]). In such methods, convergence analysis and rate statements are provided for an averaged sequence \( \bar{x}_k \) defined as \( \bar{x}_k = \sum_{t=0}^{k-1} \tilde{\gamma}_t x_t \), where

\[
\tilde{\gamma}_t \triangleq \frac{\gamma_t}{\sum_{i=0}^{k-1} \tilde{\gamma}_i}
\]

and \( x_k \) is generated via a standard SA scheme. An issue that emerges when implementing this scheme is that when the stepsize sequence \( \gamma_k \) is decreasing, the averaging weights \( \tilde{\gamma}_k \) are decreasing as well. This implies that the recent iterates \( x_k \) are assigned less weight than the original iterates. Therefore, it may make sense to consider an increasing set of weights. In fact, it is shown in [23] that by using weights of the form

\[
\tilde{\gamma}_t \triangleq \frac{\gamma_t^{-1}}{\sum_{i=0}^{k-1} \gamma_i^{-1}},
\]

the subgradient mirror-descent algorithm attains the optimal rate of convergence without requiring window-based averaging. Motivated by this result, we consider an averaged sequence of the form \( \bar{x}_k \triangleq \sum_{t=0}^{k-1} \tilde{\gamma}_{t,r} x_t \) where

\[
\tilde{\gamma}_{t,r} \triangleq \frac{\gamma_t^r}{\sum_{i=0}^{k-1} \gamma_i^r}.
\] \hspace{1cm} (2)

We proceed to show that the optimal rate of convergence \( O(1/\sqrt{k}) \) for the mean gap function is attained for choices of \( r < 1 \). Note that when \( r = 1 \), we recover the standard averaging scheme for
which window-based averaging scheme displays the optimal rate of convergence. We extend this result to the case when \( r = -1 \) and derive the optimal convergence rate. Furthermore, we discuss the improvement of finite time behavior over the window-based scheme when \( r = -1 \). We now outline our main contributions:

(a) **Convergence analysis for RSSA scheme in monotone non-Lipschitzian regimes:** We consider SVIs where the mapping is monotone and not necessarily Lipschitz continuous. A regularized smoothing SA scheme, referred to as the RSSA scheme, is developed wherein the regularization parameter, smoothing parameter, and the steplength are updated after each iteration. We proceed to show that under suitable assumptions on the smoothing, regularization, and steplength sequences, the sequence of iterates converges to the solution set of the SVI in an almost sure sense, which is in contrast with almost all available almost sure convergence results (that typically require Lipschitz continuity). We also proceed to derive a bound on the mean-squared distance of any iterate produced by the RSSA scheme to the regularized smoothed trajectory which is provably convergent to the solution set of the original SVI.

(b) **Optimal averaging schemes:** Motivated by the need to develop non-asymptotic rate statements, we first consider an averaging-based extension of the RSSA scheme, referred to as aRSSA, in which \( \bar{x}_k \) is defined as a weighted average: \( \bar{x}_k(r) \triangleq \sum_{t=0}^{k-1} \gamma_{t,r} x_t \) where \( r \in \mathbb{R} \) and \( \gamma_{t,r} \) is defined by (2). We derive the underlying conditions under which the mean gap function of the averaged sequence \( \bar{x}_k(r) \) converges to zero. Additionally, we show that the aRSSA scheme produces a sequence of iterates that converges to the solution set in an almost sure sense under prescribed conditions. When both regularization and smoothing are suppressed and \( \gamma_k = 1/\sqrt{k} \), we further show that the mean gap function diminishes to zero at the optimal rate of \( O(1/\sqrt{k}) \). Notably, when a window-based averaging sequence is employed, we show that for both \( r = -1 \) and \( r = 1 \) the averaging scheme recovers the optimal rate. We also provide a comparison for an estimate of the error bound between the case \( r = -1 \) and \( r = 1 \).

(c) **Numerics:** Preliminary numerics on a set of stochastic Nash-Cournot games support the theoretical findings. In particular, we observe that the RSSA scheme displays almost sure convergence and is relatively robust to choices of the parameter sequences. The averaged variants are seen to perform well from the standpoint of the mean gap function. Importantly, choosing \( r < 1 \) has significant benefits in terms of finite-time behavior.

The rest of the paper is organized as follows. Section 2 presents the RSSA scheme, its averaged variants, and our main assumptions. In Section 3, we prove the almost sure convergence of the RSSA scheme while in Section 4, we analyze the convergence and derive the rate for the averaged variants of the RSSA scheme. In Section 5, the performance of the proposed methods are tested on a stochastic Nash-Cournot game. The paper ends with some concluding remarks in Section 6.

**Notation:** A vector \( x \) is assumed to be a column vector, \( x^T \) denotes the transpose of a vector \( x \), and \( \|x\| \) denotes the Euclidean vector norm, i.e., \( \|x\| = \sqrt{x^T x} \). We use \( \Pi_X(x) \) to denote the Euclidean projection of a vector \( x \) on a set \( X \), i.e., \( \|x - \Pi_X(x)\| = \min_{y \in X} \|x - y\| \). We abbreviate “almost surely” as a.s. and \( E[z] \) is used to denote the expectation of a random variable \( z \). We let \( \text{dist}(s,S) \) denote the Euclidean distance of a vector \( s \in \mathbb{R}^n \) from a set \( S \subset \mathbb{R}^n \), and let \( \text{SOL}(X,F) \) denote the solution set of VI\((X,F)\) for a set \( X \) and a mapping \( F \). For brevity, when the set \( X \) and mapping \( F \) are given by (1), we let \( X^* \) denote \( \text{SOL}(X,F) \). We use \( B_n(y,\rho) \) to denote the ball centered at a point \( y \) with a radius \( \rho \), i.e., \( B_n(y,\rho) = \{ x \in \mathbb{R}^n \mid \|x - y\| \leq \rho \} \).
2 Algorithm and assumptions

We present our stochastic approximation schemes of interest in Section 2.1 while the main assumptions are outlined in Section 2.2.

2.1 Algorithm

In this section, we present the regularized smoothing stochastic approximation (RSSA) scheme for solving (1). We motivate our scheme by first defining the traditional stochastic approximation scheme for SVIs. Given an \( x_0 \in X \), the standard SA scheme generates a sequence \( \{x_k\} \):

\[
x_{k+1} := \Pi_X (x_k - \gamma_k \Phi(x_k, \xi_k)), \quad k \geq 0,
\]

where \( \{\gamma_k\} \) defines a steplength sequence, while \( x_0 \in X \) is an initial random vector independent of the random variables \( \xi_k \) and such that \( E[\|x_0\|^2] < \infty \). This SA scheme for SVIs appears to have been first studied by Jiang and Xu \[11\] where a.s. convergence statements were provided under Lipschitz continuity and strong monotonicity of the map. In deterministic variational inequality problems, Tikhonov regularization techniques have proved useful for solving merely monotone problems through the generation of increasingly accurate solutions of a sequence of a regularized VIs (cf. \[7\]). Unfortunately, in stochastic regimes, such an approach is not practical since it requires running a sequence of simulations of increasingly longer lengths. Inspired by prior work in deterministic VIs \[9, 14\], the stochastic iterative Tikhonov regularization scheme was developed subsequently \[16\]. The regularized stochastic approximation scheme (RSA) is defined as follows:

\[
x_{k+1} := \Pi_X (x_k - \gamma_k (\Phi(x_k, \xi_k) + \eta_k x_k)), \quad k \geq 0,
\]

where \( \{\eta_k\} \) denotes a regularization sequence that is driven to zero at specified rates to ensure a.s. convergence of the sequence of iterates to the least norm solution of the monotone stochastic variational inequality problem. However, the RSA scheme requires Lipschitz continuity of the map. In prior work, in the context of nonsmooth stochastic optimization \[36\], we have employed local smoothing to construct an approximate problem with a prescribed Lipschitz constant. Such a problem can then be solved via standard SA schemes. However, this avenue provides only approximate solutions. In this paper, we resolve this shortcoming by presenting a smoothed variant of the RSA scheme, referred to as the regularized smoothed SA (or RSSA) scheme under which we can recover solutions to the original problem without requiring Lipschitz continuity of the map:

\[
x_{k+1} := \Pi_X (x_k - \gamma_k (\Phi(x_k + z_k, \xi_k) + \eta_k x_k)), \quad k \geq 0,
\]

where \( z_k \in \mathbb{R}^n \) is a uniform random variable over an \( n \)-dimensional ball centered at the origin with radius \( \epsilon_k \) for any \( k \geq 0 \). To have a well defined \( \Phi \) in the RSSA scheme, we define \( X^\epsilon \) as \( \epsilon \)-enlargement of the set \( X \), i.e.,

\[
X^\epsilon \triangleq X + B_n(0, \epsilon),
\]

where \( \epsilon \) is an upper bound of the sequence \( \{\epsilon_k\} \) (which will be finite under our assumptions). Note that by introducing stochastic errors \( w_k \), the RSSA scheme is equivalent to the following method:

\[
x_{k+1} = \Pi_X (x_k - \gamma_k (F(x_k + z_k) + \eta_k x_k + w_k)), \quad k \geq 0
w_k \triangleq \Phi(x_k + z_k, \xi_k) - F(x_k + z_k), \quad k \geq 0.
\]

(RSSA)
In this representation of the RSSA scheme, \( w_k \) is the deviation between the sample \( \Phi(x, \xi_k) \) observed at the \( k \)-th iteration and the expected-value mapping \( F(x) \), at \( x = x_k + z_k \). An implicit assumption in our work is that we have access to a stochastic oracle which is able to generate random samples \( \Phi(\cdot, \xi_k) \) at a given point. Such an oracle is assumed to be an unbiased estimator, meaning that \( F(x) = E[\Phi(x, \xi)] \) for any \( x \in X^\epsilon \). The results in this paper can be extended to the case where \( \Phi \) is a biased estimator of the mapping \( F \), i.e., \( F(x) = E[\Phi(x, \xi)] + b \) for some \( b > 0 \) and all \( x \in X^\epsilon \).

The RSA and RSSA schemes in their presented forms do not easily allow for determination of non-asymptotic rates of convergence. This may be provided by constructing averaging counterparts \[27\] as well as their variational inequality counterparts \[13\]. Strictly speaking, averaging schemes are not distinct algorithmically but merely average the generated sequence of iterates. In contrast with the traditional averaging approach, we consider \textit{weighted averaging} akin to recent work in stochastic optimization \[23\] by defining the sequence \( \bar{x}_k(r) \) for \( k \geq 0 \) and \( r \in \mathbb{R} \). The \textit{averaged} variant of the RSSA scheme using the parameter \( r \) (referred to as aRSSA\(_r \)) is defined as follows:

\[
x_{k+1} := \Pi_X(x_k - \gamma_k(\Phi(x_k + z_k, \xi_k) + \eta_k x_k)),
\]

\[
\bar{x}_{k+1}(r) \triangleq \frac{\sum_{t=0}^{k} \gamma_t^r x_t}{\sum_{t=0}^{k} \gamma_t^r},
\]

Throughout the paper, when we suppress the regularization and smoothing, we refer to the aRSSA\(_r \) algorithm as aSA\(_r \). Finally, variants of the above scheme prescribe a \textit{window} over which the averaging is carried out. Denoted by aSA\(_{\ell,r} \), we define such a scheme next:

\[
x_{k+1} := \Pi_X(x_k - \gamma_k \Phi(x_k, \xi_k)),
\]

\[
\bar{x}_{k+1}(r) \triangleq \frac{\sum_{t=\ell}^{k} \gamma_t^r x_t}{\sum_{t=\ell}^{k} \gamma_t^r},
\]

where \( 0 < \ell \leq k \) and \( k \geq 1 \).

### 2.2 Assumptions

We now outline the key assumptions employed in the remainder of this paper. Let \( F_k \) denote the history of the method up to time \( k \), i.e., \( F_k = \{x_0, \xi_0, \ldots, x_{k-1}, z_1, \ldots, z_{k-1}\} \) for \( k \geq 1 \) and \( F_0 = \{x_0\} \). Our first set of assumptions is on the properties of the set \( X \), the mapping \( F \), and the non-emptiness of the solution set \( X^* \).

**Assumption 1** (Problem properties). Let the following hold:

(a) The set \( X \subset \mathbb{R}^n \) is closed, bounded, and convex;

(b) The mapping \( F(x) = E[\Phi(x, \xi)] \) is a monotone and continuous over the set \( X^\epsilon \) given in \[3\];

(c) There exists a scalar \( C > 0 \) such that \( E[\|\Phi(x, \xi)\|^2] \leq C^2 \) for any \( x \in X^\epsilon \);

(d) \( X^* \neq \emptyset \), i.e., there exists an \( x^* \in X \) such that \( (x - x^*)^T E[\Phi(x^*, \xi)] \geq 0 \) for all \( x \in X \).

**Remark 1.** Note that by using Jensen’s inequality and Assumption 1(c) we can write

\[
\|F(x)\| \leq E[\|\Phi(x, \xi)\|] = \sqrt{E[\|\Phi(x, \xi)\|^2]} \leq \sqrt{E[\|\Phi(x, \xi)\|^2]} \leq \sqrt{C^2} = C.
\]

In our analysis, we make use of the preceding inequality, i.e.

\[
\|F(x)\| \leq E[\|\Phi(x, \xi)\|] \leq C, \quad \text{for all } x \in X^\epsilon.
\]

We also use the boundedness of \( X \) by which there exists \( M > 0 \) such that

\[
\|x\| \leq M \quad \text{for all } x \in X.
\]
In the implementation of the RSSA scheme, two distinct random variables require discussion. First, the random vector $\xi$ is inherent to problem (1) while the random vector $z$ is an artificially introduced random variable. Next, we provide some assumptions on these two random variables.

**Assumption 2** (Random variables $\xi$ and $z$). Let the following hold:

(a) The random variables $\xi_j \in \mathbb{R}^d$ are identically distributed and independent for any $j \geq 0$.

(b) Random variables $z_i \in \mathbb{R}^n$ are independent and uniformly distributed in an $n$-dimensional ball with radius $\epsilon_i$ centered at the origin for any $i \geq 0$.

(c) The random variables $z_i$ and $\xi_j$ are independent from each other for any $i, j \geq 0$.

Based on this assumption, we may derive the following regarding the conditional first and second moments of $w_k$.

**Lemma 1** (Conditional first and second moments of $w_k$). Consider the RSSA scheme and suppose Assumptions 1(c) and 2 hold. Then, the stochastic error $w_k$ satisfies the following relations for any $k \geq 0$:

$$\mathbb{E}[w_k \mid \mathcal{F}_k \cup \{z_k\}] = 0 \quad \text{and} \quad \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k \cup \{z_k\}] \leq C^2.$$ 

Furthermore, for any $k \geq 0$,

$$\mathbb{E}[w_k \mid \mathcal{F}_k] = 0 \quad \text{and} \quad \mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq C^2.$$ 

**Proof.** Let us assume that $k \geq 0$ is fixed. The definition of $w_k$ in RSSA implies that

$$\mathbb{E}[w_k \mid \mathcal{F}_k \cup \{z_k\}] = \mathbb{E}[\Phi(x_k + z_k, \xi_k) \mid \mathcal{F}_k \cup \{z_k\}] - F(x_k + z_k) = F(x_k + z_k) - F(x_k + z_k) = 0,$$

where we used the independence of $z_k$ and $\xi_k$. By taking the expectation with respect to $z_k$, we immediately obtain $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$. For the term $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k \cup \{z_k\}]$ using Assumption 1(c), we may write

$$\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k \cup \{z_k\}] = \mathbb{E}[\|\Phi(x_k + z_k, \xi_k) - F(x_k + z_k)\|^2 \mid \mathcal{F}_k \cup \{z_k\}]$$

$$= \mathbb{E}[\|\Phi(x_k + z_k, \xi_k)\|^2 \mid \mathcal{F}_k \cup \{z_k\}] + \|F(x_k + z_k)\|^2$$

$$- 2\mathbb{E}[\Phi(x_k + z_k, \xi_k) \mid \mathcal{F}_k \cup \{z_k\}]^T F(x_k + z_k). \tag{5}$$

Using Assumption 1(c), we observe that Term 1 $\leq C^2$. Furthermore, we have

$$\text{Term 2} = F(x_k + z_k)^T F(x_k + z_k) = \|F(x_k + z_k)\|^2. \tag{6}$$

Therefore, from relations (5) and (6) we obtain

$$\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k \cup \{z_k\}] \leq C^2 - \|F(x_k + z_k)\|^2 \leq C^2.$$ 

Taking expectation with respect to $z_k$ implies that $\mathbb{E}[\|w_k\|^2 \mid \mathcal{F}_k] \leq C^2$. \hfill $\square$

### 3 Convergence analysis of RSSA scheme

In this section, we prove the almost sure convergence (Section 3.1) and mean-squared convergence (Section 3.2) of the RSSA scheme.
3.1 Almost sure convergence

In this subsection, we show that the sequence produced by the RSSA scheme converges to $X^*$ in an almost sure sense. There are three user-defined sequences in the RSSA scheme: the stepsize sequence denoted by $\gamma_k$, the regularization sequence denoted by $\eta_k$, and the smoothing sequence denoted by $\epsilon_k$. At each iteration, all three parameters are updated. To guarantee convergence, the tuning sequences are required to decay to zero at prescribed rates. For example, if the stepsize sequence decays to zero faster than the other two sequences, the solution iterate $x_k$ may not converge to $X^*$. Next, we present the underlying conditions on these sequences that will ultimately be used to prove a.s. convergence.

**Assumption 3.** Let the following hold:

(a) $\{\gamma_k\}, \{\eta_k\},$ and $\{\epsilon_k\}$ are positive sequences for $k \geq 0$ converging to zero;

(b) There exists $K_1 \geq 0$ such that $\frac{n}{\eta_k \epsilon_k} \leq 0.5 \left(\frac{(n-1)!!}{n!!}\right)^2$ for any $k \geq K_1$, where $n$ is the dimension of the space and $\kappa = 1$ if $n$ is odd and $\kappa = \frac{2}{\pi}$ otherwise;

(c) For any $k \geq 0$, $\epsilon_k \leq \epsilon$, where $\epsilon$ is the parameter of the set $X^*$;

(d) $\sum_k \gamma_k \eta_k = \infty$;

(e) $\sum_k \gamma_k^2 \leq \infty$;

(f) $\sum_k \frac{1}{\eta_k \gamma_k^2} \left(1 - \frac{\min\{\epsilon_k, \epsilon_{k-1}\}}{\max\{\epsilon_k, \epsilon_{k-1}\}}\right)^2 < \infty$;

(g) $\sum_k \frac{1}{\eta_k \gamma_k^2} \left(1 - \frac{\min\{\epsilon_k, \epsilon_{k-1}\}}{\max\{\epsilon_k, \epsilon_{k-1}\}}\right)^2 < \infty$;

(h) $\lim_{k \to \infty} \frac{n_k}{\epsilon_k} = 0$; (i) $\lim_{k \to \infty} \frac{1}{\eta_k \gamma_k^2} \left(1 - \frac{\min\{\epsilon_k, \epsilon_{k-1}\}}{\max\{\epsilon_k, \epsilon_{k-1}\}}\right) = 0$; (j) $\lim_{k \to \infty} \frac{1}{\eta_k \gamma_k} \left|1 - \frac{\eta_k}{\eta_{k-1}}\right| = 0$.

Assumption 3 imposes a set of conditions on the tuning sequences that ensure a.s. convergence of the RSSA scheme. The notation “!!” in the condition in (b) denotes the double factorial. Later on, in Lemma 5, we provide acceptable choices for the sequences $\{\gamma_k\}, \{\eta_k\},$ and $\{\epsilon_k\}$ that satisfy the conditions of Assumption 3.

**Remark 2.** If we neither regularize nor smooth, then our scheme reduces to the SA scheme and the necessary conditions for the almost-sure convergence would be $\sum_k \gamma_k = \infty$, $\sum_k \gamma_k^2 < \infty$, as well as the Lipschitzian property and the strong monotonicity of the mapping $F$ (cf. [11]).

In our analysis, we use a family of approximate smoothed mappings defined as follows.

**Definition 1** (Smoothed mapping). Consider mapping $F : X^\epsilon \to \mathbb{R}^n$. Let $z_k \in \mathbb{R}^n$ be a uniform random vector in $B_n(0, \epsilon_k)$ for all $k \geq 0$. The smoothed (approximate) mapping $F_k : X \to \mathbb{R}^n$ is defined by

$$F_k(x) = \mathbb{E}[F(x + z_k)] , \; \text{for any} \; x \in X.$$ 

Next, we prove the monotonicity and Lipschitz continuity of the smoothed mapping.

**Lemma 2** (Properties of smoothed mapping). Consider the smoothed mapping $F_k$ as given in Definition 1. Then, the following hold:

(a) Let $\{x_t\} \subset X$ be a convergent sequence in $X$ i.e., such that $\lim_{t \to \infty} x_t = \hat{x}$ with $\hat{x} \in X$. Also, let $F$ be continuous on the set $X^\epsilon$. If Assumption 3(c) holds and $\epsilon_t \to 0$, then

$$\lim_{t \to \infty} F_t(x_t) = F(\hat{x}).$$

(b) Let Assumptions 3(a) and 3(c) hold. For any $k \geq 0$, the mapping $F_k$ is Lipschitz continuous over the set $X$ with the parameter $\kappa = \frac{n!!}{(n-1)!!} \frac{C}{\epsilon_k}$, where $\kappa = 1$ if $n$ is odd and $\kappa = \frac{2}{\pi}$ otherwise.
(c) If the mapping $F: X^\epsilon \to \mathbb{R}^n$ is monotone over the set $X^\epsilon$, then the mapping $F_k$ is monotone over the set $X$.

Proof. Using the definition of $F_t$ and letting $c_n$ be the volume of the $n$-dimensional unit ball, i.e., $c_n = \int_{\|y\| \leq 1} dy$, we have

$$\lim_{t \to \infty} \mathbb{E}[F(x_t + z_t)] = \lim_{t \to \infty} \int_{\|z\| \leq \epsilon_t} F(x_t + z) \frac{1}{c_n \epsilon_t^d} dz.$$ 

By change of variables $y = \frac{z}{\epsilon_t}$, it follows that

$$\lim_{t \to \infty} \mathbb{E}[F(x_t + z_t)] = \frac{1}{c_n} \lim_{t \to \infty} \int_{\|y\| \leq 1} F(x_t + \epsilon_t y) dy.$$ 

By Assumption 1(c) we have that $\|F(x + z)\| \leq C$ (see Remark 1), implying that $F(x + z)$ is integrably bounded with respect to the distribution defining the random variable $z$. Thus, by appealing to Lebesgue’s dominated convergence theorem, we interchange the limit and the integral leading to the following relations:

$$\lim_{t \to \infty} \mathbb{E}[F(x_t + z_t)] = \frac{1}{c_n} \lim_{t \to \infty} \int_{\|y\| \leq 1} F(x_t + \epsilon_t y) dy = \frac{1}{c_n} \int_{\|y\| \leq 1} F(\hat{x}) dy,$$

where the last equality follows by the continuity of the mapping $F$, and $x_t \to \hat{x}$, $\epsilon_t \to 0$. Finally, we may conclude that the above integral reduces to $F(\hat{x})$ by invoking the definition of $c_n$ as the volume of $B_n(0, 1)$.

(b) Let $p_u$ denote the probability density function of the random vector $z$ and suppose $k \geq 0$ is fixed. From the definition of $F_k$, for any $x, y \in X$,

$$\|F_k(x) - F_k(y)\| = \left\| \int_{\mathbb{R}^n} F(x + z_k)p_u(z_k)dz_k - \int_{\mathbb{R}^n} F(y + z_k)p_u(z_k)dz_k \right\|.$$

By changing the integral variable in the preceding relation, we obtain

$$\|F_k(x) - F_k(y)\| = \left\| \int_{\mathbb{R}^n} (p_u(v - x) - p_u(v - y))F(v)dv \right\| \leq \int_{\mathbb{R}^n} |p_u(v - x) - p_u(v - y)||F(v)||dv \leq C \int_{\mathbb{R}^n} |p_u(v - x) - p_u(v - y)|dv,$$

where the second inequality follows from Jensen’s inequality and the last inequality is a consequence of boundedness of the mapping $F$ over $X^\epsilon$. The remainder of the proof is similar to the proof of Lemma 8 of [36].

(c) Since $F$ is monotone over $X^\epsilon$ we have that

$$(a - b)^T (F(a) - F(b)) \geq 0, \text{ for all } a, b \in X^\epsilon.$$ 

Therefore, for choice of $x + z_k$ and $y + z_k$ in $X^\epsilon$, we have that

$$((x + z_k) - (y + z_k))^T (F(x + z_k) - F(y + z_k)) \geq 0, \text{ for all } x, y \in X.$$ 

It follows that

$$(x - y)^T (F(x + z_k) - F(y + z_k)) \geq 0, \text{ for all } x, y \in X.$$ 

Now, taking expectations on both sides of the preceding relation, the monotonicity of $F_k$ follows from

$$(x - y)^T (F_k(x) - F_k(y)) \geq 0, \text{ for all } x, y \in X.$$ 

\[\square\]
Remark 3. Note that the ratio $\frac{n^2}{(n-1)^2}$ in Lemma 2(b) is of the order $\sqrt{n}$. Lemma 2(c) implies that the mapping $F_k + \eta_k I$ is strongly monotone for any $\eta_k > 0$. In view of Lemma 2(c), when $X$ is closed and convex, Theorem 2.3.3(b) of [7], page 156, ensures that $VI(X, F_k + \eta_k I)$ has a unique solution.

In the following, we define the sequence $\{s_k\}$ in which every iterate is a solution to a regularized smoothed approximation of the original SVI. This sequence forms the basis for proving the almost-sure convergence of the iterates generated by the RSSA scheme. Let $t^\ast$ denote the least norm solution of $VI(X, F)$, i.e. $t^\ast = \text{argmin}_{x \in X^\ast} \|x\|$. Note that $t^\ast$ is unique since it is the projection of the origin on the convex and bounded set $X$. Specifically, we first show that the sequence $\{s_k\}$ has its accumulation points in $X^\ast$ (Proposition 1) and identify conditions that ensure that this sequence converges to $t^\ast$, the smallest-norm solution of $VI(X, F)$. Then, we proceed to derive a bound on the difference between $x_k$ and $s_k$ (Lemma 3). By utilizing this bound, we show that $\|x_{k+1} - s_k\| \to 0$ as $k \to \infty$ in an almost sure sense (Theorem 1).

Definition 2 (Solution of the smoothed regularized problem). For each $k \geq 0$, let $s_k$ be the unique solution of $VI(X, F_k + \eta_k I)$, where $F_k : X \to \mathbb{R}^n$ is given by Definition 2 and $\eta_k > 0$ is the regularization parameter.

Recall that the $VI(X, F_k + \eta_k I)$ is a strongly monotone deterministic variational inequality since $F_k$ was shown to be a monotone map. We now present a bound on $\|s_k - s_{k-1}\|$ and proceed to prove that the sequence $\{s_k\}$ of approximate solutions has accumulation points in the set $X^\ast$.

Proposition 1 (Convergence of $\{s_k\}$). Suppose Assumptions 2 and 3 hold. Consider the sequence $\{s_k\}$ of solutions $s_k$ given in Definition 3. Then,

(a) For any $k \geq 1$,

$$\|s_k - s_{k-1}\| \leq \frac{2nC}{\eta_k - 1} \left(1 - \frac{\min \{\epsilon_k, \epsilon_{k-1}\}}{\max \{\epsilon_k, \epsilon_{k-1}\}}\right) + M \left|1 - \frac{\eta_k}{\eta_{k-1}}\right|,$$

where $M$ and $C$ are the bounds on $X$ and $F$ (see Remark 2 and Assumption 2(c), respectively).

(b) Suppose that the sequences $\{\eta_k\}$ and $\{\epsilon_k\}$ tend to zero, i.e., $\lim_{k \to \infty} \epsilon_k = 0$ and $\lim_{k \to 0} \eta_k = 0$. Then, we have the following:

(1) $\{s_k\}$ has an accumulation point and every accumulation point of $\{s_k\}$ is a solution to $VI(X, F)$;

(2) If $\lim_{k \to \infty} \frac{\epsilon_k}{\eta_k} = 0$ and $F$ is differentiable at $t^\ast$ with a bounded Jacobian in a neighborhood of $t^\ast$, then $\{s_k\}$ converges to the smallest norm solution of $VI(X, F)$.

Proof. (a) Suppose $k \geq 1$ is fixed. Since $s_k \in \text{SOL}(X, F_k + \eta_k I)$ and $s_{k-1} \in \text{SOL}(X, F_{k-1} + \eta_{k-1} I)$,

$$(s_{k-1} - s_k)^T(F_k(s_k) + \eta_k s_k) \geq 0 \text{ and } (s_k - s_{k-1})^T(F_{k-1}(s_{k-1}) + \eta_{k-1} s_{k-1}) \geq 0.$$

Adding the preceding relations, yields

$$(s_{k-1} - s_k)^T(F_k(s_k) - F_{k-1}(s_{k-1}) + \eta_k s_k - \eta_{k-1} s_{k-1}) \geq 0.$$

By adding and subtracting $(s_{k-1} - s_k)^T(F_{k-1}(s_k) + \eta_{k-1} s_k)$, we obtain that

$$(s_{k-1} - s_k)^T(F_k(s_k) - F_{k-1}(s_k) + (s_{k-1} - s_k)^T(F_{k-1}(s_k) - F_{k-1}(s_{k-1})))$$

$$+ (\eta_k - \eta_{k-1})(s_{k-1} - s_k)^T s_k - \eta_{k-1} s_k - s_{k-1} \|s_k - s_{k-1}\|^2 \geq 0.$$

By monotonicity of $F_{k-1}$, it follows that $(s_{k-1} - s_k)^T(F_{k-1}(s_k) - F_{k-1}(s_{k-1})) \leq 0$, thus implying

$$(s_{k-1} - s_k)^T(F_k(s_k) - F_{k-1}(s_k)) + (\eta_k - \eta_{k-1})(s_{k-1} - s_k)^T s_k \geq \eta_{k-1} \|s_k - s_{k-1}\|^2.$$
By the Cauchy-Schwartz inequality and by recalling that \(\|s_k\| \leq M\) (see Remark 1), we obtain

\[
\eta_{k-1} \|s_k - s_{k-1}\| \leq \|F_k(s_k) - F_{k-1}(s_k)\| + M|\eta_{k-1} - \eta_k|.
\]  

(7)

Let \(p_u\) denote the density function of the random vector \(z\) uniformly distributed over the ball \(B_n(0, \epsilon)\), i.e., \(p_u(z) \triangleq 1/(c_n \epsilon^n)\) for any \(z \in B_n(0, \epsilon)\), where \(c_n \triangleq \pi^{n/2}/\Gamma(n/2 + 1)\). To estimate the term \(\|F_k(s_k) - F_{k-1}(s_k)\|\), we consider two cases based on whether \(\epsilon_k\) is less than \(\epsilon_{k-1}\) or not.

(i) \(\epsilon_k \leq \epsilon_{k-1}\): We begin by showing that \(\|F_k(s_k) - F_{k-1}(s_k)\|\) can be expressed as follows:

\[
\|F_k(s_k) - F_{k-1}(s_k)\| = \left\| \int_{\mathbb{R}^n} F(s_k + z) p_u(z) dz - \int_{\mathbb{R}^n} F(s_{k-1} + z) p_u(z) dz \right\|
\]

where in the third equality, we note that \(\{z \in \mathbb{R}^n \mid \|z\| < \epsilon_k\} = \{z \in \mathbb{R}^n \mid \|z\| < \epsilon_k\} \cup \{z \in \mathbb{R}^n \mid \epsilon_k \leq \|z\| < \epsilon_{k-1}\}\) when \(\epsilon_k \leq \epsilon_{k-1}\). The right hand side may be further bounded as follows:

\[
\left\| \int_{\|z\| < \epsilon_k} \frac{F(s_k + z)}{c_n \epsilon_k^n} \ dz - \left( \int_{\|z\| < \epsilon_k} \frac{F(s_k + z)}{c_n \epsilon_{k-1}^n} \ dz + \int_{\epsilon_k \leq \|z\| < \epsilon_{k-1}} \frac{F(s_k + z)}{c_n \epsilon_{k-1}^n} \ dz \right) \right\|
\]

where in the last two inequalities, we use the triangle inequality and Jensen’s inequality respectively. Invoking relation (4), we obtain

\[
\|F_k(s_k) - F_{k-1}(s_k)\| \leq C \int_{\|z\| < \epsilon_k} \left| \frac{1}{c_n \epsilon_k^n} - \frac{1}{c_n \epsilon_{k-1}^n} \right| \ dz + C \int_{\epsilon_k \leq \|z\| < \epsilon_{k-1}} \frac{1}{c_n \epsilon_{k-1}^n} \ dz
\]

\[
= C \left( \frac{1}{c_n \epsilon_k^n} - \frac{1}{c_n \epsilon_{k-1}^n} \right) + C \left( \frac{1}{c_n \epsilon_{k-1}^n - c_n \epsilon_{k-1}^n} \right) = 2C \left( 1 - \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right)^n \right).
\]

Now, using relation (7), we obtain

\[
\|s_k - s_{k-1}\| \leq \frac{2C}{\eta_{k-1}} \left( 1 - \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right)^n \right) + M \left| \frac{\eta_k}{\eta_{k-1}} \right|.
\]

(8)

Since we assumed that \(\epsilon_k \leq \epsilon_{k-1}\), we may write

\[
1 - \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right)^n = 1 - \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right) + \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right) - \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right)^2 + \ldots + \left( \frac{\epsilon_k}{\epsilon_{k-1}} \right)^{n-1} \leq n \left( 1 - \frac{\epsilon_k}{\epsilon_{k-1}} \right).
\]

(9)
Therefore when \( \epsilon_k \leq \epsilon_{k-1} \), from (9) and (8), the desired inequality holds for all \( k \geq 1 \).

(ii) \((\epsilon_k \geq \epsilon_{k-1})\): Now, suppose \( \epsilon_k \geq \epsilon_{k-1} \). Following the similar steps above, one may note that that if \( \epsilon_k \geq \epsilon_{k-1} \), then \( \|F_k(s_k) - F_{k-1}(s_k)\| \leq 2C(1 - (\epsilon_{k-1}/\epsilon_k)^n) \).

Therefore, the desired equality follows by combining cases (i) and (ii) to obtain the following bound:

\[
\|F_k(s_k) - F_{k-1}(s_k)\| \leq 2nC \left( 1 - \frac{\min(\epsilon_{k-1}, \epsilon_k)}{\max(\epsilon_{k-1}, \epsilon_k)} \right).
\]

(b) We begin by considering (1). By Definition 2, the vector \( s_k \) is the solution of \( \text{VI}(X,F_k + \eta_k I) \) indicating that for all \( k \geq 0 \),

\[
(x - s_k)^T(F_k(s_k) + \eta_k s_k) \geq 0, \quad \text{for all } x \in X, \tag{10}
\]

with \( s_k \in X \). Furthermore, by Assumption 1(a), the set \( X \) is bounded and, therefore, \( \{s_k\} \) is bounded and has at least one accumulation point. Let \( \hat{s} \) denote an arbitrary accumulation point of the sequence \( \{s_k\} \), i.e. \( \lim_{k \to \infty} s_{k_i} = \hat{s} \). Observe that by Lemma 2(a), it follows that the limit \( \lim_{i \to \infty} F_{k_i}(s_{k_i}) \) exists, namely

\[
\lim_{i \to \infty} F_{k_i}(s_{k_i}) = F(\hat{s}).
\]

Thus, by taking the limit along the subsequence \( \{k_i\} \) in relation (10) and using \( \epsilon_k \to 0 \), for any \( x \in X \) we obtain

\[
(x - \lim_{i \to \infty} s_{k_i})^T \left( \lim_{i \to \infty} F_{k_i}(s_{k_i}) + \lim_{i \to \infty} \eta_k \lim_{i \to \infty} s_{k_i} \right) \geq 0
\]

\[
\implies (x - \hat{s})^TF(\hat{s}) \geq 0,
\]

showing that \( \hat{s} \) is a solution to \( \text{VI}(X,F) \). Thus, all accumulation points of \( \{s_k\} \) are solutions to \( \text{VI}(X,F) \), which proves the statement in part (1).

We now consider (2) of (b) where we show that \( \lim_{k \to \infty} x_k = t^* \). We have \( t^* \in X^* \). Therefore,

\[
(x - t^*)^TF(t^*) \geq 0, \quad \text{for any } x \in X. \tag{11}
\]

Also, we have

\[
(x - s_k)^T(F_k(s_k) + \eta_k s_k) \geq 0, \quad \text{for any } x \in X. \tag{12}
\]

Replacing \( x \) by \( s_k \) in (11) and replacing \( x \) by \( t^* \) in (12) and then summing the resulting inequalities we obtain

\[
(s_k - t^*)^T(F(t^*) - F_k(s_k) - \eta_k s_k) \geq 0
\]

\[
\implies \eta_k \|s_k\|^2 \leq \eta_k s_k^Tt^* + (s_k - t^*)^T(F(t^*) - F_k(s_k))
\]

\[
\implies \eta_k \|s_k\|^2 \leq \eta_k s_k^Tt^* + (s_k - t^*)^T(F(t^*) - F_k(t^*)) + (s_k - t^*)^T(F_k(t^*) - F_k(s_k)).
\]

By observing the nonpositivity of the third term on the right from the monotonicity of \( F_k \) and by using the Cauchy-Schwartz inequality, we have the following:

\[
\eta_k \|s_k\|^2 \leq \eta_k \|s_k\| \|t^*\| + \|s_k - t^*\| \|F(t^*) - F_k(t^*)\|
\]

\[
\implies \|s_k\| \leq \|t^*\| + \eta_k^{-1} \left( \frac{\|s_k\| + \|t^*\|}{\|s_k\|} \right) \|F(t^*) - F_k(t^*)\|.
\]
Without loss of generality, we may assume that $0 \notin X$ (since we could consider define a related problem otherwise), there exists $d > 0$ such that $\|s_k\| \geq d$. Since $X$ is bounded, then $\|t^*\| \leq M$ where $M$ is the bound on set $X$. This implies the following:

$$
\|s_k\| \leq \|t^*\| + \eta_k^{-1} \left( 1 + \frac{M}{d} \right) \left| \int (F(t^* + z_k) - F(t^*))p_u(z_k)dz \right|
$$

$$
\leq \|t^*\| + \eta_k^{-1} \left( 1 + \frac{M}{d} \right) \int_{\|z\| \leq \epsilon_k} \|F(t^* + z_k) - F(t^*)\|p_u(z_k)dz
$$

$$
\leq \|t^*\| + \eta_k^{-1} \left( 1 + \frac{M}{d} \right) \sup_{\|z\| \leq \epsilon_k} \|F(t^* + z_k) - F(t^*)\|. \quad (13)
$$

Let $J_F$ denote the Jacobian of $F$. By assumption, there exists a $\rho > 0$ where $\|J_F(x)\| \leq J_{ub}$ for any $x \in B(t^*, \rho)$. Using the mean value theorem,

$$
F(x + \delta) - F(x) = \left( \int_0^1 J_F(x + t\delta)dt \right) \delta, \quad \text{for any } \|\delta\| \leq \rho. \quad (14)
$$

Assume that $K$ is a large number such that for any $k > K, \epsilon_k < \rho$. Using the boundedness of $J_F$ and the triangle inequality, from (14) we obtain

$$
\|F(t^* + z_k) - F(t^*)\| \leq J_{ub}\|z_k\| \leq J_{ub}\epsilon_k.
$$

Note that since $b < c$, we have $\lim_{k \to \infty} \epsilon_k/\eta_k = 0$. By the preceding inequality, relation (13) and $\eta_k^{-1} \lim_{k \to \infty} \epsilon_k = 0$, we conclude that for any subsequence of $s_k$, denoted by $\{s_k\}$, has a limit point $\bar{s}$ such that $\|\bar{s}\| \leq \|t^*\|$. But from Prop. (1b), every accumulation point of $\{s_k\}$ lies in $X^*$. But, since every limit point is bounded in norm by $\|t^*\|$, it follows that every limit point of $\{s_k\}$ is $t^*$, the unique least-norm solution. It follows that $\{s_k\}$ is a convergent sequence that tends to the least-norm solution $t^*$.

**Remark 4.** We note that part (2) of (b) in the above proposition requires a local differentiability and boundedness property. This can be seen to be weaker than a global Lipschitzian requirement. We also note that without such an assumption, we may still claim that $\{s_k\}$ converges to a point in $X^*$ but cannot provide a characterization of its limit point.

Next, we establish a recursive relation that relates a bound on the difference between $x_{k+1}$ and $s_k$ with that from the prior step. Such a relation essentially captures the distance of the sequence generated by the RSSA scheme from the regularized smoothed trajectory (denoted by $s_k$) and is important in our proof of the a.s. convergence.

**Lemma 3 (A recursive relation for $\|x_{k+1} - s_k\|$).** Consider the RSSA scheme in which $\{\gamma_k\}$, $\{\eta_k\}$, and $\{\epsilon_k\}$ are sequences of positive scalars. Let Assumptions 2, 3(b), and 3(c) hold, and suppose there exists $K_2 \geq 0$ such that for any $k \geq K_2$, we have $\epsilon_k \gamma_k < 1$. Then, with $K_1$ given by Assumption 3(b), the following relation holds a.s. for any $k \geq \max\{K_1, K_2\}$:

$$
\begin{align*}
\mathbb{E}[\|x_{k+1} - s_k\|^2 | F_k] &\leq \left( 1 - \frac{1}{2}\eta_k\gamma_k \right) \|x_k - s_{k-1}\|^2 + 2C^2\gamma_k^2 + 4M^2\epsilon_k^2\gamma_k^2 \\
&+ 16n^2C^2 \left( 1 - \min\{\epsilon_k, \epsilon_{k-1}\} \right) \frac{1}{\max\{\epsilon_k, \epsilon_{k-1}\}} \frac{1}{\eta_k^2\gamma_k} + 4M^2 \left( 1 - \frac{\eta_k}{\eta_{k-1}} \right)^2 \frac{1}{\eta_k\gamma_k}. \quad (15)
\end{align*}
$$
Using the iterated expectation rule and Lemma 1, we find that
\[
- \text{term } (1 + \eta_k \gamma_k)^k \gamma_k (F(x_k + z_k) - F_k(s_k)) - \gamma_k w_k)^2 \\
= (1 - \eta_k \gamma_k)^k ||x_k - s_k||^2 + \gamma_k^2 ||F(x_k + z_k) - F_k(s_k)||^2 + \gamma_k^2 \|w_k\|^2 \\
- 2\gamma_k (1 - \eta_k \gamma_k)(x_k - s_k)^T(F(x_k + z_k) - F_k(s_k)) \\
- 2\gamma_k^2 (1 - \eta_k \gamma_k)(x_k - s_k)^T(F(x_k + z_k) - F_k(s_k)) \\
\]
(16)

Using the iterated expectation rule and Lemma 1, we find that
\[
E \left[ \left( (1 - \eta_k \gamma_k)(x_k - s_k) - \gamma_k (F(x_k + z_k) - F_k(s_k)) \right)^T w_k | F_k \right] = 0.
\]

Thus, by taking the conditional expectations conditioned on \( F_k \) in (16) and using \( E[\|w_k\|^2 | F_k] \leq C^2 \) (cf. Lemma 1), we obtain
\[
E[\|x_{k+1} - s_k\|^2 | F_k] \leq (1 - \eta_k \gamma_k)^k ||x_k - s_k||^2 + \gamma_k^2 E[\|F(x_k + z_k) - F_k(s_k)||^2 | F_k] + \gamma_k^2 C^2 \\
- 2\gamma_k (1 - \eta_k \gamma_k)(x_k - s_k)^T(E[F(x_k + z_k) | F_k] - F_k(s_k)).
\]

Noting that \( E[F(x_k + z_k) | F_k] = F_k(x_k) \) and using the monotonicity of \( F_k \) (see Lemma 2(b)), we further obtain \( (x_k - s_k)^T(E[F(x_k + z_k) | F_k] - F_k(s_k)) \geq 0 \). Since \( \eta_k \gamma_k < 1 \) for any \( k \geq K_2 \), the term \((1 - \eta_k \gamma_k)^k\) is positive implying that, for all \( k \geq K_2 \) we have almost surely,
\[
E[\|x_{k+1} - s_k\|^2 | F_k] \leq (1 - \eta_k \gamma_k)^k ||x_k - s_k||^2 + \gamma_k^2 E[\|F(x_k + z_k) - F_k(s_k)||^2 | F_k] + \gamma_k^2 C^2. \quad (17)
\]

To estimate \( E[\|F(x_k + z_k) - F_k(s_k)||^2 | F_k] \) we add and subtract \( F_k(x_k) \), which yields
\[
||F(x_k + z_k) - F_k(s_k)||^2 = ||F(x_k + z_k) - F_k(x_k)||^2 + ||F_k(x_k) - F_k(s_k)||^2 \\
+ 2(F(x_k + z_k) - F_k(x_k))^T(F_k(x_k) - F_k(s_k)) \\
\leq ||F(x_k + z_k) - F_k(x_k)||^2 + \kappa \frac{n!!}{(n-1)!! \epsilon_k} \|x_k - s_k\|^2 \\
+ 2(F(x_k + z_k) - F_k(x_k))^T(F_k(x_k) - F_k(s_k)),
\]
where the second inequality follows by the Lipschitz continuity of \( F_k \) with constant \( \kappa \frac{n!!}{(n-1)!! \epsilon_k} C \) (see Lemma 2(b)). Taking expectations conditioned on \( F_k \) and using \( F_k(x_k) = E[F(x_k + z_k) | F_k] \), we find that almost surely for all \( k \geq K_2 \),
\[
E[\|F(x_k + z_k) - F_k(s_k)||^2 | F_k] \leq E[\|F(x_k + z_k) - F_k(x_k)||^2 | F_k] + \kappa \frac{n!!}{(n-1)!! \epsilon_k} ||x_k - s_k||^2 \\
= E[\|F(x_k + z_k)||^2 | F_k] + \|F_k(x_k)||^2 + \kappa \frac{n!!}{(n-1)!! \epsilon_k} ||x_k - s_k||^2 \\
\leq C^2 + \kappa \frac{n!!}{(n-1)!! \epsilon_k} ||x_k - s_k||^2,
\]

Proof. Using the fixed point property of the projection operator at the solution \( s_k \in \text{SOL}(X, F_k + \eta_k I) \), we may write \( s_k = P_X(s_k - \gamma_k(F_k(s_k) + \eta_k s_k)) \). Employing the non-expansivity property of the projection operator, the preceding relation, and the RSSA algorithm, we obtain
\[
E \left[ \left( (1 - \eta_k \gamma_k)(x_k - s_k) - \gamma_k (F(x_k + z_k) - F_k(s_k)) \right)^T w_k | F_k \right] = 0.
\]
where the last inequality is obtained by using $\|F(x_k + z_k)\| \leq C$ (see Remark\[1\]) and by ignoring the negative term. Substituting the preceding estimate in the relation \[17\], we obtain a.s. for $k \geq K_2$,

$$E[\|x_{k+1} - s_k\|^2 | F_k] \leq \left(1 - \eta \gamma_k\right)^2 + \gamma_k^2 \left(\frac{n!!}{(n-1)!!} \frac{C}{\epsilon_k}\right)^2 \|x_k - s_k\|^2 + 2 \gamma_k^2 C^2$$

$$= \left(1 - 2 \eta \gamma_k + \gamma_k^2 \left(\frac{n!!}{(n-1)!!} \frac{C}{\epsilon_k}\right)^2\right) \|x_k - s_k\|^2 + 2 \gamma_k^2 C^2. \quad (18)$$

Using the definition of $M$ in Remark\[1\] and the triangle inequality, we may write $\|x_k - s_k\| \leq \|x_k\| + \|s_k\| \leq 2M$, which leads to the following bound on $\eta^2 \gamma_k^2 \|x_k - s_k\|^2$,

$$\eta^2 \gamma_k^2 \|x_k - s_k\|^2 \leq 4 \eta^2 \gamma_k^2 M^2.$$ 

This inequality and relation \[18\] yield a.s. for all $k \geq K_2$,

$$E[\|x_{k+1} - s_k\|^2 | F_k] \leq \left(1 - 2 \eta \gamma_k + \gamma_k^2 \left(\frac{n!!}{(n-1)!!} \frac{C}{\epsilon_k}\right)^2\right) \|x_k - s_k\|^2 + 2 \gamma_k^2 C^2 + 4 \eta^2 \gamma_k^2 M^2. \quad (19)$$

To obtain a recursion, we need to estimate the term $\|x_k - s_k\|$ in terms of $\|x_k - s_{k-1}\|$. Using the triangle inequality, we may write $\|x_k - s_k\| \leq \|x_k - s_{k-1}\| + \|s_k - s_{k-1}\|$. Therefore, we obtain

$$\|x_k - s_k\|^2 \leq \|x_k - s_{k-1}\|^2 + \|s_k - s_{k-1}\|^2 + 2 \|s_k - s_{k-1}\| \|x_k - s_{k-1}\|. \quad (20)$$

Using the relation $2ab \leq a^2 + b^2$, for $a, b \in R$, we have

$$2 \|s_k - s_{k-1}\| \|x_k - s_{k-1}\| = 2 \left(\eta \gamma_k \|s_k - s_{k-1}\|\right) \left(\frac{\|s_k - s_{k-1}\|}{\sqrt{\eta \gamma_k}}\right) \leq \eta \gamma_k \|x_k - s_{k-1}\|^2 + \frac{\|s_k - s_{k-1}\|^2}{\eta \gamma_k}.$$ 

Combining this result, Proposition\[1\](a), and (20), we obtain for all $k \geq K_2$,

$$\|x_k - s_k\|^2 \leq \left(1 + \eta \gamma_k\right) \|x_k - s_{k-1}\|^2 + 2 \left(\frac{2nC}{\eta \gamma_k} \left(1 - \min\{\epsilon_k, \epsilon_{k-1}\}\right) + M \left|1 - \frac{\eta_k}{\eta_{k-1}}\right|\right) \left(1 + \frac{1}{\eta \gamma_k}\right). \quad (21)$$

where in the last inequality we used $1 + 1/(\eta \gamma_k) < 2/\eta \gamma_k$ as a consequence of $\gamma_k \eta_k < 1$ for $k \geq K_2$. If $q_k$ is defined as

$$q_k \triangleq 1 - 2 \eta \gamma_k + \gamma_k^2 \left(\frac{n!!}{(n-1)!!} \frac{C}{\epsilon_k}\right)^2,$$

then inequalities \[19\] and \[21\] imply that for $k \geq K_2$, the following relation holds:

$$E[\|x_{k+1} - s_k\|^2 | F_k] \leq q_k \left(1 + \|x_k - s_{k-1}\|^2 + 2 \left(\frac{2nC}{\eta \gamma_k} \left(1 - \min\{\epsilon_k, \epsilon_{k-1}\}\right) + M \left|1 - \frac{\eta_k}{\eta_{k-1}}\right|\right) \left(1 + \frac{1}{\eta \gamma_k}\right)^2 \right). \quad (22)$$
By Assumption 3(b), we can write for $k \geq K_1$,

$$\frac{\gamma_k}{\eta_k \epsilon_k} \leq 0.5 \left( \frac{(n-1)!!}{n!! \kappa C} \right)^2 \Rightarrow \frac{\gamma_k^2}{\kappa (n-1)!! \epsilon_k^2} \leq \frac{\eta_k \gamma_k}{2} \Rightarrow -2 \eta_k \gamma_k + \gamma_k^2 \frac{n!!}{(n-1)!! \epsilon_k} \leq -\frac{3}{2} \eta_k \gamma_k.$$ 

Therefore, $q_k \leq 1 - \frac{3}{2} \eta_k \gamma_k$. Consequently, we may provide an upper bound on $q_k(1 + \eta_k \gamma_k)$ using the preceding relation:

$$q_k(1 + \eta_k \gamma_k) \leq (1 - \frac{3}{2} \eta_k \gamma_k)(1 + \eta_k \gamma_k) = 1 - \frac{1}{2} \eta_k \gamma_k - \frac{3}{2} \eta_k^2 \gamma_k^2 \leq 1 - \frac{1}{2} \eta_k \gamma_k.$$ 

Using relation (22) and $q_k \leq 1$ (which follows by $q_k \leq 1 - \frac{3}{2} \eta_k \gamma_k$), and $(a + b)^2 \leq 2a^2 + 2b^2$, we conclude that the desired relation holds.

The following supermartingale convergence theorem is a key in our analysis in establishing the almost sure convergence of the RSSA scheme and may be found in [26] (cf. Lemma 10, page 49).

**Lemma 4** (Robbins and Siegmund Lemma). Let $\{v_k\}$ be a sequence of nonnegative random variables, where $E[v_0] < \infty$, and let $\{\alpha_k\}$ and $\{\mu_k\}$ be deterministic scalar sequences such that $0 \leq \alpha_k \leq 1$, and $\mu_k \geq 0$ for all $k \geq 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \mu_k < \infty$, and $\lim_{k \to \infty} \frac{\mu_k}{\alpha_k} = 0$, and $E[v_{k+1} \mid v_0, \ldots, v_k] \leq (1 - \alpha_k)v_k + \mu_k$ a.s. for all $k \geq 0$. Then, $v_k \to 0$ almost surely as $k \to \infty$.

We are now ready to present the main convergence result showing that the sequence generated by the RSSA scheme has its accumulation points in the solution set $X^*$ of the original VI($F, X$) almost surely. Under the assumption that $\epsilon_k/\eta_k \to 0$ and suitable local requirements, we may further claim that the sequence converges to the smallest norm solution in $X^*$ almost surely.

**Theorem 1** (Almost sure convergence of RSSA scheme). Let Assumptions 1, 2, and 3 hold, and let $\{x_k\}$ be given by theRSSA scheme. Then the following statements hold almost surely: 

(a) $\lim_{k \to \infty} \|x_{k+1} - x_k\| = 0$ and any accumulation point of $\{x_k\}$ is a solution of VI($X, F$). 

(b) If $\lim_{k \to \infty} \frac{v_k}{\eta_k} = 0$ and $F$ is differentiable at $t^*$ with a bounded Jacobian in a neighborhood of $t^*$, then $\{x_k\}$ converges to the smallest norm solution of VI($X, F$).

**Proof.** (a) From Assumption 3(a), $\gamma_k$ and $\eta_k$ go to zero. Thus, there exists a constant $K_2 \geq 0$ such that $\gamma_k \eta_k < 1$ for any $k \geq K_2$. Let us define define sequences $\{v_k\}$, $\{\alpha_k\}$, and $\{\mu_k\}$ for $k \geq \max\{K_1, K_2\}$ given by $v_k \triangleq \|x_k - s_{k-1}\|$, $\alpha_k \triangleq \frac{1}{2} \gamma_k \eta_k$ and

$$\mu_k \triangleq 2C^2 \gamma_k^2 + 4M^2 \eta_k \gamma_k + 16n^2 C^2 \left(1 - \frac{\min\{\epsilon_k, \epsilon_k-1\}}{\max\{\epsilon_k, \epsilon_k-1\}} \right)^2 \frac{1}{\eta_k \gamma_k} + 4M^2 \left(1 - \frac{\eta_k}{\gamma_k-1} \right)^2 \frac{1}{\eta_k \gamma_k}.$$ 

Therefore, Lemma 3 implies that $E[v_{k+1} \mid F_k] \leq (1 - \alpha_k)v_k + \mu_k$, for $k \geq \max\{K_1, K_2\}$. To claim convergence of the sequence $\{x_k\}$, we show that conditions of Lemma 4 hold. The nonnegativity of $v_k$, $\alpha_k$, and $\mu_k$ for $k \geq \max\{K_1, K_2\}$ is trivial. Assumption 3(d) indicates that the condition $\sum_k \alpha_k = \infty$ is satisfied. On the other hand, positivity of $\gamma_k$ and $\eta_k$ indicates that $\alpha_k \leq 1$ holds for $k \geq \max\{K_1, K_2\}$. Since $\eta_k$ goes to zero, there exists a bound $\bar{\eta}$ such that $\eta_k \leq \bar{\eta}$. Therefore, $\mu_k \leq (2C^2 + 4M^2 \bar{\eta}^2) \gamma_k^2 + 16n^2 C^2 \text{Term 1} + 4M^2 \text{Term 2}$. Assumptions 3(e), (f), and (g) show
that \( \gamma_k^2 \). Terms 1 and 2 are summable. Therefore, we conclude that \( \mu_k \) is summable too. It remains to show that \( \lim_{k \to \infty} \frac{\mu_k}{\alpha_k} = 0 \). It suffices to show that

\[
\lim_{k \to \infty} \frac{\gamma_k^2}{\alpha_k} = 0, \quad \lim_{k \to \infty} \frac{\text{Term 1}}{\alpha_k} = 0, \quad \text{and} \quad \lim_{k \to \infty} \frac{\text{Term 2}}{\alpha_k} = 0.
\]

These three conditions hold due to Assumption 3(h), (i), and (j) respectively. In conclusion, all of the conditions of Lemma 4 hold and thus \( \| x_k - s_k \| \) goes to zero almost surely.

Since \( F \) is continuous and \( \gamma_k \) and \( \epsilon_k \) go to zero, Proposition 1(b1) implies that any limit point of the sequence \( \{ s_k \} \) converges to a solution of \( \text{VI}(X, F) \). Hence, from the result of part (a), we conclude that any accumulation point of the sequence \( \{ x_k \} \) generated by the RSSA algorithm converges to a solution of \( \text{VI}(X, F) \) almost surely.

(b) The statement in part (b) follows by part (a) and Proposition 1(b2).

A reader might question whether Assumption 3 is vacuous in that there are no set of sequences satisfying the required assumptions. We prove that this is not the case by showing that there is a set of stepsize, regularization, and smoothing sequences that satisfy the prescribed requirements.

**Lemma 5.** Suppose sequences \( \{ \gamma_k \} \), \( \{ \eta_k \} \), and \( \{ \epsilon_k \} \) are given by \( \gamma_k = \gamma_0(k+1)^{-a} \), \( \eta_k = \eta_0(k+1)^{-b} \), and \( \epsilon_k = \epsilon_0(k+1)^{-c} \) where \( a, b, \) and \( c \) satisfy the following conditions:

\[
a, b, c > 0, \quad a + 3b < 1, \quad a > b + 2c, \quad a > 0.5,
\]

and \( \gamma_0, \eta_0, \epsilon_0 \) are strictly positive scalars and \( \epsilon_0 = \epsilon \). Then, sequences \( \{ \gamma_k \} \), \( \{ \eta_k \} \), and \( \{ \epsilon_k \} \) satisfy Assumption 3.

**Proof.** We show that each part of Assumption 3 holds as follows:

(a) Assumption 3(a) holds since \( a, b, \) and \( c \) are strictly positive.

(b) To show that part (b) holds, we note that

\[
\frac{\gamma_k}{\eta_k} = \frac{\gamma_0(k+1)^{-a}}{\eta_0(k+1)^{-b}\eta_k^{2}} = (k+1)^{(a-b-2c)} \frac{\gamma_0}{\eta_0}. 
\]

Since \( a > b + 2c \), then \((k+1)^{(a-b-2c)} \to 0\). Therefore, \( \frac{\gamma_k}{\eta_k} \to 0 \) implying that there exists \( K_1 \geq 0 \) such that

\[
\frac{\gamma_k}{\eta_k} \leq 0.5 \left( \frac{n-1}{n!(k+1)} \right)^2
\]

for any \( k \geq K_1 \). This indicates that Assumption 3(b) holds.

(c) Part (c) holds because \( \epsilon_k \leq \epsilon_0 \) for any \( k \geq 0 \) and \( \epsilon_0 = \epsilon \).

(d) We have \( \sum_{k=0}^{\infty} \eta_k \gamma_k = \eta_0 \gamma_0 \sum_{k=0}^{\infty} 1/(k+1)^{a+b} \). Since \( a, b > 0 \) and \( a + 3b < 1 \), then \( a + b < 1 \). Thus, \( \sum_{k=0}^{\infty} 1/(k+1)^{a+b} = \infty \). Therefore, Assumption 3(d) is met.

(e) To show that part (e) holds, it suffices to show that \( \gamma_k^2 \) is summable. But \( \gamma_k^2 = \gamma_0^2(k+1)^{-2a} \) and \( 2a > 1 \) since \( a > 0.5 \). Therefore, \( \gamma_k^2 \) is summable.

(f) Note that sequences \( \{ \eta_k \} \) and \( \{ \epsilon_k \} \) are both decreasing. Therefore,

\[
\frac{1}{\eta_k^2} \left( 1 - \min \{ \epsilon_k, \epsilon_{k-1} \} \right)^2 = \frac{1}{\eta_k^2} \left( 1 - \frac{\epsilon_k}{\epsilon_{k-1}} \right)^2 < \frac{1}{\eta_k^2} \left( 1 - \frac{\epsilon_k}{\epsilon_{k-1}} \right)^2 \triangleq \text{Term 1}.
\]

It suffices to show that Term 1 is summable. First, we estimate \( 1 - \epsilon_k/\epsilon_{k-1} \). We have

\[
1 - \frac{\epsilon_k}{\epsilon_{k-1}} = 1 - \frac{\epsilon_0(k+1)^{-c}}{\epsilon_0k^{-c}} = 1 - \left( \frac{k}{k+1} \right)^c = 1 - \left( 1 - \frac{1}{k+1} \right)^c.
\]
Recall that the Taylor expansion of \((1 - x)^p\) for \(|x| < 1\) and any scalar \(p\) is given by

\[
(1 - x)^p = \sum_{j=0}^{\infty} (-1)^j \binom{p}{j} x^j = 1 - px + \frac{p(p-1)}{2} x^2 - \frac{p(p-1)(p-2)}{6} x^3 + \ldots.
\]

Using this expansion for \(x = \frac{1}{k+1}\) and \(p = c\), we have

\[
1 - \frac{\epsilon_k}{\epsilon_{k-1}} = 1 - \left(1 - c \frac{1}{k+1} + \frac{c(c-1)}{2} \frac{1}{(k+1)^2} - \frac{c(c-1)(c-2)}{6} \frac{1}{(k+1)^3} + \ldots\right) = \mathcal{O}(k^{-1}).
\]

Therefore, from the preceding relation, we obtain

\[
\text{Term 1} = \frac{\mathcal{O}(k^{-2})}{\eta_0 \gamma_0 (k+1)^{-3b-a}} = \mathcal{O}(k^{-(2a-3b)}).
\]

To ensure summability of Term 1, it suffices that \(2 - a - 3b > 1\) or equivalently \(a + 3b < 1\). This holds by assumption and condition (f) is met.

(g) In a similar fashion that we used in part (f), we can show that \(1 - \frac{\eta_k}{\eta_{k-1}} = \mathcal{O}(k^{-1})\). Consider Term 3 defined as follows:

\[
\text{Term 3} \equiv \frac{1}{\eta_k \gamma_k} \left(1 - \frac{\eta_k}{\eta_{k-1}}\right)^2 \frac{\mathcal{O}(k^{-2})}{\eta_0 \gamma_0 (k+1)^{-(a+b)}} = \mathcal{O}(k^{-(2a-b)}).
\]

To show that condition (g) is satisfied, it suffices to show that Term 3 is summable. From the preceding relation, we need to show that \(2 - a - b > 1\) or equivalently \(a + b < 1\). We assumed that \(a + 3b < 1\) and \(b > 0\). Thus, we have \(a + b = a + 3b - 2b < 1 - 2b < 1\). Therefore, \(\mathcal{O}(k^{-(2a-b)})\) is summable and we conclude that condition (g) is met.

(h) We have \(\gamma_k/\eta_k = \gamma_0 (k+1)^{-a}/(\eta_0 (k+1)^{-b}) = (\gamma_0/\eta_0)(k+1)^{-(a-b)}\). To show that \(\gamma_k/\eta_k\) goes to zero when \(k\) goes to infinity, we only need to show that \(a > b\). We assumed that \(a + 3b < 1\). Therefore, \(b < (1-a)/3\). Since \(a > 0.5\), the preceding relation yields \(b < 1/6\). Thus, \(b < 0.5 < a\), implying that condition (h) holds.

(i) From part (f), we have \(1 - \epsilon_k/\epsilon_{k-1} = \mathcal{O}(k^{-1})\). To show the condition (i), we write

\[
\text{Term 4} \equiv \frac{1}{\eta_k^2 \gamma_k} \left(1 - \min\{\epsilon_k, \epsilon_{k-1}\}\right) = \frac{1}{\eta_0 \gamma_0 (k+1)^{-(a-2b)}} \mathcal{O}(k^{-1}) = \mathcal{O}(k^{-(1-a-2b)}).
\]

Thus, it suffices to show that \(a + 2b < 1\). This is true since \(a + 3b < 1\) and \(b > 0\). Hence, Term 4 goes to zero implying that part (i) holds.

(j) Term 5 is defined as

\[
\text{Term 5} \equiv \frac{1}{\eta_k \gamma_k} \left|1 - \frac{\eta_k}{\eta_{k-1}}\right| = \frac{1}{\eta_0 \gamma_0 (k+1)^{-(a-b)}} \mathcal{O}(k^{-1}) = \mathcal{O}(k^{-(1-a-b)}).
\]

Since \(a + 3b < 1\) and \(b > 0\), we have \(a + b < 1\), showing that Term 5 converges to zero.

In order to satisfy the additional condition \(\lim_{k \to \infty} \frac{\epsilon_k}{\eta_k} = 0\) used in Proposition 1(b2) and Theorem 2(b), one would need an additional requirement \(b - c < 0\) in Lemma 5. As a concrete example satisfying the conditions in Lemma 5, consider the choice \(a = \frac{9}{16}\), \(b = \frac{3}{16}\), and \(c = \frac{3}{16}\). In this case we also have \(\lim_{k \to \infty} \frac{\epsilon_k}{\eta_k} = 0\) since \(b - c < 0\).
3.2 Rate of convergence to regularized smoothed trajectory

Thus far, we have discussed the convergence of the sequence \( \{x_k\} \) generated by the RSSA scheme in an almost sure sense. Naturally, one may be curious about the rate of convergence of this sequence. While the development of non-asymptotic rates of convergence have been provided in the mean in the past (either in terms of mean-squared error for solution iterates or in terms of the mean gap function), we are unaware of any statements provided in non-Lipschitzian and merely monotone regimes in terms of solution iterates. In this subsection, we provide a partial answer to this question.

Our metric of convergence rate is the \( \text{dist}(x_k, X^*) \), and the question is at what rate the error \( \text{dist}(x_k, X^*) \) will diminish to zero. We may provide a partial answer by establishing the rate at which the sequence \( \{x_k\} \) approaches the regularized smoothed trajectory \( \{s_k\} \). The idea is as follows: At step \( k \), instead of comparing the iterate \( x_k \) with a true solution \( x^* \), we want to estimate the distance between \( x_k \) and the approximate solution \( s_k \). Note that, as the algorithm proceeds, we expect \( s_k \) to be approaching to the solution set \( X^* \) (Prop. 1). The first part of this section provides such an analysis and we derive a generic bound for this dynamic error. We begin the discussion by a family of assumptions on the sequences. This set of assumptions is essential for deriving the particular rate.

**Assumption 4.** Let the following hold:

(a) There exist \( 0 < \delta < 0.5 \) and \( K_3 > 0 \) such that for any \( k \geq K_3 \)

\[ \frac{\gamma_k}{\eta_k \epsilon_k^2} \leq \frac{\gamma_{k+1}}{\eta_{k+1} \epsilon_{k+1}^2} (1 + \delta \eta_{k+1} \gamma_{k+1}); \]

(b) There exists a constant \( B_1 > 0 \) such that for any \( k \geq 0 \):

\[ \frac{\epsilon_k^2}{\eta_k^2 \eta_k^3 \gamma_k^3} \left( 1 - \frac{\min\{\epsilon_k, \epsilon_{k-1}\}}{\max\{\epsilon_k, \epsilon_{k-1}\}} \right)^2 \leq B_1; \]

(c) There exists a constant \( B_2 > 0 \) such that for any \( k \geq 0 \)

\[ \frac{\epsilon_k^2}{\eta_k^2 \gamma_k^3} \left( 1 - \frac{\eta_k}{\eta_{k-1}} \right)^2 \leq B_2. \]

The following result provides a bound on the error that relates the iterates \( \{x_k\} \) and the approximate sequence \( \{s_k\} \). This result provides us an estimate of the performance of our algorithm with respect to the iterates of the solutions to the approximated problems \( \text{VI}(X, F_k + \eta_k I) \).

**Proposition 2** (An upper bound for \( \mathbb{E}[\|x_{k+1} - s_k\|^2] \)). Consider the RSSA scheme where \( \{\gamma_k\}, \{\eta_k\}, \{\epsilon_k\} \) are strictly positive sequences. Let Assumptions \( 3(b), 3(c), \) and \( 4(a) \) hold. Suppose \( \{\eta_k\} \) is bounded by some \( \bar{\eta} > 0 \) and there exists some scalar \( K_2 \geq 0 \) such that for any \( k \geq K_2 \) we have \( \eta_k \gamma_k < 1 \). Then,

\[ \mathbb{E}[\|x_{k+1} - s_k\|^2] \leq \theta \frac{\gamma_k}{\eta_k \epsilon_k^2}, \quad \text{for any } k \geq \tilde{K}, \]

where \( \tilde{K} \equiv \max\{K_1, K_2, K_3\}, \) \( s_k \) is the unique solution of \( \text{VI}(X, F_k + \eta_k I) \), \( K_1 \) and \( K_3 \) are given by Assumptions \( 3(b), \) and \( 3(a) \) respectively. More precisely, relation (23) holds if

\[ \theta = \max \left( \frac{4M^2 \eta_k \epsilon_k^2}{\gamma_k}, \frac{2C^2 \epsilon^2 + 4M^2 \bar{\eta}^3 \epsilon^2 + 16n^2 C^2 B_1 + 4M^2 B_2}{0.5 - \delta} \right). \]

19
Proof. We begin by employing Lemma 3 and denote $E[\|x_k - s_{k-1}\|^2]$ by $e_k$ for $k \geq K + 1$. Taking expectations on both sides of (15) in Lemma 3, we obtain a recursion in terms of the mean squared error $e_k$. For any $k \geq K + 1$ we have

\[
e_{k+1} \leq \left(1 - \frac{1}{2}\eta_k \gamma_k\right) e_k + 2C^2 \gamma_k^2 + 4M^2 \eta_k^2 \gamma_k^2 + 16n^2 C^2 \left(1 - \frac{\eta_k}{\max(\eta_k, \epsilon_k)}\right) \gamma_k^2 + 4M^2 \left(1 - \frac{\eta_k}{\epsilon_k}\right)^2.
\]

(25)

To show the main result, we employ the mathematical induction on $k$. The first step is to show that the result holds for $k = K$. Using the definition of $M$ in Remark 1 and the Cauchy-Schwartz inequality, we may write

\[
e_{K+1} = E[\|x_{K+1}\|^2 - 2x_{K+1}^T s_K + \|s_K\|^2] \leq E[\|x_{K+1}\|^2 + 2\|x_{K+1}\|\|s_K\| + \|s_K\|^2]
\]

\[
\leq M^2 + 2M^2 + M^2 = \left(4M^2 \frac{\eta K}{\gamma K}\right) \frac{\gamma K}{\eta K^2}.
\]

Let us define $\theta_K = 4M^2 \eta K \epsilon K^2 / \gamma K$. Thus, the preceding relation implies that the main result holds for $k = K$ with $\theta = \theta_K$. Now, suppose $e_{t+1} \leq \theta \gamma t / (\eta e_k^2)$ for $K < t \leq k - 1$ for some finite constant $\theta > 0$. We will proceed to show that $e_{k+1} \leq \theta \gamma_k / (\eta e_k^2)$. Using the induction hypothesis, relation (25), and Assumptions 4(a) and (c) we obtain

\[
e_{k+1} \leq \left(1 - \frac{1}{2}\eta_k \gamma_k\right) \theta \frac{\gamma_k}{\eta_k \epsilon_k^2} + 2C^2 \gamma_k^2 + 4M^2 \eta_k^2 \gamma_k^2 + 16n^2 C^2 \gamma_k^2 B_1 + 4M^2 \gamma_k^2 B_2.
\]

Using Assumption 4(a), we obtain

\[
e_{k+1} \leq \left(1 - \frac{1}{2}\eta_k \gamma_k\right) \left(1 + \delta \eta_k \gamma_k\right) \theta \frac{\gamma_k}{\eta_k \epsilon_k^2} + 2C^2 \gamma_k^2 + 4M^2 \eta_k^2 \gamma_k^2 + 16n^2 C^2 \gamma_k^2 B_1 + 4M^2 \gamma_k^2 B_2.
\]

(26)

We now provide an upper bound for the term $\left(1 - \frac{1}{2}\eta_k \gamma_k\right) \left(1 + \delta \eta_k \gamma_k\right) \theta \frac{\gamma_k}{\eta_k \epsilon_k^2}$:

\[
\left(1 - \frac{1}{2}\eta_k \gamma_k\right) \left(1 + \delta \eta_k \gamma_k\right) \theta \frac{\gamma_k}{\eta_k \epsilon_k^2} = \theta \frac{\gamma_k}{\eta_k \epsilon_k^2} - \theta \left(\frac{\delta}{2}\right) \frac{\eta_k \gamma_k}{\epsilon_k^2} + \theta \eta_k \gamma_k \left(-\frac{1}{2} + \delta\right) \frac{\gamma_k}{\eta_k \epsilon_k^2} + 2C^2 \gamma_k^2.
\]

(27)

Using nonpositivity of $-\theta \left(\frac{\delta}{2}\right) \frac{\eta_k \gamma_k}{\epsilon_k^2}$ and the bound (27), the relation (26) can be expressed as follows:

\[
e_{k+1} \leq \theta \frac{\gamma_k}{\eta_k \epsilon_k^2} + \gamma_k \left[-\theta \left(\frac{1}{2} - \delta\right) + 2C^2 \epsilon^2 + 4M^2 \eta^2 \epsilon^2 + 16n^2 C^2 B_1 + 4M^2 B_2\right],
\]

(28)

where we invoke the boundedness of $\{\eta_k\}$ and $\epsilon_k$ from above by $\bar{\eta}$ and $\epsilon$, respectively (the latter follows from Assumption 3(c)). To complete the proof, it suffices to show that Term 1 is nonpositive for some $\theta > 0$. By Assumption 4(a), we have $\left(\frac{1}{2} - \delta\right) > 0$. Therefore, if

\[
\theta \geq \frac{2C^2 \epsilon^2 + 4M^2 \eta^2 \epsilon^2 + 16n^2 C^2 B_1 + 4M^2 B_2}{0.5 - \delta},
\]

then Term 1 is nonpositive. Hence, $e_{k+1} \leq \theta \gamma_k / (\eta_k \epsilon_k^2)$ and, thus, the induction argument is complete. In conclusion, if $\theta$ satisfies relation (24), then relation (23) holds for any $k \geq K$.
The following proposition states that the RSSA algorithm generates a sequence converging to the solution set of VI$(X, F)$ in a mean-square sense.

**Proposition 3** (Convergence in mean-square). Let Assumptions $[\text{A1, A2, A3, A4}]$ hold. Also, assume that $\lim_{k \to \infty} \gamma_k/(\eta_k \epsilon_k^2) = 0$, and let $\{x_k\}$ be generated by the RSSA scheme. Then, we have the following:

(a) The sequence $\{x_k\}$ converges to the solution set $X^*$ of VI$(X, F)$ in mean-squared sense, i.e.,

$$\lim_{k \to \infty} \mathbb{E} \left[ \text{dist}^2(x_k, X^*) \right] = 0.$$

(b) If in addition $\lim_{k \to \infty} \gamma_k/(\eta_k \epsilon_k^2) = 0$ and $F$ is differentiable at $t^*$ with a bounded Jacobian in a neighborhood of $t^*$, then $\{x_k\}$ converges to the smallest norm solution $t^* \in X^*$ in mean-squared sense.

**Proof.** To show part (a), using the triangle inequality and by recalling that $(a + b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we estimate $\text{dist}^2(x_{k+1}, X^*)$ from above, as follows:

$$\text{dist}^2(x_{k+1}, X^*) \leq (\text{dist}(x_{k+1}, s_k) + \text{dist}(s_k, X^*))^2 \leq 2 \text{dist}^2(x_{k+1}, s_k) + 2 \text{dist}^2(s_k, X^*).$$

Taking expectations in the preceding relation, we obtain that

$$\mathbb{E}[\text{dist}^2(x_{k+1}, X^*)] \leq 2\mathbb{E}[\|x_{k+1} - s_k\|^2] + 2\text{dist}^2(s_k, X^*). \quad (29)$$

Note that in the inequality above, the term $\text{dist}^2(s_k, X^*)$ is a deterministic quantity since $s_k$ is a (unique) solution to a deterministic problem. By Proposition 2, there exists a finite constant $\theta > 0$ such that $\mathbb{E}[\|x_{k+1} - s_k\|^2] \leq \theta \gamma_k/(\eta_k \epsilon_k^2)$. Therefore, from (29) we obtain

$$\mathbb{E}[\text{dist}^2(x_{k+1}, X^*)] \leq 2\theta \frac{\gamma_k}{\eta_k \epsilon_k^2} + 2\text{dist}^2(s_k, X^*). \quad (30)$$

Proposition [1(b1)] indicates that the term $\text{dist}^2(s_k, X^*)$ goes to zero as $k \to \infty$. Since $\lim_{k \to \infty} \gamma_k/(\eta_k \epsilon_k^2) = 0$, from relation (30), we conclude that the term $\mathbb{E}[\text{dist}^2(x_{k+1}, X^*)]$ goes to zero as $k \to \infty$.

The result in part (b) follows similarly to the preceding analysis, wherein we replace $\text{dist}(x_{k+1}, X^*)$ and $\text{dist}(s_k, X^*)$, respectively, by $\|x_{k+1} - t^*\|^2$ and $\|s_k - t^*\|^2$ with $t^*$ being the smallest norm solution, and by invoking Proposition [1(b2)].

As a counterpart of Lemma 5, the following result presents a class of the stepsize, regularization, and smoothing sequences that ensure mean-square convergence.

**Lemma 6.** Suppose sequences $\{\gamma_k\}$, $\{\eta_k\}$, and $\{\epsilon_k\}$ are given by $\gamma_k = \gamma_0(k+1)^{-a}$, $\eta_k = \eta_0(k+1)^{-b}$, and $\epsilon_k = \epsilon_0(k+1)^{-c}$ where $a$, $b$, and $c$ satisfy the following conditions:

$$a, b, c > 0, \quad a + b < 1, \quad a + b \leq \frac{2}{3}(1 + c), \quad a > b + 2c,$$

$\gamma_0$, $\eta_0$, $\epsilon_0$ are positive scalars and $\epsilon_0 \leq \epsilon$. Then, sequences $\{\gamma_k\}$, $\{\eta_k\}$, and $\{\epsilon_k\}$ satisfy Assumption [4] and $\lim_{k \to \infty} \gamma_k/(\eta_k \epsilon_k^2) = 0$. If in addition $b < c$, then $\lim_{k \to \infty} \gamma_k/(\eta_k \epsilon_k^2) = 0$.

**Proof.** The proof of this Lemma can be carried out in a similar vein to Lemma 5. We only show that part (a) is satisfied. Equivalently, we need to show that there exist $0 < \delta < 0.5$ and $K_3 \geq 0$ such that

$$\text{Term 1} \triangleq \left( \frac{\gamma_{k-1}}{\eta_{k-1} \epsilon_{k-1}^2} \right) \left( \frac{\eta_k \epsilon_k^2}{\gamma_k} \right) - 1 \leq \delta \eta_k \gamma_k, \quad \text{for any } k > K_3. \quad (31)$$
Substituting the sequences \( \{ \gamma_k \} \), \( \{ \eta_k \} \), and \( \{ \epsilon_k \} \) by their rules we obtain

\[
\text{Term 1} = \left( \frac{\gamma_0 k^{-a}}{\eta_0 k^{-b} \epsilon_0 k^{-2c}} \right) \left( \frac{\gamma_0 k + 1^{-a}}{\eta_0 k + 1^{-b} \epsilon_0 k + 1^{-2c}} \right) - 1 = \left( \frac{k + 1}{k} \right)^{a-b-2c} - 1 = \left( 1 + \frac{1}{k} \right)^{a-b-2c} - 1.
\]

Using the Taylor expansion for \((1 + x)^p\) where \( x = \frac{1}{k} \) and \( p = a - b - 2c \), it can be shown that \( \text{Term 1} = \mathcal{O}(k^{-1}) \). Suppose \( \delta \) is an arbitrary scalar in \((0,0.5)\). Multiplying and dividing by \( \delta \gamma_k \eta_k \), we obtain

\[
\text{Term 1} = \delta \gamma_k \eta_k \frac{\mathcal{O}(k^{-1})}{\delta \gamma_k \eta_k} = \delta \gamma_k \eta_k \frac{\mathcal{O}(k^{-1})}{\delta \gamma_0 \eta_0 (k+1)^{-a-b}} = \delta \gamma_k \eta_k \mathcal{O}(k^{-(1-a-b)}).
\]

Note that \( a + b < 1 \). Therefore, \( \mathcal{O}(k^{-(1-a-b)}) \to 0 \) when \( k \to 0 \). This implies that there exists some nonnegative number \( K_3 \) such that for any \( k > K_3 \), \( \mathcal{O}(k^{-(1-a-b)}) \leq 1 \). From (32) we obtain \( \text{Term 1} \leq \delta \gamma_k \eta_k \) for any \( k > K_3 \). Hence, we conclude that relation (31) holds implying that condition (a) is satisfied.

\[\text{Remark 5.} \] Figure 1 shows the feasible ranges for parameters \( a, b, \) and \( c \) when \( \gamma_k = \gamma_0 (k+1)^{-a} \), \( \eta_k = \eta_0 (k+1)^{-b} \), and \( \epsilon_k = \epsilon_0 (k+1)^{-c} \). Figure 1(a) represents the feasible set of these parameters for which the almost sure convergence is guaranteed, and Figure 1(b) shows the set for the mean-square convergence. We observe that each set is relatively large. Note that the two sets are distinct with a nonempty intersection. This corresponds well with theory in that almost-sure convergence and convergence in mean-square are not equivalent.

We conclude this section by noting that our rate statement is not altogether satisfactory in that we do not relate \( x_k \) to \( X^* \). To allow for precisely such a statement, we consider an averaging framework in the next section.

### 4 Rate of convergence analysis under weighted averaging

In the second part of this paper, our interest lies in analyzing the convergence and deriving rate statements for the averaged sequences associated with the RSSA scheme. It should be emphasized that while the underlying algorithm does not change in any way, the extracted sequence differs
in that it is a weighted average of the sequence generated by the original scheme. The $\text{aRSSA}_r$ scheme is a generalization of the classical stochastic approximation methods with averaging in two directions:

**Weighted averaging:** In the $\text{aRSSA}_r$ algorithm, the iterates $\bar{x}_k(r)$ are defined as the weighted average of $x_0, x_1, \ldots, x_k$ with the corresponding weights $\gamma_0/\sum_{t=0}^{k} \gamma_t$, $\gamma_1/\sum_{t=0}^{k} \gamma_t$, $\ldots$, $\gamma_k/\sum_{t=0}^{k} \gamma_t$. Note that when the stepsizes are decreasing, for $r > 0$ these weights are also decreasing, while for $r < 0$, the weights are increasing. When $r = 0$, $\bar{x}_k(r)$ represents the average of $x_0, x_1, \ldots, x_k$ with equal weights $\frac{1}{k}$. By allowing $r$ to be an arbitrary number, we are able to analyze the convergence rate of a class of averaging schemes. In fact, we will see that the choice of $r$ will affect the rate of convergence of a suitably defined gap function.

**Regularization and smoothing:** Similar to the first part of the paper, in the $\text{aRSSA}_r$ scheme, we employ the regularization and randomized smoothing. Using this generalization, we are able to present almost sure convergence results for the $\text{aRSSA}_r$ scheme (Prop. 5) for the sequence $\{\bar{x}_k(r)\}$ to a solution of problem (1) and also, derive the convergence rate for a gap function (Lemma 6). Note that, here we allow for the case that $\{\eta_k\}$ and $\{\epsilon_k\}$ are zero sequences (referred as $\text{aSA}_r$). In that case, the $\text{aSA}_r$ algorithm represents the classic stochastic approximation method utilizing the averaging technique.

In Section 4.1 we provide a brief background to gap functions and derive relevant bounds. We prove the almost sure convergence of the sequence derived from the $\text{aRSSA}_r$ scheme in Section 4.2. Finally, in Section 4.3, we show that the expected gap function diminishes to zero at the optimal rate of $O(1/\sqrt{k})$ and extend the result to window-based averaging.

### 4.1 An introduction to gap functions

Unlike in optimization settings where the function value provides a natural metric to measure progress, no such object naturally arises in the context of variational inequality problems. Yet, gap functions have emerged as the analog of the objective function and quantifies the optimality of a candidate solution $x$ for the problem $\text{VI}(X,F)$. It may be recalled that a function $g : X \to \mathbb{R} \cup \{-\infty, +\infty\}$ is a gap function if satisfies two properties: (i) it is restricted in sign over $X$; and (ii) $g(x) = 0$ if and only if $x$ solves $\text{VI}(X,F)$. If $g$ is a nonnegative function, then one may obtain a solution to $\text{VI}(X,F)$ by minimizing the gap function over $X$. A more expansive discussion on gap functions is provided by Larsson and Patriksson [19]. We consider a gap function that has found significant utility in the solution of monotone variational inequality problems.

**Definition 3 (Gap function).** Let $X \subseteq \mathbb{R}^n$ be a nonempty and closed set, and let the mapping $F : X \to \mathbb{R}^n$ be defined on the set $X$. Define the following function $G : X \to [0, +\infty) \cup \{+\infty\}$

$$G(x) \triangleq \sup_{y \in X} F(y)^T (x - y) \quad \text{for all } x \in X.$$  \hspace{1cm} (33)

Next, we present some properties of the described function. We make use of these relations in the convergence analysis of the scheme ($\text{aRSSA}_r$).

**Definition 4 (Weak solution).** Consider $\text{VI}(X,F)$ where the set $X \subseteq \mathbb{R}^n$ is nonempty, closed, and convex, and the mapping $F : X \to \mathbb{R}^n$ is defined on the set $X$. A vector $x^*_w \in X$ is said to be a weak solution to $\text{VI}(X,F)$ if we have

$$F(y)^T (y - x^*_w) \geq 0, \quad \text{for all } y \in X.$$  \hspace{1cm} (34)

We let $X^*_w$ denote the set of weak solutions to $\text{VI}(X,F)$.

23
Remark 6. A weak solution is considered to be a counterpart of the regular solution of $VI(X,F)$. A regular solution is $VI(X,F)$ is also referred to as a strong solution. Note that when the mapping $F$ is monotone, any strong solution of $VI(X,F)$ is also a weak solution, i.e., $X^* \subseteq X^*_w$. Moreover, when $F$ is continuous, it is known that $X^*_w \subseteq X^*$. (cf. [12]). Throughout the paper, since we assume both monotonicity and continuity of the mapping $F$, there is no distinction between a weak and strong solution.

We now derive some properties of the gap function.

Lemma 7 (Properties of $G(x)$). Consider Definition 3. We have the following properties:

(a) The function $G(x)$ given by (33) is a gap function, i.e., it satisfies the following: (i) $G(x)$ is nonnegative for any $x \in X$; and (ii) $x \in X^*$ if and only if $G(x) = 0$.

(b) Assume that the mapping $F$ is bounded over $X$, i.e., there exists a constant $C > 0$ such that $\|F(x)\| \leq C$ for any $x \in X$. Then, the following hold: (i) $G(x)$ is continuous at any $x \in X$; and (ii) If $X$ is bounded, i.e., there exists a constant $M > 0$ such that $\|x\| \leq M$ for any $x \in X$, then $G(x)$ is also bounded over $X$: $G(x) \leq 2CM$ for all $x \in X$.

4.2 Convergence analysis for the averaging schemes

Here, we derive an upper bound for the expected gap function at the averaged sequence generated by the $aRSSA$ scheme. For this, we start by providing a basic relation for the forthcoming development.

Lemma 8. Consider problem (7) and let the sequence $\{\bar{x}_k(r)\}$ be generated by the $aRSSA$ algorithm, where $\gamma_k > 0$, $\varepsilon_k \geq 0$ and $\eta_k \geq 0$ for any $k \geq 0$, and $r \in \mathbb{R}$. Suppose that Assumptions 7 and 2 hold. Then, for any $k \geq 0$ and $y \in X$ the following relation holds:

$$
\gamma_k^r F(y)^T(x_k - y) \leq \frac{1}{2} \gamma_k^{r-1} (\|x_k - y\|^2 + \|u_k - y\|^2) - \frac{1}{2} \gamma_k^{r-1} (\|x_{k+1} - y\|^2 + \|u_{k+1} - y\|^2) + \gamma_k \left(2e_k C + \eta_k M^2 + w_k^T(u_k - x_k) + \frac{\gamma_k}{2} \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k M^2 \right). \tag{35}
$$

Proof. For any $y \in X$, the non-expansivity property of the projection operator implies that

$$
\|x_{k+1} - y\|^2 = \|\Pi_X(x_k - \gamma_k (F(x_k + z_k) + \eta_k x_k + w_k)) - \Pi_X(y)\|^2 \\
\leq \|x_k - \gamma_k (F(x_k + z_k) + \eta_k x_k + w_k) - y\|^2.
$$

From the preceding relation, by noting that $F(x_k + z_k) + w_k = \Phi(x_k + z_k, \xi_k)$, we obtain

$$
\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\gamma_k (F(x_k + z_k) + \eta_k x_k + w_k)^T(x_k - y) + \gamma_k^2 \|F(x_k + z_k) + \eta_k x_k + w_k\|^2 \\
= \|x_k - y\|^2 - 2\gamma_k F(x_k + z_k)^T(x_k - y) - 2\gamma_k \eta_k x_k^T(x_k - y) - 2\gamma_k w_k^T(x_k - y) \\
+ \gamma_k^2 \|\Phi(x_k + z_k, \xi_k) + \eta_k x_k\|^2 \\
\leq \|x_k - y\|^2 - 2\gamma_k F(x_k + z_k)^T((x_k + z_k) - y) + 2\gamma_k F(x_k + z_k)^T z_k + 2\gamma_k \eta_k x_k^T y \\
- 2\gamma_k w_k^T(x_k - y) + 2\gamma_k^2 \|\Phi(x_k + z_k, \xi_k)\|^2 + 2\gamma_k^2 \eta_k M^2 \|x_k\|^2,
$$

where in the last inequality, we added and subtracted the term $2\gamma_k F(x_k + z_k)^T z_k$, dropped the term $2\gamma_k \eta_k x_k^T x_k$, and used $(a + b)^2 \leq 2a^2 + 2b^2$ to estimate the term $\|\Phi(x_k + z_k, \xi_k) + \eta_k x_k\|^2$. By
using the Cauchy-Schwartz inequality, invoking Remark 1 and by recalling \( \|z_k\| \leq \epsilon_k \), we obtain
\[
\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2 \gamma_k F(x_k + z_k)^T((x_k + z_k) - y) + 2 \gamma_k \|F(x_k + z_k)\| \|z_k\|
+ 2 \gamma_k \eta_k \|x_k\| \|y\| - 2 \gamma_k w_k^T(x_k - y) + 2 \gamma_k^2 \|\Phi(x_k + z_k, \xi_k)\|^2 + 2 \gamma_k^2 \eta_k^2 M^2
\leq \|x_k - y\|^2 - 2 \gamma_k F(x_k + z_k) - F(y)^T((x_k + z_k) - y) - 2 \gamma_k F(y)^T((x_k + z_k) - y)
+ 2 \gamma_k \epsilon_k C + 2 \gamma_k \eta_k M^2 + 2 \gamma_k w_k^T(y - x_k) + 2 \gamma_k^2 \|\Phi(x_k + z_k, \xi_k)\|^2 + 2 \gamma_k^2 \eta_k^2 M^2
\leq \|x_k - y\|^2 - 2 \gamma_k F(y)^T(x_k - y) + 2 \gamma_k \epsilon_k C + 2 \gamma_k \eta_k M^2 + 2 \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + 2 \gamma_k \eta_k^2 M^2,
\]
where in the second inequality, we add and subtract the term \( 2 \gamma_k F(y)^T((x_k + z_k) - y) \), while in the last inequality we invoke the Cauchy-Schwartz inequality to obtain \( -2 \gamma_k F(y)^T z_k \leq 2 \gamma_k \epsilon_k C \). In the last inequality, we also invoke the monotonicity property of mapping \( F \) on \( X^e \), which implies that the term \( -2 \gamma_k (F(x_k + z_k) - F(y))^T((x_k + z_k) - y) \) in the second inequality is nonpositive.

We next define an auxiliary sequence \( u_{k+1} \) as
\[
u_{k+1} = \Pi_X[u_k + \gamma_k w_k], \quad \text{for any } k \geq 0,
\]
where \( u_0 = x_0 \). By writing \( w_k^T(y - x_k) = w_k^T(u - x_k) + w_k^T(y - u_k) \), the inequality \((36)\) yields for all \( y \in X \),
\[
\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2 \gamma_k F(y)^T(x_k - y) + 4 \gamma_k \epsilon_k C + 2 \gamma_k \eta_k M^2
+ 2 \gamma_k w_k^T(u_k - x_k) + 2 \gamma_k w_k^T(y - u_k) + 2 \gamma_k^2 \|\Phi(x_k + z_k, \xi_k)\|^2 + 2 \gamma_k^2 \eta_k^2 M^2.
\]
Next, we estimate the term \( 2 \gamma_k w_k^T(y - u_k) \) by using \((37)\) to obtain for all \( y \in X \),
\[
\|u_{k+1} - y\|^2 = \|\Pi_X[u_k + \gamma_k w_k] - \Pi_X(y)\|^2 \leq \|u_k + \gamma_k w_k - y\|^2
= \|u_k - y\|^2 + 2 \gamma_k w_k^T(u_k - x_k) + \gamma_k^2 \|w_k\|^2.
\]
Therefore, we have \( 2 \gamma_k w_k^T(y - u_k) \leq \|u_k - y\|^2 - \|u_{k+1} - y\|^2 + \gamma_k^2 \|w_k\|^2 \). The preceding relation and \((38)\) imply that
\[
\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2 \gamma_k F(y)^T(x_k - y) + 4 \gamma_k \epsilon_k C + 2 \gamma_k \eta_k M^2 + 2 \gamma_k w_k^T(u_k - x_k)
+ \|u_k - y\|^2 - \|x_{k+1} - y\|^2 + \gamma_k^2 \|w_k\|^2 + 2 \gamma_k^2 \|\Phi(x_k + z_k, \xi_k)\|^2 + 2 \gamma_k^2 \eta_k^2 M^2.
\]
Rearranging the terms and multiplying both sides of the preceding inequality by \( \gamma_k^{-1}/2 \) for some constant \( r \in \mathbb{R} \), the required result follows for any \( k \geq 0 \)
\[
g_k^r F(y)^T(x_k - y) \leq \frac{1}{2} \gamma_k^{-1} \left( \|x_k - y\|^2 + \|u_k - y\|^2 \right) - \frac{1}{2} \gamma_k^{-1} \left( \|x_{k+1} - y\|^2 + \|u_{k+1} - y\|^2 \right)
+ \gamma_k \left( 2 \epsilon_k C + \eta_k M^2 + w_k^T(u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2 \right).
\]

Using Lemma 8, we next provide a generic bound for the average sequence with any \( r \in \mathbb{R} \).

**Lemma 9** (Error bounds for gap function). Consider problem \((4)\) and let the sequence \( \{\bar{x}_k(r)\}\) be produced by the \texttt{aRSSA} algorithm, where \( \gamma_k > 0, \epsilon_k \geq 0 \) and \( \eta_k \geq 0 \) for any \( k \geq 0 \), and
Consequently, by Lemma 8 (cf. relation (35)) we have for all \( k \in \mathbb{N} \),

\[
\mathbb{E}[G(\bar{x}_N(r))] \leq \frac{4M^2(\gamma_0^0 - 1 + \mathbb{I}_r \gamma_{N-1}^{-1})}{\sum_{k=0}^{N-1} \gamma_k^r} + \frac{1}{\sum_{k=0}^{N-1} \gamma_k} \left( \sum_{k=0}^{N-1} \gamma_k^r (2\epsilon_k C + \eta_k M^2 + \frac{3}{2} \gamma_k^r \eta_k^2 M^2) \right).
\]

(39)

where \( \mathbb{I}_r = 0 \) when \( r \geq 1 \) and \( \mathbb{I}_r = 1 \) when \( r < 1 \).

Proof. We consider the two cases depending on the value of \( r \), namely, \( r \geq 1 \) and \( r < 1 \).

Case of \( r \geq 1 \): Let us assume that \( r \) is an arbitrary fixed number such that \( r \geq 1 \), implying that \( r-1 \geq 0 \). Since \( \{\gamma_k\} \) is assumed to be a non-increasing sequence, it follows that \( \gamma_{k+1} \leq \gamma_k \). Consequently, by Lemma 8 (cf. relation (35)) we have for all \( k \geq 0 \) and \( y \in X \),

\[
\gamma_k^r F(y)^T(x_k - y) \leq \frac{1}{2} \gamma_0^0 - 1 (\|x_k - y\|^2 + \|u_k - y\|^2) - \frac{1}{2} \gamma_k^{r-1} (\|x_{k+1} - y\|^2 + \|u_{k+1} - y\|^2)
+ \gamma_k^r \left( 2\epsilon_k C + \eta_k M^2 + w_k^T(u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2 \right).
\]

Summing the preceding inequality from \( k = 0 \) to \( N-1 \), where \( N \geq 1 \) is a fixed number, yields

\[
\sum_{k=0}^{N-1} \gamma_k^r F(y)^T(x_k - y) \leq \frac{1}{2} \gamma_0^0 - 1 (\|x_0 - y\|^2 + \|u_0 - y\|^2) - \frac{1}{2} \gamma_N^{r-1} (\|x_N - y\|^2 + \|u_N - y\|^2)
+ \sum_{k=0}^{N-1} \gamma_k^r \left( 2\epsilon_k C + \eta_k M^2 + w_k^T(u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2 \right)
\leq 4M^2 \gamma_1^r - 1 + \sum_{k=0}^{N-1} \gamma_k \left( 2\epsilon_k C + \eta_k M^2 + w_k^T(u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2 \right),
\]

where the second inequality is a consequence of noting that \( \|x_0 - y\|^2 \leq 4M^2 \) and \( \|u_0 - y\|^2 \leq 4M^2 \), and the non-negativity of the sum \( \|x_N - y\|^2 + \|u_N - y\|^2 \). Since by the definition of \( \bar{x}_N(r) \) we have

\[
\bar{x}_N(r) = \sum_{k=0}^{N-1} \frac{\gamma_k^r}{\sum_{k=0}^{N-1} \gamma_k} x_k,
\]

we obtain for all \( y \in X \) and \( N \geq 1 \),

\[
\left( \sum_{k=0}^{N-1} \gamma_k \right) F(y)^T(\bar{x}_N(r) - y) \leq 4M^2 \gamma_1^r - 1 + \sum_{k=0}^{N-1} \gamma_k \left( 2\epsilon_k C + \eta_k M^2 + w_k^T(u_k - x_k) + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2 \right).
\]

Taking supremum over the set \( X \) with respect to \( y \) and invoking the definition of the gap function (Definition 3), we have the following inequality:

\[
\left( \sum_{k=0}^{N-1} \gamma_k \right) G(\bar{x}_N(r)) \leq 4M^2 \gamma_1^r - 1 + \sum_{k=0}^{N-1} \gamma_k \left( 2\epsilon_k C + \eta_k M^2 + \gamma_k \eta_k^2 M^2 \right)
+ \sum_{k=0}^{N-1} \gamma_k w_k^T(u_k - x_k) + \sum_{k=0}^{N-1} \gamma_k^{r+1} \left( \frac{1}{2} \|w_k\|^2 + \|\Phi(x_k + z_k, \xi_k)\|^2 \right).
\]

26
Taking expectations on both sides of the preceding inequality, we obtain

\[
\left(\sum_{k=0}^{N-1} \gamma_k^r\right) E[G(x_N(r))] \leq 4M^2\gamma_k^{r-1} + \sum_{k=0}^{N-1} \gamma_k^r \left(2\epsilon_k C + \eta_k M^2 + \gamma_k \eta_k^2 M^2\right) + \sum_{k=0}^{N-1} \gamma_k^r E[w_k^T (u_k - x_k)] \\
+ \frac{1}{2} \sum_{k=0}^{N-1} \gamma_k^{r+1} E[\|w_k\|^2] + \sum_{k=0}^{N-1} \gamma_k^{r+1} E[\|\Phi(x_k + z_k, \xi_k)\|^2].
\]  

(40)  

Term 1

\[
+ \frac{1}{2} \sum_{k=0}^{N-1} \gamma_k^{r+1} E[\|w_k\|^2] + \sum_{k=0}^{N-1} \gamma_k^{r+1} E[\|\Phi(x_k + z_k, \xi_k)\|^2].
\]

Term 2

Next, we estimate Terms 1 and 2. The ARSSA algorithm and the definition of \(u_k\) in (37) imply that \(x_k\) and \(u_k\) are both \(\mathcal{F}_k\)-measurable. Thus, the term \(u_k - x_k\) is \(\mathcal{F}_k\)-measurable. Moreover, the definition of \(w_k\) imply that \(w_k\) is \(\mathcal{F}_{k+1}\)-measurable. Therefore, for any \(k \geq 0\):

\[
E[w_k^T (u_k - x_k) | \mathcal{F}_k \cup \{z_k\}] = (u_k - x_k)^T E[w_k | \mathcal{F}_k \cup \{z_k\}] = 0,
\]

where in the last equality we have used Lemma 1. Taking expectations in the preceding equation, we obtain

\[
E[w_k^T (u_k - x_k)] = 0, \quad \text{for any } k \geq 0.
\]

(41)

Furthermore, by Lemma 1 we also have \(E[\|w_k\|^2] \leq C^2\) for all \(k \geq 0\). By Assumption 1(c) it follows that \(E[\|\Phi(x_k + z_k, \xi_k)\|^2] \leq C^2\). Substituting the preceding two upper estimates and (41) in the inequality (40), we obtain the desired inequality.

Case of \(r < 1\): Let us assume that \(r\) is an arbitrary fixed number such that \(r < 1\). Adding and subtracting the term \(0.5\gamma_k^{1-r} (\|x_k - y\|^2 + \|u_k - y\|^2)\) from the right-hand side of relation (35), we obtain the following inequality:

\[
\gamma_k^r F(y)^T (x_k - y) \leq \frac{1}{2} \gamma_k^{1-r} (\|x_k - y\|^2 + \|u_k - y\|^2) - \frac{1}{2} \gamma_k^{r-1} (\|x_k+1 - y\|^2 + \|u_k+1 - y\|^2) + \frac{1}{2} \gamma_k^{r-1} (\|x_k - y\|^2 + \|u_k - y\|^2)
\]

\[
+ \gamma_k^r \left(2\epsilon_k C + \eta_k M^2 + w_k^T (u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2\right).
\]

(35)

Since \(1 - r > 0\) and \(\{\gamma_k\}\) is non-increasing, the term \(\gamma_k^{r-1} - \gamma_k^{1-r}\) is nonnegative. Recall that we have \(\|x_k - y\|^2 \leq 2\|x_k\|^2 + \|y\|^2 \leq 4M^2\) and \(\|u_k - y\|^2 \leq 4M^2\), allowing us to claim that Term 3 \(\leq 4M^2 (\gamma_k^{r-1} - \gamma_k^{1-r})\). By using these estimates and, then, taking the summations over the resulting inequality from \(k = 1\) to \(N - 1\) for a fixed value \(N \geq 1\), and dropping the non-positive terms \(-0.5\gamma_{N-1}^{1-r} (\|x_N - y\|^2 + \|u_N - y\|^2)\) and \(-4M^2 \gamma_0^{1-r}\), we obtain the following relation for all \(y \in X\) and \(N \geq 1\),

\[
\sum_{k=1}^{N-1} \gamma_k^r F(y)^T (x_k - y) \leq \frac{1}{2} \gamma_0^{1-r} (\|x_1 - y\|^2 + \|u_1 - y\|^2) + 4M^2 \gamma_{N-1}^{r-1}
\]

\[
+ \sum_{k=1}^{N-1} \gamma_k^r \left(2\epsilon_k C + \eta_k M^2 + w_k^T (u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2\right).
\]

(42)
Consider now inequality (35) when \( k = 0 \). By adding the resulting inequality to relation (42), we obtain for all \( y \in X \) and \( N \geq 1 \),
\[
\sum_{k=0}^{N-1} \gamma_k F(y)^T (x_k - y) \leq \frac{1}{2} \gamma_0^{1-r} \left( \|x_0 - y\|^2 + \|u_0 - y\|^2 \right) + 4M^2 \gamma_{N-1}^{-1} \\
+ \sum_{k=0}^{N-1} \gamma_k \left( 2\epsilon_k C + \eta_k M^2 + w_k^T (u_k - x_k) + \frac{1}{2} \gamma_k \|w_k\|^2 + \gamma_k \|\Phi(x_k + z_k, \xi_k)\|^2 + \gamma_k \eta_k^2 M^2 \right).
\]

The remainder of the proof can be carried out in a similar fashion to that of the preceding case \((r \geq 1)\). Combining the results of both cases, we obtain the required result. \(\square\)

We now proceed to show that the expected gap function diminishes to zero as \( k \to \infty \) under suitable assumptions on the various parameter sequences. We also show that for a specific class of stepsize sequences and in the absence of smoothing and regularization, the expected gap function converges to zero at the optimal rate.

**Lemma 10** (Convergence of the expected gap function). Consider problem (1) and let sequence \( \{\bar{x}_k(r)\} \) be generated by the aRSSA algorithm. Suppose that Assumptions 1 and 2 hold. Also, assume that the sequences \( \{\gamma_k\}, \{\eta_k\}, \) and \( \{\epsilon_k\} \) are given by \( \gamma_k = \gamma_0 (k+1)^{-a} \), \( \eta_k = \eta_0 (k+1)^{-b} \), and \( \epsilon_k = \epsilon_0 (k+1)^{-c} \) with \( \gamma_0 > 0 \), \( \eta_0 \geq 0 \), \( \epsilon_0 \geq 0 \). Then, for any \( b, c > 0 \) and any \( a \) and \( r \) such that
\[
(a, r) \in \{(u, v) \mid 0 < uv \leq 1 \text{ and } v \geq 1\} \cup \{(u, v) \mid 0 < u < 1 \text{ and } v < 1\},
\]
the sequence \( \mathbb{E}[G(\bar{x}_k(r))] \) converges to zero as \( k \rightarrow \infty \).

**Proof.** We consider the two cases corresponding to the two sets defining the range of \((a, r)\).

1. Assume that \((a, r) \in \{(u, v) \mid 0 < uv \leq 1 \text{ and } v \geq 1\}\). In this case we have \( r \geq 1 \). Since \( \gamma_k = \gamma_0 (k+1)^{-a} \) is a non-increasing sequence, the conditions of Proposition 9 hold. We show that when \( 0 < ar \leq 1 \), the gap function converges to zero. Note that \( \eta_k \leq \eta_0 \) implies that \( \eta_k^2 \leq \eta_0^2 \). Let us define \( M^* \triangleq \max\{\eta_0^2 M^2, 1.5C^2\} \). From relation (39) for \( r \geq 1 \) we have
\[
\mathbb{E}[G(\bar{x}_N(r))] \leq \frac{1}{\sum_{k=0}^{N-1} \gamma_k^r} \left( 4\gamma_0^{-1} M^2 + \sum_{k=0}^{N-1} \gamma_k^r (2\epsilon_k C + \eta_k M^2 + M^* \gamma_k) \right). \tag{43}
\]

Let us define the following terms:
\[
\begin{align*}
h_N &\triangleq \left( \sum_{k=0}^{N-1} (k+1)^{-ar} \right)^{-1}, & \ell_N &\triangleq \epsilon_0 \frac{\sum_{k=0}^{N-1} (k+1)^{-(ar+c)}}{\sum_{k=0}^{N-1} (k+1)^{-ar}}, \\
m_N &\triangleq \eta_0 \frac{\sum_{k=0}^{N-1} (k+1)^{-(ar+b)}}{\sum_{k=0}^{N-1} (k+1)^{-ar}}, & p_N &\triangleq \frac{\sum_{k=0}^{N-1} (k+1)^{-(ar+a)}}{\sum_{k=0}^{N-1} (k+1)^{-ar}}. \tag{44}
\end{align*}
\]

Therefore, relation (43) implies that
\[
\mathbb{E}[G(\bar{x}_N(r))] \leq 4\gamma_0^{-1} M^2 h_N + 2C \ell_N + M^2 m_N + \gamma_0 M^* p_N. \tag{45}
\]

To show that the expectation of gap function goes to zero, it is enough to show that all the \( h_N, \ell_N, m_N, \) and \( p_N \) approach zero as \( N \rightarrow \infty \). Since we assumed that \( 0 < ar \leq 1 \), \( h_N \) goes to zero
as $N$ tends to $+\infty$. We make use of the following relations in the remainder of our analysis:

$$
\int_0^N (x+1)^{-\alpha} dx \leq \sum_{k=0}^{N-1} (k+1)^{-\alpha} \leq \int_{-1}^{N-1} (x+1)^{-\alpha} dx, \quad \text{for any } \alpha > 0, \alpha \neq 1,
$$

$$
\int_0^N (x+1)^{-\alpha} dx \leq \sum_{k=0}^{N-1} (k+1)^{-\alpha} \leq 1 + \int_0^N (x+1)^{-\alpha} dx, \quad \text{for } \alpha = 1,
$$

$$
\int_{-1}^{N-1} (x+1)^{-\alpha} dx \leq \sum_{k=0}^{N-1} (k+1)^{-\alpha} \leq \int_0^N (x+1)^{-\alpha} dx, \quad \text{for any } \alpha < 0. \quad (46)
$$

In the following, we show that $\lim_{N \to \infty} \ell_N = 0$. If $\epsilon_0 = 0$, then $\ell_N = 0$ for all $N$. Otherwise, consider the following cases:

(i) The case $ar \neq 1$ and $ar + c \neq 1$: From relation (46) we obtain the following:

$$
0 \leq \ell_N \leq \epsilon_0 \frac{\int_{-1}^{N-1} (x+1)^{-(ar+c)} dx}{\int_0^N (x+1)^{-ar} dx} \leq \epsilon_0 \frac{N^{1-(ar+c)-0}}{1-(ar)} \frac{(N+1)^{1-(ar)-1}}{(N+1)^{1-ar}} = 0.
$$

(ii) The case $ar \neq 1$ and $ar + c = 1$: Since $c > 0$, we have $ar < 1$. Relation (46) yields the following:

$$
0 \leq \ell_N \leq 1 + \epsilon_0 \frac{\int_{-1}^{N-1} (x+1)^{-(ar+c)} dx}{\int_0^N (x+1)^{-ar} dx} \leq \epsilon_0 \frac{1 + \ln(N)}{(N+1)^{1-ar}} \frac{2 - ar}{N(N+1)^2} = 0.
$$

(iii) The case $ar = 1$ and $ar + c \neq 1$: Since $c > 0$, we have $ar + c > 1$, and relation (46) implies

$$
0 \leq \ell_N \leq \epsilon_0 \frac{\int_{-1}^{N-1} (x+1)^{-(ar+c)} dx}{\int_0^N (x+1)^{-ar} dx} \leq \epsilon_0 \frac{N^{1-(ar+c)-0}}{1-(ar)} \frac{\ln(N+1)}{\ln(N+1)} = 0.
$$

In conclusion, when $r \geq 1$, when $0 < ar \leq 1$, we have $\lim_{N \to \infty} \ell_N = 0$. A similar limit can be derived for $m_N$ and $p_N$. Therefore, using relation (45) we conclude that $\lim_{N \to \infty} E[G(\bar{x}_N(r))] = 0$.

(2) Assume now that $(a, r) \in \{(u, v) \mid 0 < u < 1 \text{ and } v < 1\}$. In this case $r < 1$. From relation (39) and the definition of $M^*$ in the first part of this proof, we have

$$
E[G(\bar{x}_N(r))] \leq \frac{1}{\sum_{k=0}^{N-1} \gamma_k^r} \left( \frac{4M^2}{\gamma_0^{1-r}} + \frac{4M^2}{\gamma_1^{1-r}} + \sum_{k=0}^{N-1} \gamma_k^r (2\epsilon_k C + \eta_k M^2 + \gamma_k M^*) \right). \quad (47)
$$
Consider the definitions given by the following:

\[ q_N \triangleq \left( \sum_{k=0}^{N-1} (k + 1)^{-ar} \right)^{-1}, \quad s_N \triangleq \epsilon_0 \frac{\sum_{k=0}^{N-1} (k + 1)^{-(ar+c)}}{\sum_{k=0}^{N-1} (k + 1)^{-ar}}, \quad (48) \]

\[ t_N \triangleq \eta_0 \frac{\sum_{k=0}^{N-1} (k + 1)^{-(ar+b)}}{\sum_{k=0}^{N-1} (k + 1)^{-ar}}, \quad u_N \triangleq \epsilon_0 \frac{\sum_{k=0}^{N-1} (k + 1)^{-(ar+a)}}{\sum_{k=0}^{N-1} (k + 1)^{-ar}}, \quad (49) \]

\[ v_N \triangleq \frac{N^{a(1-r)}}{\sum_{k=0}^{N-1} (k + 1)^{-ar}}. \quad (50) \]

Relation (47) implies that

\[ E[G(\bar{x}_N(r))] \leq 2\gamma_0^{-2} M^2 q_N + Cs_N + M^2 t_N + \gamma_0 M^2 u_N + \frac{2M^2}{\gamma_0} v_N. \quad (51) \]

If \(0 < r < 1\), then using relation (46) and the definition of \(v_N\), we have

\[ 0 \leq v_N \leq \frac{N^{a(1-r)}}{\int_0^N (x + 1)^{-ar} dx} = \frac{N^{a(1-r)}}{(N+1)^{1-(ar)-1}} \quad (52) \]

\[ \implies 0 \leq \lim_{N \to \infty} v_N \leq \lim_{N \to \infty} N^{-1-a} = 0. \quad (53) \]

If \(r < 0\), then using relation (46), the definition of \(v_N\), and by noting that \(ar < 1\) since \(0 < a < 1\) and \(r < 1\), we may deduce the following:

\[ 0 \leq v_N \leq \frac{N^{a(1-r)}}{\int_{N-1}^{-1} (x + 1)^{-ar} dx} = \frac{N^{a(1-r)}}{\frac{N^{1-(ar)}}{1-(ar)} - 0} \quad (54) \]

\[ \implies 0 \leq \lim_{N \to \infty} v_N \leq \lim_{N \to \infty} N^{-1-a} = 0. \quad (55) \]

Since \(r < 1\) and \(a > 0\), we have \(1 \leq N^{a(1-r)}\) implying that \(q_N \leq v_N\) for all \(N\). Therefore, \(q_N\) tends to zero as \(N \to \infty\). To show that \(u_N\) tends to zero as \(N \to +\infty\), we consider the following cases:

(i) The case that \(a(1+r) \neq 1\) and \(r > 0\):

\[ 0 \leq u_N \leq \frac{\int_{1}^{-1} (x + 1)^{-a(1+r)} dx}{\int_0^N (x + 1)^{-ar} dx} \leq \frac{\frac{N^{-a} - 0}{1-(ar+c)}}{\frac{(N+1)^{1-(ar)-1}}{1-(ar)}} \]

\[ \implies 0 \leq \lim_{N \to \infty} u_N \leq \lim_{N \to \infty} N^{-a} = 0. \]

(ii) The case that \(a(1+r) \neq 1\) and \(-1 \leq r < 0\):

\[ 0 \leq u_N \leq \frac{\int_{-1}^{N-1} (x + 1)^{-a(1+r)} dx}{\int_{-1}^{N-1} (x + 1)^{-ar} dx} \leq \frac{\frac{N^{-a} - 0}{1-(ar+c)}}{\frac{N^{1-(ar)-1}}{1-(ar)}} \]

\[ \implies 0 \leq \lim_{N \to \infty} u_N \leq \lim_{N \to \infty} N^{-a} = 0. \]

(iii) The case that \(a(1+r) \neq 1\) and \(r < -1\):

\[ 0 \leq u_N \leq \frac{\int_0^N (x + 1)^{-a(1+r)} dx}{\int_{-1}^{N-1} (x + 1)^{-ar} dx} \leq \frac{\frac{N^{1-(ar)-1}}{1-(ar)}}{\frac{N^{-a} - 0}{1-(ar)}} \]

\[ \implies 0 \leq \lim_{N \to \infty} u_N \leq \lim_{N \to \infty} N^{-a} = 0. \]
(iv) The case that $a(1+r) = 1$: Since $a < 1$, we have $r > 0$. Thus,

$$0 \leq \frac{1 + \int_0^N (x + 1)^{-a(1+r)} dx}{\int_0^N (x + 1)^{-ar} dx} \leq \frac{1 + \ln(N)}{(N+1)^{1-(ar)}}$$

$$\implies \lim_{N \to \infty} u_N \leq \lim_{N \to \infty} \frac{\ln(N)}{(N+1)^{1-(ar)}} = \lim_{N \to \infty} \frac{2 - ar}{N(N+1)^{2-(ar)}} = 0.$$ 

In a similar fashion to the preceding analysis, one can show that $\lim_{N \to \infty} s_N = \lim_{N \to \infty} t_N = 0$.

In conclusion, in the case that $r < 1$, when $0 < a < 1$, we have $q_N, s_N, t_N, u_N, v_N$ tend to zero as $N \to +\infty$. Therefore, using relation (47) we conclude that $\lim_{N \to \infty} E[G(\bar{x}_N(r))] = 0$. □

In the following, we analyze the convergence of the averaged sequence $\bar{x}_k(r)$ to the solution set of problem (I). First, we present conditions under which a subsequence of the averaged sequence converges to the solution set almost surely.

**Proposition 4** (Almost sure convergence of subsequences of $\bar{x}_k(r)$). Consider problem (I) and suppose the conditions of Lemma 10 are satisfied. Then, the following relations hold almost surely:

(i) $\liminf_{k \to \infty} G(\bar{x}_k(r)) = 0$ and (ii) $\liminf_{k \to \infty} \text{dist}(\bar{x}_k(r), X^*) = 0$. (56)

**Proof.** (i) Since the conditions of Lemma 10 hold, we have $\lim_{k \to \infty} E[G(\bar{x}_k(r))] = 0$. Invoking Lemma 7(a) yields $G(\bar{x}_k(r)) \geq 0$ for any $k \geq 1$. Using Fatou’s lemma, we conclude that

$$\liminf_{k \to \infty} G(\bar{x}_k(r)) = 0$$

holds almost surely.

To prove (ii) in (56), we note that every accumulation point of the sequences produced by this scheme lies in $X$ by the definition of the algorithm and by the closed-ness of $X$. It follows that at every accumulation point, the gap function is nonnegative. We now proceed by contradiction and assume the result is false. Consequently, we have that

$$\liminf_{k \to \infty} \text{dist}(\bar{x}_k(r), X^*) > 0$$

with a positive probability.

Consequently, along any sequence produced by the algorithm, with positive probability, there exists no subsequence that converges to the solution set. In other words, with positive probability, we have that the gap function tends to a positive number (since the limit point lies in $X$) along every such subsequence associated with this sequence, i.e.,

$$\liminf_{k \to \infty} G(\bar{x}_k(r)) > 0$$

But this contradicts the fact that $\liminf_{k \to \infty} G(\bar{x}_k(r)) = 0$ a.s. and, hence, the result follows. □

In Proposition 4 we proved the convergence of the averaged sequence in a subsequential sense. However, in the absence of regularization and smoothing, there is no guarantee that the entire sequence $\bar{x}_k(r)$ is convergent. Motivated by this shortcoming, in sequel, we present a class of stepsize, regularization and smoothing sequences such that the entire averaging sequence is convergent in an almost-sure sense to the least norm solution of the problem. Subsequently, we also provide a rate analysis for the expected gap function when almost sure convergence is attained. We make use of the following well-known result in our analysis.
Lemma 11. Let \( \{u_t\} \subset \mathbb{R}^n \) be a convergent sequence of vectors with the limit point \( \hat{u} \in \mathbb{R}^n \). Suppose that \( \{\alpha_k\} \) is a sequence of positive numbers where \( \sum_{k=0}^{\infty} \alpha_k = \infty \). Consider the average sequence \( \{v_k\} \) given by
\[
v_k = \frac{\sum_{t=0}^{k-1} \alpha_t u_t}{\sum_{t=0}^{k-1} \alpha_t} \quad \text{for all } k \geq 1.
\]
Then, we have \( \lim_{k \to \infty} v_k = \hat{u} \).

Remark 7. Note that when \( \{x_k\} \) is a convergent sequence, from Lemma 11, the condition \( \sum_{t=0}^{\infty} \gamma_t^r = \infty \) needs to be met so that the averaging sequence \( \bar{x}_k(r) \) converges to the same limit point. When the stepsize \( \gamma_k \) is of the form \( \frac{\alpha_0}{(k+1)^p} \), this condition is equivalent to \( ar \leq 1 \). For example, when \( 0.5 < a < 1 \), \( r \) has to lie in \((-\infty, 2)\) while \( r \geq 2 \) leads to a violation of this requirement.

Proposition 5 (Almost sure convergence of the sequence \( \{\bar{x}_k(r)\} \)). Consider problem (1) and let sequence \( \{\bar{x}_k(r)\} \) be generated by the anKSSA algorithm. Suppose that Assumptions 1 and 2 hold. Also, assume that sequences \( \{\gamma_k\} \), \( \{\eta_k\} \), and \( \{\epsilon_k\} \) are given by \( \gamma_k = \gamma_0(k+1)^{-a} \), \( \eta_k = \eta_0(k+1)^{-b} \), and \( \epsilon_k = \epsilon_0(k+1)^{-c} \) with positive constants \( \gamma_0, \eta_0, \epsilon_0 \). Moreover, assume that mapping \( F \) is differentiable at \( t^* \) and its Jacobian is bounded in a neighborhood of \( t^* \). Suppose that \( (a, b, c, r) \) are chosen such that the following hold:
\[
a, b, c > 0, \quad a > 0.5, \quad a + 3b < 1, \quad b + 2c < a, \quad b < c \quad \text{and} \quad r \leq \frac{1}{a}.
\] (57)
Then, almost surely, \( \lim_{k \to \infty} \bar{x}_k(r) = t^* \) where \( t^* \) is the least norm solution of VI(\( X, F \)).

Proof. Since the conditions of Lemma 5 are satisfied, we may invoke Theorem 1b. This ensures that \( \{x_k\} \) tends to \( t^* \) in an a.s. sense. Since we assumed \( ar \leq 1 \), Lemma 11 (see Remark 7) implies that \( \lim_{k \to \infty} \bar{x}_k(r) = \lim_{k \to \infty} x_k = t^* \) almost surely. \( \square \)

4.3 Rate analysis for the gap function

In this subsection, we analyze the convergence rate of the expected gap function.

Proposition 6 (Convergence rate of gap function). Suppose the conditions of Proposition 5 are satisfied. Then, for any given \( 0 < \delta < \frac{1}{6} \), there exist some \( a, b, c, \) and \( r \) satisfying (57) for which the term \( \mathbb{E}[G(\bar{x}_k(r))] \) converges to zero with the order \( O(k^{-1/6}) \). More precisely, let \( 0 < \delta < \frac{1}{6} \) be a given number and choose \( \delta' \) such that \( 0 < \delta' < \min \left( \frac{1}{\sqrt{3}}, 1 \right) \). Suppose \( a = 0.5 + 3(\delta - \delta') \) and \( b = \frac{1}{6} - \delta, \ c = \frac{1}{6}, \) and \( r < \frac{0.5 - 3(\delta - \delta')}{3(\delta - \delta)} \). Then, \( \mathbb{E}[G(\bar{x}_k(r))] \) converges to zero with the order \( O(k^{-1/6}) \).

Proof. Note that since \( ar = 0.5 - 3(\delta - \delta') \) and \( 0 < \delta' < \delta < \frac{1}{6} \), we have \( 0 < ar < 0.5 \) and \( 0 < r < 1 \). Therefore, the inequality (39) holds for \( r < 1 \). Let us define \( M^* \triangleq \max \{a_0^2 M^2, 1.5C^2\} \) and suppose \( q_N, s_N, t_N, u_N, \) and \( v_N \) are defined by (48)–(50). It follows that
\[
\mathbb{E}[G(\bar{x}_N(r))] \leq 2 \gamma_0^{-2} M^2 q_N + C s_N + M^2 t_N + \gamma_0 M^* u_N + \frac{2 M^2}{\gamma_0} v_N.
\] (58)

From relations (46), we can write
\[
\int_{-1}^{N-1} (x+1)^{-\alpha} dx \leq \sum_{k=0}^{N-1} (k+1)^{-\alpha} \leq 1 + \int_{0}^{N} (x+1)^{-\alpha} dx, \quad \text{for any } \alpha \in \mathbb{R}.
\]
Thus, we have

\[ q_N \leq \frac{1}{N^{1-(ar)}-0} \quad \implies q_N = O(N^{1-ar}) = O(N^{0.5+3(\delta-\delta')}) , \]

\[ s_N \leq 1 + \frac{(N+1)^{(a+(b+c)-1)} N^{1-(ar)-0}}{1-(ar)} \quad \implies s_N = O(N^{-c}) = O(N^{-\frac{1}{2}}) , \]

\[ t_N \leq 1 + \frac{(N+1)^{(a+(b+c)-1)} N^{1-(ar)-0}}{1-(ar)} \quad \implies t_N = O(N^{-b}) = O(N^{-\left(\frac{1}{2}-\delta\right)}) , \]

\[ u_N \leq 1 + \frac{(N+1)^{(a+(b+c)-1)} N^{1-(ar)-0}}{1-(ar)} \quad \implies u_N = O(N^{-a}) = O(N^{-\left(0.5+3(\delta-\delta')\right)}) , \]

\[ v_N \leq \frac{N^{a(1-r)}}{N^{1-(ar)-0}} \quad \implies v_N = O(N^{-(1-a)}) = O(N^{-\left(0.5-3(\delta-\delta')\right)}) . \]

Note that since \( \delta' < \frac{1-6\delta}{9} \), it follows that \( \frac{1}{6} - \delta < 0.5 - 3(\delta - \delta') \). Therefore, from (58), we obtain

\[ \mathbb{E}[G(\bar{x}_N(r))] = O(N^{-\left(\frac{1}{2}-\delta\right)}) . \]

In the next set of results, we set the regularization and smoothing parameters to zero i.e., \( \eta_k = \epsilon_k = 0 \) for all \( k \geq 0 \). It follows that the aRSSA_r algorithm without regularization and smoothing reduces to a modified aSA_r algorithm given by:

\[
\bar{x}_{k+1}(r) \triangleq \sum_{t=0}^{k} \gamma_t^r x_t ,
\]

\[ x_{k+1} = \Pi_X (x_k - \gamma_k \Phi(x_k, \xi_k)) , \quad \text{(aSA_r)} \]

First, we show that the expected gap function of the averaged sequence generated by the aSA_r algorithm converges to zero at the optimal rate of \( O(1/\sqrt{k}) \) for \( r < 1 \).

**Proposition 7** (Optimal rate of convergence for \( \bar{x}_k(r) \)). Consider problem \( \mathcal{P} \) and let sequence \( \{\bar{x}_k(r)\} \) be generated by the aSA_r algorithm and suppose that Assumptions \( \mathcal{I} \) and \( \mathcal{A}(a) \) hold. Then, we have the following results:

(a) Suppose \( \gamma_k = \gamma_0(k+1)^{-a} \) with \( \gamma_0 > 0 \) and with any \( a \) and \( r \) such that

\[
(a,r) \in \{(u,v) \mid 0 < u < 1 \text{ and } v < 1 \text{ and } u(1+v) \neq 1\} .
\]

Then, \( \mathbb{E}[G(\bar{x}_k(r))] \) converges to zero as \( k \to \infty \) at the rate \( O(k^{-\min\{a,1-a\}}) \).

(b) Suppose \( \gamma_k = \frac{\gamma_0}{\sqrt{k+1}} \) with \( \gamma_0 > 0 \) and \( r < 1 \). Then, \( \mathbb{E}[G(\bar{x}_k(r))] \) converges to zero as \( k \to \infty \) at the rate \( O\left(\frac{1}{\sqrt{k}}\right)\).

**Proof.** (a) Since the conditions of Lemma 10 hold, we have

\[
\lim_{k \to \infty} \mathbb{E}[G(\bar{x}_k(r))] = 0, \quad \text{for any } \{(u,v) \mid 0 < u < 1 \text{ and } v < 1\}. \]

Therefore,

\[
\lim_{k \to \infty} \mathbb{E}[G(\bar{x}_k(r))] = 0, \quad \text{for any } \{(u,v) \mid 0 < u < 1 \text{ and } v < 1 \text{ and } u(1+v) \neq 1\} .
\]
Consider the definitions of $q_N, s_N, t_N, u_N$ and $v_N$ in (48)–(50) in the proof of Lemma 10. Since $\eta_k = \epsilon_k = 0$ for all $k \geq 0$, it follows that $s_N = t_N = 0$. From $r < 1$ and $a > 0$, we have $1 \leq N^{a(1-r)}$ implying that $q_N \leq v_N$. Therefore, relation (51) can be rewritten as

$$E[G(\bar{x}_N(r))] \leq \left(4\gamma_0^r - 2M^2 + \frac{4M^2}{\gamma_0}\right) v_N + \gamma_0 M^s u_N.$$  

When $(a, r) \in \{(u, v) | 0 < u < 1 \text{ and } v < 1 \text{ and } u(1+v) \neq 1\}$, from the preceding relation and the proof of Lemma 10(b), we obtain the following inequality:

$$E[G(\bar{x}_N(r))] \leq O\left(N^{-a}\right) + O\left(N^{-(1-a)}\right) = O\left(N^{-\min\{a,1-a\}}\right).$$

(b) In this case, $a = 0.5$. Since $r < 1$, we have $a(1+r) = 0.5(1+r) < 1$ implying that $(0.5, r) \in \{(u, v) | 0 < u < 1 \text{ and } v < 1 \text{ and } u(1+v) \neq 1\}$. From part (a), we have

$$E[G(\bar{x}_N(r))] \leq O\left(N^{-\min\{0.5,1-a\}}\right) = O\left(N^{-0.5}\right).$$

Comparing this result with the more standard averaging scheme that uses $r = 1$ (cf. [24]) supports the idea of using $r < 1$ for the averaging sequence $\bar{x}_k(r)$. Specially, when $r < 0$ and the sequence $\gamma_k$ is decreasing, the weights in the averaging sequence grow implying that more recently generated iterates are attributed more weight. A more general form of the aSA algorithm is when the average sequence is calculated using a window-based formula given by Algorithm aSA$_{\ell,r}$. The following result is derived using Proposition 9. Note that its proof is similar to that of Proposition 9 assuming $\eta_k = \epsilon_k = 0$ for all $k \geq 0$. Also, note that in the absence of regularization, the multiplier of $C^2$ changes from 1.5 to 1.

**Corollary 1.** Consider problem (1) and let the sequence $\{\bar{x}_k(r)\}$ be generated by the aSA$_{\ell,r}$ algorithm, where $\gamma_k > 0$ and $r \in \mathbb{R}$. Suppose Assumptions 4 and 5 hold, and let the stepsize sequence $\{\gamma_k\}$ be non-increasing. Then,

$$E\left[G(\bar{x}_N^\ell(r))\right] \leq \frac{1}{\sum_{k=\ell}^{N-1} \gamma_k} \left(4M^2(\gamma_{\ell-1}^{r-1} + \gamma_{N-1}^{r-1}\mathbb{I}_r) + C^2 \sum_{k=\ell}^{N-1} \gamma_k^{r+1}\right),$$  

(59)

where $0 \leq \ell \leq N - 1$, $N \geq 1$, $\mathbb{I}_r = 0$ when $r \geq 1$ and $\mathbb{I}_r = 1$ when $r < 1$.

**Remark 8.** The above result generalizes the bound in [24] in two directions. First, instead of assuming $r = 1$, we allow for $r$ to be a real number, leading to the addition of the term $\gamma_{N-1}^{r-1}\mathbb{I}_r$. Second, we derive this bound for the gap function of monotone variational inequality problems, while the bound in [24] addresses convex stochastic optimization problems. Our generalization leads to a slightly different bound; specifically, in that $C^2$ in (59) is replaced by $0.5C^2$ in the optimization setting.

Next, we develop a window-based diminishing stepsize rule and provide an associated rate result.

**Proposition 8** (A generalized window-based averaging scheme). Consider problem (1) and let the sequence $\{\bar{x}_N^\ell(r)\}$ be generated by the aSA$_{\ell,r}$ algorithm, where $r \in \{1, -1\}$ and $\ell = \lceil N\rceil$ for a fixed $\lambda \in (0, 1)$ with $N > \frac{1}{\frac{1}{1+\lambda}}$. Suppose that Assumptions 4 and 5 hold and the stepsize sequence $\{\gamma_N\}$ is given by

$$\gamma_N = \frac{2M\sqrt{1+\mathbb{I}_r}}{C\sqrt{N+1}}, \quad \text{for all } N \geq 0.$$  

Then, $E\left[G(\bar{x}_N^\ell(r))\right] = O\left(\frac{1}{\sqrt{N}}\right)$.  

34
Proof. Note that \( \gamma_0 = \frac{2M\sqrt{1+\ell}}{C} \) implying \( \gamma_N = \frac{\gamma_0}{\sqrt{N+1}} \). We consider two cases: \( r = 1 \) and \( r = -1 \).

(i) The case of \( r = 1 \): From (59) we obtain
\[
E\left[G(\bar{x}_N^\ell(1))\right] \leq \frac{4M^2 + \gamma_0^2 N^2 \sum_{k=1}^{N-1} \frac{1}{k+1}}{\gamma_0} \leq \frac{4M^2}{\gamma_0} \left( 1 + \frac{1}{\sqrt{N+1}} \right) \leq \frac{4M^2}{\gamma_0} \left( 1 + \frac{1}{\sqrt{N+1}} \right) \leq 2MC \left( \frac{2 + \ln \left( \frac{N}{r+1} \right)}{2(\sqrt{N+1} - \sqrt{r+1})} \right) \leq UB1,
\]
where in the second inequality, we bound the summation by the integral. Using the definition of \( \lambda \), we may write
\[
UB1 \leq M\left(\frac{1 + \lambda^{-1}}{\sqrt{N+1} - \sqrt{\lambda N} + 2}\right),
\]
implying that \( E[G(\bar{x}_N^\ell(1))] = O\left(\frac{1}{\sqrt{N}}\right) \).

(ii) The case of \( r = -1 \): We have \( \gamma_0 = \frac{2\sqrt{2}M}{C} \). From (59) we obtain
\[
E\left[G(\bar{x}_N^\ell(-1))\right] \leq \frac{4M^2 \gamma_0^2 (\ell + 1 + N - 1) + C^2 \sum_{k=\ell}^{N-1} 1}{\gamma_0^2 \sum_{k=\ell}^{N-1} \sqrt{k+1}} \leq 2MC \left( \frac{N+\ell+1}{\sqrt{N}} + \sqrt{2}(N-\ell) \right) \leq 2MC \left( \frac{1}{\sqrt{N+1}} \right) \leq UB2.
\]
We can write
\[
UB2 \leq M\left(\frac{3\sqrt{2}(3N-\ell-1)}{2(\sqrt{N} - \sqrt{\ell})(N+\ell + \sqrt{N}\ell)}\right) \leq M\left(\frac{3\sqrt{2}(3 - \lambda + \frac{1}{N})}{2(\sqrt{N} - \sqrt{\lambda N} + 1)(3\lambda)}\right),
\]
implying that \( E[G(\bar{x}_N^\ell(-1))] = O\left(\frac{1}{\sqrt{N}}\right) \). \( \square \)

Remark 9. Proposition 8 shows that using the window-based averaging scheme for both \( r = 1 \) and \( r = -1 \), the expected gap function converges to zero at the optimal rate. However, one natural question is which value of \( r \) is better to choose. To address this question, we compare the estimates \( UB1 \) and \( UB2 \) with each other and we show that when \( N \geq 2 \), \( UB2 < UB1 \) for any choice of \( \lambda \). To simplify the analysis assume \( \lambda N \) is integer, implying that \( \ell = \lambda N \). We can write
\[
\frac{UB1}{UB2} = \left(\frac{1 + \frac{N}{\ell}}{1.5\sqrt{2}(3N-\ell+1)}\right) \left(\frac{N^{1.5} - \ell^{1.5}}{\sqrt{N+1} - \ell}\right) \geq \left(\frac{1 + \frac{1}{N}}{1.5\sqrt{2}(3N-\ell+1)}\right) \left(N + \ell + \sqrt{N}\ell\right)
\]
\[
= \left(\frac{1 + \frac{1}{\lambda}}{1.5\sqrt{2}(3 - \lambda + \frac{1}{N})}\right) \left(\frac{1 + \frac{1}{\lambda}}{1.5\sqrt{2}(3 - \lambda + \frac{1}{N})}\right) \geq \left(\frac{1 + \frac{1}{\lambda}}{1.5\sqrt{2}(3 - \lambda + \frac{1}{N})}\right) \triangleq h(\lambda).
\]
Minimizing the convex function \( h(\lambda) \) for \( 0 < \lambda < 1 \) using Matlab, the global minimizer is \( \lambda^* = 0.602 \) and \( h(\lambda^*) = 1.029 \) implying that when \( N \geq 2 \), \( UB2 < UB1 \) for any choice of \( 0 < \lambda < 1 \).
5 Numerical Results

In this section, we compare the performance of our schemes through a set of computational experiments conducted on a stochastic Nash-Cournot game. In Section 5.1, we introduce the stochastic Nash-Cournot game and derive the (sufficient) equilibrium conditions, which are compactly stated as a stochastic variational inequality. In Section 5.2, the simulation results for the RSSA scheme are presented and support the asymptotic a.s. and mean-square convergence results from Theorem 1 and Proposition 3, respectively. Next, in Section 5.3, we provide the simulation results the aRSSA scheme across different values of parameter \( r \). Throughout this section, we use the gap function’s evaluation as the metric for our comparisons. To calculate the gap value, we use the commercial solver KNITRO.

5.1 A networked stochastic Nash-Cournot game

A classical example of a Nash game is a networked Nash-Cournot game \([22, 14]\). In this problem, there are \( I \) firms that compete over a network of \( J \) nodes in selling a product. Each firm \( i \) wants to maximize profit by choosing nodal production at every node \( j \), denoted by \( g_{ij} \), and the level of sales at node \( j \), denoted by \( s_{ij} \). Let \( \bar{s}_j = \sum_{i=1}^{I} s_{ij} \) denote the aggregate sales at node \( j \). By the Cournot structure, we assume that the price at node \( j \), denoted by \( P_j(\bar{s}_j, \xi) \), is a nonlinear stochastic function of the form \( a_j - b_j \bar{s}_j^\gamma \), where \( a_j \) is a uniform random variable drawn from \([lb_{aj}, ub_{aj}]\), and \( b_j \) and \( \gamma \) are constants. Furthermore, we assume the firm \( i \)’s cost of production at node \( j \) is denoted by the \( C_{ij}(g_{ij}) \triangleq c_j g_{ij} + d_j \), where \( c_j \) and \( d_j \) are constants. Other than the nonnegativity constraints for \( s_{ij} \) and \( g_{ij} \), there are two types of constraints. Firm \( i \)’s production at node \( j \) is capacitated by \( \text{cap}_{ij} \). Also, the aggregated level of sales of each firm is equal to the aggregated level of production. Therefore, firm \( i \)’s optimization problem is given by the following (Note that we assume transportation costs are zero):

\[
\min_{x_i \in X_i} \mathbb{E}[f_i(x, \xi)],
\]

where \( x = (x_1; \ldots; x_I) \) with \( x_i = (g_i; s_i), g_i = (g_{i1}; \ldots; g_{iJ}), s_i = (s_{i1}; \ldots; s_{iJ}) \),

\[
f_i(x, \xi) \triangleq \sum_{j=1}^{J} (C_{ij}(g_{ij}) - P_j(\bar{s}_j, \xi)s_{ij}),
\]

and \( X_i \triangleq \left\{ (g_i, s_i) \mid \sum_{j=1}^{J} g_{ij} = \sum_{j=1}^{J} s_{ij}, \quad g_{ij}, s_{ij} \geq 0, \quad g_{ij} \leq \text{cap}_{ij}, \text{ for all } j = 1, \ldots, J \right\} \).

Applying the interchange between the expectation and the derivative operator, the resulting equilibrium conditions of the preceding stochastic Nash-Cournot game can be compactly captured by the stochastic variational inequality \( \text{VI}(X,F) \) where \( X \triangleq \prod_{i=1}^{I} X_i \) and \( F(x) = (F_1(x); \ldots; F_I(x)) \) with \( F_i(x) = \mathbb{E}[\nabla_{x_i} f_i(x, \xi)] \). Note that it can be shown that when \( 1 < \sigma \leq 3 \) and \( I \leq \frac{3\sigma-1}{\sigma-1} \), or \( \sigma = 1 \), the mapping \( F \) is strictly monotone. We consider a Cournot competition with 5 firms and 4 nodes, i.e., \( I = 5 \) and \( J = 4 \). We assume \( \sigma = 1, [lb_{aj}, ub_{aj}] = [49.5, 50.5], \text{cap}_{ij} = 300, b_j = 0.05, c_j = 1.5 \) for all \( i \) and \( j \). Throughout this section, we assume the mean and the standard deviation of the gap function is calculated using a sample of size 50. Also, we assume the starting point of algorithms is the origin, unless stated otherwise.

Throughout this section we use the following notation: \( N \) denotes the simulation length in the scheme, \( x_0 \) denotes the starting point of the algorithm. Furthermore, the gap function is given by
Defintion[3]. We examine both the RSSA scheme, its averaged variant given by aRSSA, for different values of r as well as the window-based variant denoted by aRSSA_{ℓ,r}. In the aRSSA_{ℓ,r} scheme, ℓ is assumed to be equal to [λN] where 0 < λ < 1 is a constant. I denotes the number of firms and J denotes the number of nodes.

5.2 Convergence of the RSSA scheme

In this section, we present the simulation results for the RSSA scheme and report the performance of the algorithm using the sample mean and sample standard deviation of the gap function. Table 2 shows the results for 4000 iterations. S(i) refers to the setting of parameters a, b, and c. Recall that these are the parameters of the stepsize γ_k, regularization η_k, and the smoothing sequence ε_k. More precisely, we assumed γ_k = γ_0(k + 0.1N)^−a, η_k = η_0(k + 1)^−b, and ε_k = ε_0(k + 1)^−c where γ_0 = 1, η_0 = 10^{−4} and ε_0 = 10^{−2}. Note that the term 0.1N is added to stabilize the performance of the SA scheme. Furthermore, we chose η_0 and ε_0 to be smaller when b and c are small, respectively. In the first 9 settings, our goal is to study the sensitivity of the RSSA algorithm with respect to the parameters a, b, and c. In these settings, the values of the parameters given in the table satisfy conditions of both Lemma 5 and Lemma 6. In the third group, we increase a and keep b and c unchanged. In the second group, b is increasing, while in the third group c is increasing. We observe that increasing a slows down the convergence of the gap function, but increasing b or c speeds up the convergence of the gap function slightly. This makes sense because the optimal rate of convergence is attained at a = 0.5. On the other hand, by making b or c, larger the regularization and smoothing sequences decay to zero faster implying that the perturbations introduced in the SA algorithm due to regularization and smoothing techniques are fading out. We also observe that the average value of the gap function is more sensitive to the change in the parameter a while being more robust to the changes in b or c. In setting S(10), the parameters ensure convergence in the mean-squared sense provided by Lemma 6 but do not suffice in ensuring almost sure convergence provided in Lemma 5. In the setting S(11), the converse holds. Figure 2 illustrates the sample mean of gap function over the set of simulations for settings S(10) and S(11). Both plots show the sample mean for the gap function. The round dots in these plots represent the observed gap values for each of the 50 sample paths at every 100 iterations. We observe from Figure 2(a), that although the average gap function is approaching zero, the variance across sample paths is relatively large. This suggests that there may be sample paths that may not converge to the solution set. This observation is aligned with the knowledge that the choice of (a, b, c) do not guarantee almost sure convergence for S(10). However, in Figure 2(b) the conditions of almost sure convergence are met, we observe that the variance of the observed data is far smaller and all of the 50 trajectories remain close to the sample mean suggesting that almost all trajectories converge to the solution set.

Table 2: RSSA algorithm with different settings of parameters a, b, and c

| S(i) | Parameters | Gap function |
|------|------------|--------------|
|      | a          | b            | c            | mean        | std         |
| 1    | 0.504      | 0.099       | 0.290       | 0.120e−3   | 2.68e−3    |
| 2    | 0.600      | 0.099       | 0.290       | 1.03e−1    | 9.48e−3    |
| 3    | 0.700      | 0.099       | 0.290       | 5.90e+1    | 1.74e−1    |
| 4    | 0.504      | 0.166       | 0.167       | 8.56e−4    | 3.56e−4    |
| 5    | 0.501      | 0.130       | 0.167       | 4.89e−3    | 2.68e−3    |
| 6    | 0.501      | 0.166       | 0.167       | 4.33e−3    | 2.11e−3    |
| 7    | 0.501      | 0.166       | 0.130       | 5.05e−3    | 2.05e−3    |
| 8    | 0.501      | 0.166       | 0.100       | 5.05e−3    | 2.42e−3    |
| 9    | 0.501      | 0.166       | 0.167       | 4.33e−3    | 2.11e−3    |
| 10   | 0.401      | 0.200       | 0.190       | 7.75e−3    | 9.04e−3    |
| 11   | 0.600      | 0.133       | 0.099       | 9.42e−2    | 7.56e−3    |
5.3 Convergence of the aRSSA<sub>r</sub> and aRSSA<sub>ℓ,r</sub> schemes

Our goal lies in comparing the performance of the averaging schemes across different values of r. Motivated by Prop. 8, the stepsize used in our analysis is assumed to be of the form $\gamma_k = \frac{2M}{C\sqrt{k+1}}$ where M is the bound on the Euclidean norm of $x \in X$ and C represents the bound on the norm on mapping F over the set X. Note that here we use identical stepsizes for $r = 1$ and $r = -1$, to allow for using the same iterates generated by the SA algorithm for both schemes. In Table 3, we report the sample mean of the gap function over 50 samples. Note that in this table, the rows correspond to the value of the parameter $\lambda$ which changes from 0 to 1 in steps of 0.1. Note that $\lambda = 0$ implies that $\ell = 0$, implying that this is the aRSSA<sub>r</sub> scheme. Moreover, $\lambda = 1$ corresponds to the SA scheme without averaging since $\ell = N$ in this case. Cases when $\lambda$ is between 0 and 1 correspond to the aRSSA<sub>ℓ,r</sub> scheme. The columns in the table are sorted based on the iteration number $N$ from 1000 to 4000. Each column includes the results for the case that $r = -1$ and the standard choice $r = +1$. Naturally, when $\lambda = 1$, there is no averaging and in the last row of the table, the gap values is identical for both $r = -1$ and $r = +1$. Importantly, we see that for any value of $N$ and $\lambda$ (except for a few cases), both the aRSSA<sub>r</sub> scheme and the aRSSA<sub>ℓ,r</sub> scheme have lower gaps when $r = -1$. Specifically, when $\lambda$ is small, this difference becomes significant. For example, the case that $\lambda = 0$ and $N = 1000$, the gap value for aSA<sub>ℓ,r</sub> scheme with $r = -1$ is about 39, while this gap is nearly two orders of magnitude larger at 1190 when $r = +1$. We show this difference in Figure 3 and Figure 4. It can be seen that when comparing both averaging schemes,
both aRSSA_r and aRSSA_{\ell,r} have a lower gap for r = -1 vs r = 1 for any of the examined values of N. Recall that the original motivation of averaging schemes lay in developing a higher level of robustness to the underlying randomness; The smaller the value of \( \lambda \), the more iterates of \( x_k \) are used in the averaging schemes, implying the more robustness of the SA scheme. Therefore, there is a trade-off between increasing \( \lambda \) and the robustness of SA scheme. When \( \lambda \) is large, although the difference between the performance of \( r = -1 \) and \( r = +1 \) becomes small, however as shown in Table 3 the case \( r = -1 \) almost always has a smaller gap value than the case \( r = 1 \).

![Figure 3: Gap function: aRSSA_r scheme (l) and aRSSA_{\ell,r} schemes with \( \lambda = 0.1 \) (r)](image)

![Figure 4: Gap function: aRSSA_{r,\ell} scheme with \( \lambda = 0.2 \) (l) and \( \lambda = 0.3 \)).](image)

One question that may arise here is what value of \( \lambda \) is the best choice for the aSA_{\ell,r} scheme. We observe that the answer to this question depends on N. When N is small, in this case 1000, the SA scheme has the minimum error implying that \( \lambda = 1 \) performs the best among other values. However, for larger values of N, the minimum error occurs for smaller \( \lambda \) and the larger N, the smaller the value of \( \lambda \). For example, at N = 4000, \( \lambda = 0.5 \) has the smallest error.

39
5.4 Sensitivity analysis

In this section we investigate the performance of the averaging schemes when some parameters of the Cournot game change. First, we increase the number of firms from 5 to 15 and maintain other parameters fixed. Table 4 shows the simulation results for the new problem with 15 firms. We observe that the results are similar to the case where $I = 5$. Importantly, for almost any $N$ and any $\lambda < 1$, the averaging schemes perform better with $r = -1$ than with $r = +1$. Specifically, when $\lambda$ is small this difference is significant. Next, we assume that $x_0$ for every sample path is a point $r = 0$.

Table 4: Comparison of Gap function for $I = 15$

| Scheme | $N = 1000$ | $N = 2000$ | $N = 4000$ | $N = 6000$ |
|--------|------------|------------|------------|------------|
| $r$    | $\lambda$ | $r = -1$   | $r = +1$   | $r = -1$   | $r = +1$   | $r = -1$   | $r = +1$   |
| ---    | ---        | ---        | ---        | ---        | ---        | ---        | ---        |
| aSA   | 0.1        | 0.14       | 0.14       | 0.14       | 0.14       | 0.14       | 0.14       |
|       | 0.2        | 0.28       | 0.28       | 0.28       | 0.28       | 0.28       | 0.28       |
|       | 0.3        | 0.42       | 0.42       | 0.42       | 0.42       | 0.42       | 0.42       |
|       | 0.4        | 0.56       | 0.56       | 0.56       | 0.56       | 0.56       | 0.56       |
|       | 0.5        | 0.70       | 0.70       | 0.70       | 0.70       | 0.70       | 0.70       |
|       | 0.6        | 0.84       | 0.84       | 0.84       | 0.84       | 0.84       | 0.84       |
|       | 0.7        | 0.98       | 0.98       | 0.98       | 0.98       | 0.98       | 0.98       |
|       | 0.8        | 1.12       | 1.12       | 1.12       | 1.12       | 1.12       | 1.12       |
|       | 0.9        | 1.26       | 1.26       | 1.26       | 1.26       | 1.26       | 1.26       |
|       | 1.0        | 1.40       | 1.40       | 1.40       | 1.40       | 1.40       | 1.40       |

where $s_{ij} = g_{ij} = 150$ for any $i, j$, rather than the origin. Table 5 provides the simulation results for this case. The performance of all the schemes is similar to the original setting. We observe that the averaging schemes have a smaller expected gap function when $r = -1$ than when $r = +1$. Lastly, we are interested in observing the performance of the averaging schemes for other choices of $r$. In Prop. 7(b) we showed that the optimal rate of convergence is attained when $r < 1$. Our goal is to compare the case $r = +1$ with two other cases where $r = -0.5$ and $r = +0.5$. In this study, we used the original settings of parameters. Table 6 presents the results of this simulation. Interestingly, comparing these results with those in Table 5 we see that both cases $r = -0.5$ and $r = +0.5$ have a superior performance to $r = +1$. It is worth noting that when $\lambda \leq 0.6, r = -0.5$ tends to perform better than $r = 0.5$. A natural question that emerges is the best choice of $r$. While one may conjecture that that when $r < 1$, the performance of the averaging scheme improves as $r$ tends to $-\infty$, this is not true. Consider a setting when $r$ goes to $-\infty$. Consequently, $\bar{x}_N$ tends to $x_{N-1}$ implying the SA scheme represents the case with $r = -\infty$. However, for example in Table 3 when $r = -1$ and $\lambda = 0.5$ the aRSSA scheme performs better than the SA scheme. Therefore, decreasing $r$ would not necessarily speed up the convergence of the gap function. Finding the best choice for $r$ requires more analysis and remains the subject of future research.

6 Concluding remarks

We consider a stochastic variational inequality problem with monotone and possibly non-Lipschitzian maps over a closed, convex, and compact set. Much of the past research aimed at deriving almost
Table 6: Comparison of gap function: $r = -0.5$ and $r = +0.5$

| Scheme       | $N=1000$ | $N=2000$ | $N=3000$ | $N=4000$ |
|--------------|----------|----------|----------|----------|
|              | $r = -0.5$ | $r = +0.5$ | $r = -0.5$ | $r = +0.5$ | $r = -0.5$ | $r = +0.5$ | $r = -0.5$ | $r = +0.5$ |
| aSA          | 7.45e+1  | 4.13e+2  | 1.40e+1  | 1.48e+2  | 5.13e+0  | 8.00e+1  | 2.58e+0  | 5.24e+1  |
| aSA$_r$      | 0.1      | 1.23e+1  | 1.90e+1  | 6.53e-1  | 1.10e+0  | 7.58e-2  | 1.37e-1  | 1.34e-2  |
|              | 0.2      | 1.28e+1  | 2.68e+1  | 6.53e-1  | 1.10e+0  | 7.58e-2  | 1.37e-1  | 1.34e-2  |
|              | 0.3      | 6.84e+0  | 8.78e+0  | 2.38e-1  | 3.31e-1  | 2.04e-2  | 2.96e-2  | 3.01e-3  |
|              | 0.4      | 3.98e+0  | 4.66e+0  | 1.02e-1  | 1.26e-1  | 7.16e-3  | 8.99e-3  | 1.23e-3  |
|              | 0.5      | 2.47e+0  | 2.72e+0  | 4.93e-2  | 5.60e-2  | 3.33e-3  | 3.75e-3  | 8.61e-4  |
|              | 0.6      | 1.61e+0  | 1.70e+0  | 2.64e-2  | 2.83e-2  | 2.09e-3  | 2.19e-3  | 8.58e-4  |
|              | 0.7      | 1.09e+0  | 1.12e+0  | 1.57e-2  | 1.63e-2  | 1.72e-3  | 1.74e-3  | 1.09e-3  |
|              | 0.8      | 7.66e-1  | 7.74e-1  | 1.04e-2  | 1.06e-2  | 1.76e-3  | 1.77e-3  | 1.26e-3  |
|              | 0.9      | 5.53e-1  | 5.54e-1  | 7.65e-3  | 7.68e-3  | 1.84e-3  | 1.84e-3  | 1.60e-3  |
| SA          | 4.12e-1  | 4.12e-1  | 6.64e-3  | 6.64e-3  | 2.97e-3  | 2.97e-3  | 2.36e-3  | 2.36e-3  |

Appendix

Proof of Lemma 7

Proof. (a) We start by showing part (i). Let $x \in X$ be an arbitrary fixed vector in the set $X$. We have

$$G(x) = \sup_{y \in X} F(y)^T (x - y) \geq F(z)^T (x - z), \quad \text{for any } z \in X.$$

For $z = x$, the preceding inequality implies that $G(x) \geq F(x)^T (x - x) = 0$. Therefore, the gap function (33) is nonnegative for any $x \in X$, thus showing part (i). To prove part (ii), assume that $x^* \in X_w^*$. Relation (34) implies that

$$F(y)^T (x_w^* - y) \leq 0, \quad \text{for any } y \in X.$$

Invoking the definition of $G$ in (33), from preceding inequality we have

$$G(x_w^*) \leq 0, \quad \text{for any } y \in X.$$

However, since $x_w^* \in X$, Lemma 7(a) indicates that $G(x_w^*) \geq 0$. Therefore, we conclude that for any $x^* \in X_w^*$, we have $G(x^*) = 0$. Now assume that $G(x) = 0$ for some $x \in X$. Therefore,
sup_{y \in X} F(y)^T(x - y) = 0 implying that \( F(y)^T(x - y) \leq 0 \) for any \( y \in X \). Equivalently, we have \( F(y)^T(y - x) \geq 0 \) for any \( y \in X \). This implies that \( x \in X^*_u \).

(b)(i) Let \( \{u_k\} \subset \mathbb{R}^n \) be an arbitrary sequence in \( X \) such that \( \lim_{k \to \infty} u_k = u_0 \). Since \( X \) is a closed set, we have \( u_0 \in X \). We want to show that \( \lim_{k \to \infty} G(u_k) = G(u_0) \). We show this relation in two steps. First, using relation (33), for any \( k \geq 0 \), we have

\[
G(u_k) = \sup_{y \in X} F(y)^T(u_k - y) = \sup_{y \in X} F(y)^T(u_k - u_0 + u_0 - y)
\]

\[
= \sup_{y \in X} \left( F(y)^T(u_0 - y) + F(y)^T(u_k - u_0) \right)
\]

\[
\leq \sup_{y \in X} F(y)^T(u_0 - y) + \sup_{y \in X} F(y)^T(u_k - u_0),
\]

where in the second relation we add and subtract \( u_0 \), and in the last relation we used the well-known inequality \( \sup_{A} (f + g) \leq \sup_{A} f + \sup_{A} g \) for any two real valued functions \( f \) and \( g \) defined on the set \( A \). Using the Cauchy-Schwarz inequality and relation (60), we obtain for any \( k \geq 0 \),

\[
G(u_k) \leq \sup_{y \in X} F(y)^T(u_0 - y) + \sup_{y \in X} (\|F(y)\| \|u_k - u_0\|)
\]

\[
= \sup_{y \in X} F(y)^T(u_0 - y) + \|u_k - u_0\| \sup_{y \in X} \|F(y)\| \leq \sup_{y \in X} F(y)^T(u_0 - y) + C \|u_k - u_0\|,
\]

where in the last inequality we used the boundedness assumption of the mapping \( F \) over the set \( X \). Taking limits on both sides of the preceding inequality, we obtain

\[
\lim_{k \to \infty} G(u_k) \leq \lim_{k \to \infty} \left( \sup_{y \in X} F(y)^T(u_0 - y) + C \|u_k - u_0\| \right)
\]

\[
= \sup_{y \in X} F(y)^T(u_0 - y) + C \lim_{k \to \infty} \|u_k - u_0\| = G(u_0),
\]

where the last relation is obtained by recalling that \( u_0 \) is the limit point of the sequence \( \{u_k\} \). In the second step of the proof for continuity of \( G(x) \), using relation (33), for any \( y \in X \) and any \( k \geq 0 \), we have \( G(u_k) \geq F(y)^T(u_k - y) \). Let \( v \in \mathbb{R}^n \) be an arbitrary fixed vector in \( X \). Therefore, the preceding inequality holds for \( y = v \), i.e.,

\[
G(u_k) \geq F(v)^T(u_k - v).
\]

Taking limit from both sides of the preceding inequality when \( k \) goes to infinity, we have

\[
\lim_{k \to \infty} G(u_k) \geq \lim_{k \to \infty} F(v)^T(u_k - v) = F(v)^T \left( \lim_{k \to \infty} (u_k) - v \right) = F(v)^T(u_0 - v).
\]

Since the preceding relation holds for any arbitrary \( v \in X \), taking supremum from the right-hand side and using the relation (33) we obtain

\[
\lim_{k \to \infty} G(u_k) \geq \sup_{v \in X} F(v)^T(u_0 - v) = G(u_0).
\]

From (61) and (62), we conclude that the gap function \( G(x) \) is continuous at any \( x \in X \).

(b)(ii) For any \( x, y \in X \) we have

\[
F(y)^T(x - y) \leq \|F(y)\|\|x - y\| \leq \|F(y)\| (\|x\| + \|y\|) \leq 2CM,
\]

where the first, second, and third inequalities follow from the Cauchy-Schwarz inequality, the triangle inequality, and the boundedness assumption on the mapping \( F \) and the set \( X \). Taking the supremum over \( y \in X \) in the preceding relation and by using (33), we obtain the desired result.
Proof of Lemma 11.

Proof. Let $\epsilon > 0$ be an arbitrary scalar. Since $\lim_{t \to \infty} u_t = \hat{u}$, there exists an integer $T_1$ such that
\[ \|u_t - \hat{u}\| < \frac{\epsilon}{2}, \quad \text{for any } t > T_1. \]  
(63)
Moreover, since $\lim_{k \to \infty} \alpha_t = \infty$, there exists an integer $T_2$ such that
\[ \sum_{t=0}^{T_1} \alpha_t > \frac{2}{\epsilon} \left\| \sum_{t=0}^{T_1-1} \alpha_t (u_t - \hat{u}) \right\|, \quad \text{for any } t > T_2. \]  
(64)
Let $k$ be an integer such that $k > \max\{T_1, T_2\}$. We can write
\[ \|v_k - \hat{u}\| = \left\| \left( \sum_{t=0}^{k-1} \alpha_t \right)^{-1} \sum_{t=0}^{k-1} \alpha_t u_t - \hat{u} \right\| = \left( \sum_{t=0}^{k-1} \alpha_t \right)^{-1} \left\| \sum_{t=0}^{k-1} \alpha_t u_t - \hat{u} \right\| \]
\[ \leq \left( \sum_{t=0}^{k-1} \alpha_t \right)^{-1} \left\| \sum_{t=0}^{T_1-1} \alpha_t (u_t - \hat{u}) \right\| + \left( \sum_{t=0}^{k-1} \alpha_t \right)^{-1} \left\| \sum_{t=T_1}^{k-1} \alpha_t (u_t - \hat{u}) \right\|, \]  
(65)
where the inequality is a consequence of invoking the triangle inequality. The second term in (65) can be seen to be less than $\epsilon/2$ by invoking (63), while (64) implies that the first term is less than $\epsilon/2$, implying the following:
\[ \|v_k - \hat{u}\| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \quad \text{for any } k > \max\{T_1, T_2\}. \]
Therefore, it follows that $\lim_{k \to \infty} v_k = \hat{u}$. \qed

References

[1] D. P. Bertsekas, *Stochastic optimization problems with nondifferentiable cost functionals*, Journal of Optimization Theory and Applications 12 (1973), no. 2, 218–231.

[2] J. R. Birge and F. Louveaux, *Introduction to stochastic programming: Springer series in operations research*, Springer, 1997.

[3] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM Journal on Optimization 9 (1999), 877–900.

[4] D. Cicek, M. Broadie, and A. Zeevi, *General bounds and finite-time performance improvement for the Kiefer-Wolfowitz stochastic approximation algorithm*, Operations Research 59 (2011), 1211–1224.

[5] J. C. Duchi, P. L. Bartlett, and Martin J. Wainwright, *Randomized smoothing for stochastic optimization*, SIAM Journal on Optimization (SIOPT) 22 (2012), no. 2, 674–701.

[6] Y. M. Ermoliev, *Stochastic quasigradient methods and their application to system optimization*, Stochastics 9 (1983), 1–36.
[7] F. Facchinei and J.-S. Pang, Finite-dimensional variational inequalities and complementarity problems. Vols. I,II, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[8] S. Ghadimi and G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, part I: a generic algorithmic framework, SIAM Journal on Optimization 22 (2012), no. 4, 1469–1492.

[9] E. G. Golštein and N. V. Tretyakov, Modified Lagrangians and monotone maps in optimization, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Inc., New York, 1996, Translated from the 1989 Russian original by Tretyakov, A Wiley-Interscience Publication. MR MR1386892 (97a:90001)

[10] E. Gustavsson, M. Patriksson, and A.-B. Stromberg, Primal convergence from dual subgradient methods for convex optimization, Mathematical Programming (2014), 1–26 (English).

[11] H. Jiang and H. Xu, Stochastic approximation approaches to the stochastic variational inequality problem, IEEE Transactions in Automatic Control 53 (2008), no. 6, 1462–1475.

[12] A. Juditsky, A. Nemirovski, and C. Tauvel, Solving variational inequalities with stochastic mirror-prox algorithm, Stochastic Systems 1 (2011), no. 1, 17–58.

[13] _____, Solving variational inequalities with stochastic mirror-prox algorithm, Stochastic Systems (2011), DOI: 10.1214/10–SSY011, 17–58.

[14] A. Kannan and U. V. Shanbhag, Distributed computation of equilibria in monotone Nash games via iterative regularization techniques, SIAM Journal of Optimization 22 (2012), no. 4, 1177–1205.

[15] A. Kannan, U. V. Shanbhag, and H. M. Kim, Addressing supply-side risk in uncertain power markets: stochastic Nash models, scalable algorithms and error analysis, Optimization Methods and Software 28 (2013), no. 5, 1095–1138.

[16] J. Koshal, A. Nedić, and U. V. Shanbhag, Regularized iterative stochastic approximation methods for variational inequality problems, IEEE Transactions on Automatic Control 58(3) (2013), 594–609.

[17] H. J. Kushner and G. G. Yin, Stochastic approximation and recursive algorithms and applications, Springer, New York, 2003.

[18] H. Lakshmanan and D. Farias, Decentralized recourse allocation in dynamic networks of agents, SIAM Journal on Optimization 19 (2008), no. 2, 911–940.

[19] T. Larsson and M. Patriksson, A class of gap functions for variational inequalities, Mathematical Programming 64 (1994), 53–79.

[20] S. Lu, Symmetric confidence regions and confidence intervals for normal map formulations of stochastic variational inequalities, To appear.

[21] S. Lu and A. Budhiraja, Confidence regions for stochastic variational inequalities, Math. Oper. Res. 38 (2013), no. 3, 545–568.

[22] C. Metzler, B.F. Hobbs, and J.-S. Pang, Nash-cournot equilibria in power markets on a linearized dc network with arbitrage: Formulations and properties, Networks and Spatial Theory 3 (2003), no. 2, 123–150.
A. Nedić and S. Lee, *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM Journal on Optimization **24** (2014), no. 1, 84–107.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization **19** (2009), no. 4, 1574–1609.

V. I. Norkin, *The analysis and optimization of probability functions*, Tech. report, International Institute for Applied Systems Analysis technical report, 1993, WP-93-6.

B.T. Polyak, *Introduction to optimization*, Optimization Software, Inc., New York, 1987.

B.T. Polyak and A.B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim. **30** (1992), no. 4, 838–855.

U. Ravat and U. V. Shanbhag, *On the characterization of solution sets in smooth and nonsmooth stochastic convex Nash games*, SIAM Journal of Optimization **21** (2011), no. 3, 1046–1081.

H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statistics **22** (1951), 400–407.

R.T. Rockafellar and R.J-B Wets, *Variational analysis*, Springer, Berlin, 1998.

A. Shapiro, *Monte carlo sampling methods*, Handbook in Operations Research and Management Science, vol. 10, Elsevier Science, Amsterdam, 2003, pp. 353–426.

A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, The society for industrial and applied mathematics and the mathematical programming society, Philadelphia, USA, 2009.

V. A. Steklov, *Sur les expressions asymptotiques de certaines fonctions définies par les équations différentielles du second ordre et leurs applications au problème du développement d'une fonction arbitraire en séries procédant suivant les diverses fonctions*, Comm. Charkov Math. Soc. **2** (1907), no. 10, 97–199.

M. Wang and D. P. Bertsekas, *Incremental constraint projection methods for variational inequalities*, Mathematical Programming (2014), DOI 10.1007/s10107–014–0769–x.

H. Xu, *Sample average approximation methods for a class of stochastic variational inequality problems*, Asia-Pacific Journal of Operational Research **27** (2010), no. 1, 103–119.

F. Yousefian, A. Nedić, and U. V. Shanbhag, *On stochastic gradient and subgradient methods with adaptive steplength sequences*, Automatica **48** (2012), no. 1, 56–67, An extended version of the paper available at: http://arxiv.org/abs/1105.4549.

F. Yousefian, A. Nedić, and U. V. Shanbhag, *A regularized smoothing stochastic approximation (RSSA) algorithm for stochastic variational inequality problems*, Proceedings of the 2013 Winter Simulation Conference (2013), 933–944.

F. Yousefian, A. Nedić, and U. V. Shanbhag, *Self-tuned stochastic approximation schemes for nonsmooth stochastic multi-user optimization and Nash games*, Under review, an extended version of the paper is available at http://arxiv.org/abs/1301.1711, 2014.