Variance-Reduced Splitting Schemes for Monotone Stochastic Generalized Equations

Shisheng Cui and Uday V. Shanbhag

Abstract—We consider monotone inclusion problems with expectation-valued operators, a class of problems that subsumes convex stochastic optimization problems with possibly smooth expectation-valued constraints as well as subclasses of stochastic variational inequality and equilibrium problems. A direct application of splitting schemes is complicated by the need to resolve problems with expectation-valued maps at each step, a concern addressed via sampling. Accordingly, we propose an avenue for addressing uncertainty in the mapping: variance-reduced stochastic modified forward-backward splitting scheme. We consider structured settings when the map can be decomposed into an expectation-valued map and a maximal monotone map in a tractable resolution. We show that the proposed schemes are equipped with a.s. convergence guarantees, linear (strongly monotone $A$) and $O(1/k)$ (monotone $A$) rates of convergence while achieving optimal oracle complexity bounds. The rate statements in monotone regimes appear to be among the first and leverage the Fitzpatrick gap function for monotone inclusions. Furthermore, the schemes rely on weaker moment requirements on noise and allow for weakening unbiasedness requirements on oracles in strongly monotone regimes. Preliminary numerics on a class of two-stage stochastic variational inequality problems reflect these findings and show that the variance-reduced schemes outperform stochastic approximation schemes and sample-average approach approaches. The benefits of achieving deterministic rates of convergence become even more salient when resolvent computation is expensive.

Index Terms—Splitting schemes, stochastic approximation, stochastic generalized equations.

I. INTRODUCTION

The generalized equation (GE) (alternately referred to as the inclusion problem) represents a crucial mathematical object in decision and control theory, representing a set-valued generalization to the more standard root-finding problem, which requires solving $F(x) = 0$, where $F: \mathbb{R}^n \to \mathbb{R}^n$ is single-valued. Specifically, if $T$ is a set-valued map, defined as $T: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and $T$ is characterized by a distinct structure in that it can be cast as the sum of two operators $A$ and $B$, then the GE takes the form

$$0 \in T(x) \triangleq A(x) + B(x). \quad \text{(GE)}$$

Here $A : \mathbb{R}^n \to \mathbb{R}^n$ is a single-valued map and $B : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued map. While such objects have a storied history, an excellent overview was first provided by Robinson [1]. GE have been extensively examined since the 70s when Rockafellar [2] developed a proximal point scheme for a GE characterized by monotone operators, subsuming schemes such as the augmented Lagrangian method [3], Douglas–Rachford splitting [4], among others. A large class of optimization and equilibrium problems can be modeled as (GE), including the necessary conditions of nonlinear programming problems, variational inequality and complementarity problems, and a range of equilibrium problems (cf. [1]). Under suitable requirements on $A$ and $B$, a range of splitting methods can be developed, representing a vibrant area of research over the last two decades [4, 5, 6, 7].

Here, we address the stochastic counterpart of GE, a problem class recently studied via sample-average approximation (SAA) techniques [7]. Formally, the stochastic GE requires an $x \in \mathbb{R}^n$ such that

$$0 \in \mathbb{E}[T(x, \xi(\omega))] \triangleq \mathbb{E}[A(x, \xi(\omega))] + B(x) \quad \text{(SGE)}$$

where the components of the map $A$ are denoted by $A_i$, $i = 1, \ldots, n$, $\xi : \Omega \to \mathbb{R}^d$ is a random variable, $A_i : \mathbb{R}^n \times \Omega \rightrightarrows \mathbb{R}^n$ is a set-valued map, $\mathbb{E}[\cdot]$ denotes the expectation, and the associated probability space is given by $(\Omega, \mathcal{F}, \mathbb{P})$. In the remainder of this article, we refer to $A(x, \xi(\omega))$ by $A(x, \omega)$. The expectation of a set-valued map leverages the Aumann integral [8] and is formally defined as $\mathbb{E}[A_i(x, \xi(\omega))] = \left\{ \int v_i(\omega)dP(\omega) \mid v_i(\omega) \in A_i(x, \xi(\omega)) \right\}$. Consequently, the expectation $\mathbb{E}[A(x, \omega)]$ can be defined as a Cartesian product of the sets $\mathbb{E}[A_i(x, \omega)]$, defined as $\mathbb{E}[A_i(x, \omega)] = \prod_{i=1}^n \mathbb{E}[A_i(x, \omega)]$. We motivate (SGE) by considering some examples. Consider the stochastic convex optimization problem given by $\min_{x \in \chi} \mathbb{E}[g(x, \omega)]$, where $g(\bullet, \omega)$ is a differentiable convex function for every $\omega$ and $\chi$ is a closed and convex set. Such a problem can be equivalently stated as $0 \in T(x) \triangleq \mathbb{E}[G(x, \omega)] + N_\chi(x)$, where $G(x, \omega) = \nabla_x g(x, \omega)$ and $N_\chi(x)$ denotes the normal cone of $\chi$ at $x$. In fact, single-valued stochastic variational inequality problems [9] can be cast as stochastic inclusions as well as seen by $0 \in T(x) \triangleq \mathbb{E}[F(x, \omega)] + N_\chi(x)$, where $F(\bullet, \omega)$ is a realization of the mapping. This introduces...
a pathway for examining stochastic analogs of traffic equilibrium [10] and Nash equilibrium problems [11] as well as a myriad of other problems subsumed by variational inequality problems [12].

We now provide a brief review of prior research.

1) **Stochastic operator splitting schemes:** The regime where the maps are expectation valued has seen relatively less study [15]. Stochastic proximal gradient schemes [16] are an instance of stochastic operator splitting techniques, where $A$ is either the gradient or the subdifferential operator. In the context of monotone inclusions, when $A$ is a more general monotone operator, a.s. convergence of the iterates has been proven in [13] when $A$ is Lipschitz and expectation-valued while $B$ is maximal monotone. In fact, in our preliminary work [17], we prove a.s. convergence and derive an optimal rate in terms of the gap function in the context of stochastic variational inequality problems with structured operators. Stability analysis [18] and two-timescale variants [19] have also been examined. A rate statement of $O(\frac{1}{k})$ in terms of mean-squared error has also been provided when $A$ is additionally strongly monotone in [13]. A comparison of rate statements for stochastic operator-splitting schemes is provided in Table I from which we note that (vr-SMBFS) is equipped with deterministic (optimal) rate statements, optimal or near-optimal sample-complexity, a.s. convergence guarantees, and does not require conditionally unbiased oracles in strongly monotone regimes. We believe our rate statements here and in [17] are among the first in maximal monotone settings (to the best of our knowledge).

2) **Other related schemes:** A natural approach for resolving SGEs is via SAA [20]. It proves a.s. convergence and establishes the rate of convergence in expectation under a strong monotonicity assumption.

### A. Gaps and Resolution

There are several gaps in the literature and we provide the corresponding resolution in this paper as follows.

1) **Poorer empirical performance when residuals are costly:** Deterministic schemes for strongly monotone and monotone GEs display linear and $O(1/k)$ rate in resolvent operations while stochastic analogs display rates of $O(1/k)$ and $O(1/\sqrt{k})$, respectively, leading to poorer practical behavior when the resolvent is challenging to compute, e.g., in strongly monotone regimes, worst case complexity in resolvent operations increases from $O(\log(1/\epsilon))$ to $O(1/\epsilon)$. Our scheme (vr-SMBFS) achieves deterministic rates of convergence with optimal/near-optimal oracle complexities in both monotone and strongly monotone regimes, allowing for run times comparable to deterministic counterparts.

2) **Absence of rate statements for monotone operators:** To the best of our knowledge, there appear to be no nonasymptotic rate statements available in monotone regimes. In (vr-SMBFS), rate statements are now provided.

3) **Biased oracles:** Often, conditional unbiasedness of the oracle may not hold. Our proposed scheme allows for possibly biased oracles in some select settings.

4) **State-dependent bounds on subgradients and second moments:** Many subgradient and stochastic approximation (SA) schemes impose bounds of the form $E[\|G(x,\omega)\|^2] \leq M^2$ where $G(x,\omega) \in T(x,\omega)$ or $E[\|w\|^2 | x] \leq \sigma^2$ where $w = A(x,\omega) - \mathbb{E}[A(x,\omega)]$. Both sets of assumptions are often challenging to impose in non-compact regimes. Our scheme can accommodate state-dependent bounds to allow for non-compact domains.

### B. Outline and Contributions

We now articulate our contributions. In Section III, we consider the resolution of monotone inclusions in structured regimes where the map can be expressed as the sum of two maps, facilitating the use of splitting. In this context, when one of the maps is expectation valued, we prove a.s. convergence and derive an optimal rate when $A$ is additionally strongly monotone in [13]. A comparison of rate statements for stochastic operator-splitting schemes is provided in Table I from which we note that (vr-SMBFS) is equipped with deterministic (optimal) rate statements, optimal or near-optimal sample-complexity, a.s. convergence guarantees, and does not require conditionally unbiased oracles in strongly monotone regimes. We believe our rate statements here and in [17] are among the first in maximal monotone settings (to the best of our knowledge).

Before proceeding, we comment on the term **variance-reduced (VR).**

1) **Terminology and applicability:** The moniker “variance-reduced” reflects the usage of increasingly accurate approximations of the expectation-valued map, as opposed to noisy sampled variants that are used in single sample
schemes. The resulting schemes are often referred to as \textit{minibatch} SA schemes and often achieve deterministic rates of convergence. This work draws inspiration from the early work by Friedlander and Schmidt [21] and Byrd et al. [22], which demonstrates how increasing sample sizes can enable achieving deterministic rates of convergence. Similar findings regarding the nature of sampling rates have been presented in [23]. This avenue has proven particularly useful in developing accelerated gradient schemes for both smooth [24] and non-smooth [25] convex/nonconvex stochastic optimization, VR quasi-Newton schemes [26], among others. Schemes such as SVRG [27] and SAGA [28] also achieve deterministic rates of convergence but are customized for finite-sum problems, unlike minibatch schemes that can process expectations over general probability spaces. Unlike in minibatch schemes where increasing batch sizes are employed, in schemes such as SVRG, the entire set of samples is periodically employed for computing a step.

2) \textit{Weaker assumptions and stronger statements}: The proposed VR framework has several crucial benefits, often unavailable in single-sample regimes. For instance, rate statements are derived in monotone regimes, which have hitherto been unavailable. Second, a.s. convergence guarantees are obtained. Our results often rely on weaker moment and bias assumptions. Finally, the schemes allow for deterministic rates, leading to superior practical behavior as the numerics reveal.

3) \textit{Sampling requirements}: Naturally, VR schemes can generally be employed only when sampling is relatively cheap compared to the main computational step (such as computing a projection or a prox.) In terms of overall sample-complexity, the proposed schemes are near optimal. As \( k \) becomes large, one might question how one might contend with \( N_k \) tending to \( +\infty \). This issue does not arise given our interest in \( \epsilon \)-approximations; e.g., if \( \epsilon = 10^{-3} \), then such a scheme requires approximately \( \mathcal{O}(10^3) \) steps (in monotone settings). Since \( N_k \approx [k^\alpha] \) and \( \alpha > 1 \), we require approximately \( \mathcal{O}(10^3)^{1+\alpha} \) samples in total. In a setting where multicore architecture is ubiquitous, such requirements are not terribly onerous particularly since computational costs have been reduced from \( \mathcal{O}(10^6) \) (single-sample) to \( \mathcal{O}(10^3) \). It is worth noting that finite-sum problems routinely have \( 10^6 \) or more samples and competing schemes such as SVRG require evaluating the full batch-size intermittently; i.e., such schemes use at least \( \mathcal{O}(10^9) \) samples to achieve similar accuracies as our scheme.

II. \textbf{BACKGROUND}

In this section, we provide some background on splitting schemes, building a foundation for the subsequent sections. Consider the \textit{GE}

\[
0 \in T(x) \triangleq A(x) + B(x). \quad \text{(GE)}
\]

If the resolvent of either \( A \) or \( B \) (or both) is tractable, then splitting schemes assume relevance. Notable instances include Douglas–Rachford splitting [4], [6], Peaceman–Rachford splitting [5], [6], and forward–backward splitting [6], [29].

1) \textit{Douglas–Rachford splitting}: In this scheme, the resolvent of \( A \) and \( B \) can be separately evaluated to generate a sequence defined as follows:

\[
x_{k+\frac{1}{2}} := (I + \gamma A)^{-1}(x_k)
\]

\[
x_{k+1} := (I + \gamma B)^{-1}(2x_{k+\frac{1}{2}} - x_k) \quad \text{(DRS)}
\]

\[
+ x_k - x_{k+\frac{1}{2}}.
\]

2) \textit{Peaceman–Rachford splitting}: In contrast, in the Peaceman–Rachford splitting method, the roles of \( A \) and \( B \) are exchanged in each iteration, given by the following:

\[
x_{k+\frac{1}{2}} := (I + \gamma B)^{-1}(I - \gamma A)(x_k)
\]

\[
x_{k+1} := (I + \gamma A)^{-1}(I - \gamma B)(x_{k+\frac{1}{2}}). \quad \text{(PRS)}
\]

3) \textit{Forward–backward splitting}: Moreover, if the resolvent of \( B \) is easier to evaluate and \( A \) and \( B \) are maximal monotone, the forward–backward splitting method was applied to convex optimization in [30]

\[
x_{k+1} := (I + \gamma B)^{-1}(I - \gamma A)(x_k). \quad \text{(FBS)}
\]

In [14], a stochastic variant of the FBS method, developed for strongly monotone maps, is equipped with a rate of \( \mathcal{O}(1/k) \) while in [13], maximal monotone regimes are examined and a.s. convergence statements are provided. A drawback of (FBS) is the requirement of either a strong monotonicity assumption on \( A^{-1} \) or that \( A \) be Lipschitz continuous on \( \text{dom}(A) = \mathbb{R}^n \) and \( T \) be strongly monotone; this motivated the modified FBS (MFBS) scheme where convergence was proven when \( A \) is monotone and Lipschitz [31]

\[
x_{k+\frac{1}{2}} := (I + \gamma B)^{-1}(I - \gamma A)(x_k)
\]

\[
x_{k+1} := x_{k+\frac{1}{2}} - \gamma (A(x_{k+\frac{1}{2}}) - A(x_k)). \quad \text{(MFBS)}
\]

In Section III, we develop a VR stochastic MFBS scheme where \( A \) is Lipschitz and monotone, \( A(x) \triangleq \mathbb{E}[A(x, \omega)] \), and \( B \) is maximal monotone with a tractable resolvent; we derive linear and sublinear convergence under strongly monotone and merely monotone \( A \), respectively, achieving deterministic rates of convergence.

A. \textit{Two Motivating Examples}

1) \textit{Subclass of Stochastic Multileader Multifollower Games}: Consider a class of multileader multifollower games [32] with \( N \) leaders, denoted by \( \{1, \ldots, N\} \) and \( M \) followers, given by \( \{1, \ldots, M\} \). In general, this class of games is challenging to analyze since the player problems are nonconvex, and early existence statements have relied on eliminating follower-level decisions, leading to a noncooperative game with convex nonsmooth player problems. Adopting a similar approach in examining a stochastic generalization of a quadratic setting examined in [33] with a single follower where \( M = 1 \),
suppose the follower problem is
\[ \min_{y_i \in l_i(x_i)} \frac{1}{2} y_i^T Q_i y_i - b_i(x_i)^T y_i \quad \text{(Follow}_i(x_i)) \]
where \( Q_i \) is a positive definite and diagonal matrix, \( b_i(\bullet) \) and \( l_i(\bullet) \) are affine functions. Suppose the leaders compete in a Cournot game in which the \( i \)th leader solves
\[ \min_{x_i \in X_i} c_i(x_i) - E[p(X, \omega)x_i] + a_i y_i(x_i) \quad \text{(Leader}_i(x_{-i})) \]
where \( c_i : X_i \to \mathbb{R} \) is a smooth convex function, the inverse-demand function \( p(\bullet) \) is defined as \( p(X) = d(\omega) - r(\omega)X \), \( d(\omega), r(\omega) > 0 \) for every \( \omega \in \Omega \), \( X \triangleq \sum_{i=1}^N x_i \), \( y_i(x_i) \) denotes a best response of follower \( i \), and \( X_i \) is a closed and convex set in \( \mathbb{R} \). Follower \( i \)'s best response \( y_i(x) \), given leader-level decisions \( x \), can be derived by considering the necessary and sufficient conditions of optimality
\[ y_i(x_i) = \max\{Q_i^{-1}b_i(x_i), l_i(x_i)\}. \]
Consequently, we may eliminate the follower-level decision in the leader-level problem, leading to a nonsmooth stochastic Nash equilibrium problem given by the following:
\[ \min_{x_i \in X_i} c_i(x_i) - E[p(X, \omega)x_i] + a_i y_i(x_i) \quad \text{(Leader}_i(x_{-i})) + a^T \max\{Q_i^{-1}b_i(x_i), l_i(x_i)\}. \]
Under convexity of \( b_i(\bullet) \) and \( l_i(\bullet) \), and suitable assumptions on \( Q_i \) and \( a_i \), the expression \( a_i^T \max\{Q_i^{-1}b_i(x_i), l_i(x_i)\} \) is a convex function in \( x_i \), a fact that follows from observing that this term is a scaling of the maximum of two convex functions. Consequently, the necessary and sufficient equilibrium conditions of this game are given by \( 0 \in \nabla c_i(x_i) + E[r(\omega)(X + x_i) - d(\omega)] + \partial h_i(x_i) + N_{\mathcal{X}_i}(x_i) \) for \( i = 1, \ldots, N \) where \( h_i(\bullet) \) defined as \( h_i(x_i) \triangleq a_i^T \max\{Q_i^{-1}b_i(x_i), l_i(x_i)\} \) is a convex function in \( x_i \). Then, the necessary and sufficient equilibrium conditions are given by
\[ 0 \in T(x) \triangleq A(x) + B(x) \quad \text{(SGE mil)} \]
where \( A(x) \triangleq G(x) + R(x), B(x) \triangleq D(x) + N_{\mathcal{X}_i}(x) \).
Here, \( G(x) \triangleq (\nabla c_i(x_i))_{i=1}^N \), \( R(x) \triangleq E[r(\omega)(X + x) - d(\omega)] \), and \( D(x) \triangleq \sum_{i=1}^N E[\partial h_i(x_i)] \). We observe that \( G(x) \) is a monotone map while \( D(x) \) is the Cartesian product of the expectations of subdifferentials of convex functions, implying that \( D(x) \) is also monotone. Furthermore, \( R(x) \) is monotone since \( \nabla_x R(x) = E[r(\omega)(1 + 11T)] \geq 0 \). Since \( N_{\mathcal{X}_i} \) is a normal cone of a convex set, which is a monotone map \([12], T \) is monotone.

2) Model-Predictive Control With Probabilistic and Risk Constraints: Model-predictive control (MPC) is a framework for the control of complex systems \([34]\). It obviates the challenging derivation/computation of a feedback control law with a repeated resolution of a finite-horizon-constrained optimization problem. Contending with uncertainty has prompted the development of several approaches. (i) Robust approaches. Robust frameworks for MPC \([35], [36]\) often require bounded and deterministic descriptions of uncertainty, a property inherited from robust optimization \([37]\); (ii) Probabilistic framework. Under a probabilistic representation of the uncertainty, a chance-constrained MPC framework \([38]\) can be adopted, allowing for shaping the probability distribution of system states. Such avenues have assumed relevance in climate control, process control, power systems operation, and vehicle path planning (cf. \([39]\) for a survey). Suppose the dynamics are captured by a linear discrete-time system, defined as
\[ x_{t+1} = A(\delta)x_t + B_w(\delta)u_t + B_w(\delta)w_t \]
where \( x_0 \) is given, \( x_t \in \mathcal{X} \) denotes the state of the system at time \( t, u_t \in \mathcal{U} \subseteq \mathbb{R}^m \) represents the control input vector at time \( t \), and \( w_t \in \mathbb{R}^p \) is an unmeasurable disturbance signal at time \( t \). In addition, \( \mathcal{X} \) and \( \mathcal{U} \) represent the set of states and controls, respectively, while the random matrices \( A(\delta), B_w(\delta), B_w(\delta) \) lie in \( \mathbb{R}^{n \times n}, \mathbb{R}^{n \times m} \), and \( \mathbb{R}^{m \times p} \), respectively. We assume access to the distributions governing \( \delta \) and \( w_k \). Suppose \( \pi \triangleq \{\pi_0(\cdot), \ldots, \pi_{N-1}(\cdot)\} \) represents feedback-control policy where \( \pi_i : \mathbb{R}^n \to \mathcal{U} \subseteq \mathbb{R}^m \) denotes the state feedback control law for \( i = 0, 1, \ldots, N - 1 \). We may then formally define the value function \( V_N(x, \pi) \triangleq E_{x_0}[\sum_{i=1}^{N-1} J(x_i, u_i) + J_N(x_N)] \) where \( E_{x_0}[\bullet] \triangleq E[\bullet | x_0] \). Here, \( J_N(\bullet) \) represents the terminal cost function while \( J(x_i, u_i) \) represents the stage-wise cost of taking control \( u_i \) while at state \( x_i \). In addition, suppose \( \mathcal{X}_c \) denotes a set of undesirable outcomes. The resulting chance-constrained stochastic control problem requires determining the feedback-control law \( \pi \) that minimizes \( V_N(x, \pi) \) subject to the prescribed dynamics and probabilistic requirements on the state. This problem is challenging, motivating the construction of a finite-horizon open-loop counterpart. To this end, we define \( x_{t:T} \) and \( u_{t:T} \) as \( x_{t:T} \triangleq \{x_t, \ldots, x_{t+T-1}\} \) and \( u_{t:T} \triangleq \{u_t, \ldots, u_{t+T-1}\} \), respectively, while the finite-horizon value function at the \( t \)th step looking \( T \) periods ahead, denoted by \( V_t(x_{t:T}, u_{t:T}) \), is defined as \( V_t(x_{t:T}, u_{t:T}) \triangleq E_{x_t}\left[\sum_{i=t}^{t+T-1} J(x_i, u_i)\right] \) where \( T < N \). Given a horizon \( T \), the resulting MPC framework \([40]\) requires minimizing \( V_t(x_{t:T}, u_{t:T}) \) subject to the prescribed dynamics and the probabilistic state-constraints, given \( x_t \). A formal definition of the chance-constrained stochastic control problem (CC-SC) and its finite-horizon counterpart (CC-MPC(2)) is provided next.

The decision control \( u_t \) is obtained from resolving (CC-MPC(2)) and is then applied to the system after which the window is moved ahead. The resulting problem (CC-MPC(T)) is then resolved when \( t + T < N \) (alternatively, the horizon \( T \) is reduced appropriately). This formulation is relatively flexible and can be used to address diverse types of objectives and constraints. In general, the problem (CC-MPC(T)) is challenging, owing to the presence of the chance constraint. The probability function can be recast as an expectation of an indicator function over a set but this leads to discontinuous integrands. Recently, the second author has developed avenues where under prescribed

Authorized licensed use limited to the terms of the license agreement with IEEE. Restrictions apply.
assumptions under which the following holds [41]:
\[ P[\zeta \in K | \zeta \in K(x)] = E_{\zeta}[F(x, \zeta)], \]
where \( K \) is a set in \( \mathbb{R}^n \) symmetric about the origin, \( K(x) \) is defined as \( K(x) \triangleq \{ \zeta : c(x, \zeta) \geq 0 \} \), \( T \in \mathbb{R}^{d \times n} \), and
\[
c(x, \zeta) \triangleq \begin{cases} 1 - |c^T x|^m, & \text{Setting A} \\ T^T - \zeta, & \text{Setting B}. \end{cases}
\]

The integrand \( F(\bullet, \zeta) \) is defined appropriately in Settings A and B where in each case, it is shown that \( \zeta \). In fact, we can then show that a composition of \( E[F(\bullet, \zeta)] \) is convex; e.g., in Setting A, \( 1/E[F(\bullet, \zeta)] \) is convex. For expositional ease, we may recast (CC-MPC) as the following chance-constrained problem (CCP) and provide its necessary and sufficient optimality conditions in (SGE CCP).

(CCP) \( \min_{x} E_{\omega}[h(x, \omega)] \) (Reg. conds)
\[ \text{s.t. } g(x) \leq 0, \quad \lambda \]

where \( g(x) \triangleq \frac{1}{E|x|} - \frac{1}{(1-\epsilon)}, \quad \lambda \triangleq \chi \times \mathbb{R}^+, \quad N_{\lambda}(z) \) denotes the normal cone of \( \lambda \) at \( z \), \( H(x, \lambda, \lambda) \triangleq \{ \nabla_x f(x) + \lambda \partial_x g(x) \} \times \{ g(x) \} \), We close by noting that monotone inclusions with expectation-valued operators are relevant in decision and control problems, providing motivation for addressing their tractable resolution.

III. STOCHASTIC MFBS SCHEMES

In this section, we analyze stochastic (operator) splitting schemes. When \( A(x) \triangleq \mathbb{E}[A(x, \omega)], A : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}^n \), and \( B \) has a cheap resolvent, we develop a VR splitting framework. In Section III-A, we provide some background and assumptions and derive convergence theory for monotone and strongly monotone settings in Sections III-B–III-C, respectively.

A. Background and Assumptions

Akin to other settings that employ SA, we assume the presence of a stochastic first-order oracle for operator \( A \) that produces a sample \( A(x, \omega) \) given a vector \( x \). Such a sample is core to developing a VR-modified forward–backward splitting (VR-SMFBS) scheme reliant on \( \sum_{\omega \in \Omega} A(x, \omega) \) to approximate \( \mathbb{E}[A(x, \omega)] \) at iteration \( k \). Given an \( x_0 \in \mathbb{R}^n \), we formally define such a scheme next
\[
x_{k+1} := (I + \gamma k A)^{-1}(x_k - \gamma k A_k) \quad \text{(vr – SMFBS)}
\]
where \( A_k \triangleq \frac{\sum_{\omega \in \Omega} A(x, \omega)}{N_k} \), and \( \gamma \) are estimators of \( A(x_k) \) and \( A(x_{k+1}) \), respectively. We assume the following on operators \( A \) and \( B \).

Assumption 1: The operator \( A \) is single-valued, monotone and \( L \)-Lipschitz on \( \mathbb{R}^n \), i.e., \( \forall x, y \in \mathbb{R}^n, \| A(x) - A(y) \| \leq L \| x - y \| \) and \( (A(x) - A(y))(x - y) \geq 0 \); the operator \( B \) is maximal monotone on \( \mathbb{R}^n \), i.e., no monotone map \( B' \) exists such that \( \text{gph} \ B \subset \text{gph} \ B' \), where \( \text{gph} \ B \) is defined as \( \text{gph} \ B \triangleq \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in B(x)\} \).

Suppose \( F_k \) denotes the history up to iteration \( k \)
\[
F_k \triangleq \{ x_0, \{ \omega_j, j \}_1 \forall j, \{ \omega_j, j \}_1 \forall j, \ldots, \{ \omega, j, k-1 \}_1 \forall j \}
\]

where \( F_0 = \{ x_0 \} \). Suppose \( w_{j, k} \triangleq A(x_k, \omega_{j, k}) - A(x_k) \), \( \bar{w}_k \triangleq \sum_{j=1}^{n_k} w_{j, k} \), \( \bar{w}_{j, k} \triangleq A(x_k, \omega_{j, k}) - A(x_{k+1}) \), and \( \bar{w}_{k+1} \triangleq \sum_{j=1}^{n_k} w_{j, k+1} \), where \( N_k \) denotes the batch size of samples \( A(x, \omega_{j, k}) \) at iteration \( k \). We impose the following bias and moment assumptions on \( \bar{w}_k \) and \( \bar{w}_{k+1} \). Note that Assumption 2(ii) is weakened in the strongly monotone regime, allowing for biased oracles.

Assumption 2: At iteration \( k \), the following hold in an a.s. sense: (i) The conditional means \( \mathbb{E}[w_{j, k} | F_k \] and \( \mathbb{E}[w_{j, k} | F_k] \) are zero for all \( k \) in an a.s. sense. (ii) The conditional second moments are bounded in an a.s. sense as follows, i.e., there exists \( \nu_1, \nu_2 \) such that \( \mathbb{E}[w_{j, k}^2 | F_k] \leq \nu_1 \| x_k \|^2 + \nu_2^2 \) and \( \mathbb{E}[w_{j, k} + 1 | F_k] \leq \nu_1^2 \| x_{k+1} \|^2 + \nu_2^2 \) for all \( k \) in an a.s. sense. In addition, we assume \( \{ \omega_{j, k}, j \}_j \) are i.i.d samples for \( k \geq 0 \).

When the set \( \chi \) is possibly unbounded, assuming that the conditional second moment of \( w_{j, k} \) is uniformly bounded a.s. is often a stringent requirement. Instead, we impose a state-dependent assumption on \( w_{j, k} \). We conclude this section by defining a residual function for a GE.

Lemma 1 (Residual function for (GE)); Suppose \( \gamma > 0, T = A + B, and \)
\[
\gamma_r(x) \triangleq \| x - (I + \gamma B)^{-1}(x - \gamma A(x)) \|
\]

Then, \( \gamma_r \) is a residual function for (GE).

Proof: By definition, \( \gamma_r(0) = 0 \) if and only if \( x = (I + \gamma B)^{-1}(x - \gamma A(x)) \). This can be interpreted as follows, leading to the conclusion that \( x \in T^{-1}(0) \)
\[
x = ((I + \gamma B)^{-1}(x - \gamma A(x)) \leftarrow x - \gamma A(x) \in (I + \gamma B)(x) \Leftarrow 0 \in A(x) + B(x).
\]

We conclude this section with two lemmas [42] crucial for proving claims of almost sure convergence.

Lemma 2: Let \( \nu_k, \delta_k, \delta_k \), and \( \psi_k \) be nonnegative random variables adapted to \( \sigma \)-algebra \( F_k \), and let the following relations hold almost surely:
\[
\mathbb{E}[v_{k+1} | F_k] \leq (1 + u_k) \psi_k - \delta_k + \psi_k \forall k
\]
\[
\sum_{k=0}^{\infty} u_k < \infty, \text{ and } \sum_{k=0}^{\infty} \psi_k < \infty
\]

Then, a.s., \( \lim_{k \rightarrow \infty} v_k = v \) and \( \sum_{k=0}^{\infty} \delta_k < \infty \), where \( v \geq 0 \) is a random variable.

Lemma 3: Consider a sequence \( v_k \) of nonnegative random variables adapted to the \( \sigma \)-algebra \( F_k \) and satisfying \( \mathbb{E}[v_{k+1} | F_k] \leq (1 - a_k) v_k + b_k \) for \( k \geq 0 \) where \( a_k \in \]
B. Convergence Analysis Under Monotone Monotone A

In this section, we derive a.s. convergence guarantees and rate statements. First, we prove the a.s. convergence of the sequence generated by this scheme. We start with a lemma.

Lemma 4: Consider a sequence \( \{x_k\} \) generated by (vr-SMFBS). Let Assumptions 1 and 2 hold. Suppose \( \gamma_k \equiv \gamma < \frac{1}{\sqrt{2L}} \). Then, for any \( k \geq 0 \)

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] \leq \left( 1 + \frac{a_1}{\gamma_k} \right) \|x_k - x^*\|^2
\]

where \( a_1 \equiv 2\nu_1^2\gamma^2(8(2\nu_1^2\gamma^2 + 1 + (1 + \gamma L)^2) + 5 - 2\gamma^2 L^2) \), \( a_2 \equiv 2\nu_1^2\gamma^2(4(4\nu_1^2\gamma^2 + 1) + 5 - 2\gamma^2 L^2) \), and \( a_3 \equiv \gamma^2(4(\nu_1^2\gamma^2 + 1) + 5 - 2\gamma^2 L^2)\). 2:

Proof: From the definition of \( x_{k+\frac{1}{2}} \) and \( x_{k+1} \), we have

\[
x_{k+\frac{1}{2}} = x_k - \gamma(u_k + \bar{w}_k)
\]

\[
x_{k+1} = x_{k+\frac{1}{2}} - \gamma(u_{k+\frac{1}{2}} + \bar{w}_{k+\frac{1}{2}} - u_k - \bar{w}_k)
\]

where \( u_k = A(x_k), u_{k+\frac{1}{2}} = A(x_{k+\frac{1}{2}}), v_{k+\frac{1}{2}} \in B(x_{k+\frac{1}{2}}) \). From \( 0 \in A(x^*) + B(x^*) \), \( u^* + v^* = 0 \), where \( u^* = A(x^*) \) and \( v^* \in B(x^*) \), respectively. We have the following equality by leveraging the above definitions:

\[
\|x_k - x^*\|^2 = \|x_k - x_{k+\frac{1}{2}} + x_{k+\frac{1}{2}} - x_{k+1} + x_{k+1} - x^*\|^2
\]

\[
= \|x_k - x_{k+\frac{1}{2}}\|^2 + \|x_{k+\frac{1}{2}} - x_{k+1}\|^2 + \|x_{k+1} - x^*\|^2
\]

\[
+ 2(x_k - x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}} - x_{k+1})
\]

\[
+ 2(x_k - x_{k+\frac{1}{2}})^T(x_{k+\frac{1}{2}} - x^*)
\]

\[
+ 2(x_{k+\frac{1}{2}} - x_{k+1})^T(x_{k+1} - x^*)
\]

By Lemma 1, \( r_\gamma \) is a residual function for (GE), defined as \( r_\gamma(x) \equiv \|x - (I + \gamma B)^{-1}(x - \gamma A(x))\| \). It follows that

\[
r_\gamma^2(x_k) = \|x_k - (I + \gamma B)^{-1}(x_k - \gamma A(x_k))\|^2
\]

\[
\leq 2\|x_k - x_{k+\frac{1}{2}}\|^2 + 2\|(I + \gamma B)^{-1}(x_k - \gamma A(x_k)) - (I + \gamma B)^{-1}(x_k - \gamma A(x_k))\|^2
\]

\[
\leq 2\|x_k - x_{k+\frac{1}{2}}\|^2 + 2\|\gamma B(x_k - x_{k+\frac{1}{2}})\|^2
\]

\[
\|x_k - x_{k+\frac{1}{2}}\|^2 \leq -\frac{1}{2}r_\gamma^2(x_k) + \gamma^2\|\bar{w}_k\|^2.
\]

(3)

Following (2), we have

\[
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \|x_k - x_{k+\frac{1}{2}}\|^2
\]

\[
+ 2\|\gamma B(x_k - x_{k+\frac{1}{2}})\|^2 + 2\|\gamma B(x_k - x_{k+\frac{1}{2}})\|^2
\]

\[
\|x_k - x_{k+\frac{1}{2}}\|^2 \leq -\frac{1}{2}r_\gamma^2(x_k) + \gamma^2\|\bar{w}_k\|^2.
\]

(3)

\[
\leq -\frac{1}{2}r_\gamma^2(x_k) + \gamma^2\|\bar{w}_k\|^2.
\]

(4)

Taking expectations conditioned on \( \mathcal{F}_k \) and the tower law, we obtain the following bound:

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k]
\]

\[
\leq \|x_k - x^*\|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_\gamma^2(x_k)
\]

\[
+ \mathbb{E}[\|2\gamma \|\bar{w}_k\|^2 | \mathcal{F}_k + \frac{1}{2}] \| \mathcal{F}_k
\]

\[
+ \mathbb{E}[\|5 - 2\gamma^2 L^2\|\|^2 \|\bar{w}_k\|^2 | \mathcal{F}_k]
\]

\[
- \mathbb{E}[\|2\gamma \|\bar{w}_k\|^2 | \mathcal{F}_k + \frac{1}{2}] \| \mathcal{F}_k
\]
\[ \leq \| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k) + \frac{4\gamma^2 (r_k^2 \| x_k + \frac{1}{2} k \| F_k + \nabla^2) + (5 - 2\gamma^2 L^2) \gamma^2 r_k^2 \| x_k \|^2 + \nu_k^2 \| F_k \| + (5 - 2\gamma^2 L^2) \gamma^2 r_k^2 \| x_k \|^2 + \nu_k^2 \| F_k \|}{N_k} + \frac{\| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k)}{N_k} \]

Before proceeding, we need the following relation. By definition of \( x_{k+\frac{1}{2}} \), it follows that

\[ \mathbb{E}[\| x_{k+\frac{1}{2}} - x^* \|^2 | F_k] \leq 2 \mathbb{E}[ (I + \gamma B)^{-1} (x_k - \gamma A_k) - (I + \gamma B)^{-1} (x_k - \gamma A(x_k))] + 2 \| (I + \gamma B)^{-1} \| (x^* - \gamma A(x^*)) \|^2 \]

Using (6) in (5), we obtain

\[ \mathbb{E}[\| x_{k+1} - x^* \|^2 | F_k] \leq \| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k) + \frac{4\gamma^2 (4r_k^2 \gamma^2 \| x_k \|^2 + 4\gamma^2 (1 + \gamma L^2) \| x_k - x^* \|^2 + 2\gamma^2 \| x^* \|^2 + \nu_k^2)}{N_k} + \frac{\| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k)}{N_k} \]

Noting \( \mathbb{E}[\| x_{k+1} - x^* \|^2 | F_k] \leq \frac{\| x_k \|^2 + \nu_k^2}{N_k} \), a.s.

\[ \mathbb{E}[\| x_{k+1} - x^* \|^2 | F_k] \leq \| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k) + \frac{4\gamma^2 (4r_k^2 \gamma^2 \| x_k \|^2 + 4\gamma^2 (1 + \gamma L^2) \| x_k - x^* \|^2 + 2\gamma^2 \| x^* \|^2 + \nu_k^2)}{N_k} + \frac{\| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k)}{N_k} \]

Using Lipschitz continuity of \( A \) and \( \gamma_k \leq \min\{\frac{1}{4\gamma L^2}, \frac{1}{\gamma \| x^* \|^2}\} \) for all \( k \), it follows that

\[ \| x_{k+1} - x^* \|^2 \leq \| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k) + \frac{4\gamma^2 (4r_k^2 \gamma^2 \| x_k \|^2 + 4\gamma^2 (1 + \gamma L^2) \| x_k - x^* \|^2 + 2\gamma^2 \| x^* \|^2 + \nu_k^2)}{N_k} + \frac{\| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k)}{N_k} \]

where \( a_1, a_2, \) and \( a_3 \) are defined appropriately.

**Theorem 1 (a.s. convergence of (vr-SMFBS))** Consider a sequence \( \{x_k\} \) generated by (vr-SMFBS). Let Assumptions 1 and 2 hold. Suppose \( \gamma_k \equiv \gamma < \frac{1}{\sqrt{2} L^2} \), \( \{N_k\} \) is a nondecreasing sequence, and \( \sum_{k=0}^{\infty} \frac{1}{N_k} < M \). Then, for any \( x_0 \in \mathbb{R}^n \), \( \{x_k\} \) converges to a solution \( x^* \in \mathcal{A}^* \) in an a.s. sense.

**Proof:** We may now apply Lemma 2, which allows us to claim that \( \{\| x_k - x^* \|^2\} \) is convergent for any \( x^* \in \mathcal{A}^* \) and \( \sum_{k=0}^{\infty} r_k(x_k)^2 < \infty \) in an a.s. sense. Therefore, in an a.s. sense, we have \( \lim_{k \to \infty} r_k(x_k) = 0 \). Since \( \{\| x_k - x^* \|^2\} \) is a convergent sequence in an a.s. sense, \( \{x_k\} \) is bounded a.s. and has a convergent subsequence. Consider any convergent subsequence of \( \{x_k\} \) with an index set denoted by \( K \) and suppose its limit point is denoted by \( \bar{x} \). We have that \( \lim_{k \in K} r_k(x_k) = r_k(\bar{x}) = 0 \) a.s. since \( r_k(.) \) is a continuous function. It follows that \( \bar{x} \) is a solution to \( 0 \in T(\bar{x}) \). Consequently, some convergent subsequence of \( \{x_k\}_{k \geq 0} \), denoted by \( \bar{x} \), satisfies \( \lim_{k \to \infty} x_k = \bar{x} \) a.s. Since \( \{\| x_k - x^* \|^2\} \) is convergent a.s. for any \( x^* \in \mathcal{A}^* \), it follows that \( \{\| x_k - \bar{x} \|^2\} \) is convergent a.s. and its unique limit point is zero. Thus, every subsequence of \( \{x_k\} \) converges a.s. to \( \bar{x} \), which leads to the claim that the entire sequence \( \{x_k\} \) converges a.s. to a point \( \bar{x} \in \mathcal{A}^* \).

When the sampling process is computationally expensive (i.e., such as in the queueing systems), we prove the following corollary regarding (vr-SMFBS) with \( N_k = 1 \) for every \( k \).

**Corollary 1 (a.s. convergence under single sample):** Consider a sequence \( \{x_k\} \) generated by (vr-SMFBS). Let Assumptions 1 and 2 hold. Suppose \( \sum_{k=0}^{\infty} \gamma_k = \infty \), \( \sum_{k=0}^{\infty} \| x_k \|^2 < \infty \) and \( N_k = 1 \) for every \( k \in \mathbb{N} \). In addition, suppose \( A \) is coecoerive with constant \( c \) and strictly monotone. Furthermore, suppose \( \gamma_k \leq \min\{\frac{1}{4\gamma L^2}, \frac{1}{\gamma \| x^* \|^2}\} \) for \( k \geq 0 \). Then, \( \{x_k\} \) converges to a solution \( x^* \in \mathcal{A}^* \) in an a.s. sense.

**Proof:** Following (4) and using the coercive property of \( A \) and \( \gamma_k \leq \min\{\frac{1}{4\gamma L^2}, \frac{1}{\gamma \| x^* \|^2}\} \) for all \( k \), it follows that

\[ \| x_{k+1} - x^* \|^2 \leq \| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k) + \frac{4\gamma^2 (4r_k^2 \gamma^2 \| x_k \|^2 + 4\gamma^2 (1 + \gamma L^2) \| x_k - x^* \|^2 + 2\gamma^2 \| x^* \|^2 + \nu_k^2)}{N_k} + \frac{\| x_k - x^* \|^2 - \frac{1}{2} \left( 1 - 2\gamma^2 L^2 \right) r_k^2(x_k)}{N_k} \]

Following in the same fashion as in Lemma 4, we get

\[ \mathbb{E}[\| x_{k+1} - x^* \|^2 | F_k] \leq (1 + \gamma \| b_1 \|) \| x_k - x^* \|^2 + \gamma \| b_2 \| \| x^* \|^2 + \gamma \| b_3 \| \| x_k - u_k \|^2 - \gamma \| b_3 \| \| x_k - u_k \|^2 \]

By the nonsummability of \( \{\gamma_k\} \) whereby \( \sum_{k=0}^{\infty} \gamma_k = \infty \) allows us to claim \( \lim_{k \to \infty} \| x_k - u_k \|^2 = 0 \). This implies that some subsequence of \( \{x_k\} \) converges to a point in an a.s. sense. Let the point be \( \bar{x} \) and we have \( A(\bar{x}) = A(x^*) \). Recall that \( A \) is strictly monotone, thus \( (A(\bar{x}) - A(x^*))^T (\bar{x} - x^*) > 0 \) holds at \( x^* \) and it is clear that \( \bar{x} = x^* \). By Lemma 2, \( \{\| x_k - x^* \|^2\} \) is a
convergent sequence in an a.s. sense, the entire sequence \( \{x_k\} \) converges to \( x^* \) in an a.s. sense.

To establish the rate under monotonicity, we need to introduce a metric for ascertaining progress. In strongly monotone regimes, the mean-squared error serves as such a metric while the function value represents such a metric in optimization regimes. In merely monotone variational inequality problems, a special case of monotone inclusion problems, the gap function has proved useful (cf. [12], [43]). When considering the more general monotone inclusion problem, Borwein and Dutta presented a gap function [44], inspired by the Fitzpatrick function [45], [46].

**Definition 1 (Gap function):** Given a set-valued mapping \( T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n \), then the gap function \( G \) associated with the inclusion problem \( 0 \in T(x) \) is defined as

\[
G(x) \triangleq \sup_{y \in \text{dom}(T)} \sup_{z \in T(y)} z^T(x - y) \quad \forall x \in \mathbb{R}^n.
\]

The gap function is nonnegative for all \( x \in \mathbb{R}^n \) and is zero if and only if \( 0 \in T(x) \). To derive the convergence rate under maximal monotonicity, we require boundedness of the domain of \( T \) as formalized by the next assumption.

**Assumption 3:** The domain of \( T \) is bounded, i.e., there exists \( D_T > 0 \) such that \( \|x\| \leq D_T \quad \forall x \in \{v \in \mathbb{R}^n \mid T(v) \neq \emptyset\} \).

Clearly, from the definition, a convex gap function can be extended-valued and its domain is contingent on the boundedness properties of \( \text{dom} \ T \). When \( \text{dom} \ T \) is bounded, the gap function is globally defined but when \( \text{dom} \ T \) is unbounded, one resolution is based on the notion of restricted merit functions, first introduced in [47]. In this approach, the gap function is defined on a bounded set, which belongs to \( \text{dom} \ T \). In such instances, a local rate of convergence can be obtained.

We begin by establishing an intermediate result.

**Lemma 5:** Let Assumptions 1 and 2 hold. Suppose \( \{x_k\} \) denotes a sequence generated by \((\text{vr-SMBSF})\). Then, for all \( y \in \text{dom}(T), z \in T(y) \) and \( k \geq 0 \)

\[
2 \gamma z^T(x_k + \frac{1}{2} T(y) - x_k - y) \leq \|x_k - y\|^2 - \|x_{k+1} - y\|^2
- (1 - 2 \gamma^2 L^2) \|x_k - x_{k+1}\|^2 + 2 \gamma^2 \|w_k + \frac{1}{2} T(y) - w_k\|^2
+ 2 \gamma^2 \|w_k + \frac{1}{2} T(y) - x_k\|^2.
\]

**Proof:** According to (4) and replacing \( x^* \) with \( y \in \text{dom}(T) \)

\[
2 \gamma z^T(x_k + \frac{1}{2} T(y) - x_k - y) \leq \|x_k - y\|^2 - \|x_{k+1} - y\|^2
- (1 - 2 \gamma^2 L^2) \|x_k - x_{k+1}\|^2 + 2 \gamma^2 \|w_k + \frac{1}{2} T(y) - w_k\|^2
- 2 \gamma^2 \|w_k + \frac{1}{2} T(y) - w_k\|^2
\leq \|x_k - y\|^2 - \|x_{k+1} - y\|^2(1 - 2 \gamma^2 L^2) \|x_k - x_{k+1}\|^2 + 2 \gamma^2 \|w_k + \frac{1}{2} T(y) - x_k\|^2
+ 2 \gamma^2 \|w_k + \frac{1}{2} T(y) - x_k - y\|^2.
\] (8)

Invoking Lemma 5, we derive a rate statement for \( \bar{x}_K \), an average of the iterates \( \{x_k + \frac{1}{2} T(y)\} \) generated by \((\text{vr-SMFBS})\) over the window constructed from \( 0 \) to \( K - 1 \), defined as

\[
\bar{x}_K \triangleq \frac{\sum_{k=0}^{K-1} x_k + \frac{1}{2} T(y)}{K}.
\] (9)

**Proposition 1 (Rate statement under monotonicity):** Consider the \((\text{vr-SMFBS})\) scheme. Suppose \( x_0 \in \mathbb{R}^n \) and let \( \{\bar{x}_K\} \) be defined in (9). Let Assumptions 1–3 hold. Suppose \( \gamma_k \equiv \gamma < \frac{1}{\sqrt{2L}} \), \( \{N_k\} \) is a nondecreasing sequence, and \( \sum_{k=0}^{\infty} \frac{1}{N_k} < M \).

(a) For any \( K \geq 1 \), \( E[\bar{G}(\bar{x}_K)] = O\left(\frac{1}{K}\right) \).

(b) Suppose \( N_k = [k^\alpha] \), for \( \alpha > 1 \). Then the oracle complexity to compute an \( \bar{x}_{K+1} \) such that \( E[G(\bar{x}_{K+1})] \leq \epsilon \) is bounded as

\[
\sum_{k=0}^{\infty} \frac{1}{N_k} \leq O\left(\frac{1}{\epsilon^{1/\alpha}}\right).
\]

**Proof:** (a) We first define a sequence \( \{u_k\} \) such that

\[
u_{k+1} := u_k - \gamma \tilde{w}_{k+\frac{1}{2}}.
\]

where \( u_0 \in \mathbb{R}^n \). We may then express the last term on the right in (8) as follows:

\[
2 \gamma \tilde{w}_{k+\frac{1}{2}}^T(y - x_{k+\frac{1}{2}}) = 2 \gamma \tilde{w}_{k+\frac{1}{2}}^T(y - u_k)
+ 2 \gamma \tilde{w}_{k+\frac{1}{2}}^T(u_k - x_{k+\frac{1}{2}})
= \|u_k - y\|^2 - \|u_{k+1} - y\|^2 + \gamma^2 \|\tilde{w}_{k+\frac{1}{2}}\|^2
+ 2 \gamma \tilde{w}_{k+\frac{1}{2}}^T(u_k - x_{k+\frac{1}{2}}).
\] (10)

Invoking Lemma 5 and summing over \( k \), we have

\[
\sum_{k=0}^{K-1} 2 \gamma z^T(x_{k+\frac{1}{2}} - y) \leq \|x_0 - y\|^2 + 2 \gamma \sum_{k=0}^{K-1} \|\tilde{w}_k - \tilde{w}_{k+\frac{1}{2}}\|^2
+ \sum_{k=0}^{K-1} 2 \gamma \tilde{w}_{k+\frac{1}{2}}^T(y - x_{k+\frac{1}{2}}).
\] (11)

Dividing (11) by \( K \), we obtain the following:

\[
\frac{1}{K} \sum_{k=0}^{K-1} 2 \gamma z^T(x_{k+\frac{1}{2}} - y) \leq \frac{1}{K} \|x_0 - y\|^2
+ \frac{2 \gamma \sum_{k=0}^{K-1} \|\tilde{w}_k - \tilde{w}_{k+\frac{1}{2}}\|^2}{K}
+ \frac{2 \gamma \sum_{k=0}^{K-1} \tilde{w}_{k+\frac{1}{2}}^T(y - x_{k+\frac{1}{2}})}{K}.
\] (12)

Using (10) in (12) and invoking (9), it follows that

\[
\gamma \bar{z}^T(\bar{x}_K - y) \leq \frac{1}{\sqrt{2L}} \|x_0 - y\|^2
+ \frac{2 \gamma \sum_{k=0}^{K-1} \|\tilde{w}_k - \tilde{w}_{k+\frac{1}{2}}\|^2}{2K}
+ \frac{2 \gamma \sum_{k=0}^{K-1} \tilde{w}_{k+\frac{1}{2}}^T(y - x_{k+\frac{1}{2}})}{2K}.
\]

Taking supremum over \( z \in T(y) \) and \( y \in \text{dom}(T) \) and leveraging the compactness of \( \text{dom}(T) \), we obtain the inequality

\[
\gamma \sup_{y \in \text{dom}(T)} \sup_{z \in T(y)} z^T(\bar{x}_K - y) \leq \frac{2 \gamma \|x_0\|^2 + \|u_0\|^2}{K}
+ \frac{\gamma \sum_{k=0}^{K-1} \|\tilde{w}_k - \tilde{w}_{k+\frac{1}{2}}\|^2}{2K}
+ \frac{\gamma \sum_{k=0}^{K-1} \tilde{w}_{k+\frac{1}{2}}^T(u_k - x_{k+\frac{1}{2}})}{2K}.
\]
By invoking the definition of $G(x)$ and letting $D \triangleq 2D_f^2 + \|x_0\|^2 + \|u_0\|^2$, we obtain the following relation:

$$
\gamma G(x_K) \leq D^2 + \gamma^2 \sum_{k=0}^{K-1} (2\|w_k - \bar{w}_k + \frac{1}{2}\|^2 + \|w_{k+\frac{1}{2}}\|^2) + \frac{\gamma^2 \sum_{k=0}^{K-1} \|w_k - \bar{w}_k + \frac{1}{2}\|^2}{2} + \frac{\gamma^2 \sum_{k=0}^{K-1} \|w_k - \bar{w}_k + \frac{1}{2}\|^2}{2}.
$$

(13)

Before proceeding, we establish bounds for $x_k$ and $x_{k+\frac{1}{2}}$. From Proposition 1, we know $\{x_k\}$ converges to $x^*$, which indicates that $\|x_k - x^*\|$ is bounded. We denote this bound by $\|x_k - x^*\| \leq D_T$. Following (6)

$$
\mathbb{E}[\|x_k + \frac{1}{2} - x^*\|^2 | F_k] \leq 2\gamma^2 (\nu_1^2 \|x_k\|^2 + \nu_2^2) + 2(1+\gamma L)^2 D_T^2.
$$

Taking expectations on (13), leads to the following bound:

$$
\mathbb{E}[\gamma G(x_K)] \leq D^2 + \gamma^2 \sum_{k=0}^{K-1} \frac{\nu_1^2 (4\|x_k\|^2 + 6\|x_{k+\frac{1}{2}}\|^2) + 10\nu_2^2}{2} \leq \frac{2D^2 + \gamma^2 \sum_{k=0}^{K-1} \nu_2^2 ((8+24\gamma^2)\|x_k\|^2 + 12(1+\gamma L)^2D_T^2) + 2\nu_1^2 (2\|w_k\|^2 + \|w_{k+\frac{1}{2}}\|^2) + 10\nu_2^2}{2K}.
$$

(14)

by defining $\hat{C} \triangleq (2D^2 + \gamma^2 M((8+24\gamma^2)\|x_{k+\frac{1}{2}}\|^2 + 12(1+\gamma L)^2D_T^2) + 2\nu_1^2 (2\|w_{k+\frac{1}{2}}\|^2 + \|w_k\|^2)/2$. It follows that $\mathbb{E}[G(x_K)] \leq \frac{\hat{C}}{K} (1/K)$. (b) For $\epsilon$ sufficiently small and when $\hat{C}$ is an appropriate constant, the result follows:

$$
\sum_{k=0}^{K} N_k \leq \sum_{k=0}^{K} \frac{\hat{C}}{\epsilon} (1+\alpha)^{\alpha+1} \leq \frac{\hat{C}}{\epsilon} (1+\alpha)^{\alpha}.
$$

Comment: A rate statement for the last iterate can also be derived as well as shown in [48]. Let $K_\epsilon \triangleq \inf\{k \geq 1 : \mathbb{E}[r^2(x_k)] \leq \epsilon\}$ where finiteness of $K_\epsilon$ can be shown a finite number, allowing for showing that $\mathbb{E}[r^2(x_K)] \leq O(1/K_\epsilon)$. Therefore, for $K \geq K_\epsilon$ iterations, we obtain a rate $\mathbb{E}[r^2(x_K)] \leq \epsilon$. Furthermore, the scheme proposed in [48] is aimed at stochastic variational inequality problems. Such problems are a special case of the stochastic GE considered herein that the map $B$ is chosen to be the normal cone associated with a closed and convex set $X$. The algorithm (vr-SMBFS) employs the resolvent of $B$ in the update (rather than the standard projection operator in [48]). This generalization requires introducing a distinct residual function for deriving convergence guarantees. In addition, our analysis leads to both a.s. convergence and linear convergence under strong monotonicity of the operator as shown next. Notably, a little care allows for claiming such statements under biased oracles. By contrast, no results for strongly monotone operator are presented in [48].

C. Convergence Analysis Under Strongly Monotone $A$

Here, we conduct an analysis under strong monotonicity.

**Assumption 4**: The mapping $A$ is $\sigma$-strongly monotone, i.e.,

$$
(A(x) - A(y))^T(x - y) \geq \sigma \|x - y\|^2, \forall x, y \in \mathbb{R}^n.
$$

The next lemma is essential to our rate analysis.

**Lemma 6**: Let Assumptions 1 and 4 hold. Then, the following holds for every $k$

$$
\|x_{k+1} + \frac{1}{2} - x^*\|^2 \leq (1 - \sigma \gamma + \gamma^2)\|x_k - x^*\|^2 - (1 - 2\gamma^2 (L^2 + \frac{1}{2} - 2\sigma \gamma)\|x_k - x_{k+\frac{1}{2}}\|^2 + (4\gamma^2 + 2)\|\bar{w}_{k+\frac{1}{2}}\|^2 + 4\gamma^2 \|\bar{w}_k\|^2.
$$

(15)

**Proof**: According to Assumption 4, we have

$$
-2\gamma (u_{k+\frac{1}{2}} + v_{k+\frac{1}{2}})^T (x_{k+\frac{1}{2}} - x^*) \leq -2\gamma \sigma \|x_{k+\frac{1}{2}} - x^*\|^2.
$$

(16)

Using (16) in (4), we deduce

$$
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + 2\gamma^2 \|\bar{w}_{k+\frac{1}{2}} - \bar{w}_k\|^2 - 2\gamma \sigma \|x_{k+\frac{1}{2}} - x^*\|^2 + 2\gamma^2 \|\bar{w}_{k+\frac{1}{2}} - \bar{w}_k\|^2 - 2\gamma \bar{w}_T^T (x_{k+\frac{1}{2}} - x^*).
$$

Theorem 2 (a.s. convergence without unbiasedness): Let Assumptions 1, 2(i), and 4 hold. Consider a sequence $\{x_k\}$ generated by (vr-SMBFS). Suppose $\gamma_k \equiv \gamma < \min\{\frac{\sigma}{2}, \frac{\gamma}{20}, \frac{\gamma^2}{4L}\}$, $N_0 \geq \frac{2(2(2\gamma + \gamma^2)^2)}{\gamma^2}$, $\{N_k\}$ is a nondecreasing sequence, and
where \( \delta_k \equiv \frac{(24\gamma^2 + 8)\nu_2^2}{N_0} \|x_k\|^2 + \frac{(12\gamma^2 + 4)\nu_2^2}{N_0} \) and the final inequality follows from \( N_0 \geq \frac{2(24\gamma^2 + 8)^2}{\sigma_\gamma} \). We observe that

\[
(1 - \frac{1}{2}\sigma_\gamma + \gamma^2) = (1 - (\gamma - \frac{\sigma_\gamma}{\gamma}) < \frac{\gamma}{2} < 1
- \frac{1}{2}\sigma_\gamma + \gamma^2) > (1 - 1 + \gamma^2) > 0
- \frac{(1 - 2\gamma^2 \tilde{L}^2 - \frac{3}{2}\sigma_\gamma)}{\|x_k - x^*\|^2} < \frac{q}{\gamma} - \left( \frac{\gamma}{2} - 2\gamma^2 \tilde{L}^2 \right)^2 < \frac{q}{\gamma}.
\]

In other words, if \( \gamma < \min\{\frac{q}{\gamma}, \frac{1}{2\gamma} \}, (17) \) can be further bounded by \( (1 - a_k)\|x_k - x^*\|^2 + \delta_k \) where \( a_k = \frac{\gamma}{\gamma} \) for all \( k \) and \( \delta_k, \sigma\) satisfy Lemma 3. Consequently, \( \|x_k - x^*\|^2 \to 0 \) in an a.s. sense as \( k \to \infty \).

Besides strong monotonicity, Theorem 2 analyzes the scheme with possible biased oracles. Often, conditional unbiasedness of the oracle may be harder to impose and one may need to impose weaker assumptions. Our proposed schemes allow for possibly biased oracles in some select settings. Next, we provide rate and complexity statements involving (vr-SMBFS) under geometrically increasing \( N_k \).

**Proposition 2** (Linear convergence): Let Assumptions 1, 2(ii), and 4 hold. Consider a sequence \( \{x_k\} \) generated by (vr-SMBFS). Suppose \( \gamma_k \equiv \gamma < \min\{\frac{q}{\gamma}, \frac{1}{2\gamma}, \frac{\gamma^2}{\gamma} \} \), \( N_0 \geq \frac{2(24\gamma^2 + 8)^2}{\sigma_\gamma} \), \( \tilde{L}^2 = L^2 + \frac{2}{\gamma} \), \( \|x^0 - x^*\| \leq D_0 \) and \( N_k \equiv N_0 |\rho^{-k+1}| \) for all \( k > 0 \). Then, the following holds,

(a) Suppose \( q \equiv (1 - \frac{\gamma^2}{\gamma}) < 1 \). Then, \( \mathbb{E}[\|x_k - x^*\|^2] \leq \tilde{D} \rho^k \) where \( \tilde{D} > 0, \tilde{\rho} = \max\{q, \rho\} \) if \( \rho 
eq \gamma \) and \( \tilde{\rho} \equiv (q, 1) \) if \( \rho = q \).

(b) Suppose \( x_{K+1} \) is such that \( \mathbb{E}[\|x_{K+1} - x^*\|^2] \leq \epsilon \). Then, the oracle complexity is \( \sum_{k=0}^{K} N_k \leq O(\frac{1}{\epsilon}) \).

**Proof:** (a) By taking unconditional expectations on (17)

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q\mathbb{E}[\|x_k - x^*\|^2] + \frac{\rho}{N_k} \tag{18}
\]

where \( \rho \equiv (24\gamma^2 + 8)\nu_2^2 \|x^0\|^2 + (12\gamma^2 + 4)\nu_2^2 \) and \( q \equiv (1 - \frac{\gamma^2}{\gamma}) \). Recall that \( N_k \) can be bounded as seen next

\[
N_k = N_0 |\rho^{-k+1}| \geq N_0 |\frac{1}{2} |\rho^{-k+1}| | \geq N_0 |\frac{\rho}{2} |\rho^{-k+1}| \tag{19}
\]

We now consider three cases.

(i) \( q < \rho < 1 \). Using (19) in (18) and defining \( \tilde{D} \equiv \frac{2\rho}{N_0} \) and \( \tilde{D} \equiv D_0 + \tilde{D} \), we obtain

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q\mathbb{E}[\|x_k - x^*\|^2] + \frac{\rho}{N_k} \leq q^{k+1} \|x_0 - x^*\| + D \sum_{j=1}^{k+1} q^{k+1-j} \rho^j \leq D_0 q^{k+1} + D \rho^{k+1} \sum_{j=1}^{k+1} \frac{q^{k+1-j} \rho^j}{\rho} \leq \tilde{D} \rho^{k+1}.
\]

(ii) \( \rho < q < 1 \). Akin to (i) and defining \( \tilde{D} \) appropriately,

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \tilde{D} q^{k+1}.
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
(iii) $\rho = q < 1$. If $\tilde{p} \in (q, 1)$ and $\tilde{D} > \frac{1}{m(\rho/q)^+}$, then

$$E[\|x_{k+1} - x^*\|^2] \leq q^{k+1}E[\|x_0 - x^*\|^2] + D \sum_{j=1}^{k+1} q^{j-1}$$

$$\leq D_0 q^{k+1} + D \sum_{j=1}^{k+1} q^{j-1} = D_0 q^{k+1} + D(k + 1)q^{k+1}$$

$$\leq D\tilde{p}^{k+1},$$

where $\tilde{D} \triangleq (D_0 + D \cdot \tilde{D})$.

Thus, $\{x_k\}$ converges linearly in an expected-value sense.

(b) Case (i): If $q < \rho < 1$. From (a), it follows that

$$E[\|x_{k+1} - x^*\|^2] \leq \tilde{D} p^{k+1} \leq \epsilon \Longrightarrow K \geq \log_{1/\rho}(\tilde{D}/\epsilon) - 1.$$ 

If $K = \lceil \log_{1/\rho}((\tilde{D}/\epsilon)) \rceil - 1$, then (vr-SMFBS) requires $\sum_{k=0}^{K} N_k$ evaluations. Since $N_k = N_0 \rho^{-k+1} \lesssim N_0 \rho^{-k+1}$, then we have

$$\sum_{k=0}^{K} N_0 \rho^{-k+1} = \sum_{k=1}^{\lceil \log_{1/\rho}(\tilde{D}/\epsilon) \rceil} N_0 \rho^{-k} \leq \frac{N_0}{\rho\left(\frac{1}{\rho}\right)} \left(\frac{1}{\rho}\right)^{\lceil \log_{1/\rho}(\tilde{D}/\epsilon) \rceil + 1}.$$

We omit cases (ii) and (iii), which lead to similar complexities.

Remark: We comment on our findings next.

(a) Rates and asymptotics. We believe that the findings fill important gaps in terms of providing rate statements for monotone inclusions. In particular, the rate statement in monotone regimes utilizes a lesser-known gap function while the VR schemes achieve deterministic rates of convergence. In addition, the oracle complexities are near-optimal. Furthermore, optimal rates and guarantees in strongly monotone regimes obviate the use of conditional unbiasedness of the oracle.

(b) Algorithm parameters: Akin to more traditional first-order schemes, these schemes rely on utilizing constant steplengths and leverage problem parameters such as Lipschitz and strong monotonicity constants. We believe that by using diminishing steplength sequences, we may be able to derive weaker rate statements that do not rely on problem parameters.

(c) Expectation-valued $B$: We may consider a setting where $B$ is expectation-valued and the resolvent operation is approximated via SA.

IV. NUMERICAL RESULTS

In this section, we apply the proposed schemes on a 2-stage SVI problem described in [50, Sec. 5.1].

Problem parameters for 2-stage SVI: We generate a set of i.i.d samples $\{\xi_i\}_{i=1}^J$, where $\xi_i \sim U[-5, 0]$. Suppose $h_i(\omega) = \xi_i$ for $i = 1, \ldots, J$. In addition, $c_i(x) = \frac{1}{2} m_i x_i^2 + \xi_i x_i$, $M \in \mathbb{R}^{J \times J}$ is a diagonal matrix with nonnegative elements $M = \text{diag}(m_1, \ldots, m_J)$ while $\ell = [\ell_1, \ldots, \ell_J]^T \in \mathbb{R}^J$ where $\ell_i \in U(2, 3)$. Furthermore, the inverse demand function $p$ is defined as $p(X) = d - rX$ where $d = 1$ and $r = 1$. Thus, $G(x)$, as defined in [50, Sec. 5.1], can be simplified as $G(x) = Mx + f$. In this setting, $A(x) = Mx + \ell + R(x) + D^T(x)$ and $B(x) = X(x)$, where $X = \mathbb{R}_+^J$. The Lipschitz of $A$ is given by $L = L_B + L_R + L_D$ where $L_B = \max_i m_i$, $L_R = r[1 + 11]^2$ and $L_D = \frac{1}{2}$. Implementations are in MATLAB on a 16GB-PC with a 6-Core Intel Core i7 processor (2.6 GHz).

We describe the schemes being compared and specify their algorithm parameters. Solution quality is compared by estimating the residual function $\text{res}(x) = \|x - \Pi_X (x - \gamma A(x))\|$.

A. Algorithm Specifications

We consider the following two schemes in this example.

1. (SA): This SA scheme utilizes the following update where $x_0$ is randomly generated in $[0, 1]^J$.

$$x_{k+1} := \Pi_X [x_k - \gamma_k A(x_k, \omega_k)] \quad \text{(SA)}$$

where $A(x_k) = E[A(x_k, \omega_k)]$, and $\gamma_k \triangleq \frac{1}{4\gamma}$. Since the steplength assumption and we consider $N_k = \lfloor k^{1.01} \rfloor$ for merely monotone problems, $N_k = \lfloor k^{1.01} \rfloor$ for strongly monotone problems.

2. (vr-SMFBS): Variance-reduction stochastic modified forward–backward scheme. We choose a constant $\gamma = \frac{1}{4\gamma}$, which satisfies the steplength assumption and we consider $N_k = \lfloor k^{1.01} \rfloor$ for merely monotone problems, $N_k = \lfloor k^{1.01} \rfloor$ for strongly monotone problems.

B. Performance Comparison and Insights

In Fig. 1, we compare both schemes under mere monotonicity and strong monotonicity, respectively, and examine sensitivities to the sample growth rate. Standard SA schemes may struggle when the problem is ill-conditioned and we examine the performance of the schemes in such regimes and provide the results in for merely monotone and strongly monotone settings in Tables II and III, respectively.

Key findinds: (vr-SMFBS) trajectories are characterized by significantly smaller empirical errors than (SA). There is little impact on (vr-SMFBS) when varying the sample growth rate.
TABLE III

| Key | (vr-SMFBS) With (SA) |
|-----|---------------------|
| n   | 
| 100 | 3.8 + 2.5 | 3.5 + 2.3 |
| 2000 | 5.6 + 2.8 | 5.3 + 2.6 |
| 4000 | 7.6 + 4.1 | 7.3 + 3.9 |
| 10000 | 11.7 + 5.5 | 11.4 + 5.2 |

We compare the performance of our scheme with the (SA) scheme [51]. Let \( (\omega_1, \ldots, \omega_J) \) denote independent identically distributed (i.i.d.) samples. Then, with (SA), we solve the following for \( i = 1, \ldots, J \):

\[
\begin{align*}
0 &\leq x_i - c_i(x_i) + r_i \cdot (X + x_i) - d + \frac{1}{\nu} \sum_{i=1}^{J} \lambda_i(x_i) \geq 0 \\
0 &\leq \gamma_i(x_i) + \lambda_i(x_i) + \lambda_i(x_i) \geq 0 \\
0 &\leq \lambda_i(x_i) + x_i - \gamma_i(x_i) \geq 0, \quad \forall \nu \in \mathbb{R} \setminus \{0\}
\end{align*}
\]

This problem is cast as a linear complementarity problem, allowing for utilizing PATH [52] to compute a solution. We compare (SA) with (vr-SMFBS) in Table IV. From the results, we observe that although the empirical errors of both schemes are similar, the (SA) scheme takes far longer than (vr-SMFBS) when using a large number of samples. In fact, (vr-SMFBS) scales well with overall number of evaluations.

V. CONCLUDING REMARKS

Monotone inclusions represent an important class of problems and their stochastic counterpart subsumes a large class of stochastic optimization and equilibrium problems. Such objects arise in optimization, game-theoretic, and MPC problems afflicted by uncertainty. We propose a VR splitting framework for resolving such problems when the map is structured. Under suitable assumptions on the sample size, we prove that the scheme displays a.s. convergence guarantees and achieves optimal linear and sublinear rates in strongly monotone and monotone regimes while achieving either optimal or near-optimal sample-complexities. By incorporating state-dependent bounds on noise and weakening unbiasedness requirements (in strongly monotone settings), we develop techniques that can accommodate far more general settings. Preliminary numerics on a class of two-stage stochastic variational inequality problems suggest that the scheme outperforms SA schemes, as well as SAA approaches.

### References

1. S. M. Robinson, *Generalized Equations*. Berlin, Germany: Springer, 1983, pp. 346–367.
2. R. T. Rockafellar, “Monotone operators and the proximal point algorithm,” *SIAM J. Control Optim.*, vol. 14, no. 5, pp. 877–898, 1976.
3. R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, vol. 9. Philadelphia, PA, USA: SIAM, 1989.
4. J. Douglas and H. H. Rachford, “On the numerical solution of heat conduction problems in two and three space variables,” *Trans. Amer. Math. Soc.*, vol. 82, no. 2, pp. 421–439, 1956.
5. D. W. Peaceman and H. H. Rachford Jr., “The numerical solution of parabolic and elliptic differential equations,” *J. Soc. Ind. Appl. Math.*, vol. 3, no. 1, pp. 28–41, 1955.
6. P.-L. Lions and B. Mercier, “Splitting algorithms for the sum of two nonlinear operators,” *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, 1979.
7. X. Chen, A. Shapiro, and H. Sun, “Convergence analysis of sample average approximation of two-stage stochastic generalized equations,” *SIAM J. Optim.*, vol. 29, no. 1, pp. 135–161, 2019.
8. R. J. Aumann, “Integrals of set-valued functions,” *J. Math. Anal. Appl.*, vol. 12, no. 1, pp. 1–12, 1965.
9. H. Jiang and H. Xu, “Stochastic approximation approaches to the stochastic variational inequality problem,” *IEEE Trans. Autom. Control*, vol. 53, no. 6, pp. 1462–1475, Jul. 2008.
10. U. Ravat and U. V. Shanbhag, “On the existence of solutions to stochastic quasi-variational inequality and complementarity problems,” *Math. Program.*, vol. 165, no. 1, pp. 291–330, 2017.
11. U. Ravat and U. V. Shanbhag, “On the characterization of solution sets of smooth and nonsmooth convex stochastic Nash games,” *SIAM J. Optim.*, vol. 21, no. 3, pp. 1168–1199, 2011.
12. F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems* (Vol. I). Springer Series in Operations Research). New York, NY, USA: Springer, 2003.
13. L. Rosasco, S. Villa, and B. C. V"u, “Stochastic forward-backward splitting for monotone inclusions,” *J. Optim. Theory Appl.*, vol. 169, no. 2, pp. 288–406, 2016.
14. P. L. Combettes and J.-C. Pesquet, “Stochastic approximations and perturbations in forward-backward splitting for monotone operators,” *Pure Appl. Funct. Anal.*, vol. 1, no. 1, pp. 13–37, 2016.
15. A. Ruszczyński, “Decomposition methods in stochastic programming,” *Math. Program.*, vol. 79, no. 1–3, pp. 333–353, 1997.
16. M. Schmidt, N. L. Roux, and F. R. Bach, “Convergence rates of inexact proximal-gradient methods for convex optimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1458–1466.
17. S. Cui and U. V. Shanbhag, “On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems,” in *Proc. IEEE 55th Conf. Decis. Control*, 2016, pp. 4510–4515.
18. V. G. Yaj and S. Bhatnagar, “Analysis of stochastic approximation schemes with set-valued maps in the absence of a stability guarantee and their stabilization,” *IEEE Trans. Autom. Control.*, vol. 65, no. 3, pp. 1100–1115, Mar. 2020.
19. V. G. Yaj and S. Bhatnagar, “Stochastic recursive inclusions in two timescales with nonadditive iterate-dependent Markov noise,” *Math. Oper. Res.*, vol. 45, no. 4, pp. 1405–1444, 2020.
20. X. Chen, R. T.-B. Wets, and Y. Zhang, “Stochastic variational inequalities: Residual minimization smoothing sample average approximations,” *SIAM J. Optim.*, vol. 22, no. 2, pp. 649–673, 2012.
21. M. P. Friedlander and M. Schmidt, “Hybrid deterministic-stochastic methods for data fitting,” *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. A1380–A1405, 2012.
22. R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, “Sample size selection in optimization methods for machine learning,” *Math. Program.*, vol. 134, no. 1, pp. 127–155, 2012.
23. R. Pasupathy, P. Glynn, S. Ghosh, and F. S. Hashemi, “On sampling rates in simulation-based recursions,” *SIAM J. Optim.*, vol. 28, no. 1, pp. 45–73, 2018.
[24] A. Jofré and P. Thompson, “On variance reduction for stochastic smooth convex optimization with multiplicative noise,” Math. Program., vol. 174, no. 1/2, pp. 253–292, 2019.

[25] A. Jalilzadeh, U. V. Shanbhag, J. Blanchet, and P. W. Glynn, “Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs,” Stochastic Syst., vol. 12, no. 4, pp. 373–410, 2022.

[26] A. Jalilzadeh, A. Nedić, U. V. Shanbhag, and F. Yousefian, “A variable sample-size stochastic quasi-Newton method for smooth and nonsmooth stochastic convex optimization,” Math. Operations Res., vol. 47, no. 1, pp. 690–719, 2022.

[27] R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” in Proc. 26th Int. Conf. Neural Inf. Process. Syst., 2013, pp. 315–323.

[28] A. Defazio, F. Bach, and S. Lacoste-Julien, “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., 2014, pp. 1646–1654.

[29] G. B. Passty, “Ergodic convergence to a zero of the sum of monotone operators in Hilbert space,” J. Math. Anal. Appl., vol. 72, no. 2, pp. 383–390, 1979.

[30] S.-P. Han and G. Lou, “A parallel algorithm for a class of convex programs,” SIAM J. Control Optim., vol. 26, no. 2, pp. 345–355, 1988.

[31] P. Tseng, “A modified forward-backward splitting method for maximal monotone mappings,” SIAM J. Control Optim., vol. 38, no. 2, pp. 431–446, 2000.

[32] H. D. Sherali, “A multiple leader Stackelberg model and analysis,” Operations Res., vol. 32, no. 2, pp. 390–404, 1984.

[33] A. Herty, S. Steffensen, and A. Thünen, “Solving quadratic-multi-leader-follower games by smoothing the follower’s best response,” Optim. Methods Softw., vol. 37, pp. 772–799, 2020.

[34] J. B. Rawlings, D. Q. Mayne, and M. Diehl. Model Predictive Control: Theory, Computation, and Design. Nob Hill, CA, USA: Nob Hill Publishing Madison, 2017, vol. 2.

[35] D. A. Allan, C. N. Bates, M. J. Rischbeck, and J. B. Rawlings, “On the inherent robustness of optimal and suboptimal nonlinear MPC,” Syst. Control Lett., vol. 106, pp. 68–78, 2017.

[36] S. Højstüttma, C. M. Lagoa, and F. Dabbene, “Identification of switched autoregressive exogenous systems from large noisy datasets,” Int. J. Robust Nonlinear Control, vol. 30, no. 15, pp. 5777–5801, 2020.

[37] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, Robust Optimization. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[38] P. K. Mishra, S. S. Diwale, C. N. Jones, and D. Chatterjee, “Reference tracking stochastic model predictive control over unreliable channels and bounded control actions,” Automatica J. IFAC, vol. 127, 2021, Art. no. 109512.

[39] A. Mesbah, “Stochastic model predictive control: An overview and perspectives for future research,” IEEE Control Syst. Mag., vol. 36, no. 6, pp. 30–44, Dec. 2016.

[40] E. Camacho and C. Alba, Model Predictive Control. Advanced Textbooks in Control and Signal Processing. Berlin, Germany: Springer, 2013.

[41] I. E. Bardakci, A. Jalilzadeh, C. Lagoa, and U. V. Shanbhag, “Probability maximization via Minkowski functionals: Convex representations and tractable resolution,” Math. Program., vol. 199, no. 1/2, pp. 595–637, 2023.

[42] B. T. Polyak, Introduction to Optimization. New York, NY, USA: Optim. Software New York, 1987.

[43] T. Larsson and M. Patriksson, “A class of gap functions for variational inequalities,” Math. Program., vol. 64, no. 1, pp. 53–79, 1994.

[44] J. M. Borwein and J. D. Vanderwerff, Convex Functions: Constructions, Characterizations and Counterexamples. Cambridge, MA, USA: Cambridge Univ. Press, 2010, vol. 109.

[45] S. Fitzpatrick, “Representing monotone operators by convex functions,” in Proc. Workshop/Miniconference Funct. Anal. Optim., 1988, pp. 59–65.

[46] Y. Nesterov, “Dual extrapolation and its applications to solving variational inequalities and related problems,” Math. Program., vol. 109, no. 2, pp. 319–344, 2007.

[47] R. I. Boţ, P. Mertikopoulos, M. Staudigl, and P. T. Vuong, “Minibatch forward-backward-forward methods for solving stochastic variational inequalities,” Stochastic Syst., vol. 11, no. 2, pp. 112–139, 2021.

[48] H. Ahmadi, “On the analysis of data-driven and distributed algorithms for convex optimization problems,” Ph.D. dissertation, The Pennsylvania State Univ., University Park, PA, USA, 2016.

[49] S. Cui, U. V. Shanbhag, M. Staudigl, and P. Vuong, “Stochastic relaxed inertial forward-backward-forward splitting for monotone inclusions in Hilbert spaces,” Comput. Optim. Appl., vol. 83, no. 2, pp. 465–524, 2022.

[50] J. Jiang, Y. Shi, X. Wang, and X. Chen, “Regularized two-stage stochastic variational inequalities for Cournot-Nash equilibrium under uncertainty,” J. Comput. Math., vol. 37, no. 6, pp. 813–842, 2019.

[51] S. P. Dirkse and M. C. Ferris, “The PATH solver: A monomtomone stabilization scheme for mixed complementarity problems,” Optim. Methods Softw., vol. 5, no. 2, pp. 123–156, 1995.

Shisheng Cui received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2009, the M.S. degree in computer science from Stanford University, Stanford, CA, USA, in 2011, and the Ph.D. degree in industrial engineering from Pennsylvania State University, University Park, PA, USA, in 2019. He is currently a Professor with the School of Automation, Beijing Institute of Technology, Beijing. His current research interests include optimization, variational inequality problems, and inclusion problems complicated by nonsmoothness and uncertainty. 

Uday V. Shanbhag received the undergraduate degree from IIT, Mumbai, Mumbai, India, in 1993, the master’s degree from MIT, Cambridge, MA, USA, in 1998, and the Ph.D. degree in management science and engineering (OR) from Stanford University, Stanford, CA, USA, in 2006. He has been holding the Cary and Sheila Bello Chaired Professorship in the department of industrial and manufacturing engineering with Penn State University, State College, PA, USA, since Nov. 2017 and has been with Penn State University since Fall 2012, prior to which he was an Assistant and subsequently a tenured Associate Professor with the University of Illinois at Urbana-Champaign from 2006 to 2012.