Mathematical Programming
https://doi.org/10.1007/s10107-022-01893-6

FULL LENGTH PAPER

Series B

Complexity guarantees for an implicit smoothing-enabled method for stochastic MPECs

Shisheng Cui[1] · Uday V. Shanbhag[1] · Farzad Yousefian[2]

Received: 15 April 2021 / Accepted: 15 August 2022

Abstract

Mathematical programs with equilibrium constraints (MPECs) represent a class of hierarchical programs that allow for modeling problems in engineering, economics, finance, and statistics. While stochastic generalizations have been assuming increasing relevance, there is a pronounced absence of efficient first/zeroth-order schemes with non-asymptotic rate guarantees for resolving even deterministic variants of such problems. We consider a subclass of stochastic MPECs (SMPECs) where the parametrized lower-level equilibrium problem is given by a deterministic/stochastic variational inequality problem whose mapping is strongly monotone, uniformly in upper-level decisions. Under suitable assumptions, this paves the way for resolving the implicit problem with a Lipschitz continuous objective via a gradient-free zeroth-order method by leveraging a locally randomized spherical smoothing framework. Efficient algorithms for resolving the implicit problem allow for leveraging any convexity property possessed by the implicit problem, which in turn facilitates the computation of approximate global minimizers. In this setting, we present schemes for single-stage and two-stage stochastic MPECs when the upper-level problem is either convex or nonconvex in an implicit sense. **(I). Single-stage SMPECs.** In single-stage SMPECs, in convex regimes, our proposed inexact schemes are characterized by a complexity in

U. V. Shanbhag: Shanbhag gratefully acknowledges the support from NSF CMMI-1538605, DOE ARPA-E award DE-AR0001076, and ONR grant N00014-22-1-2589.
F. Yousefian: Yousefian gratefully acknowledges the support from NSF CAREER grant ECCS-1944500 and ONR grant N00014-22-1-2757.

✉ Farzad Yousefian
farzad.yousefian@rutgers.edu

Shisheng Cui
suc256@psu.edu

Uday V. Shanbhag
udaybag@psu.edu

[1] Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, State College, PA 16802, USA

[2] Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854, USA

Published online: 07 October 2022

upper-level projections, upper-level samples, and lower-level projections of $\mathcal{O}(\frac{1}{\epsilon^2})$, $\mathcal{O}(\frac{1}{\epsilon^2})$, and $\mathcal{O}(\frac{1}{\epsilon^2}\ln(\frac{1}{\epsilon}))$, respectively. Analogous bounds for the nonconvex regime are $\mathcal{O}(\frac{1}{\epsilon})$, $\mathcal{O}(\frac{1}{\epsilon^2})$, and $\mathcal{O}(\frac{1}{\epsilon^3})$, respectively. **(II). Two-stage SMPECs.** In two-stage SMPECs, in convex regimes, our proposed inexact schemes have a complexity in upper-level projections, upper-level samples, and lower-level projections of $\mathcal{O}(\frac{1}{\epsilon^2})$, $\mathcal{O}(\frac{1}{\epsilon^2})$, and $\mathcal{O}(\frac{1}{\epsilon^2}\ln(\frac{1}{\epsilon}))$ while the corresponding bounds in the nonconvex regime are $\mathcal{O}(\frac{1}{\epsilon})$, $\mathcal{O}(\frac{1}{\epsilon^2})$, and $\mathcal{O}(\frac{1}{\epsilon^2}\ln(\frac{1}{\epsilon}))$, respectively. In addition, we derive statements for accelerated schemes in settings where the exact solution of the lower-level problem is available. Preliminary numerics suggest that the schemes scale with problem size, are relatively robust to modification of algorithm parameters, show distinct benefits in obtaining near-global minimizers for convex implicit problems in contrast with competing solvers, and provide solutions of similar accuracy in a fraction of the time taken by sample-average approximation (SAA).

Mathematics Subject Classification 65K15 · 90C15 · 90C30 · 90C33

1 Introduction

In this paper, we consider the resolution of variants and stochastic generalizations of the mathematical program with equilibrium constraints (MPEC), given by

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & f(\mathbf{x},\mathbf{y}) \\ \text{subject to} \quad & \mathbf{y} \in \text{SOL}(\mathcal{Y}, F(\mathbf{x},\bullet)), \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{MPEC}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a real-valued function, $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$ represents a real-valued mapping, $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ denote closed and convex sets, and $\text{SOL}(\mathcal{Y}, F(\mathbf{x},\bullet))$ denotes the solution set of the parametrized variational inequality problem $\text{VI}(\mathcal{Y}, F(\mathbf{x},\bullet))$, given an upper-level decision $\mathbf{x}$. Recall that the variational inequality problem $\text{VI}(\mathcal{Y}, F(\mathbf{x},\bullet))$ requires a vector $\mathbf{y}$ in the set $\mathcal{Y}$ such that

$$(\tilde{\mathbf{y}} - \mathbf{y})^T F(\mathbf{x},\mathbf{y}) \;\geq\; 0, \qquad \forall\, \tilde{\mathbf{y}} \;\in\; \mathcal{Y}. \tag{VI($\mathcal{Y}$, $F(\mathbf{x},\bullet)$)}$$

MPECs have a broad range of applications arising in hierarchical optimization, frictional contact problems, power systems [31], traffic equilibrium problems [45], and Stackelberg equilibrium problems [74]. A comprehensive survey of models, analysis, and algorithms can be found in [50] while a subsequent monograph emphasized the implicit framework [60].

The MPEC is an ill-posed generalization of a nonconvex and nonlinear program, an observation that follows from considering the setting where $\mathcal{Y}$ is given by $\mathbb{R}^m_+$. In such an instance, (MPEC) reduces to a mathematical program with complementarity constraints (MPCC) since $\mathbf{y}$ solves $\text{VI}(\mathcal{Y}, F(\mathbf{x},\bullet))$ if and only if $\mathbf{y}$ solves $\text{CP}(\mathcal{Y}, F(\mathbf{x},\bullet))$, defined as the problem of finding a vector $\mathbf{y}$ such that

$$\mathcal{Y} \ni \mathbf{y} \ \perp \ F(\mathbf{x}, \mathbf{y}) \in \mathcal{Y}^*, \qquad (\mathrm{CP}(\mathcal{Y}, F(\mathbf{x}, \bullet)))$$

where $\mathcal{Y}^* \triangleq \{u \mid y^T u \geq 0,\ y \in \mathcal{Y}\}$. When $\mathcal{Y}$ is the nonnegative orthant, then (MPEC) reduces to the following MPCC, which can be cast as an ill-posed nonlinear program.

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & f(\mathbf{x}, \mathbf{y}) \\ \text{subject to} \quad & 0 \leq \mathbf{y} \perp F(\mathbf{x}, \mathbf{y}) \geq 0, \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \qquad (\mathrm{MPCC})$$

Ill-posedness of (MPCC) arises from noting that standard constraint qualifications (such as the Mangasarian–Fromovitz constraint qualification) fail to hold at any feasible point. This has led to a concerted effort in developing weaker stationarity conditions for MPECs [70] as well as a host of regularization [2, 25, 36, 46, 66] and penalization [32] schemes.

Yet an enduring gap persists in the development of algorithms for such problems. Despite a wealth of developments in the field of zeroth and first-order algorithms for deterministic and stochastic convex and nonconvex optimization, there are no available non-asymptotic rate guarantees for either zeroth or first-order schemes for MPECs or their stochastic variants. In particular, our interest lies in the study of two distinct stochastic variants presented as follows.

1.1 Problems of interest

We focus on the problem (MPEC) where the lower-level map $F(\mathbf{x}, \bullet)$ is strongly monotone over $\mathcal{Y}$ uniformly in $\mathbf{x}$. This ensures that the solution of $\mathrm{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$ is a singleton for every $\mathbf{x} \in \mathcal{X}$. We consider two settings.

(i) *Single-stage SMPECs*.[1] Single-stage MPECs capture a class of stochastic MPECs with constraints given by parametrized variational inequality problems with expectation-valued maps. Such problems assume relevance in modeling a range of stochastic equilibrium problems; more specifically, such problems represent the necessary and sufficient equilibrium conditions of smooth stochastic convex optimization problems and smooth stochastic convex Nash equilibrium problems [37, 38]. They can also be employed for modeling settings in power systems [4, 22], structural optimization [19], and transportation science [52, 63]. More formally, suppose the variational inequality problem $\mathrm{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$ is characterized by a map F whose components are expectation-valued, i.e.,

$$F(\mathbf{x}, \mathbf{y}) \triangleq \begin{pmatrix} \mathbb{E}[G_1(\mathbf{x}, \mathbf{y}, \xi(\omega))] \\ \vdots \\ \mathbb{E}[G_m(\mathbf{x}, \mathbf{y}, \xi(\omega))] \end{pmatrix}, \qquad (1)$$

[1] In some of the literature on stochastic programming, this class of problems is also known as *one-stage* SMPEC. However, inspired by this paper [68] and for expository reasons, we have adopted the term *single-stage* SMPEC.

where $G_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ and $\xi : \Omega \to \mathbb{R}^d$ denotes a random variable associated with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Note that the expectations in (1) are taken with respect to the probability distribution $\mathbb{P}$. For the ease of presentation, throughout the paper, we refer to the integrand $G_i(\mathbf{x}, \mathbf{y}, \xi(\omega))$ by $G_i(\mathbf{x}, \mathbf{y}, \omega)$. In effect, the lower-level problem is a stochastic variational inequality problem [37, 82]. In addition, the objective may also be expectation-valued and the pessimistic version of the resulting problem is defined as follows.

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & f(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}, \omega)] \\ \text{subject to} \quad & \mathbf{y} \in \text{SOL}(\mathcal{Y}, \mathbb{E}[G(\mathbf{x}, \bullet, \omega)]), \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{SMPEC$^{\mathbf{1s}}$}$$

An instance where (SMPEC$^{\mathbf{1s}}$) emerges arises when the lower-level equilibrium problem captures the equilibrium conditions of a convex stochastic optimization problem, given by

$$\min_{\mathbf{y} \in \mathcal{Y}} \mathbb{E}[h(\mathbf{x}, \mathbf{y}, \omega)], \tag{2}$$

where $F(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{E}[\nabla_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}, \omega)]$. A more general instance is when a solution to the lower-level equilibrium problem is a Nash equilibrium of a noncooperative game with expectation-valued objectives, as given by

$$\min_{\mathbf{y}_i \in \mathcal{Y}_i} \mathbb{E}[h_i(\mathbf{x}, \mathbf{y}_i, \mathbf{y}_{-i}, \omega)], \tag{3}$$

where $i \in \{1, \ldots, N\}$, N denotes the number of players, $\mathbf{y}_i \in \mathcal{Y}_i$ and $h_i(\mathbf{x}, \mathbf{y}_i, \mathbf{y}_{-i}, \omega)$ denote the strategy set and the cost function of player $i \in \{1, \ldots, N\}$, respectively, and $\mathbf{y}_{-i} = (y_j)_{j \neq i}$. Under some mild conditions, it is known that the equilibrium conditions corresponding to a Nash equilibrium can be characterized as $\text{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet))$ where $\mathcal{Y} \triangleq \prod_{i=1}^N \mathcal{Y}_i$ and $F(\mathbf{x}, \mathbf{y}) \triangleq \left(\mathbb{E}[\nabla_{\mathbf{y}_i} h_i(\mathbf{x}, \mathbf{y}_i, \mathbf{y}_{-i}, \omega)]\right)_{i=1}^N$ (cf. Chap. 1 in [21]). An alternate approach for modeling uncertainty in MPECs is provided in the next model, where the lower-level problem constraints are imposed in an almost sure (a.s.) sense [16].

(ii) *Two-stage SMPECs.* Two-stage stochastic MPECs are characterized by equilibrium constraints $\text{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet, \omega))$ for almost every $\omega \in \Omega$. We provide motivation by considering the following two-stage leader-follower game in which the follower makes a *second-stage* decision $\mathbf{y}$ contingent on the leader's decision $\mathbf{x}$ and the realization of uncertainty is denoted by ω. Consequently, the leader's first-stage problem requires minimizing her expected cost $\mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\omega), \omega)]$ where $\mathbf{y}(\omega)$ represents follower's second-stage (i.e., recourse) decision, given $\mathbf{x}$ and ω. A pessimistic version of this problem can be compactly represented as (SMPEC$^{\mathbf{2s}}$),

defined next where "a.e." means "almost every".

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}(\omega)} \quad & \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\omega), \omega)] \\ \text{subject to} \quad & \mathbf{y}(\omega) \in \mathrm{SOL}(\mathcal{Y}(\mathbf{x}, \omega), G(\mathbf{x}, \bullet, \omega)), \text{ for a.e. } \omega \in \Omega \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{SMPEC$^{\mathbf{2s}}$}$$

In regimes where $\mathrm{VI}(\mathcal{Y}(\mathbf{x}, \omega), G(\mathbf{x}, \bullet, \omega))$ has a unique solution for any $\mathbf{x} \in \mathcal{X}$ and any $\omega \in \Omega$, the pessimistic and optimistic versions of the SMPECs coincide and we may recast (SMPEC$^{\mathbf{2s}}$) as the following *implicit* stochastic optimization problem where $\mathbf{y} : \mathcal{X} \times \Omega \to \mathbb{R}^m$ denotes a single-valued solution map of $\mathrm{VI}(\mathcal{Y}, F(\mathbf{x}, \bullet, \omega))$.

$$\begin{aligned} \min_{\mathbf{x}} \quad & f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)] \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{SMPEC$^{\mathbf{imp,2s}}$}$$

The implicit counterpart of (SMPEC$^{\mathbf{1s}}$), denoted by (SMPEC$^{\mathbf{imp,1s}}$), is defined analogously.

1.2 Gaps and contributions

The lower-level parametrized variational inequality problem can often be recast as a parametrized complementarity problem (e.g., when the VI admits a suitable regularity condition [50]). The MPEC then reduces to a mathematical program with complementarity constraints (MPCC). Nonlinear programming (NLP) approaches aligned around sequential quadratic programming [25] and interior-point schemes [2, 46, 66] have been applied for resolving MPCCs (See [50] for a survey). This represents a dominant algorithmic thread for resolving MPECs while a second lies in implicit programming approaches [1, 30, 39, 42, 43, 50, 53]. Yet, there are some key shortcomings of such avenues in such regimes, motivating the present research.

(a) *Limited convergence guarantees for existing NLP/regularization/penalization schemes.* Most interior-point [2, 46, 66], sequential quadratic programming (SQP) [25], and penalization/regularization schemes [2, 15, 46] for resolving MPECs are characterized by convergence to strong-stationary or C-stationary points in the full space of upper and lower-level decisions with rate guarantees only available in a local sense. Such schemes do not leverage any convexity properties in obtaining stronger guarantees. In particular, there appear to be no efficient schemes that can provide convergence guarantees to global minimizers (in an implicit sense) in either deterministic or stochastic regimes.

(b) *Implementability concerns with existing implicit approaches.* Existing implicit programming approaches (cf. [1, 7, 30, 39, 42, 43, 53]) require exact resolution of the lower-level problem (precluding the resolution of lower-level stochastic variational inequality problems), can generally not accommodate uncertainty in their

lower/upper-level, and are not equipped with non-asymptotic rate and complexity guarantees, particularly when the implicit problem is nonconvex.

(c) *Lack of efficient first/zeroth-order schemes.* While there have been tremendous advances in providing non-asymptotic rate guarantees for efficient first/zeroth-order algorithms for convex and nonconvex optimization problems [12, 24, 27, 58, 59], the resolution of MPECs via such avenues has been largely ignored. In fact, we are unaware of any efficient first/zeroth-order scheme for deterministic MPECs even under strong monotonicity assumptions at the lower-level.

(d) *Lack of scalability and convergence of schemes for stochastic MPECs.* Sample-average approximation [10, 49, 72] and smoothing schemes [47] for (SMPEC$^{\mathbf{2s}}$) have been studied extensively. While SAA schemes provide an avenue for approximation, the SAA problems become increasingly difficult to solve since the number of constraints grows linearly with the sample-size. Absent such sampling, such avenues can generally contend with finite sample-spaces.

Collectively, these gaps motivate the development of tools and techniques for this challenging class of stochastic nonconvex problems. To this end, we develop a zeroth-order algorithmic framework equipped with convergence rate guarantees that is applied on the implicit formulation of the problem. In this formulation, the objective function is viewed as a function in terms of the variable $\mathbf{x}$. While the implicit programming approach has been utilized before [47, 50, 78], several challenges arise when considering the development of iterative solution methods: (i) a closed-form characterization for $\mathbf{y}(\bullet)$ (or $\mathbf{y}(\bullet, \omega)$) is possibly unavailable which in turn, precludes the applicability of the standard first-order schemes; (ii) the implicit function is possibly nondifferentiable and nonconvex in $\mathbf{x}$ which complicates the convergence analysis and, in particular, the derivation of rate statements. In fact, one cannot compute subgradients or Clarke generalized gradients easily in such settings; (iii) in inexact regimes where there is a lack of access to an oracle for computing $\mathbf{y}(\bullet)$ (or $\mathbf{y}(\bullet, \omega)$), standard zeroth-order methods may not be directly applied. This is primarily because an inexact value of $\mathbf{y}(\bullet)$ may lead to a biased zeroth-order gradient approximation for the implicit function and the level of bias may even grow undesirably, as the parameters are updated iteratively; (iv) finally, in settings where the implicit problem is convex, asymptotically convergent accelerated schemes with rate statements are unavailable.

Contributions. In this paper, we aim at addressing these challenges through the development of a locally randomized zeroth-order scheme where the gradient of the implicit function is approximated at perturbed and possibly inexact evaluations of $\mathbf{y}(\bullet)$ (single-stage) and $\mathbf{y}(\bullet, \omega)$ (two-stage). Tables 1 and 2 provide the new complexity statements derived in this work for single-stage and two-stage SMPECs, respectively. The contributions in different regimes are as follows.

(1) Single-stage SMPECs. We consider the single-stage problem (SMPEC$^{\mathbf{1s}}$) in Sect. 3.

(1-i) *Inexact convex settings:* We develop (ZSOL$^{\mathbf{1s}}_{\mathrm{cnvx}}$), defined in Algorithm 1 where we employ a zeroth-order method for minimizing the implicit function. In the inexact variant of this method, to solve the stochastic VI at the lower-level and approximate $\mathbf{y}(\bullet)$, we employ a variance-reduced stochastic approximation method presented by Algorithm 2. In Theorem 1, we derive non-asymptotic con-

vergence rates and also obtain an overall iteration complexity of $\mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}\right)$ and $\mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2} \ln\left(n^2 L_0 \tilde{L}_0^2 \epsilon^{-1}\right)\right)$ for the number of projections on the set $\mathcal{X}$ and $\mathcal{Y}$, respectively, where L_0 and $\tilde{L}_0$ are defined by Assumption 1. Importantly, both the stepsize and smoothing parameters are updated iteratively using prescribed rules allowing for establishing convergence to an optimal solution of the original single-stage SMPEC.

(1-ii) *Exact convex settings:* The convergence statements for the exact variant of ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) are provided in Corollary 1. In particular, we derive the iteration complexity of $\mathcal{O}\left(n^2 L_0^2 \epsilon^{-2}\right)$. This implies that to obtain an ϵ-solution, the number of calls to the solution oracle of the lower-level variational inequality problem is at most $\mathcal{O}\left(n^2 L_0^2 \epsilon^{-2}\right)$.

(1-iii) *Inexact nonconvex settings:* In the case where the implicit function is nonconvex, we develop ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{ncvx}}$), defined in Algorithm 3. We analyze the convergence properties of this zeroth-order scheme under a constant stepsize and smoothing parameter. In Theorem 2, to obtain an ϵ-solution (characterized by mean norm-squared of a residual mapping) to the smoothed approximate SMPEC, we derive non-asymptotic convergence rates for solving the smoothed implicit problem and obtain an overall iteration complexity of $\mathcal{O}\left(n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1}\right)$ and $\mathcal{O}\left(n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2}\right)$ for the number of projections on the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively.

(1-iv) *Exact nonconvex settings:* In Corollary 2 we provide the results for the exact variant of ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{ncvx}}$). To obtain an ϵ-solution (characterized by mean norm-squared of a residual mapping), we derive an iteration complexity of $\mathcal{O}\left(n^2 L_0^2 \epsilon^{-1}\right)$ for solving the smoothed approximate SMPEC. The number of calls to the solution oracle of the lower-level variational inequality problem is at most $\mathcal{O}\left(n^4 L_0^4 \epsilon^{-2}\right)$.

(2) Two-stage SMPECs. We consider the two-stage problem (SMPEC$^{\mathbf{2s}}$) in Sect. 4.

(2-i) *Inexact convex settings*: We present ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx}}$), defined in Algorithm 5, for addressing two-stage SMPECs with a convex implicit objective function. In Theorem 3, for the inexact setting, we derive an overall iteration complexity of $\mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}\right)$ and $\mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2} \ln\left(n^2 L_0 \tilde{L}_0^2 \epsilon^{-1}\right)\right)$ for the projections on the set $\mathcal{X}$ and $\mathcal{Y}$, respectively. These statements are similar to those obtained in the single-stage model. However, unlike in the single-stage case, the inexact variant of ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx}}$) does not require any new samples in solving the lower-level problem, i.e., in Algorithm 6, a parametrized deterministic variational inequality problem is solved.

(2-ii) *Exact convex settings*: In Corollary 4, we provide the iteration complexity of $\mathcal{O}\left(n^2 L_0^2 \epsilon^{-2}\right)$, similar to that of the single-stage counterpart. This implies that the number of calls to the solution oracle of the lower-level variational inequality problem is at most $\mathcal{O}\left(n^2 L_0^2 \epsilon^{-2}\right)$.

(2-ii-a) *Accelerated exact convex settings*: We develop a variance-reduced accelerated zeroth-order scheme called ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx,acc}}$), formally specified by Algorithm 7. In Proposition 5, we improve the complexity to $\mathcal{O}(1/\epsilon)$ in terms of upper-level projection steps while the number of lower-level variational inequality problems is no worse than $\mathcal{O}(1/\epsilon^{2+\delta})$ for $\delta > 0$.

Table 1 Complexity guarantees for solving single-stage SMPECs

Single-stage SMPECs	Convex implicit		Nonconvex implicit	
	Inexact	Exact	Inexact	Exact
Upper level				
# Projections	$n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}$	$n^2 L_0^2 \epsilon^{-2}$	$n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1}$	$n^2 L_0^2 \epsilon^{-1}$
# Samples	$n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}$	$n^2 L_0^2 \epsilon^{-2}$	$n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2}$	$n^4 L_0^4 \epsilon^{-2}$
Lower level				
# Projections	$\frac{n^4 L_0^2 \tilde{L}_0^4}{\epsilon^2} \ln\left(\frac{n^2 L_0 \tilde{L}_0^2}{\epsilon}\right)$	–	$n^6 L_0^6 \tilde{L}_0^6 \epsilon^{-3}$	–
# Samples	$n^{4\bar{\tau}} L_0^{2\bar{\tau}} \tilde{L}_0^{4\bar{\tau}} \epsilon^{-2\bar{\tau}}$	–	$n^6 L_0^6 \tilde{L}_0^6 \epsilon^{-3}$	–

(2-iii) *Inexact nonconvex settings*: In addressing two-stage models with a nonconvex implicit objective function, we develop ($\texttt{ZSOL}_{\text{ncnvx}}^{\mathbf{2s}}$), a variance-reduced zeroth-order method. This scheme is presented by Algorithm 8. In Theorem 4 we obtain non-asymptotic convergence rates for solving the smoothed implicit problem and derive an overall iteration complexity of $\mathcal{O}\left(n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1}\right)$ and $\mathcal{O}\left(n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2}\right)$ for the projections on the set $\mathcal{X}$ and $\mathcal{Y}$, respectively. These results are similar to those we obtained for the single-stage counterpart. However, in computing an approximate $\mathbf{y}(\bullet, \omega)$ in the lower-level problem in Algorithm 6, unlike in the single-stage regime, we solve a deterministic variational inequality problem.

(2-iv) *Exact nonconvex settings:* Lastly, in Corollary 4, we consider the exact variant of ($\texttt{ZSOL}_{\text{ncnvx}}^{\mathbf{2s}}$). Similar to the single-stage case, to obtain an ϵ-solution (characterized by mean norm-squared of a residual mapping), we derive an iteration complexity of $\mathcal{O}\left(n^2 L_0^2 \epsilon^{-1}\right)$ for solving the smoothed approximate SMPEC. The number of calls to the solution oracle of the lower-level variational inequality problem is at most $\mathcal{O}\left(n^4 L_0^4 \epsilon^{-2}\right)$.

(3) Comprehensive numerics. In Sect. 5, we provide a comprehensive set of numerics where we provide empirical support for the scalability and convergence claims for inexact schemes for single and two-stage SMPECs. Such investigations also suggest the limited scalability of SAA schemes as well as the ability of the proposed schemes to compute near-global solutions under convexity of the implicit problems, in contrast with their SAA counterparts. Finally, the benefits of acceleration in terms of accuracy is observed as promised by theoretical claims.

To the best of our knowledge, all the above-mentioned rate and complexity results in addressing both the single-stage and two-stage SMPECs appear to be novel.

Notation. Throughout, we use the following notation and definitions. We let $\mathcal{X}^*$ and f^* denote the optimal solution set and the optimal objective value of a corresponding implicit problem, respectively. We define $D_{\mathcal{X}} \triangleq \frac{1}{2}\sup_{\mathbf{x}\in\mathcal{X}} \text{dist}^2(\mathbf{x}, \mathcal{X}^*)$. We let $\mathbb{B}$ denote the unit ball defined as $\mathbb{B} \triangleq \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}$ and $\mathbb{S}$ denote the surface of the ball $\mathbb{B}$, i.e., $\mathbb{S} \triangleq \{v \in \mathbb{R}^n \mid \|v\| = 1\}$. Given a set $\mathcal{X} \subseteq \mathbb{R}^n$ and a scalar $\eta > 0$, we let $\mathcal{X}_\eta$ denote the expanded set $\mathcal{X} + \eta\mathbb{B}$. Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and a set $\mathcal{X} \subseteq \mathbb{R}^n$, we write $f \in C^{0,0}(\mathcal{X})$ if f is Lipschitz continuous on the set

Table 2 Complexity guarantees for solving two-stage SMPECs

Two-stage SMPECs	Convex implicit			Nonconvex implicit	
	Inexact	Exact	Accelerated	Inexact	Exact
Upper level					
# Projections	$n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}$	$n^2 L_0^2 \epsilon^{-2}$	ϵ^{-1}	$n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1}$	$n^2 L_0^2 \epsilon^{-1}$
# Samples	$n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}$	$n^2 L_0^2 \epsilon^{-2}$	$\epsilon^{-(2+\delta)}$	$n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2}$	$n^4 L_0^4 \epsilon^{-2}$
Lower level					
# Projections	$\frac{n^4 L_0^2 \tilde{L}_0^4}{\epsilon^2} \ln\left(\frac{n^2 L_0 \tilde{L}_0^2}{\epsilon}\right)$	–	–	$\frac{n^4 L_0^4 \tilde{L}_0^4}{\epsilon^2} \ln\left(\frac{n^2 L_0^2 \tilde{L}_0^2}{\epsilon}\right)$	–

$\mathcal{X}$, i.e., $|f(\mathbf{x}) - f(\tilde{\mathbf{x}})| \leq L_0 \|\mathbf{x} - \tilde{\mathbf{x}}\|$ for all $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$ and some $L_0 > 0$. In the case where f is globally Lipschitz, i.e., $\mathcal{X} = \mathbb{R}^n$, we write $f \in C^{0,0}$. Given a continuously differentiable function and a set $\mathcal{X} \subseteq \mathbb{R}^n$, we write $f \in C^{1,1}(\mathcal{X})$ if ∇f is Lipschitz continuous on the set $\mathcal{X}$, i.e., $\|\nabla f(\mathbf{x}) - \nabla f(\tilde{\mathbf{x}})\| \leq L_1 \|\mathbf{x} - \tilde{\mathbf{x}}\|$ for all $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$ and some $L_1 > 0$. Similarly, we write $f \in C^{1,1}$ to denote that ∇f is globally Lipschitz. We denote the Euclidean projection of a vector $\mathbf{x}$ on a set $\mathcal{X}$ by $\Pi_{\mathcal{X}}(\mathbf{x})$, i.e., $\|\mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{x})\| = \min_{\bar{\mathbf{x}} \in \mathcal{X}} \|\mathbf{x} - \bar{\mathbf{x}}\|$. Throughout, unless otherwise specified, for the ease of presentation, we use $\mathbb{E}[\bullet]$ to denote the expectation with respect to all the random variables under discussion.

2 Preliminaries

In this section, we begin by outlining the key assumptions imposed on (SMPEC$^{\mathbf{1s}}$) and (SMPEC$^{\mathbf{2s}}$) in Sect. 2.1. Our treatment and analysis differ based on whether the implicit function $f^{\mathbf{imp}}$ is either convex or nonconvex. In the latter case, the resulting problem reduces to a nonsmooth nonconvex program with possibly expectation-valued objectives. In such settings, we provide a brief discussion of stationarity conditions in Sect. 2.2 while a discussion of locally randomized spherical smoothing techniques is presented in Sect. 2.3.

2.1 Problem definition

Throughout this paper, we assume that in the case of (SMPEC$^{\mathbf{1s}}$), the set $\mathcal{Y}$ is closed and convex in $\mathbb{R}^m$ and the parametrized map $F(\mathbf{x}, \bullet)$ is strongly monotone on $\mathcal{Y}$ uniformly in $\mathbf{x}$. An analogous assumption for (SMPEC$^{\mathbf{2s}}$) requires that $G(\mathbf{x}, \bullet, \omega)$ is strongly monotone on $\mathcal{Y}$ for every $\omega \in \Omega$. Since the lower-level problem is strongly monotone, the solution map of the lower-level problem is single-valued. Consequently, we may recast (SMPEC$^{\mathbf{2s}}$) as the following implicit program in $\mathbf{x}$.

$$\min_{\mathbf{x} \in \mathcal{X}} \; f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)], \tag{SMPEC$^{\mathbf{imp,2s}}$}$$

where $f^{\mathbf{imp}}(\bullet)$ is assumed to be Lipschitz continuous on a closed and convex set $\mathcal{X}$. Note that such a property on $f^{\mathbf{imp}}$ holds if $f^{\mathbf{imp}}$ is locally Lipschitz on a compact set. In the case of (SMPEC$^{\mathbf{1s}}$), the implicit problem reduces to

$$\min_{\mathbf{x}\in\mathcal{X}} \; f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)], \tag{SMPEC$^{\mathbf{imp,1s}}$}$$

where $\mathbf{y}(\mathbf{x})$ represents the solution to a variational inequality problem VI$(\mathcal{Y}, F(\mathbf{x}, \bullet))$. Note that this problem subsumes (SMPEC$^{\mathbf{1s}}$) by suppressing the expectation in the upper-level. We now formalize the assumptions on the problems of interest.

Assumption 1 (*Properties of* $f, F, \mathcal{X}, \mathcal{Y}$)
(a) Consider the problem (SMPEC$^{\mathbf{imp,1s}}$).
(a.i) For some $\eta_0 > 0$, $\tilde{f}(\bullet, \mathbf{y}(\bullet), \omega)$ is $L_0(\omega)$-Lipschitz continuous on $\mathcal{X} + \eta_0\mathbb{B}$ for every $\omega \in \Omega$, where $L_0 \triangleq \sqrt{\mathbb{E}[L_0^2(\omega)]} < \infty$. $\tilde{f}(\mathbf{x}, \bullet, \omega)$ is $\tilde{L}_0(\omega)$-Lipschitz for all $\mathbf{x} \in \mathcal{X} + \eta_0\mathbb{B}$ and for every $\omega \in \Omega$, where $\tilde{L}_0 \triangleq \sqrt{\mathbb{E}[\tilde{L}_0^2(\omega)]} < \infty$.
(a.ii) $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are nonempty, closed, bounded, and convex sets.
(a.iii) $F(\mathbf{x}, \bullet)$ is a μ_F-strongly monotone and L_F-Lipschitz continuous map on $\mathcal{Y}$ uniformly in $\mathbf{x} \in \mathcal{X}$.
(b) Consider the problem (SMPEC$^{\mathbf{imp,2s}}$).
(b.i) For some $\eta_0 > 0$, $\tilde{f}(\bullet, \mathbf{y}(\bullet, \omega), \omega)$ is $L_0(\omega)$-Lipschitz continuous on $\mathcal{X} + \eta_0\mathbb{B}$ for every $\omega \in \Omega$, where $L_0 \triangleq \sqrt{\mathbb{E}[L_0^2(\omega)]} < \infty$. $\tilde{f}(\mathbf{x}, \bullet, \omega)$ is $\tilde{L}_0(\omega)$-Lipschitz for all $\mathbf{x} \in \mathcal{X} + \eta_0\mathbb{B}$ and for every $\omega \in \Omega$, where $\tilde{L}_0 \triangleq \sqrt{\mathbb{E}[\tilde{L}_0^2(\omega)]} < \infty$.
(b.ii) $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are nonempty, closed, bounded, and convex sets.
(b.iii) $G(\mathbf{x}, \bullet, \omega)$ is a $\mu_F(\omega)$-strongly monotone and $L_F(\omega)$-Lipschitz continuous map on $\mathcal{Y}$ uniformly in $\mathbf{x} \in \mathcal{X}$ for every $\omega \in \Omega$, and there exist scalars $\mu_F, L_F \in (0, +\infty)$ such that $\inf_{\omega\in\Omega} \mu_F(\omega) \geq \mu_F$ and $\sup_{\omega\in\Omega} L_F(\omega) \leq L_F$. □

Remark 1 As outlined in Assumption 1, throughout we assume that the mapping in the lower-level parametrized by $\mathbf{x}$ is strongly monotone on $\mathcal{Y}$ uniformly in $\mathbf{x}$. The assumption is inherent to most implicit methods for resolving MPECs and our proposed schemes inherit that characteristic. When considering sample-average approximation schemes in the context of SMPECs, we observe that similar assumptions have been adopted in a subset of prior work including [47, 71, 79]. In fact, lower-level uniqueness is by no means a rarely seen phenomenon. It is inherent to a host of problems in practice [16, 54, 74, 76] and there is a significant body of research on implicit methods for solving MPECs in a range of settings [1, 7, 30, 39, 42, 43, 53]. In the current work, we intend to assess the fundamental gaps on the performance under a requirement on lower-level uniqueness but we allow for far more generality in the lower-level problem (e.g., in terms of accommodating expectation-valued maps) and either convexity or nonconvexity in terms of the upper-level problem.

We observe that the requirement that f is Lipschitz continuous on $\mathcal{X} + \eta_0\mathbb{B}$ (rather than $\mathcal{X}$) is a consequence of employing a smoothed approximation of f in our algorithm development. A natural question is whether the Lipschitz continuity of the

objective f over $\mathcal{X}$ in the implicit problem follows under reasonable conditions. The next result addresses precisely such a concern.

Proposition 1 *Consider the problem* (SMPEC$^{\mathbf{1s}}$). *Let Assumption* 1(a.ii, a.iii) *hold. Suppose* $\tilde{f}(\bullet, \bullet, \omega)$ *is continuously differentiable on* $\mathcal{C} \times \mathbb{R}^m$ *where* $\mathcal{C}$ *is an open set containing* $\mathcal{X}$. *Then the function* $f^{\mathbf{imp}}$, *defined as* $f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)]$, *is Lipschitz and directionally differentiable on* $\mathcal{X}$.

Proof This result follows from invoking [64, Cor. 4.2] together with the compactness of $\mathcal{X}$. □

Proposition 2 *Consider the problem* (SMPEC$^{\mathbf{2s}}$). *Let Assumption* 1(b.ii, b.iii) *hold. Suppose* $\tilde{f}(\bullet, \bullet, \omega)$ *is continuously differentiable on* $\mathcal{C} \times \mathbb{R}^m$ *where* $\mathcal{C}$ *is an open set containing* $\mathcal{X}$. *Then the function* $f^{\mathbf{imp}}$, *defined as* $f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)]$, *is Lipschitz and directionally differentiable on* $\mathcal{X}$.

Proof This result follows from invoking [64, Cor. 4.3] together with the compactness of $\mathcal{X}$. □

In a subset of regimes, $f^{\mathbf{imp}}$ is captured by the next assumption.

Assumption 2 (*Convexity of* f *in implicit problem*) Consider any of the implicit problems (SMPEC$^{\mathbf{imp,2s}}$) or (SMPEC$^{\mathbf{imp,1s}}$). Then the implicit function $f^{\mathbf{imp}}$ is convex on $\mathcal{X}$.

We note that there has been extensive study of conditions under which the implicit function $f^{\mathbf{imp}}$ is indeed convex (for example, see [16, 64, 78]). In fact, the convexity of the implicit function can be proven in MPECs arising in a host of application-driven regime [16, 73, 74, 76, 78], there appear to be no explicit conditions to the best of our knowledge.

2.2 Stationarity conditions

While the implicit function $f^{\mathbf{imp}}$ can be shown to be convex in some specific settings, the function $f^{\mathbf{imp}}$ is Lipschitz continuous on $\mathcal{X}$ in more general settings. Consequently, the problem can be compactly stated as

$$\min_{\mathbf{x} \in \mathcal{X}} \; h(\mathbf{x}) \triangleq f^{\mathbf{imp}}(\mathbf{x}). \tag{4}$$

We observe that h is a nonsmooth and possibly nonconvex function on $\mathcal{X}$. In the remainder of this subsection, we recap some of the concepts of Clarke's nonsmooth calculus that will facilitate the development of stationarity conditions. We begin by defining the directional derivative, a key object necessary in addressing nonsmooth and possibly nonconvex optimization problems (*cf.* [11]).

Definition 1 The directional derivative of h at $\mathbf{x}$ in a direction v is defined as

$$h^{\circ}(\mathbf{x}, v) \triangleq \limsup_{\mathbf{y} \to \mathbf{x}, t \downarrow 0} \left(\frac{h(\mathbf{y} + tv) - h(\mathbf{y})}{t} \right). \tag{5}$$

The Clarke generalized gradient at $\mathbf{x}$ can then be defined as

$$\partial h(\mathbf{x}) \triangleq \left\{ \zeta \in \mathbb{R}^n \mid h^{\circ}(\mathbf{x}, v) \geq \langle \zeta, v \rangle, \quad \forall v \in \mathbb{R}^n \right\}. \tag{6}$$

In other words, $h^{\circ}(\mathbf{x}, v) = \sup_{g \in \partial h(\mathbf{x})} \langle g, v \rangle$. □

If h is continuously differentiable at $\mathbf{x}$, we have that the Clarke generalized gradient reduces to the standard gradient, i.e., $\partial h(\mathbf{x}) = \nabla_{\mathbf{x}} h(\mathbf{x})$. If $\mathbf{x}$ is a minimal point of h, then we have that $0 \in \partial h(\mathbf{x})$. For purposes of completeness, we recap some properties of ∂h. Recall that if h is locally Lipschitz on an open set $\mathcal{C}$ containing $\mathcal{X}$, then h is differentiable almost everywhere on $\mathcal{C}$ by Rademacher's theorem [11]. Suppose $\mathcal{C}_h$ denotes the set of points where h is not differentiable. We may then recall some properties of Clarke generalized gradients.

Proposition 3 (Properties of Clarke generalized gradients [11]) *Suppose h is Lipschitz continuous on $\mathbb{R}^n$. Then the following hold.*

(i) *$\partial h(\mathbf{x})$ is a nonempty, convex, and compact set and $\|g\| \leq L$ for any $g \in \partial h(\mathbf{x})$.*
(ii) *h is differentiable almost everywhere.*
(iii) *$\partial h(\mathbf{x})$ is an upper semicontinuous map defined as*

$$\partial h(\mathbf{x}) = conv \left\{ g \mid g = \lim_{k \to \infty} \nabla_{\mathbf{x}} h(\mathbf{x}_k), \mathcal{C}_h \not\ni \mathbf{x}_k \to \mathbf{x} \right\}.$$

We may also define the δ-generalized gradient [28] as

$$\partial_{\delta} h(\mathbf{x}) \triangleq \text{conv} \left\{ \zeta : \zeta \in \partial h(\mathbf{y}), \|\mathbf{x} - \mathbf{y}\| \leq \delta \right\}. \tag{7}$$

Under the assumption that h is globally bounded from below and Lipschitz continuous on $\mathcal{X}$, in nonconvex regimes, our interest lies in developing techniques for computing an *approximate* stationary point. For instance, when h is L-smooth, then computing an approximate stationary point in unconstrained regimes such that $\|\nabla_{\mathbf{x}} h(\mathbf{x})\| \leq \epsilon$ requires at most $\mathcal{O}(1/\epsilon^2)$ gradient steps. Much of the prior work in the computation of stationary points of nonconvex and nonsmooth functions is either asymptotic [8, 9] or relies on some structure [6, 48, 80] where the nonconvex part is smooth while the convex part may be closed and proper. However, the question of computing approximate stationary points for functions that are both nonconvex and nonsmooth has been less studied.

2.3 Properties of spherical smoothing of *f*

We consider an iterative smoothing approach in this paper where a smoothed approximation of h is minimized and the smoothing parameter is progressively reduced. This

avenue has a long history, beginning with the efforts by Steklov [75] leading to significant efforts in both convex [18, 44, 81] and nonconvex [59] regimes. In this paper, we consider the following smoothing of h, given by h_η where

$$h_\eta(\mathbf{x}) \triangleq \mathbb{E}_{u\in\mathbb{B}}[h(\mathbf{x}+\eta u)], \tag{8}$$

where u is a random vector in the unit ball $\mathbb{B}$, defined as $\mathbb{B} \triangleq \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}$. Throughout, we let $\mathbb{S}$ denote the surface of the ball $\mathbb{B}$, i.e., $\mathbb{S} \triangleq \{v \in \mathbb{R}^n \mid \|v\| = 1\}$. We also let $\eta\mathbb{B}$ and $\eta\mathbb{S}$ denote the ball with radius η and its surface, respectively. Recall that if h is locally Lipschitz over a compact set $\mathcal{X}$, it is globally Lipschitz on $\mathcal{X}$. We may derive the following properties on h_η.

Lemma 1 (Properties of spherical smoothing)[2]

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $\eta > 0$ is a given scalar. Let h_η be defined as (8). Then the following hold.

(i) *The smoothed function h_η is continuously differentiable over $\mathcal{X}$. In particular, for any $\mathbf{x} \in \mathcal{X}$, we have that*

$$\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) = \left(\frac{n}{\eta}\right) \mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x}+v)\frac{v}{\|v\|}\right]. \tag{9}$$

Suppose $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0. For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have that (ii)–(iv) hold.

(ii) $|h_\eta(\mathbf{x}) - h_\eta(\mathbf{y})| \leq L_0\|\mathbf{x}-\mathbf{y}\|$.

(iii) $|h_\eta(\mathbf{x}) - h(\mathbf{x})| \leq L_0\eta$.

(iv) $\|\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) - \nabla_{\mathbf{x}} h_\eta(\mathbf{y})\| \leq \frac{L_0 n}{\eta}\|\mathbf{x}-\mathbf{y}\|$.

(v) *If h is convex and $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0, then h_η is convex and satisfies the following for any $\mathbf{x} \in \mathcal{X}$.*

$$h(\mathbf{x}) \leq h_\eta(\mathbf{x}) \leq h(\mathbf{x}) + \eta L_0. \tag{10}$$

(vi) *If h is convex and $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0, then $\nabla_x h_\eta(\mathbf{x}) \in \partial_\delta h(\mathbf{x})$ where $\delta \triangleq \eta L_0$.*

(vii) *If $h \in C^{1,1}(\mathcal{X}_\eta)$ with constant L_1, then $\|\nabla_x h_\eta(\mathbf{x}) - \nabla_x h(\mathbf{x})\| \leq \eta L_1 n$.*

(viii) *Suppose $h \in C^{0,0}(\mathcal{X}_\eta)$ with parameter L_0. Let us define for $v \in \eta\mathbb{S}$*

$$g_\eta(\mathbf{x}, v) \triangleq \left(\frac{n}{\eta}\right)\frac{(h(\mathbf{x}+v)-h(\mathbf{x}))v}{\|v\|}.$$

Then, for any $\mathbf{x} \in \mathcal{X}$, we have that $\mathbb{E}_{v\in\eta\mathbb{S}}[\|g_\eta(\mathbf{x}, v)\|^2] \leq L_0^2 n^2$.

[2] We note that while spherical smoothing has apparently been studied in [56], we did not have access to this text. Part (i) of our lemma is inspired by Flaxman et al. [24] while other parts either follow in a fashion similar to Gaussian smoothing [59] or are directly proven.

Proof (i) We elaborate on the proof sketch provided in [24]. By definition, we have that

$$h_\eta(\mathbf{x}) = \mathbb{E}_{u\in\eta\mathbb{B}}[h(\mathbf{x}+u)] = \int_{\eta\mathbb{B}} h(\mathbf{x}+u)\,p(u)du.$$

Let $p(u)$ denote the probability density function of u. Since u is uniformly distributed in the ball $\eta\mathbb{B}$, we have that $p(u) = \frac{1}{\mathrm{Vol}(\eta\mathbb{B})}$ for any $u \in \eta\mathbb{B}$. Consequently,

$$h_\eta(\mathbf{x}) = \int_{\eta\mathbb{B}} h(\mathbf{x}+u)\,p(u)du = \frac{\int_{\eta\mathbb{B}} h(\mathbf{x}+u)du}{\mathrm{Vol}_n(\eta\mathbb{B})}.$$

We may then compute the derivative $\nabla_{\mathbf{x}} h_\eta(\mathbf{x})$ by leveraging Stoke's theorem and by defining $\tilde{p}(v) = \frac{1}{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}$ for all v.

$$\begin{aligned}
\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) &= \nabla_{\mathbf{x}}\left[\frac{\int_{\eta\mathbb{B}} h(\mathbf{x}+u)du}{\mathrm{Vol}_n(\eta\mathbb{B})}\right] \overset{\text{Stoke's theorem}}{=} \left[\frac{\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}dv}{\mathrm{Vol}_n(\eta\mathbb{B})}\right] \\
&= \left[\frac{\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}dv}{\mathrm{Vol}_n(\eta\mathbb{B})}\right]\frac{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}{\mathrm{Vol}_{n-1}(\eta\mathbb{S})} \\
&= \left[\frac{\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}dv}{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}\right]\frac{\mathrm{Vol}_{n-1}(\eta\mathbb{S})}{\mathrm{Vol}_n(\eta\mathbb{B})} = \left[\int_{\eta\mathbb{S}} h(\mathbf{x}+v)\frac{v}{\|v\|}\tilde{p}(v)dv\right]\frac{n}{\eta} \\
&= \frac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x}+v)\frac{v}{\|v\|}\right].
\end{aligned}$$

(ii) We have

$$\begin{aligned}
|h_\eta(\mathbf{x}) - h_\eta(\mathbf{y})| &= |\mathbb{E}_{u\in\mathbb{B}}[h(\mathbf{x}+\eta u)] - \mathbb{E}_{u\in\mathbb{B}}[h(\mathbf{y}+\eta u)]| \\
&\overset{\text{Jensen's ineq.}}{\leq} \mathbb{E}_{u\in\mathbb{B}}[|h(\mathbf{x}+\eta u) - h(\mathbf{y}+\eta u)|] \\
&\overset{h\in C^{0,0}(\mathcal{X}_\eta)}{\leq} \mathbb{E}_{u\in\mathbb{B}}[L_0\|\mathbf{x}-\mathbf{y}\|] = L_0\|\mathbf{x}-\mathbf{y}\|.
\end{aligned}$$

(iii) Next, we show that $|h_\eta(\mathbf{x}) - h(\mathbf{x})|$ can be bounded in terms of η and L_0.

$$\begin{aligned}
|h_\eta(\mathbf{x}) - h(\mathbf{x})| &= \left|\int_{\eta\mathbb{B}} (h(\mathbf{x}+u) - h(\mathbf{x}))p(u)du\right| \\
&\leq \int_{\eta\mathbb{B}} |(h(\mathbf{x}+u) - h(\mathbf{x}))|\,p(u)du \\
&\leq L_0\int_{\eta\mathbb{B}} \|u\|\,p(u)du \leq L_0\eta\int_{\eta\mathbb{B}} p(u)du = L_0\eta.
\end{aligned}$$

(iv) Note that we have $\mathcal{X} + \eta\mathbb{S} \subseteq \mathcal{X} + \eta\mathbb{B}$. Thus, from the definition of $\mathcal{X}_\eta$ and $h \in C^{0,0}(\mathcal{X}_\eta)$, we have $h \in C^{0,0}(\mathcal{X} + \eta\mathbb{S})$. As such, we have

$$\begin{aligned}\left\|\nabla_{\mathbf{x}} h_\eta(\mathbf{x}) - \nabla_{\mathbf{x}} h_\eta(\mathbf{y})\right\| &= \left\| \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x}+v)\tfrac{v}{\|v\|}\right] - \tfrac{n}{\eta}\mathbb{E}_{v\in\mathbb{S}}\left[h(\mathbf{y}+v)\tfrac{v}{\|v\|}\right]\right\| \\ &\leq \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left\|(h(\mathbf{x}+v)-h(\mathbf{y}+v))\tfrac{v}{\|v\|}\right\|\right] \\ &\leq \tfrac{L_0 n}{\eta}\|\mathbf{x}-\mathbf{y}\|\mathbb{E}_{v\in\eta\mathbb{S}}\left[\tfrac{\|v\|}{\|v\|}\right] = \tfrac{L_0 n}{\eta}\|\mathbf{x}-\mathbf{y}\|.\end{aligned}$$

(v) First, note that from $h \in C^{0,0}(\mathcal{X}_\eta)$, we have that $h \in C^{0,0}(\text{int}(\mathcal{X}_\eta))$. Noting that $\text{int}(\mathcal{X}_\eta)$ is an open set, from part (b) of Theorem 3.61 in [6], we have that $\|\tilde{g}\| \leq L_0$ for all $\mathbf{x} \in \text{int}(\mathcal{X}_\eta)$ and $\tilde{g} \in \partial h(\mathbf{x})$. The desired statements then follow from part (a) and part (b) of Lemma 2 [83].

(vi) From part (v), function h_η is convex and $h(\mathbf{y}) + \eta L_0 \geq h_\eta(\mathbf{y})$ for any $\mathbf{y} \in \mathcal{X}$. Thus, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$h(\mathbf{y}) + \eta L_0 \geq h_\eta(\mathbf{y}) \geq h_\eta(\mathbf{x}) + \nabla h_\eta(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \geq h(\mathbf{x}) + \nabla h_\eta(\mathbf{x})^T(\mathbf{y}-\mathbf{x}).$$

The result follows by choosing $\delta = \eta L_0$.

(vii) Note that we can show that $\int_{\eta\mathbb{S}} vv^T p_v(v)dv = \frac{\eta^2}{n}\mathbf{I}$. We may then express $\nabla_x h(x)$ as

$$\begin{aligned}\nabla_x h(\mathbf{x}) &= \tfrac{n}{\eta^2}\left(\int_{\eta\mathbb{S}} vv^T p_v(v)dv\right)\nabla_x h(\mathbf{x}) = \tfrac{n}{\eta^2}\left(\int_{\eta\mathbb{S}} v^T\nabla_x h(\mathbf{x})vp_v(v)dv\right) \\ &= \tfrac{n}{\eta}\left(\int_{\eta\mathbb{S}} v^T\nabla_x h(\mathbf{x})\tfrac{v}{\|v\|}p_v(v)dv\right) = \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left(\nabla_x h(\mathbf{x})^T v\right)\tfrac{v}{\|v\|}\right],\end{aligned}$$

where the third inequality follows from $\|v\| = \eta$ for $v \in \eta\mathbb{S}$. From this relation, part (i), and by recalling that $\frac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[h(\mathbf{x})\frac{v}{\|v\|}\right] = 0$, we can write

$$\begin{aligned}\|\nabla_x h_\eta(\mathbf{x}) - \nabla_x h(\mathbf{x})\| &= \left\|\tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[(h(\mathbf{x}+v)-h(\mathbf{x}))\tfrac{v}{\|v\|}\right]\right. \\ &\quad \left. -\tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left(\nabla h(\mathbf{x})^T v\right)\tfrac{v}{\|v\|}\right]\right\| \\ &\leq \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[\left|h(\mathbf{x}+v)-h(\mathbf{x})-\nabla h(\mathbf{x})^T v\right|\tfrac{\|v\|}{\|v\|}\right] \\ &\leq \tfrac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[L_1\|v\|^2\right] = n\eta L_1,\end{aligned}$$

where the two inequalities follow from Jensen's inequality, the Cauchy-Schwarz inequality, and L_1-smoothness of h.

(viii) We observe that for any $\mathbf{x}$, $\mathbb{E}_{v\in\eta\mathbb{S}}[\|g_\eta(\mathbf{x}, v)\|^2]$ may be bounded as follows.

$$\mathbb{E}_{v\in\eta\mathbb{S}}[\|g_\eta(\mathbf{x},v)\|^2] = \frac{n^2}{\eta^2}\int_{\eta\mathbb{S}} \frac{\|(h(\mathbf{x}+v)-h(\mathbf{x}))v\|^2}{\|v\|^2} p_v(v)dv$$
$$\leq \frac{n^2}{\eta^2}\int_{\eta\mathbb{S}} L_0^2\|v\|^2 p_v(v)dv \leq n^2\int_{\eta\mathbb{S}} p_v(v)dv = n^2L_0^2.$$

□

Remark 2 (*Local vs global smoothing*) Gaussian smoothing as employed in [59] allows for unbounded random variables as part of the smoothing process. However, this precludes contending with compact regimes which we may require to impose Lipschitzian assumptions. Furthermore, in many settings, the domain of the function is compact and Gaussian smoothing cannot be adopted. Instead, local smoothing requires that the smoothing random variable have compact support. In [81, 83], we examine smoothing schemes based on random variables defined on either a cube or a sphere. However, most of the results of the previous lemma are novel with respect to [83].

We intend to develop schemes for computing approximate stationary points of (4) by an iterative smoothing scheme. However, this needs formalizing the relationship between the original problem and its smoothed counterpart. Before proceeding, we define the δ-Clarke generalized gradient of h, denoted by $\partial_\delta h(\mathbf{x})$ at $\mathbf{x}$, as follows [28].

$$\partial_\delta h(\mathbf{x}) \triangleq \text{conv}\,\{\zeta \mid \zeta \in \partial h(\mathbf{y}), \|\mathbf{y}-\mathbf{x}\| \leq \delta\}. \tag{11}$$

It was first shown by Goldstein [28] that $\partial_\delta h(\mathbf{x})$ is a nonempty, compact, and convex set.

Proposition 4 *Consider the problem* (4) *where h is a locally Lipschitz continuous function and $\mathcal{X}$ is a closed, convex, and bounded set in $\mathbb{R}^n$.*

(i) *For any $\eta > 0$ and any $\mathbf{x} \in \mathbb{R}^n$, $\nabla h_\eta(\mathbf{x}) \in \partial_{2\eta} h(\mathbf{x})$. Furthermore, if $0 \notin \partial h(\mathbf{x})$, then there exists an η such that $\nabla_{\mathbf{x}} h_{\tilde{\eta}}(\mathbf{x}) \neq 0$ for $\tilde{\eta} \in (0, \eta]$.*
(ii) *For any $\eta > 0$ and any $\mathbf{x} \in \mathcal{X}$,*

$$\left[0 \in \nabla_{\mathbf{x}} h_\eta(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})\right] \implies \left[0 \in \partial_{2\eta} h(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})\right]. \tag{12}$$

Proof (i) and (ii) represent a constrained counterparts of [51, Prop. 2.2 and Cor. 2.1]. □

Lemma 1(v) provides a statement that relates the true objective to its smoothed counterpart in convex regimes. This provides an avenue for developing finite-time schemes for computing approximate solutions to the *original problem*. Proposition 4(ii) provides a relationship in settings where h is locally Lipschitz; in particular, it is shown that if $\mathbf{x}$ satisfies stationarity of the η-smoothed problem, it satisfies a suitable $2\eta-$stationarity property for the original problem.

3 Zeroth-order methods for single-stage SMPECs

In this section, we present a zeroth-order framework for contending with (SMPEC$^{\mathbf{imp,1s}}$). The remainder of this section is organized as follows. In Sect. 3.1, we introduce an implicit zeroth-order scheme that can allow for constructing a smoothed zeroth-order gradient through leveraging inexact solutions of the lower-level problem. To address settings where the implicit problem is convex, we derive rate and complexity guarantees for an iteratively smoothed gradient framework in Sect. 3.2 when the lower-level problem is either inexactly or exactly resolved. In these settings, the smoothing parameter is progressively reduced at each iteration. Lastly in Sect. 3.3, we derive the iteration complexity in addressing the nonconvex case under a constant smoothing parameter.

3.1 An implicit zeroth-order scheme

Since the implicit function is merely Lipschitz continuous, we employ a zeroth-order framework that relies on computing a zeroth-order approximation of the gradient. Consider the implicit problem (SMPEC$^{\mathbf{imp,1s}}$). Given the function $f^{\mathbf{imp}}$ and a scalar η, we consider a spherical smoothing denoted by $f_\eta^{\mathbf{imp}}$ based on (8), defined as

$$f_\eta^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}_{u\in\mathbb{B}}[f^{\mathbf{imp}}(\mathbf{x}+\eta u)] = \mathbb{E}_{u\in\mathbb{B}}[\mathbb{E}[\tilde{f}(\mathbf{x}+\eta u, \mathbf{y}(\mathbf{x}+\eta u), \omega)]], \tag{G-Smooth$^{\mathbf{1s}}$}$$

where u is uniformly distributed in the unit ball $\mathbb{B}$. Let $g_\eta(\mathbf{x})$ denote a zeroth-order approximation of the gradient of $f_\eta^{\mathbf{imp}}(\mathbf{x})$. Invoking Lemma 1, one choice for g_η is given as follows for any $\mathbf{x}$.

$$g_\eta(\mathbf{x}) = \left(\frac{n}{\eta}\right)\mathbb{E}_{v\in\eta\mathbb{S}}\left[\frac{\left(f^{\mathbf{imp}}(\mathbf{x}+v) - f^{\mathbf{imp}}(\mathbf{x})\right)v}{\|v\|}\right]. \tag{13}$$

In general, given the presence of the expectation, $g_\eta(\mathbf{x})$ is challenging to evaluate and a common approach has been in utilizing an unbiased estimate given by $g_\eta(\mathbf{x}, v, \omega)$ defined as

$$g_\eta(\mathbf{x}, v, \omega) \triangleq \left(\frac{n}{\eta}\right)\left[\frac{\left(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)\right)v}{\|v\|}\right]. \tag{14}$$

Given a vector $\mathbf{x}_0 \in \mathcal{X}$, we may employ (14) in constructing a sequence $\{\mathbf{x}_k\}$ where $\mathbf{x}_k$ satisfies the following projected stochastic gradient update.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_\eta(\mathbf{x}_k, v_k, \omega_k)\right]. \tag{15}$$

Motivated by the development of the stochastic approximation (SA) scheme [67], the projected stochastic gradient and gradient-free schemes have been studied exten-

sively in convex and nonconvex regimes (e.g., see [26, 27, 55, 81] and the references therein). Recall that in the SA schemes, the standard requirements on the stepsize sequence include $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. The scheme (15) has been studied for addressing nonsmooth convex and nonconvex optimization problems [59] while unconstrained nonconvex regimes were examined in [26]. In particular, in the work by Nesterov and Spokoiny [59], zeroth-order randomized smoothing gradient schemes are proposed under a single sample with a fixed smoothing parameter η with the assumption that the smoothing random variable v has a Gaussian distribution. Importantly, a direct adoption of such smoothing schemes to address the hierarchical problems studied in this work is afflicted by several challenges.

(i) *Lack of asymptotic guarantees.* When $\eta > 0$, the scheme generates a sequence that is convergent to an approximate solution, at best. In addition, the choice of η is contingent on accurate estimates of other problem parameters (such as L_0), in the absence of which, η may be chosen to be extremely small. This often afflicts the practical behavior of the scheme. Moreover, employing a fixed η precludes asymptotic convergence to the true counterpart. Instead, in most of our schemes, we employ a mini-batch approximation of $g_\eta(\mathbf{x})$, denoted by $g_{\eta,N}(\mathbf{x})$ and defined as

$$g_{\eta,N}(\mathbf{x}) \triangleq \frac{\sum_{j=1}^{N} g_\eta(\mathbf{x}, v_j, \omega_j)}{N}. \tag{16}$$

Furthermore, we replace a fixed η by a diminishing sequence $\{\eta_k\}$, the resulting iterative smoothing scheme being articulated as follows.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k,N_k}(\mathbf{x}_k)\right]. \tag{17}$$

(ii) *Unavailability of exact solutions of* $\mathbf{y}(\mathbf{x})$. Even if $\mathbf{y}(\bullet)$ is a single-valued map requiring the solution of a strongly monotone lower-level problem, computing a solution to this problem is not necessarily cheap. As a consequence, our scheme needs to account for random errors in the computation of $g_{\eta_k}(\mathbf{x}_k)$, denoted by $\tilde{b}_k$. As a consequence, the resulting scheme is defined as follows.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k(g_{\eta_k,N_k}(\mathbf{x}_k) + \tilde{b}_k)\right], \quad \text{for all} \quad k \geq 0. \tag{18}$$

In particular, when considering problems (SMPEC$^{\mathbf{1s}}$), exact solutions of $\mathbf{y}(\mathbf{x}_k)$ are generally unavailable in finite time. Instead, one can take t_k steps of a standard projection scheme.

$$\mathbf{y}_{t+1} := \Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \beta_t \bar{F}(\mathbf{x}_k, \mathbf{y}_t)\right], \qquad t = 0, \ldots, t_k - 1, \tag{19}$$

where $\bar{F}(\mathbf{x}_k, \mathbf{y}_t) \triangleq \frac{\sum_{\ell=1}^{M_t} G(\mathbf{x}_k, \mathbf{y}_t, \omega_{\ell,t})}{M_t}$. In such a variance-reduced scheme, when M_t grows at a geometric rate, $\ln\left(\frac{1}{\epsilon_k}\right)$ steps of (19) are required to obtain an ϵ_k-solution of $\mathbf{y}_k$ [33].

(iii) *Bias in* $\tilde{b}_k$. A key issue that arises from (ii) emerges in the form of bias. In particular, $g_{\eta_k,N_k}(\mathbf{x}_k) + \tilde{b}_k$ is not necessarily an unbiased estimator of $g_{\eta_k}(\mathbf{x}_k)$. Further, it remains unclear how the bias and variance of $g_{\eta_k,N_k}(\mathbf{x}_k)+\tilde{b}_k$ propagate through this framework (18)–(19) as γ_k, η_k, and N_k are updated iteratively in the outer loop (18). Consequently, in the development of the inexact smoothing scheme (18)–(19), it remains critical to design prescribed stepsize, smoothing, and sample-size sequences to control the accuracy of the estimator $g_{\eta_k,N_k}(\mathbf{x}_k)+\tilde{b}_k$ and consequently, ascertain the convergence of the generated iterate to an optimal solution of the underlying MPEC. This concern will be examined in detail in the subsequent sections.

3.2 Convex single-stage regimes

In this subsection, we consider resolving the implicit formulations when the implicit function is convex. As pointed out earlier, the convexity of the implicit problem often holds in practice (cf. [16, 64, 78]). We first consider the inexact case where the exact value of $\mathbf{y}(\bullet)$ is not necessarily available. We then specialize our statements to settings where exact solutions of lower-level problems can be employed.

3.2.1 An inexact zeroth-order scheme

We now delve into developing and analyzing an inexact zeroth-order method for resolving the implicit variant (SMPEC$^{\mathbf{imp,1s}}$). We begin by providing the general setup and assumptions. Then, we provide some key results and algorithms. Before proceeding, we consider the following assumption.

Assumption 3 Given a sequence $\{\eta_k\}$, let $\{v_k\} \in \mathbb{R}^n$ be iid replicates uniformly distributed on $\eta_k\mathbb{S}$ for all $k \geq 0$. Also, let $\{\omega_k\}$ be iid replicates.

Remark 3 Throughout the paper, for the ease of presentation, we assume that there exists an oracle that returns the replicates of ω in the upper-level. The function $\tilde{f}(\bullet, \bullet, \omega)$ can then be evaluated using a second oracle. Note that this assumption is without loss of any generality and an alternative approach is to assume that there exists an oracle that generates the random realizations of $\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)$ and $\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega)$ directly.

Consider the implicit form of (SMPEC$^{\mathbf{1s}}$), i.e., (SMPEC$^{\mathbf{imp,1s}}$) where the lower-level problem is complicated by the presence of expectation-valued maps, i.e., F is defined as (1) and satisfies Assumption 1 (a.iii). In such an instance, obtaining $\mathbf{y}(\mathbf{x})$ is impossible in finite time unless the expectation can be tractably resolved. Instead, by employing stochastic approximation methods for addressing the lover-level problem, we consider the case where we have access to an approximate solution $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k)$ such that the following holds a.s.

$$\mathbb{E}[\|\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k) - \mathbf{y}(\mathbf{x}_k)\|^2 \mid \mathbf{x}_k] \leq \tilde{\epsilon}_k, \quad \text{where} \quad \mathbf{y}(\mathbf{x}_k) \in \mathrm{SOL}(\mathcal{Y}, F(\mathbf{x}_k, \bullet)). \tag{20}$$

As a consequence, we may define an inexact zeroth-order gradient mapping $g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)$ as follows.

$$g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) \triangleq \frac{n(\tilde{f}(\mathbf{x}+v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}), \omega))v}{\|v\|\eta}, \tag{21}$$

where $v \in \eta\mathbb{S}$ and $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k)$ is an output of a variance-reduced stochastic approximation scheme. The outline of the proposed zeroth-order solver ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) is presented in Algorithm 1 while an inexact solution of $\mathbf{y}(\mathbf{x})$ is computed by Algorithm 2. We impose the following assumptions on the lower-level evaluations $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})$ in Algorithm 2.

Assumption 4 Consider Algorithm 2. Let the following hold for all $k \geq 0$, $t \geq 0$, $\hat{\mathbf{x}}_k \in \mathcal{X}$, $\mathbf{y}_t \in \mathcal{Y}$, and $1 \leq \ell \leq M_t$ where M_t denotes the batch size at iteration t.

(a) The replicates $\{G(\bullet, \bullet, \omega_{\ell,t})\}_{\ell=1}^{M_t}$ are generated randomly and are iid.
(b) $\mathbb{E}[G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] = F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$ holds almost surely.
(c) $\mathbb{E}[\|G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) - F(\hat{\mathbf{x}}_k, \mathbf{y}_t)\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] \leq \nu_{\mathbf{y}}^2\|\mathbf{y}_t\|^2 + \nu_G^2$ holds almost surely for some deterministic scalars $\nu_{\mathbf{y}} \geq 0$ and $\nu_G > 0$.

Algorithm 1 $\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$: Zeroth-order method for convex (SMPEC[1s])

1: **input:** Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize sequence $\{\gamma_k\}$, smoothing parameter sequence $\{\eta_k\}$, inexactness sequence $\{\tilde{\epsilon}_k\}$, $r \in [0, 1)$, and $S_0 := \gamma_0^r$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Generate iid replicates $\omega_k \in \Omega$ and $v_k \in \eta_k\mathbb{S}$
4: Do one of the following, depending on the type of the scheme.

- Inexact scheme: Call Algorithm 2 twice to obtain $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k)$ and $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k + v_k)$
- Exact scheme: Evaluate $\mathbf{y}(\mathbf{x}_k)$ and $\mathbf{y}(\mathbf{x}_k + v_k)$

5: Evaluate the inexact or exact zeroth-order gradient approximation as follows.

$$g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k) := \frac{n\left(\tilde{f}(\mathbf{x}_k+v_k, \mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k+v_k), \omega_k) - \tilde{f}(\mathbf{x}_k, \mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k), \omega_k)\right)v_k}{\|v_k\|\eta_k} \tag{Inexact}$$

$$g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k) := \frac{n\left(\tilde{f}(\mathbf{x}_k+v_k, \mathbf{y}(\mathbf{x}_k+v_k), \omega_k) - \tilde{f}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k), \omega_k)\right)v_k}{\|v_k\|\eta_k}. \tag{Exact}$$

6: Update $\mathbf{x}_k$ as follows.

$$\mathbf{x}_{k+1} := \begin{cases} \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k)\right] & \text{(Inexact)} \\ \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k)\right] & \text{(Exact)} \end{cases}$$

7: Update the averaged iterate as follows. $S_{k+1} := S_k + \gamma_{k+1}^r$ and $\bar{\mathbf{x}}_{k+1} := \frac{S_k\bar{\mathbf{x}}_k + \gamma_{k+1}^r\mathbf{x}_{k+1}}{S_{k+1}}$
8: **end for**

Algorithm 2 Variance-reduced SA method for lower-level of convex (SMPEC$^{\mathbf{1s}}$)

input: An arbitrary $\mathbf{y}_0 \in \mathcal{Y}$, vector $\hat{\mathbf{x}}_k$ (that is either $\mathbf{x}_k$ or $\mathbf{x}_k + v_k$ from Alg. 1), scalar $\rho \in (0,1)$, stepsize $\alpha > 0$, mini-batch sequence $\{M_t\}$ with $M_t := \lceil M_0 \rho^{-t} \rceil$, integer k, and scalars $M_0, \tau > 0$ (see Def. (2))
Compute $t_k := \lceil \tau \ln(k+1) \rceil$
for $t = 0, 1, \ldots, t_k - 1$ **do**
  Generate random realizations of the stochastic mapping $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})$ for $\ell = 1, \ldots, M_t$
  Update $\mathbf{y}_t$ as follows. $\mathbf{y}_{t+1} := \Pi_{\mathcal{Y}} \left[\mathbf{y}_t - \alpha \frac{\sum_{\ell=1}^{M_t} G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})}{M_t} \right]$
end for
Return $\mathbf{y}_{t_k}$

Before analyzing ($\mathtt{ZSOL}^{\mathbf{1s}}_{\mathbf{cnvx}}$), we review the properties of the exact zeroth-order stochastic gradient denoted by $g_\eta(\mathbf{x}, v, \omega)$ and show that it is an unbiased estimator of the gradient of the smoothed implicit function. We then derive a bound on the second moment of this stochastic gradient under the assumption that the implicit stochastic function is Lipschitz.

Remark 4 Throughout, we use the definition $g_\eta(\mathbf{x}, v) \triangleq \left(\frac{n}{\eta}\right) \frac{\left(f^{\mathbf{imp}}(\mathbf{x}+v) - f^{\mathbf{imp}}(\mathbf{x})\right) v}{\|v\|}$, where $f^{\mathbf{imp}}(\bullet)$ is the implicit function defined by (SMPEC$^{\mathbf{imp,1s}}$) or (SMPEC$^{\mathbf{imp,2s}}$) .

Lemma 2 (Properties of the single-stage exact zeroth-order gradient) *Suppose Assumption* 1(a) *holds. Consider* (SMPEC$^{\mathbf{imp,1s}}$). *Given* $\mathbf{x} \in \mathcal{X}$ *and* $\eta > 0$, *consider the stochastic zeroth-order mapping* $g_\eta(\mathbf{x}, v, \omega)$ *defined by* (14) *for* $v \in \eta\mathbb{S}$ *and* $k \geq 0$, *where* v *and* ω *are independent.*

Then, $\nabla f^{\mathbf{imp}}_\eta(\mathbf{x}) = \mathbb{E}[g_\eta(\mathbf{x}, v, \omega) \mid \mathbf{x}]$ *and* $\mathbb{E}[\|g_\eta(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}] \leq L_0^2 n^2$ *almost surely for all* $k \geq 0$.

Proof From (14) and that $f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)]$ we can write

$$\begin{aligned}\mathbb{E}[g_\eta(\mathbf{x}, v, \omega) \mid \mathbf{x}] &= \mathbb{E}_{v \in \eta\mathbb{S}}\left[\left(\frac{n}{\eta}\right) \frac{\left(f^{\mathbf{imp}}(\mathbf{x}+v) - f^{\mathbf{imp}}(\mathbf{x})\right) v}{\|v\|} \mid \mathbf{x}\right] \\ &= \left(\frac{n}{\eta}\right) \mathbb{E}_{v \in \eta\mathbb{S}}\left[f^{\mathbf{imp}}(\mathbf{x}+v) \frac{v}{\|v\|} \mid \mathbf{x}\right] \overset{\text{Lemma 1}(i)}{=} \nabla f^{\mathbf{imp}}_\eta(\mathbf{x}).\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}[\|g_\eta(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}, \omega] &= \left(\frac{n}{\eta}\right)^2 \mathbb{E}\left[\left\| \frac{\left(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)\right) v}{\|v\|} \right\|^2 \mid \mathbf{x}, \omega\right] \\ &= \left(\frac{n}{\eta}\right)^2 \int_{\eta\mathbb{S}} \frac{\left\| \left(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega)\right) v \right\|^2}{\|v\|^2} p_v(v) dv \\ &\overset{\text{Assumption 1(a.i)}}{\leq} \frac{n^2}{\eta^2} \int_{\eta\mathbb{S}} L_0^2(\omega) \|v\|^2 p_v(v) dv\end{aligned}$$

$$\leq n^2 L_0^2(\omega) \int_{\eta\mathbb{S}} p_v(v) dv = n^2 L_0^2(\omega).$$

Taking expectations with respect to ω on both sides of the preceding inequality and invoking $L_0^2 \triangleq \mathbb{E}[L_0^2(\omega)] < \infty$, we obtain the desired bound. □

We are now ready to present the properties of the inexact zeroth-order gradient mapping.

Lemma 3 (Properties of the single-stage inexact zeroth-order gradient) *Consider* (SMPEC$^{\mathbf{imp,1s}}$). *Suppose Assumption* 1(a) *holds. Let* $g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)$ *be defined as* (21) *for* $\omega \in \Omega$ *and* $v \in \eta\mathbb{S}$ *for* $\eta, \tilde{\epsilon} > 0$. *Suppose* $\mathbb{E}[\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|^2 \mid \mathbf{x}, \omega] \leq \tilde{\epsilon}$ *almost surely for all* $\mathbf{x} \in \mathcal{X}$. *Then, the following hold for the single-stage model for any* $\mathbf{x} \in \mathcal{X}$.

(a) $\mathbb{E}[\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}] \leq 3n^2 \left(\frac{2\tilde{L}_0^2 \tilde{\epsilon}}{\eta^2} + L_0^2 \right)$, *almost surely.*

(b) $\mathbb{E}\left[\left\| g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega) \right\|^2 \mid \mathbf{x} \right] \leq \frac{4\tilde{L}_0^2 n^2 \tilde{\epsilon}}{\eta^2}$, *almost surely.*

Proof **(a)** Adding and subtracting $g_\eta(\mathbf{x}, v, \omega)$, we obtain from (21)

$$\begin{aligned}
&\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\| \\
&= \left\| \frac{n(\tilde{f}(\mathbf{x}+v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega))v}{\|v\|\eta} \right. \\
&\quad \left. + g_\eta(\mathbf{x}, v, \omega) + \frac{n(\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}), \omega))v}{\|v\|\eta} \right\| \\
&\leq \left\| \frac{n(\tilde{f}(\mathbf{x}+v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega))v}{\|v\|\eta} \right\| + \left\| g_\eta(\mathbf{x}, v, \omega) \right\| \\
&\quad + \left\| \frac{n(\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}), \omega))v}{\|v\|\eta} \right\| \\
&\leq \frac{\|\tilde{f}(\mathbf{x}+v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega)\| n \|v\|}{\|v\|\eta} + \left\| g_\eta(\mathbf{x}, v, \omega) \right\| \\
&\quad + \frac{\|\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}), \omega)\| n \|v\|}{\eta \|v\|} \\
&\leq \frac{\tilde{L}_0(\omega) \|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v) - \mathbf{y}(\mathbf{x}+v)\| n}{\eta} + \left\| g_\eta(\mathbf{x}, v, \omega) \right\| + \frac{\tilde{L}_0(\omega) \|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}) - \mathbf{y}(\mathbf{x})\| n}{\eta}.
\end{aligned}$$

Invoking Lemma 2, we may then bound the conditional second moment of $\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\|$ as follows.

$$\mathbb{E}[\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}] \leq 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2(\omega) n^2 \|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v) - \mathbf{y}(\mathbf{x}+v)\|^2}{\eta^2} \right) \mid \mathbf{x} \right]$$

$$
\begin{aligned}
&+ 3\mathbb{E}\left[\left\|g_\eta(\mathbf{x}, v, \omega)\right\|^2 \mid \mathbf{x}\right] \\
&+ 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2(\omega) n^2 \|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|^2}{\eta^2}\right) \mid \mathbf{x}\right] \\
&\leq 6\left(\frac{\tilde{L}_0^2 n^2 \tilde{\epsilon}}{\eta^2}\right) + 3L_0^2 n^2, \text{ a.s.}
\end{aligned} \tag{22}
$$

(b) We first derive a bound on $\left\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega)\right\|$.

$$
\begin{aligned}
&\left\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega)\right\| \\
&= \left\| \frac{n(\tilde{f}(\mathbf{x}+v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}), \omega))v}{\|v\|\eta} \right. \\
&\quad \left. - \frac{n(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega))v}{\|v\|\eta} \right\| \\
&\leq \left\| \frac{n(\tilde{f}(\mathbf{x}+v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v), \omega) - \tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v), \omega))v}{\|v\|\eta} \right\| \\
&\quad + \left\| \frac{n(\tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}), \omega))v}{\|v\|\eta} \right\| \\
&\leq \frac{\tilde{L}_0 n \|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v) - \mathbf{y}(\mathbf{x}+v)\|}{\eta} + \frac{\tilde{L}_0 n \|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}) - \mathbf{y}(\mathbf{x})\|}{\eta},
\end{aligned}
$$

where in the last inequality we use the definition of $\tilde{L}_0$ in Assumption 1 (a.i). It follows that $\mathbb{E}\left[\left\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega)\right\|^2 \mid \mathbf{x}\right] \leq \frac{4\tilde{L}_0^2 n^2 \tilde{\epsilon}}{\eta^2}$ holds almost surely. □

We make use of the following result in the convergence and rate analysis.

Lemma 4 (Lemma 2.11 in [40]) *Let* $\{\bar{\mathbf{x}}_k\}$ *be generated by Algorithm 1. Let* $\alpha_{k,N} \triangleq \frac{\gamma_k^r}{\sum_{j=0}^N \gamma_j^r}$ *for* $k \in \{0, \ldots, N\}$ *and* $N \geq 0$. *Then, for any* $N \geq 0$, *we have* $\bar{\mathbf{x}}_N = \sum_{k=0}^N \alpha_{k,N} \mathbf{x}_k$. *Furthermore, if* $\mathcal{X}$ *is a convex set, then* $\bar{\mathbf{x}}_N \in \mathcal{X}$.

Remark 5 Lemma 4 allows for representing $\bar{\mathbf{x}}_k$ in Algorithm 1 as a weighted average of the generated iterates $\{\mathbf{x}_k\}$. The term γ_k^r in the last step of ($\texttt{ZSOL}_{\text{cnvx}}^{\textbf{ls}}$) is employed to build the weights $\frac{\gamma_k^r}{\sum_{j=0}^N \gamma_j^r}$ where $0 \leq r < 1$ is a fixed parameter that can be arbitrarily chosen. This averaging scheme was studied earlier [40, 82] and allows for achieving the best convergence rate for SA methods.

We are now in a position to develop rate and complexity statements for Algorithms 1–2. The parameters for both schemes are defined next and the main result is presented by Theorem 1.

Definition 2 (*Parameters for Algorithms* 1–2) Let the stepsize and smoothing sequence in Algorithm 1 be given by $\gamma_k := \frac{\gamma_0}{(k+1)^a}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$, respectively for all $k \geq 0$ where γ_0, η_0, a, and b are strictly positive. In Algorithm 2, suppose $\alpha \leq \frac{\mu_F}{2L_F^2}$, $M_t := \lceil M_0 \rho^{-t} \rceil$ for $t \geq 0$ for some $0 < \rho < 1$ where $M_0 \geq \frac{2v_{\mathbf{y}}^2}{L_F^2}$. Let $t_k := \lceil \tau \ln(k+1) \rceil$ where $\tau \geq \frac{-2(a+b)}{\ln(\max\{1-\mu_F\alpha, \rho\})}$. Finally, suppose $r \in [0, 1)$ is an arbitrary scalar.

Theorem 1 (Rate and complexity statements and almost sure convergence for inexact $\texttt{ZSOL}_{\text{cnvx}}^{\mathbf{1s}}$) *Consider the sequence* $\{\bar{\mathbf{x}}_k\}$ *generated by applying Algorithm* 1 *on* (SMPEC$^{\mathbf{imp},\mathbf{1s}}$). *Suppose Assumptions* 1–4 *hold and algorithm parameters are defined by Definition* 2.
(a) *Suppose* $\hat{\mathbf{x}}_k \in \mathcal{X} + \eta_k \mathbb{S}$ *and let* $\{\mathbf{y}_{t_k}\}$ *be the sequence generated by Algorithm* 2. *Then for suitably defined* $\tilde{d} < 1$ *and* $B(\hat{\mathbf{x}}_k) > 0$, *the following holds for* $t_k \geq 1$.

$$\mathbb{E}[\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq \tilde{\epsilon}_k \triangleq B(\hat{\mathbf{x}}_k)\tilde{d}^{t_k}.$$

(b) *Let* $a = 0.5$ *and* $b \in [0.5, 1)$ *and* $0 \leq r < 2(1-b)$. *Then, for all* $K \geq 2^{\frac{1}{1-r}} - 1$ *we have*

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq (2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{2\theta_0(\hat{\mathbf{x}}_k)\gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}} + (2-r)\left(\frac{\eta_0 L_0}{1-0.5r-b}\right)\frac{1}{(K+1)^b},$$

where $\theta_0(\hat{\mathbf{x}}_k) \triangleq D_{\mathcal{X}} + \frac{(2+3\gamma_0^2)n^2\tilde{L}_0^2 B}{\eta_0^2\gamma_0^2} + 1.5n^2L_0^2$. *In particular, when* $b := 1-\delta$ *and* $r = 0$, *where* $\delta > 0$ *is a small scalar, we have for all* $K \geq 1$

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq 2\left(\frac{D_{\mathcal{X}}}{\gamma_0} + 2\theta_0(\hat{\mathbf{x}}_k)\gamma_0\right)\frac{1}{\sqrt{K+1}} + \left(\frac{2\eta_0 L_0}{\delta}\right)\frac{1}{(K+1)^{1-\delta}}.$$

(c) *Suppose* $\gamma_0 := \mathcal{O}(\frac{1}{L_0})$, $a := 0.5$, $b := 0.5$, *and* $r := 0$. *Let* $\epsilon > 0$ *be an arbitrary scalar and* K_ϵ *be such that* $\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_{K_\epsilon})\right] - f^* \leq \epsilon$. *Then,*
(c-1) *the total number of upper-level projection steps on* $\mathcal{X}$ *is* $K_\epsilon = \mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}\right)$.
(c-2) *the sample complexity of upper-level function evaluations is* $\mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2}\right)$.
(c-3) *the total number of lower-level projection steps on* $\mathcal{Y}$ *is*
$\mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2} \ln\left(n^2 L_0 \tilde{L}_0^2 \epsilon^{-1}\right)\right)$.
(c-4) *the sample complexity of lower-level evaluations of the mapping is*
$\mathcal{O}\left(n^{4\bar{\tau}} L_0^{2\bar{\tau}} \tilde{L}_0^{4\bar{\tau}} \epsilon^{-2\bar{\tau}}\right)$ *where* $\bar{\tau} \geq 1 - \tau \ln(\rho)$.
(d) *For any* $a \in (0.5, 1]$ *and* $b > 1-a$, *there exists* $\mathbf{x}^* \in \mathcal{X}^*$ *such that* $\lim_{k\to\infty} \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 = 0$ *almost surely.*

Proof **(a)** Let the error Δ_t be defined as $\Delta_t \triangleq \bar{F}(\hat{\mathbf{x}}_k, \mathbf{y}_t) - F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$ for $t \geq 0$. Next, we estimate a bound on the term $\mathbb{E}[\|\Delta_t\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t]$. From Assumption 4, we have that

the following holds a.s.

$$\begin{aligned}\mathbb{E}[\|\Delta_t\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] &= \mathbb{E}\left[\left\|\frac{\sum_{\ell=1}^{M_t}(G(\hat{\mathbf{x}}_k,\mathbf{y}_t,\omega_{\ell,t})-F(\hat{\mathbf{x}}_k,\mathbf{y}_t))}{M_t}\right\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t\right] \\ &= \tfrac{1}{M_t^2}\mathbb{E}\left[\sum_{\ell=1}^{M_t}\left\|G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t}) - F(\hat{\mathbf{x}}_k, \mathbf{y}_t)\right\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t\right] \\ &\leq \tfrac{\nu_{\mathbf{y}}^2\|\mathbf{y}_t\|^2+\nu_G^2}{M_t}. \end{aligned} \tag{23}$$

From $\mathbf{y}(\hat{\mathbf{x}}_k) \in \mathrm{SOL}(\mathcal{Y}, F(\hat{\mathbf{x}}_k, \bullet))$, $\mathbf{y}(\hat{\mathbf{x}}_k) = \Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k) - \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k))\right]$ for any $\alpha > 0$. It follows that

$$\begin{aligned}\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 &= \|\Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha\bar{F}(\hat{\mathbf{x}}_k, \mathbf{y}_t)\right] - \Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k) - \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k))\right]\|^2 \\ &\leq \|\mathbf{y}_t - \alpha\bar{F}(\hat{\mathbf{x}}_k, \mathbf{y}_t) - \mathbf{y}(\hat{\mathbf{x}}_k) + \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k))\|^2 \\ &= \|\mathbf{y}_t - \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}_t) - \alpha\Delta_t - \mathbf{y}(\hat{\mathbf{x}}_k) + \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k))\|^2 \\ &= \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha^2\|F(\hat{\mathbf{x}}_k, \mathbf{y}_t) - F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k))\|^2 + \alpha^2\|\Delta_t\|^2 \\ &\quad - 2\alpha(\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T(F(\hat{\mathbf{x}}_k, \mathbf{y}_t) - F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k))) \\ &\quad - 2\alpha(\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k) - \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}_t) + \alpha F(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k)))^T\Delta_t. \end{aligned}$$

Taking conditional expectations in the preceding relation, using (23), and invoking the strong monotonicity and Lipschitzian property of the mapping F in Assumption 1, we obtain

$$\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] \leq \left(1 - 2\mu_F\alpha + \alpha^2L_F^2\right)\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \left(\tfrac{\nu_{\mathbf{y}}^2\|\mathbf{y}_t\|^2+\nu_G^2}{M_t}\right)\alpha^2.$$

Taking expectations on both sides, we obtain

$$\begin{aligned}\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] &\leq \left(1 - 2\mu_F\alpha + \alpha^2L_F^2\right)\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \\ &\quad + \left(\tfrac{\nu_{\mathbf{y}}^2\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}(\hat{\mathbf{x}}_k)+\mathbf{y}(\hat{\mathbf{x}}_k)\|^2]+\nu_G^2}{M_t}\right)\alpha^2 \\ &\leq \left(1 - 2\mu_F\alpha + \alpha^2L_F^2 + \tfrac{2\nu_{\mathbf{y}}^2}{M_0}\alpha^2\right)\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \\ &\quad + \left(\tfrac{2\nu_{\mathbf{y}}^2\|\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\nu_G^2}{M_t}\right)\alpha^2. \end{aligned}$$

Let $\lambda \triangleq 1 - 2\mu_F\alpha + \alpha^2L_F^2 + \frac{2\nu_{\mathbf{y}}^2}{M_0}\alpha^2$ and $\Lambda_t(\hat{\mathbf{x}}_k) \triangleq \frac{2\nu_{\mathbf{y}}^2\|\mathbf{y}(\hat{\mathbf{x}}_k)\|^2+\nu_G^2}{M_t}\alpha^2$ for $t \geq 0$. Note that since $M_0 \geq \frac{2\nu_{\mathbf{y}}^2}{L_F^2}$ and that $\alpha \leq \frac{\mu_F}{2L_F^2}$, we have $\lambda \leq 1 - \mu_F\alpha < 1$. We obtain for

any $t \geq 0$

$$\begin{aligned}\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] &\leq \lambda^{t+1}\|\mathbf{y}_0 - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \sum_{j=0}^{t} \lambda^{t-j}\Lambda_j(\hat{\mathbf{x}}_k) \\ &\leq \lambda^{t+1}\|\mathbf{y}_0 - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \Lambda_0(\hat{\mathbf{x}}_k)(\max\{\lambda, \rho\})^t \sum_{j=0}^{t} \left(\tfrac{\min\{\lambda,\rho\}}{\max\{\lambda,\rho\}}\right)^{t-j} \\ &\leq \lambda^{t+1}\|\mathbf{y}_0 - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \tfrac{\Lambda_0(\hat{\mathbf{x}}_k)(\max\{\lambda,\rho\})^t}{1-(\min\{\lambda,\rho\}/\max\{\lambda,\rho\})} \leq B(\hat{\mathbf{x}}_k)\tilde{d}^{t+1},\end{aligned}$$

where $\tilde{d} \triangleq \max\{\lambda, \rho\}$ and $B(\hat{\mathbf{x}}_k) \triangleq \sup_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y} - \mathbf{y}_0\|^2 + \frac{\Lambda_0(\hat{\mathbf{x}}_k)}{\max\{\lambda,\rho\}-\min\{\lambda,\rho\}}$. Note that in view of compactness of $\mathcal{Y}$, $B(\hat{\mathbf{x}}_k) < \infty$. Also, without loss of generality, we assume that $\rho \neq \lambda$.

(b) Let us define $\bar{F}(\hat{\mathbf{x}}_k, \mathbf{y}_t) \triangleq \frac{\sum_{\ell=1}^{M_t} G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_{\ell,t})}{M_t}$ for $t \geq 0$ and $k \geq 0$. Note that from the compactness of the set $\mathcal{X}$ and the continuity of the implicit function, the set $\mathcal{X}^*$ is nonempty. Let $\mathbf{x}^* \in \mathcal{X}$ be an arbitrary optimal solution. We have that

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \left\|\Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k)\right] - \Pi_{\mathcal{X}}\left[\mathbf{x}^*\right]\right\|^2 \\ &\leq \left\|\mathbf{x}_k - \gamma_k g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k) - \mathbf{x}^*\right\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k) \\ &\quad + \gamma_k^2\|g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k)\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T (g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k) + w_k) \\ &\quad + \gamma_k^2\|g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k)\|^2,\end{aligned}$$

where we define $w_k \triangleq g_{\eta_k,\tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k) - g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k)$. Taking conditional expectations on the both sides, and invoking Lemma 2 and Lemma 3 (a), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathbf{x}_k\right] &\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T \nabla f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k) \\ &\quad - 2\gamma_k\mathbb{E}\left[(\mathbf{x}_k - \mathbf{x}^*)^T w_k \mid \mathbf{x}_k\right] + 3n^2\gamma_k^2\left(\tfrac{2\tilde{L}_0^2\tilde{\epsilon}_k}{\eta_k^2} + L_0^2\right).\end{aligned}$$

Invoking the convexity of $f_{\eta_k}^{\mathbf{imp}}$, bounding $-2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T w_k$, and rearranging the terms, we obtain

$$\begin{aligned}2\gamma_k\left(f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k) - f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}^*)\right) &\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathbf{x}_k\right] \\ &\quad + \gamma_k^2\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \mathbb{E}\left[\|w_k\|^2 \mid \mathbf{x}_k\right] \\ &\quad + 3n^2\gamma_k^2\left(\tfrac{2\tilde{L}_0^2\tilde{\epsilon}_k}{\eta_k^2} + L_0^2\right).\end{aligned}$$

From Lemma 3 (b) we obtain

$$2\gamma_k\left(f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)-f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}^*)\right)\leq\|\mathbf{x}_k-\mathbf{x}^*\|^2-\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\mid\mathbf{x}_k\right]+\gamma_k^2\|\mathbf{x}_k-\mathbf{x}^*\|^2+\frac{4\tilde{L}_0^2n^2\tilde{\epsilon}_k}{\eta_k^2}+3n^2\gamma_k^2\left(\frac{2\tilde{L}_0^2\tilde{\epsilon}_k}{\eta_k^2}+L_0^2\right).$$

From Lemma 1 (v) we have that $f^{\mathbf{imp}}(\mathbf{x}_k)\leq f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)$ and $f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}^*)\leq f^*+\eta_kL_0$. From the preceding inequalities we obtain

$$2\gamma_k\left(f^{\mathbf{imp}}(\mathbf{x}_k)-f^*\right)\leq\|\mathbf{x}_k-\mathbf{x}^*\|^2-\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\mid\mathbf{x}_k\right]+\gamma_k^2\|\mathbf{x}_k-\mathbf{x}^*\|^2+(4+6\gamma_0^2)\frac{\tilde{L}_0^2n^2\tilde{\epsilon}_k}{\eta_k^2}+2\gamma_k\eta_kL_0+3n^2L_0^2\gamma_k^2.$$

Next, we derive a bound on $\frac{\tilde{\epsilon}_k}{\eta_k^2}$. From part (a) and the update rule of η_k we have

$$\frac{\tilde{\epsilon}_k}{\eta_k^2}=\left(\frac{\tilde{\epsilon}_k}{\eta_k^2\gamma_k^2}\right)\gamma_k^2=\left(\frac{(\max\{\lambda,\rho\})^{t_k}B(\hat{\mathbf{x}}_k)(k+1)^{2(a+b)}}{\eta_0^2\gamma_0^2}\right)\gamma_k^2.\tag{24}$$

Note that from $\alpha\leq\frac{\mu_F}{2L_F^2}$ and $M_0\geq\frac{2\nu_{\mathbf{y}}^2}{L_F^2}$, we have $\lambda\leq1-\mu_F\alpha$. Thus, we have $\tau\geq\frac{-2(a+b)}{\ln(\max\{1-\mu_F\alpha,\rho\})}\geq\frac{-2(a+b)}{\ln(\max\{\lambda,\rho\})}$. From $t_k:=\lceil\tau\ln(k+1)\rceil\geq\tau\ln(k+1)$ and $\tau\geq\frac{-2(a+b)}{\ln(\max\{\lambda,\rho\})}$ we have that

$$\begin{aligned}(\max\{\lambda,\rho\})^{t_k}(k+1)^{2(a+b)}&\leq\left((\max\{\lambda,\rho\})^{\tau}\,\mathrm{e}^{2(a+b)}\right)^{\ln(k+1)}\\&\leq(\max\{\lambda,\rho\})^{\tau}\,\mathrm{e}^{2(a+b)}\leq1.\end{aligned}$$

This relation and (24) imply that $\frac{\tilde{\epsilon}_k}{\eta_k^2}\leq\left(\frac{B(\hat{\mathbf{x}}_k)}{\eta_0^2\gamma_0^2}\right)\gamma_k^2$. Also, note that since $\mathcal{X}$ is bounded, there exists a scalar $D_{\mathcal{X}}\triangleq\frac{1}{2}\sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}-\mathbf{x}^*\|^2$ such that $D_{\mathcal{X}}<\infty$. Therefore, we obtain

$$2\gamma_k\left(f^{\mathbf{imp}}(\mathbf{x}_k)-f^*\right)\leq\|\mathbf{x}_k-\mathbf{x}^*\|^2-\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\mid\mathbf{x}_k\right]+2\gamma_k^2\theta_0(\hat{\mathbf{x}}_k)+2\gamma_k\eta_kL_0,\tag{25}$$

where $\theta_0(\hat{\mathbf{x}}_k)\triangleq D_{\mathcal{X}}+\frac{(2+3\gamma_0^2)n^2\tilde{L}_0^2B(\hat{\mathbf{x}}_k)}{\eta_0^2\gamma_0^2}+1.5n^2L_0^2<\infty$. Taking expectations on the both sides and multiplying both sides by $\frac{\gamma_k^{r-1}}{2}$, we have that

$$\gamma_k^r\left(\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_k)\right]-f^*\right)\leq\frac{\gamma_k^{r-1}}{2}\left(\mathbb{E}\left[\|\mathbf{x}_k-\mathbf{x}^*\|^2\right]-\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2\right]\right)$$

$$+ \gamma_k^{1+r} \theta_0(\hat{\mathbf{x}}_k) + \gamma_k^r \eta_k L_0. \tag{26}$$

Adding and subtracting the term $\frac{\gamma_{k-1}^{r-1}}{2} \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right]$, we obtain

$$\begin{aligned}
&\gamma_k^r \left(\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_k) \right] - f^* \right) \\
&\leq \frac{\gamma_{k-1}^{r-1}}{2} \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \frac{\gamma_k^{r-1}}{2} \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] \\
&\quad + \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}\right) D_{\mathcal{X}} + \theta_0(\hat{\mathbf{x}}_k) \gamma_k^{1+r} + \gamma_k^r \eta_k L_0.
\end{aligned}$$

Summing the inequality from $k = 1, \ldots, K$ we obtain

$$\begin{aligned}
\sum_{k=1}^{K} \gamma_k^r \left(\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_k) \right] - f^* \right) &\leq \frac{\gamma_0^{r-1}}{2} \mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}^*\|^2\right] + \left(\gamma_K^{r-1} - \gamma_0^{r-1}\right) D_{\mathcal{X}} \\
&\quad + \theta_0(\hat{\mathbf{x}}_k) \sum_{k=1}^{K} \gamma_k^{1+r} + L_0 \sum_{k=1}^{K} \gamma_k^r \eta_k.
\end{aligned}$$

Rewriting (26) when $k := 0$, we obtain

$$\begin{aligned}
\gamma_0^r \left(\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_0) \right] - f^* \right) &\leq \frac{\gamma_0^{r-1}}{2} \left(\mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}^*\|^2\right] \right) \\
&\quad + \theta_0(\hat{\mathbf{x}}_k) \gamma_0^{1+r} + \gamma_0^r \eta_0 L_0.
\end{aligned}$$

Adding the preceding two relations together and using the definition of $D_{\mathcal{X}}$, we obtain

$$\sum_{k=0}^{K} \gamma_k^r \left(\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_k) \right] - f^* \right) \leq D_{\mathcal{X}} \gamma_K^{r-1} + \theta_0(\hat{\mathbf{x}}_k) \sum_{k=0}^{K} \gamma_k^{1+r} + L_0 \sum_{k=0}^{K} \gamma_k^r \eta_k.$$

From the definition of $\bar{\mathbf{x}}_K \triangleq \sum_{k=0}^{K} \alpha_{k,K} \mathbf{x}_k$ (see Lemma 4) and by applying the convexity of the implicit function, we have for all $K \geq 2^{\frac{1}{1-r}} - 1$

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K) \right] - f^* \leq \frac{D_{\mathcal{X}} \gamma_K^{r-1} + \theta_0(\hat{\mathbf{x}}_k) \sum_{k=0}^{K} \gamma_k^{1+r} + L_0 \sum_{k=0}^{K} \gamma_k^r \eta_k}{\sum_{k=0}^{K} \gamma_k^r}.$$

Substituting $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$, we obtain the following by invoking Lemma 13

$$\begin{aligned}
&\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K) \right] - f^* \\
&\leq \frac{D_{\mathcal{X}} \gamma_0^{r-1} (K+1)^{0.5(1-r)} + \theta_0(\hat{\mathbf{x}}_k) \gamma_0^{1+r} \frac{(K+1)^{1-0.5(1+r)}}{1-0.5(1+r)} + \gamma_0^r \eta_0 L_0 \frac{(K+1)^{1-0.5r-b}}{1-0.5r-b}}{\gamma_0^r \frac{(K+1)^{1-0.5r}}{2-r}}
\end{aligned}$$

$$\leq (2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0}+\frac{2\theta_0(\hat{\mathbf{x}}_k)\gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}}+(2-r)\left(\frac{\eta_0 L_0}{1-0.5r-b}\right)\frac{1}{(K+1)^b}.$$

(c) The results in (c-1) and (c-2) follow directly from part (b) by substituting γ_0 and r. To show part (c-3), note that in Algorithm 1, we have $t_k := \lceil \tau \ln(k+1)\rceil$. From part (b), we require that the total number of iterations of the SA scheme is bounded as follows.

$$\begin{aligned}
2\sum_{k=0}^{K_\epsilon} t_k &= 2\sum_{k=0}^{K_\epsilon}\lceil \tau \ln(k+1)\rceil \leq 2\sum_{k=0}^{K_\epsilon}(1+\tau\ln(k+1))\\
&\leq 2(K_\epsilon+1)+2\tau\ln(1)+2\tau\sum_{k=1}^{K_\epsilon}\ln(k+1)\\
&\leq 2(K_\epsilon+1)+2\tau\int_2^{K_\epsilon+1}\ln(u)du \leq 2(K_\epsilon+1)+2\tau(K_\epsilon+2)\ln(K_\epsilon+2)\\
&\leq 4\max\{\tau,1\}(K_\epsilon+2)\ln(K_\epsilon+2).
\end{aligned}$$

The bound in (c-3) follows from the preceding inequality and the bound on K_ϵ in (c-1). To show (c-4), note that the total samples used in the lower-level is as follows.

$$\begin{aligned}
2\sum_{k=0}^{K_\epsilon}\sum_{t=0}^{t_k} M_t &= 2\sum_{k=0}^{K_\epsilon}\sum_{t=0}^{t_k}\lceil M_0\rho^{-t}\rceil \leq 4M_0\sum_{k=0}^{K_\epsilon}\sum_{t=0}^{t_k}\rho^{-t}\\
&= \mathcal{O}\left(\sum_{k=0}^{K_\epsilon}\frac{\rho^{-t_k}}{\ln(\frac{1}{\rho})}\right)=\mathcal{O}\left(\sum_{k=0}^{K_\epsilon}\frac{\rho^{-\tau\ln(k+1)}}{\ln(\frac{1}{\rho})}\right)\\
&\leq \mathcal{O}\left(\sum_{k=0}^{K_\epsilon}\frac{e^{(\bar{\tau}-1)\ln(k+1)}}{\ln(\frac{1}{\rho})}\right)=\mathcal{O}\left(\sum_{k=0}^{K_\epsilon}\frac{(k+1)^{\bar{\tau}-1}}{\ln(\frac{1}{\rho})}\right)\leq\mathcal{O}\left(\frac{K_\epsilon^{\bar{\tau}}}{\ln(\frac{1}{\rho})}\right),
\end{aligned}$$

where $\bar{\tau}\geq 1+\tau\ln(\frac{1}{\rho})$. The bound in (c-4) follows from the preceding inequality and the bound on K_ϵ in (c-1).

(d) Consider relation (25). Rearranging the terms, for all $k\geq 0$ we have

$$\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 \mid \mathbf{x}_k\right]\leq \|\mathbf{x}_k-\mathbf{x}^*\|^2-2\gamma_k\left(f^{\mathbf{imp}}(\mathbf{x}_k)-f^*\right)+2\gamma_k^2\theta_0(\hat{\mathbf{x}}_k)+2\gamma_k\eta_k L_0.$$

Note that $\sum_{k=0}^{\infty}\gamma_k^2<\infty$ and $\sum_{k=0}^{\infty}\gamma_k\eta_k<\infty$ since $b>0.5$. Thus, in view of Lemma 15, we have that $\{\|\mathbf{x}_k-\mathbf{x}^*\|^2\}$ is a convergent sequence in an almost sure sense and $\sum_{k=0}^{\infty}\gamma_k(f^{\mathbf{imp}}(\mathbf{x}_k)-f^*)<\infty$ almost surely. The former statement implies that $\{\mathbf{x}_k\}$ is a bounded sequence in an a.s. sense. Further, the latter statement and $\sum_{k=0}^{\infty}\gamma_k=\infty$ imply that $\liminf_{k\to\infty} f^{\mathbf{imp}}(\mathbf{x}_k)=f^*$ in an a.s. sense. Thus, from the continuity of the implicit function, there is a subsequence of $\{\mathbf{x}_k\}_{k\in\mathcal{K}}$ with a limit point denoted by $\hat{\mathbf{x}}$ such that $\hat{\mathbf{x}}\in\mathcal{X}^*$. Since $\{\|\mathbf{x}_k-\mathbf{x}^*\|^2\}$ is a convergent sequence for

all $\mathbf{x}^* \in \mathcal{X}^*$, we have $\{\|\mathbf{x}_k - \hat{\mathbf{x}}\|^2\}$ is a convergent sequence. But we have shown that $\lim_{k\to\infty,\, k\in\mathcal{K}} \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2 = 0$ almost surely. Hence $\lim_{k\to\infty} \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2 = 0$ almost surely where $\hat{\mathbf{x}} \in \mathcal{X}^*$. Next, we show that $\lim_{k\to\infty} \|\bar{\mathbf{x}}_k - \hat{\mathbf{x}}\|^2 = 0$. In view of Lemmas 4 and 14, it suffices to have $\sum_{k=0}^{\infty} \gamma_k^r = \infty$ or equivalently, we must have $ar \leq 1$. This is already satisfied as a consequence of $a \in (0.5, 1]$ and $r \in [0, 1)$. □

Remark 6 (*Variance-reduction schemes*)

(i) In Algorithm 2 we employ a variance-reduced (VR) scheme in computing an ϵ-solution of the parametrized VI at the lower-level. This is crucial since it allows for computing an ϵ-solution in $\ln(1/\epsilon)$ steps while in a non-VR regime, it would have taken $\mathcal{O}(1/\epsilon)$ steps. Variance-reduction on strongly monotone VIs has been studied in [13, 33, 34], amongst others.

(ii) In addressing single-stage SMPECs, while employing a VR scheme in either lower-level or upper-level is possible, but sometimes this approach may not be advisable to be adopted at the both levels simultaneously. For instance, in ($\texttt{ZSOL}_{\text{cnvx}}^{\mathbf{1s}}$), employing a VR scheme in the upper-level would lead to requiring an increasing number of inexact solutions of a lower-level stochastic VI at each iteration, where each of these solutions would require a VR scheme to be employed in the lower-level. Consequently, this may render the scheme impractical.

Remark 7 (*Definition of history*) We conclude this subsection with a brief remark regarding the formal definition of the $\sigma-$algebra for Algorithms 1–2. First, $\mathcal{F}_{0,0} \triangleq \{\mathbf{x}_0\}$. In addition, $\mathcal{F}_{k,0}$ is defined as

$$\begin{aligned}
&\mathcal{F}_{1,0} = \mathcal{F}_{0,0} \cup \{\omega_0, v_0\} \cup \mathcal{F}_{0,t_0}^1 \cup \mathcal{F}_{0,t_0}^2, \quad \text{where} \\
&\mathcal{F}_{0,t}^1 \triangleq \left\{ \left\{ G(\mathbf{x}_0, \mathbf{y}_0, \omega_{\ell,0}) \right\}_{\ell=1}^{M_0}, \ldots, \left\{ G(\mathbf{x}_0, \mathbf{y}_{t-1}, \omega_{\ell,t-1}) \right\}_{\ell=1}^{M_0} \right\} \quad \text{and} \\
&\mathcal{F}_{0,t}^2 \triangleq \left\{ \left\{ G(\mathbf{x}_0 + v_0, \mathbf{y}_0, \omega_{\ell,0}) \right\}_{\ell=1}^{M_0}, \ldots, \left\{ G(\mathbf{x}_0 + v_0, \mathbf{y}_{t-1}, \omega_{\ell,t-1}) \right\}_{\ell=1}^{M_0} \right\} \\
&\qquad \text{for} \quad t = 0, \ldots, t_0 - 1.
\end{aligned}$$

At the kth iteration with $k > 0$, we have that

$$\begin{aligned}
&\mathcal{F}_{k,0} = \mathcal{F}_{k-1,0} \cup \{\omega_k, v_k\} \cup \mathcal{F}_{k,t_k}^1 \cup \mathcal{F}_{k,t_k}^2, \quad \text{where} \\
&\mathcal{F}_{k,t}^1 \triangleq \left\{ \left\{ G(\mathbf{x}_k, \mathbf{y}_0, \omega_{\ell,0}) \right\}_{\ell=1}^{M_t}, \ldots, \left\{ G(\mathbf{x}_k, \mathbf{y}_{t-1}, \omega_{\ell,t-1}) \right\}_{\ell=1}^{M_t} \right\} \quad \text{and} \\
&\mathcal{F}_{k,t}^2 \triangleq \left\{ \left\{ G(\mathbf{x}_k + v_k, \mathbf{y}_0, \omega_{\ell,t}) \right\}_{\ell=1}^{M_t}, \ldots, \left\{ G(\mathbf{x}_k + v_k, \mathbf{y}_{t-1}, \omega_{\ell,t-1}) \right\}_{\ell=1}^{M_t} \right\} \\
&\qquad \text{for} \quad t = 0, \ldots, t_k - 1.
\end{aligned}$$

In particular, at the tth iteration of the SA scheme at the kth upper-level step, we may define $\mathcal{F}_{k,t}$ as

$$\begin{aligned}
&\mathcal{F}_{k,t} \triangleq \mathcal{F}_{k,0} \cup \left\{ \left\{ G(\hat{\mathbf{x}}_k, \mathbf{y}_0, \omega_{\ell,0}) \right\}_{\ell=1}^{M_0}, \ldots, \left\{ G(\hat{\mathbf{x}}_k, \mathbf{y}_{t-1}, \omega_{\ell,t-1}) \right\}_{\ell=1}^{M_{t-1}} \right\}, \\
&\qquad \text{for} \quad t = 0, \ldots, t_k - 1.
\end{aligned}$$

Furthermore, at the tth step of the lower-level SA scheme associated with the kth iteration, the history is denoted by $\mathcal{F}^1_{k-1,t}$ and $\mathcal{F}^2_{k-1,t}$, defined as

$$\mathcal{F}^1_{k-1,t} \triangleq \mathcal{F}_{k-1,0} \cup \{v_k, \omega_k\} \cup \mathcal{F}^1_{k,t-1} \quad \text{and} \quad \mathcal{F}^2_{k-1,t} \triangleq \mathcal{F}_{k-1,0} \cup \{v_k, \omega_k\} \cup \mathcal{F}^2_{k,t-1}.$$

Naturally, one can employ these histories in constructing the conditional expectations; specifically, at the kth iteration, we may use $\mathcal{F}_{k-1,0}$ while at the tth step of the lower-level SA scheme at the kth iteration, we may use $\mathcal{F}_{k-1,t-1}$. For expository ease, we use the iterate as a proxy in constructing the conditional expectation, as the reader will observe. Note that for expository ease, we employ $\mathbf{y}_t$ at iteration k as a proxy for the history (rather than $\mathbf{y}_{k,t}$).

3.2.2 An exact zeroth-order scheme

In this subsection, we consider the case where an exact solution of the lower-level problem is available. This case is particularly relevant when the lower-level problem is a deterministic variational inequality problem and highly accurate solutions are available. We develop a zeroth-order method where the gradient mapping is approximated using two evaluations of the implicit function. Similar to the inexact setting, we allow for iterative smoothing and provide the convergence analysis in addressing the original implicit problem. In the following, we derive non-asymptotic convergence rate statements and also, show an almost sure convergence result for the proposed zeroth-order method in the exact regimes.

Corollary 1 (Rate and complexity statements and a.s. convergence for exact ($\mathtt{ZSOL}^{\mathbf{1s}}_{\mathrm{cnvx}}$)) *Consider the problem* (SMPEC$^{\mathbf{imp},\mathbf{1s}}$). *Suppose Assumptions 1–3 hold. Let* $\{\bar{\mathbf{x}}_k\}$ *denote the sequence generated by Algorithm 1 (exact variant) in which the stepsize and smoothing sequences are defined as* $\gamma_k := \frac{\gamma_0}{(k+1)^a}$ *and* $\eta_k := \frac{\eta_0}{(k+1)^b}$, *respectively, for all* $k \geq 0$ *where* γ_0 *and* η_0 *are strictly positive. Then, the following statements hold.*

(a) *Let* $a = 0.5$ *and* $b \in [0.5, 1)$ *and* $0 \leq r < 2(1-b)$. *Then, for all* $K \geq 2^{\frac{1}{1-r}} - 1$ *we have*

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq (2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{L_0^2 n^2 \gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}} + (2-r)\left(\frac{\eta_0 L_0}{1-0.5r-b}\right)\frac{1}{(K+1)^b}.$$

In particular, when $b := 1-\delta$ *and* $r = 0$, *where* $\delta > 0$ *is a small scalar, we have for all* $K \geq 1$

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq 2\left(\frac{D_{\mathcal{X}}}{\gamma_0} + L_0^2 n^2 \gamma_0\right)\frac{1}{\sqrt{K+1}} + \left(\frac{2\eta_0 L_0}{\delta}\right)\frac{1}{(K+1)^{1-\delta}}.$$

(b) *Let* $a := 0.5$, $b = 0.5$, $r = 0$, $\gamma_0 := \frac{\sqrt{D_{\mathcal{X}}}}{nL_0}$, *and* $\eta_0 \leq \sqrt{D_{\mathcal{X}}}n$. *Then, the iteration complexity in projection steps on* $\mathcal{X}$ *as well as the total sample complexity of*

upper-level evaluations, for achieving $\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_{K_\epsilon})\right] - f^* \leq \epsilon$ *for some* $\epsilon > 0$ *is given by* K_ϵ *where* K_ϵ *is bounded as follows.*

$$K_\epsilon \geq \frac{64n^2 L_0^2 D_{\mathcal{X}}}{\epsilon^2}.$$

(c) *For any* $a \in (0.5, 1]$ *and* $b > 1 - a$*, there exists* $\mathbf{x}^* \in \mathcal{X}^*$ *such that* $\lim_{k\to\infty} \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 = 0$ *almost surely.*

Proof **(a)** Let $\mathbf{x}^* \in \mathcal{X}^*$ be an arbitrary optimal solution. We may expand $\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2$ as

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \left\|\Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k)\right] - \Pi_{\mathcal{X}}\left[\mathbf{x}^*\right]\right\|^2 \\ &\leq \left\|\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k) - \mathbf{x}^*\right\|^2 \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k) \\ &\quad + \gamma_k^2 \|g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k)\|^2.\end{aligned}$$

Taking conditional expectations on the both sides and invoking Lemma 2, we obtain

$$\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathbf{x}_k\right] \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k(\mathbf{x}_k - \mathbf{x}^*)^T \nabla f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k) + \gamma_k^2 L_0^2 n^2.$$

Invoking the convexity of f_{η_k}, we obtain

$$\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \mid \mathbf{x}_k\right] \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma_k\left(f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k) - f_{\eta_k}^{\mathbf{imp}}(x^*)\right) + \gamma_k^2 L_0^2 n^2. \tag{27}$$

Taking expectations from both sides of the preceding relation and rearranging the terms, we obtain

$$\begin{aligned}2\gamma_k\left(\mathbb{E}\left[f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)\right] - f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}^*)\right) \leq \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] \\ + \gamma_k^2 L_0^2 n^2.\end{aligned}$$

From the Lipschitzian property of the implicit function and Lemma 1 (v), we have that

$$f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}^*) \leq f^* + \eta_k L_0. \tag{28}$$

From the preceding two inequalities and that $f^{\mathbf{imp}}(\mathbf{x}_k) \leq f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)$, we obtain

$$\begin{aligned}2\gamma_k\left(\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_k)\right] - f^*\right) \leq \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right] + \gamma_k^2 L_0^2 n^2 \\ + 2\gamma_k \eta_k L_0.\end{aligned}$$

The rest of the proof follows in a similar fashion to that of Theorem 1 (b).

(b) Under the specified setting, from part (a) we have

$$\begin{aligned}\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* &\leq 2\left(\tfrac{D_{\mathcal{X}}}{\gamma_0} + L_0^2 n^2 \gamma_0\right)\tfrac{1}{\sqrt{K+1}} + \left(\tfrac{2\eta_0 L_0}{0.5}\right)\tfrac{1}{\sqrt{K+1}} \\ &= 2(nL_0\sqrt{D_{\mathcal{X}}} + nL_0\sqrt{D_{\mathcal{X}}})\tfrac{1}{\sqrt{K+1}} + \left(4nL_0\sqrt{D_{\mathcal{X}}}\right)\tfrac{1}{\sqrt{K+1}} \\ &= \tfrac{8nL_0\sqrt{D_{\mathcal{X}}}}{\sqrt{K+1}} \leq \epsilon.\end{aligned}$$

This implies the desired bound.

(c) The proof follows in a similar vein to that of Theorem 1 (d).

□

3.3 Nonconvex single-stage SMPEC

In this subsection, in addressing (SMPEC$^{\mathbf{imp,1s}}$) in the nonconvex case, we consider a smoothed implicit problem given by the following.

$$\begin{aligned}\min \quad & f_\eta^{\mathbf{imp}}(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X},\end{aligned} \tag{29}$$

where $f_\eta^{\mathbf{imp}}$ is defined by (G-Smooth$^{\mathbf{1s}}$) for a given $\eta > 0$.

3.3.1 An inexact zeroth-order scheme

In this subsection, we consider the case where an exact solution of the lower-level problem is unavailable. The outline of the proposed zeroth-order scheme is given by Algorithms 3–4. We make the following assumptions in these algorithms.

Assumption 5 Consider Algorithm 3. Given a mini-batch size of N_k and a smoothing parameter $\eta > 0$, let $\{v_{j,k}\}_{j=1}^{N_k} \in \mathbb{R}^n$ be N_k iid replicates generated at epoch k from the uniform distribution on $\eta\mathbb{S}$ for all $k \geq 0$. Also, let the random realizations $\{\omega_{j,k}\}_{j=1}^{N_k}$ be iid replicates.

Assumption 6 Consider Algorithm 4. Let the following hold and for all $k \geq 0$, $t \geq 0$, $\hat{\mathbf{x}}_k \in \mathcal{X} + \eta_k\mathbb{S}$, and $\mathbf{y}_t \in \mathcal{Y}$.

(a) The replicates $\{G(\bullet, \bullet, \omega_t)\}_{t=0}^{\infty}$ are generated randomly and are iid.
(b) $\mathbb{E}[G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t) \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] = F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$ holds almost surely.
(c) $\mathbb{E}[\|G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t) - F(\hat{\mathbf{x}}_k, \mathbf{y}_t)\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] \leq \nu_G^2$ holds almost surely for some $\nu_G > 0$.

Assumption 6 provides standard conditions on the first and second moment of the stochastic oracle. Such conditions have been assumed in the literature of the SA schemes extensively (e.g., see [55, 81]). We utilize the following definition and lemma in the analysis in this subsection.

Definition 3 (*The residual mappings*) Suppose Assumption 1 holds. Given a scalar $\beta > 0$ and a smoothing parameter $\eta > 0$, for any $\mathbf{x} \in \mathbb{R}^n$, let the residual mapping $G_{\eta,\beta}$ and its error-afflicted counterpart $\tilde{G}_{\eta,\beta}$ be defined as

$$G_{\eta,\beta}(\mathbf{x}) \triangleq \beta\Big(\mathbf{x} - \Pi_{\mathcal{X}}\Big[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x})\Big]\Big), \tag{30}$$

$$\tilde{G}_{\eta,\beta}(\mathbf{x}) \triangleq \beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}) + \tilde{e})\right]\right), \tag{31}$$

where $\tilde{e} \in \mathbb{R}^n$ is an arbitrary given vector.

It may be observed that $G_{\eta,\beta}$ is a residual for stationarity for the minimization of smooth nonconvex objectives over convex sets (cf. [6]). In fact, the first part of (32) is a consequence of the well known result relating the residual function $G_{\eta,\beta}(\mathbf{x})$ to the standard stationarity condition (cf. [5, Thm. 9.10]) while the second implication in (32) is Prop. 4.

Lemma 5 *Consider the problem* (29). *Then the following holds for any* $\eta, \beta > 0$.

$$\begin{aligned}\big[G_{\eta,\beta}(\mathbf{x}) = 0\big] &\iff \Big[0 \in \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})\Big] \\ &\implies \Big[0 \in \partial_{2\eta} f^{\mathbf{imp}}(\mathbf{x}) + \mathcal{N}_{\mathcal{X}}(\mathbf{x})\Big].\end{aligned} \tag{32}$$

Consequently, a zero of the residual of the η-smoothed problem satisfies an η-approximate stationarity property for the original problem. The residual $\tilde{G}_{\eta,\beta}$ represents the counterpart of $G_{\eta,\beta}$ when employing an error-afflicted estimate of the gradient. In fact, since our framework relies on sampling, leading to error, we obtain bounds on $\tilde{G}_{\eta,\beta}$. But it is still necessary to derive bounds on the original residual $G_{\eta,\beta}$ but this can be provided in terms of $\tilde{G}_{\eta,\beta}$ and $\tilde{e}$, the error in the gradient.

Lemma 6 *Let Assumption 1 hold. Then the following holds for any* $\beta, \eta > 0$, *and* $\mathbf{x} \in \mathbb{R}^n$.

$$\|G_{\eta,\beta}(\mathbf{x})\|^2 \leq 2\|\tilde{G}_{\eta,\beta}(\mathbf{x})\|^2 + 2\|\tilde{e}\|^2.$$

Proof From Definition 3, we may bound $G_{\eta,\beta}(\mathbf{x})$ as follows.

$$\begin{aligned}\|G_{\eta,\beta}(\mathbf{x})\|^2 &= \left\|\beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x})\right]\right)\right\|^2 \\ &= \Big\|\beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}) + \tilde{e})\right]\right) \\ &\quad + \beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}) + \tilde{e})\right] - \beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x})\right]\Big\|^2 \\ &\leq 2\left\|\beta\left(\mathbf{x} - \Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}) + \tilde{e})\right]\right)\right\|^2 \\ &\quad + 2\left\|\beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}(\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}) + \tilde{e})\right] - \beta\Pi_{\mathcal{X}}\left[\mathbf{x} - \tfrac{1}{\beta}\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x})\right]\right\|^2 \\ &\leq 2\|\tilde{G}_{\eta,\beta}(\mathbf{x})\|^2 + 2\|\tilde{e}\|^2,\end{aligned}$$

where the last inequality is a consequence of the non-expansivity of the Euclidean projector. □

The proposed scheme can be compactly represented as

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma\left(\nabla_{\mathbf{x}} f_{\eta}^{\mathbf{imp}}(\mathbf{x}_k) + e_k\right)\right], \tag{33}$$

where $k \geq 0$ and we define the stochastic errors $e_k \triangleq g_{\eta,N_k,\tilde{\epsilon}_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f_{\eta}^{\mathbf{imp}}(\mathbf{x}_k)$ for $k \geq 0$. We make use of the following result in the convergence analysis.

Algorithm 3 $\texttt{ZSOL}_{\text{ncvx}}^{\mathbf{1s}}$: Variance-reduced zeroth-order method for nonconvex (SMPEC$^{\mathbf{1s}}$)

input: Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize $\gamma > 0$, smoothing parameter $\eta > 0$, mini-batch sequence $\{N_k\}$ such that $N_k := k+1$, an integer K, a scalar $\lambda \in (0,1)$, and an integer R randomly selected from $\{\lceil \lambda K \rceil, \ldots, K\}$ using a uniform distribution

for $k = 0, 1, \ldots, K-1$ **do**

Do one of the following, depending on the type of the scheme.

- Inexact scheme: Call Algorithm 4 to obtain $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k)$
- Exact scheme: Evaluate $\mathbf{y}(\mathbf{x}_k)$

for $j = 1, \ldots, N_k$ **do**

Generate $v_{j,k} \in \eta\mathbb{S}$

Do one of the following.

- Inexact scheme: Call Algorithm 4 to obtain $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k + v_{j,k})$
- Exact scheme: Evaluate $\mathbf{y}(\mathbf{x}_k + v_{j,k})$

Evaluate the inexact or exact zeroth-order gradient approximation as follows.

$$g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) := \frac{n\left(\tilde{f}(\mathbf{x}_k+v_{j,k},\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k+v_{j,k}),\omega_{j,k}) - \tilde{f}(\mathbf{x}_k,\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k),\omega_{j,k})\right)v_{j,k}}{\|v_{j,k}\|\eta} \quad \text{(Inexact)}$$

$$g_{\eta}(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) := \frac{n\left(\tilde{f}(\mathbf{x}_k+v_{j,k},\mathbf{y}(\mathbf{x}_k+v_{j,k}),\omega_{j,k}) - \tilde{f}(\mathbf{x}_k,\mathbf{y}(\mathbf{x}_k),\omega_{j,k})\right)v_{j,k}}{\|v_{j,k}\|\eta} \quad \text{(Exact)}$$

end for

Evaluate the mini-batch zeroth-order gradient.

$$g_{\eta,N_k,\tilde{\epsilon}_k}(\mathbf{x}_k) := \frac{\sum_{j=1}^{N_k} g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k} \quad \text{(Inexact)}$$

$$g_{\eta,N_k}(\mathbf{x}_k) := \frac{\sum_{j=1}^{N_k} g_{\eta}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k} \quad \text{(Exact)}$$

Update $\mathbf{x}_k$ as follows.

$$\mathbf{x}_{k+1} := \begin{cases} \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma g_{\eta,N_k,\tilde{\epsilon}_k}(\mathbf{x}_k)\right] & \text{(Inexact)} \\ \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma g_{\eta,N_k}(\mathbf{x}_k)\right] & \text{(Exact)} \end{cases}$$

end for

Return $\mathbf{x}_R$

Algorithm 4 SA method for lower-level of nonconvex (SMPEC[1s])

1: **input:** An arbitrary $\mathbf{y}_0 \in \mathcal{Y}$, vector $\hat{\mathbf{x}}_k$, and initial stepsize $\alpha_0 > \frac{1}{2\mu_F}$
2: Set $t_k := k+1$
3: **for** $t = 0, 1, \ldots, t_k - 1$ **do**
4: Generate a random realization of the stochastic mapping $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t)$
5: Update $\mathbf{y}_t$ as follows. $\mathbf{y}_{t+1} := \Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha_t G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t)\right]$
6: Update the stepsize using $\alpha_{t+1} := \frac{\alpha}{t+\Gamma}$
7: **end for**
8: Return $\mathbf{y}_{t_k}$

Lemma 7 *Let Assumption 1 hold. Suppose $\mathbf{x}_k$ is generated by Algorithm 3 in which $\gamma \in (0, \frac{\eta}{nL_0})$ for a given $\eta > 0$. Then, we have for any k,*

$$f_\eta^{\mathbf{imp}}(\mathbf{x}_{k+1}) \le f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + \left(-1 + \tfrac{nL_0\gamma}{\eta}\right)\tfrac{\gamma}{4}\|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma\|e_k\|^2.$$

Proof Note that by Lemma 1 (iv), $\nabla f_\eta^{\mathbf{imp}}(\bullet)$ is Lipschitz with parameter $L \triangleq \frac{nL_0}{\eta}$. By the descent lemma, we have that

$$\begin{aligned} f_\eta^{\mathbf{imp}}(\mathbf{x}_{k+1}) &\le f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)^T(\mathbf{x}_{k+1} - \mathbf{x}_k) + \tfrac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + \left(\nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + e_k\right)^T(\mathbf{x}_{k+1} - \mathbf{x}_k) \\ &\quad - e_k^T(\mathbf{x}_{k+1} - \mathbf{x}_k) + \tfrac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \end{aligned}$$

From the properties of the Euclidean projection, we have that

$$\begin{aligned} &(\mathbf{x}_k - \gamma(\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + e_k)) - \mathbf{x}_{k+1})^T(\mathbf{x}_k - \mathbf{x}_{k+1}) \le 0 \\ &\Longrightarrow (\nabla_x f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + e_k))^T(\mathbf{x}_{k+1} - \mathbf{x}_k) \le -\tfrac{1}{\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \end{aligned}$$

In addition, for any $u, v \in \mathbb{R}^n$ we can write $u^T v \le \frac{1}{2}\left(\gamma\|u\|^2 + \frac{\|v\|^2}{\gamma}\right)$. Thus, we have that

$$-e_k^T(\mathbf{x}_{k+1} - \mathbf{x}_k) \le \tfrac{\gamma}{2}\|e_k\|^2 + \tfrac{1}{2\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.$$

Consequently, from the preceding three inequalities we have that

$$\begin{aligned} f_\eta^{\mathbf{imp}}(\mathbf{x}_{k+1}) &\le f_\eta^{\mathbf{imp}}(\mathbf{x}_k) - \tfrac{1}{\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 + \tfrac{1}{2\gamma}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad + \tfrac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + \left(-\tfrac{1}{2\gamma} + \tfrac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \tfrac{\gamma}{2}\|e_k\|^2. \end{aligned}$$

From $\gamma < \frac{1}{L}$, we have

$$f_\eta^{\mathbf{imp}}(\mathbf{x}_{k+1}) \le f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + \left(-\tfrac{1}{2\gamma} + \tfrac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 + \tfrac{\gamma}{2}\|e_k\|^2$$

$$
\begin{aligned}
&= f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + \left(-\tfrac{1}{2\gamma} + \tfrac{L}{2}\right)\gamma^2\|\tilde{G}_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\
&= f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + (-1+L\gamma)\,\tfrac{\gamma}{2}\|\tilde{G}_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\
&\overset{\text{Lemma 6}}{\leq} f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + (-1+L\gamma)\,\tfrac{\gamma}{4}\|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 \\
&\quad + (1-L\gamma)\,\tfrac{\gamma}{2}\|e_k\|^2 + \tfrac{\gamma}{2}\|e_k\|^2 \\
&= f_\eta^{\mathbf{imp}}(\mathbf{x}_k) + (-1+L\gamma)\,\tfrac{\gamma}{4}\|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 + \left(1-\tfrac{L\gamma}{2}\right)\gamma\|e_k\|^2.
\end{aligned}
$$

Substituting $L := \frac{nL_0}{\eta}$ we obtain the desired inequality. □

We make use of the following result in the convergence analysis.

Lemma 8 *Let $\{e_k\}$ be a non-negative sequence such that for an arbitrary non-negative sequence $\{\gamma_k\}$, the following relation is satisfied.*

$$e_{k+1} \leq (1-\alpha\gamma_k)e_k + \beta\gamma_k^2, \quad \textit{for all} \quad k \geq 0. \tag{34}$$

where α and β are positive scalars. Suppose $\gamma_k = \frac{\gamma}{k+\Gamma}$ for any $k \geq 0$, where $\gamma > \frac{1}{\alpha}$ and $\Gamma > 0$. Then, we have

$$e_k \leq \frac{\max\left\{\frac{\beta\gamma^2}{\alpha\gamma-1}, \Gamma e_0\right\}}{k+\Gamma}, \quad \textit{for all} \quad k \geq 0. \tag{35}$$

Next, we present the rate and complexity result for the proposed inexact method for addressing the nonconvex case.

Theorem 2 (Rate and complexity statements for inexact ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{ncvx}}$)) *Consider Algorithms 3–4 for solving (SMPEC$^{\mathbf{imp,1s}}$) and suppose Assumptions 1, 5 and 6 hold.*
(a) *Given $\hat{\mathbf{x}}_k \in \mathcal{X}$, let $\mathbf{y}(\hat{\mathbf{x}}_k)$ denote the unique solution of VI($\mathcal{Y}$, $F(\hat{\mathbf{x}}_k, \bullet)$). Let $\mathbf{y}_{t_k}$ be generated by Algorithm 4 where $t_k := k+1$. Let us define $C_F \triangleq \max_{\mathbf{x}\in X,\ \mathbf{y}\in\mathcal{Y}} \|F(\mathbf{x},\mathbf{y})\|$. Then for all $t_k \geq 0$, we have*

$$\mathbb{E}[\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq \tilde{\epsilon}_k \triangleq \frac{\max\left\{\frac{(C_F^2+v_G^2)\alpha^2}{2\alpha\mu_F-1}, \Gamma \sup_{\mathbf{y}\in\mathcal{Y}}\|\mathbf{y}-\mathbf{y}_0\|^2\right\}}{t_k+\Gamma}.$$

(b) *The following holds for any $\gamma < \frac{\eta}{nL_0}$, $\ell \triangleq \lceil \lambda K \rceil$, and all $K > \frac{2}{1-\lambda}$.*

$$
\begin{aligned}
&\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \\
&\leq \frac{n^2\gamma(1-2\ln(\lambda))\left(1-\frac{nL_0\gamma}{2\eta}\right)\left(\frac{8\tilde{L}_0^2(C_F^2+v_G^2)}{\eta^2\mu_F^2} + L_0^2\right) + \mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right] - f^* + 2L_0\eta}{\left(1-\frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1-\lambda)K}.
\end{aligned}
$$

(c) *Suppose $\gamma = \frac{\eta}{2nL_0}$ and $\eta = \frac{1}{L_0}$. Let $\epsilon > 0$ be an arbitrary scalar and K_ϵ be such that $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \epsilon$. Then,*

(c-1) *the total number of upper-level projection steps on* $\mathcal{X}$ *is* $K_\epsilon = \mathcal{O}\left(n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1}\right)$.
(c-2) *the sample complexity of upper-level function evaluations is* $\mathcal{O}\left(n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2}\right)$.
(c-3) *the total number of lower-level projection steps on* $\mathcal{Y}$ *is* $\mathcal{O}\left(n^6 L_0^6 \tilde{L}_0^6 \epsilon^{-3}\right)$.
(c-4) *the sample complexity of lower-level evaluations of the mapping is* $\mathcal{O}\left(n^6 L_0^6 \tilde{L}_0^6 \epsilon^{-3}\right)$.

Proof **(a)** Let the error Δ_t be defined as $\Delta_t \triangleq G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t) - F(\hat{\mathbf{x}}_k, \mathbf{y}_t)$ for $t \geq 0$. We have

$$\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{x}_k)\|^2 &= \|\Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha_t G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t)\right] - \Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k)\right]\|^2 \\
&\leq \|\mathbf{y}_t - \alpha_t G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_t) - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \\
&= \|\mathbf{y}_t - \alpha_t F(\hat{\mathbf{x}}_k, \mathbf{y}_t) - \alpha_t \Delta_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \\
&= \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha_t^2 \|F(\hat{\mathbf{x}}_k, \mathbf{y}_t)\|^2 + \alpha_t^2 \|\Delta_t\|^2 \\
&\quad - 2\alpha_t (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\hat{\mathbf{x}}_k, \mathbf{y}_t) \\
&\quad - 2\alpha_t (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k) - \alpha_t F(\hat{\mathbf{x}}_k, \mathbf{y}_t))^T \Delta_t.
\end{aligned}$$

Taking conditional expectations from the preceding relation and invoking Assumption 6, we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] \leq \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha_t^2 (C_F^2 + \nu_G^2) \\
- 2\alpha_t (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\hat{\mathbf{x}}_k, \mathbf{y}_t).
\end{aligned}$$

From strong monotonicity of mapping $F(\hat{\mathbf{x}}_k, \bullet)$ uniformly in $\hat{\mathbf{x}}_k$ and the definition of $\mathbf{y}(\hat{x}_k)$, we have

$$\begin{aligned}
(\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\hat{\mathbf{x}}_k, \mathbf{y}_t) &\geq (\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k))^T F(\mathbf{y}(\hat{x}_k), \hat{\mathbf{x}}_k) + \mu_F \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \\
&\geq \mu_F \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2.
\end{aligned}$$

From the preceding relations, we obtain

$$\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 \mid \hat{\mathbf{x}}_k, \mathbf{y}_t] \leq (1 - 2\mu_F \alpha_t) \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2 + \alpha_t^2 (C_F^2 + \nu_G^2).$$

Taking expectations from both sides, we have

$$\mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq (1 - 2\mu_F \alpha_t) \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] + \alpha_t^2 (C_F^2 + \nu_G^2).$$

Noting that in Algorithm 4 we have $\alpha_0 > \frac{1}{2\mu_F}$, using Lemma 8, we obtain that

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k)\|^2] \leq \frac{\max\left\{\frac{(C_F^2 + \nu_G^2)\alpha^2}{2\alpha\mu_F - 1}, \Gamma \sup_{\mathbf{y} \in \mathcal{Y}} \|\mathbf{y} - \mathbf{y}_0\|^2\right\}}{t + \Gamma}, \quad \text{for all} \quad t \geq 0.$$

(b) We can write

$$\begin{aligned}
\mathbb{E}\left[\|e_k\|^2 \mid \mathbf{x}_k\right] &= \mathbb{E}\left[\left\|g_{\eta,N_k,\tilde{\epsilon}_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\
&= \mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k,v_{j,k},\omega_{j,k})}{N_k} - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\
&\leq 2\mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k,v_{j,k},\omega_{j,k})}{N_k} - \frac{\sum_{j=1}^{N_k} g_\eta(\mathbf{x}_k,v_{j,k},\omega_{j,k})}{N_k}\right\|^2 \mid \mathbf{x}_k\right] \\
&\quad + 2\mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_\eta(\mathbf{x}_k,v_{j,k},\omega_{j,k})}{N_k} - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\
&\leq \frac{2\sum_{j=1}^{N_k}\mathbb{E}\left[\left\|g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k,v_{j,k},\omega_{j,k}) - g_\eta(\mathbf{x}_k,v_{j,k},\omega_{j,k})\right\|^2 | \mathbf{x}_k\right]}{N_k} \\
&\quad + \frac{2\sum_{j=1}^{N_k}\mathbb{E}\left[\left\|g_\eta(\mathbf{x}_k,v_{j,k},\omega_{j,k}) - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 | \mathbf{x}_k\right]}{N_k^2} \\
&\leq \frac{8\tilde{L}_0^2 n^2 \tilde{\epsilon}_k}{\eta^2} + \frac{2\sum_{j=1}^{N_k}\left(\mathbb{E}\left[\left\|g_\eta(\mathbf{x}_k,v_{j,k},\omega_{j,k})\right\|^2 | \mathbf{x}_k\right] - \left\|\nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2\right)}{N_k^2} \\
&\leq \frac{8\tilde{L}_0^2 n^2 \tilde{\epsilon}_k}{\eta^2} + \frac{2n^2 L_0^2}{N_k}, \qquad (36)
\end{aligned}$$

where in the second inequality, the first term is implied by the relation $\left\|\sum_{i=1}^m u_i\right\|^2 \leq m\sum_{i=1}^m \|u_i\|^2$ for any $u_i \in \mathbb{R}^n$ for all $i = 1,\ldots,m$. The second term in the second inequality is implied by noting that from Lemma 2, $g_\eta(\mathbf{x}_k, v)$ is an unbiased estimator of $\nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)$. The third inequality is obtained using Lemma 3. From Lemma 7 we have

$$\left(1 - \tfrac{nL_0\gamma}{\eta}\right)\tfrac{\gamma}{4}\|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 \leq f_\eta^{\mathbf{imp}}(\mathbf{x}_k) - f_\eta^{\mathbf{imp}}(\mathbf{x}_{k+1}) + \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma\|e_k\|^2.$$

Let $f_\eta^{\mathbf{imp},*} \triangleq \inf_{x\in\mathcal{X}} f_\eta^{\mathbf{imp}}(\mathbf{x})$. Summing the preceding relation from $k = \ell,\ldots,K-1$ where $\ell \triangleq \lceil \lambda K \rceil$, we have that

$$\left(1 - \tfrac{nL_0\gamma}{\eta}\right)\tfrac{\gamma}{4}\sum_{k=\ell}^{K-1}\|G_{\eta,1/\gamma}(\mathbf{x}_k)\|^2 \leq f_\eta^{\mathbf{imp}}(\mathbf{x}_\ell) - f_\eta^{\mathbf{imp}}(\mathbf{x}_K) + \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma\sum_{k=\ell}^{K-1}\|e_k\|^2.$$

Taking expectations on both sides, it follows that

$$\begin{aligned}
&\left(1 - \tfrac{nL_0\gamma}{\eta}\right)\tfrac{\gamma}{4}(K-\ell)\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \\
&\leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma\sum_{k=\ell}^{K-1}\mathbb{E}\left[\|e_k\|^2\right] + \mathbb{E}\left[f_\eta^{\mathbf{imp}}(\mathbf{x}_\ell)\right] - f_\eta^{\mathbf{imp},*}
\end{aligned}$$

$$
\begin{aligned}
&= \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \mathbb{E}\left[\|e_k\|^2\right] + \mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell) + f_\eta^{\mathbf{imp}}(\mathbf{x}_\ell) - f^{\mathbf{imp}}(\mathbf{x}_\ell)\right] \\
&\quad - f_\eta^{\mathbf{imp},*} + f^* - f^* \\
&\leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \mathbb{E}\left[\|e_k\|^2\right] \\
&\quad + \mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right] - f^* + \mathbb{E}\left[\left|f_\eta^{\mathbf{imp}}(\mathbf{x}_\ell) - f^{\mathbf{imp}}(\mathbf{x}_\ell)\right|\right] + \left|f^* - f_\eta^{\mathbf{imp},*}\right| \\
&\leq \left(1 - \tfrac{nL_0\gamma}{2\eta}\right)\gamma \sum_{k=\ell}^{K-1} \left(\frac{8\tilde{L}_0^2 n^2 \tilde{\epsilon}_k}{\eta^2} + \frac{2n^2 L_0^2}{N_k}\right) + \mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right] - f^* + 2L_0\eta,
\end{aligned}
$$

where the preceding relation is implied by invoking the bound on $\mathbb{E}\left[\|e_k\|^2\right]$ and Lemma 1 (iii). Note that from part (a), we have $\tilde{\epsilon}_k = \frac{2(C_F^2+\nu_G^2)}{\mu_F^2 t_k}$ where $t_k := k+1$. Also, $N_k := k+1$. Note that $K > \frac{2}{1-\lambda}$ implies $\ell \leq K-1$. From Lemma 13, using $\ell \geq 1$ we have $\sum_{k=\ell}^{K-1} \frac{1}{k+1} \leq \frac{1}{\ell+1} + \ln\left(\frac{K}{\ell+1}\right) \leq 0.5 + \ln\left(\frac{N}{\lambda N+1}\right) \leq 0.5 - \ln(\lambda)$. Also, $K - \ell \geq K - \lambda K = (1-\lambda)K$. Thus, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \\
&\leq \frac{\left(1 - \tfrac{nL_0\gamma}{2\eta}\right)2n^2\gamma\left(\frac{8\tilde{L}_0^2(C_F^2+\nu_G^2)}{\eta^2\mu_F^2} + L_0^2\right)(0.5 - \ln(\lambda)) + \mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right] - f^* + 2L_0\eta}{\left(1 - \tfrac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1-\lambda)K}.
\end{aligned}
$$

(c) To show (c-1), using the relation in part (b) and substituting $\gamma = \frac{\eta}{2nL_0}$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \\
&\leq \frac{6n^2(1 - 2\ln(\lambda))\left(\frac{8\tilde{L}_0^2(C_F^2+\nu_G^2)}{\eta^2\mu_F^2} + L_0^2\right) + \frac{16nL_0}{\eta}(\sup_{\mathbf{x}\in\mathcal{X}} f^{\mathbf{imp}}(\mathbf{x}) - f^*) + 32nL_0^2}{(1-\lambda)K}.
\end{aligned}
$$

Further, from $\eta = \frac{1}{L_0}$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \\
&\leq \frac{6n^2L_0^2(1 - 2\ln(\lambda))\left(\frac{8\tilde{L}_0^2(C_F^2+\nu_G^2)}{\mu_F^2} + 1\right) + 16nL_0^2(\sup_{\mathbf{x}\in\mathcal{X}} f^{\mathbf{imp}}(\mathbf{x}) - f^*) + 32nL_0^2}{(1-\lambda)K}.
\end{aligned}
$$

This implies that $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{\mathcal{O}\left(n^2L_0^2\tilde{L}_0^2\right)}{K}$ and thus, we obtain $K_\epsilon = \mathcal{O}\left(n^2L_0^2\tilde{L}_0^2\epsilon^{-1}\right)$. Next, we show (c-2). The overall sample complexity of upper-level evaluations is as follows.

$$\sum_{k=0}^{K_\epsilon} N_k = \sum_{k=0}^{K_\epsilon}(k+1) = \mathcal{O}(K_\epsilon^2) = \mathcal{O}\left(n^4L_0^4\epsilon^{-2}\right).$$

To show (c-3), note that the total number of lower-level projection steps is given by

$$\sum_{k=0}^{K_\epsilon}(1+N_k)t_k = \sum_{k=0}^{K_\epsilon}(k+1)(k+2) = \mathcal{O}(K_\epsilon^3) = \mathcal{O}\left(n^6L_0^6\epsilon^{-3}\right).$$

Noting that at each iteration in Algorithm 4 a single sample is taken, we obtain the bound in (c-4). □

Remark 8 (*Variance-reduction and smoothing schemes in the nonconvex case*)

(i) Unlike in ($\texttt{ZSOL}_{\text{cvx}}^{\mathbf{1s}}$), in ($\texttt{ZSOL}_{\text{ncvx}}^{\mathbf{1s}}$) we employ a variance-reduction scheme in the upper-level. This is mainly because, in contrast with the convex case, the use of the Euclidean projection in ($\texttt{ZSOL}_{\text{ncvx}}^{\mathbf{1s}}$) leads to the presence of the persistent error term $\left(1-\frac{nL_0\gamma}{2\eta}\right)\gamma\|e_k\|^2$ (see Lemma 7). The use of variance-reduction helps with contending with this error in establishing the convergence and rate results.

(ii) Unlike in ($\texttt{ZSOL}_{\text{cvx}}^{\mathbf{1s}}$), in ($\texttt{ZSOL}_{\text{ncvx}}^{\mathbf{1s}}$) we employ a constant smoothing parameter. This is because assuming an iteratively updating smoothing parameter η_k in the nonconvex case does not appear to allow for constructing a recursive error bound. For this reason, in the nonconvex case we limit our study to the case when the smoothing parameter is constant.

3.3.2 An exact zeroth-order scheme

In this subsection, we present the rate and complexity results for the exact variant of Algorithm 3.

Corollary 2 (Rate and complexity statements for exact ($\texttt{ZSOL}_{\text{ncvx}}^{\mathbf{1s}}$)) *Consider Algorithms 3 (exact variant) for solving* (SMPEC$^{\mathbf{imp,1s}}$) *and suppose Assumptions 1 and 5 hold.*

(a) *The following holds for any* $\gamma < \frac{\eta}{nL_0}$, $\ell \triangleq \lceil \lambda K \rceil$, *and all* $K > \frac{2}{1-\lambda}$.

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{n^2L_0^2\gamma(0.5-\ln(\lambda))\left(1-\frac{nL_0\gamma}{2\eta}\right)+\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right]-f^*+2L_0\eta}{\left(1-\frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1-\lambda)K}.$$

(b) *Suppose* $\gamma = \frac{\eta}{2nL_0}$ *and* $\eta = \frac{1}{L_0}$. *Let* $\epsilon > 0$ *be an arbitrary scalar and* K_ϵ *be such that* $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \epsilon$. *Then the following hold.*

(b-1) *The total number of upper-level projection steps on* $\mathcal{X}$ *is* $K_\epsilon = \mathcal{O}\left(n^2 L_0^2 \epsilon^{-1}\right)$.
(b-2) *The total sample complexity of upper-level is* $\mathcal{O}\left(n^4 L_0^4 \epsilon^{-2}\right)$.

Proof The proof can be carried out in a similar vein to that of Theorem 2 by noting that $\tilde{\epsilon}_k := 0$ in the exact variant. The main difference lies in establishing the upper bound on $\mathbb{E}\left[\|e_k\|^2 \mid \mathbf{x}_k\right]$ in (36). To be precise, we derive this bound as follows.

$$\begin{aligned}
\mathbb{E}\left[\|e_k\|^2 \mid \mathbf{x}_k\right] &= \mathbb{E}\left[\left\|g_{\eta,N_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\
&= \mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_\eta(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k} - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\
&\leq \frac{\sum_{j=1}^{N_k} \mathbb{E}\left[\left\|g_\eta(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) - \nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 | \mathbf{x}_k\right]}{N_k^2} \\
&\leq \frac{\sum_{j=1}^{N_k} \left(\mathbb{E}\left[\left\|g_\eta(\mathbf{x}_k, v_{j,k}, \omega_{j,k})\right\|^2 | \mathbf{x}_k\right] - \left\|\nabla_{\mathbf{x}} f_\eta^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2\right)}{N_k^2} \leq \frac{n^2 L_0^2}{N_k}, \quad \text{almost surely.}
\end{aligned}$$

□

4 Zeroth-order methods for two-stage SMPECs

In this section, we extend the zeroth-order schemes from the previous section to allow for accommodating two-stage model (SMPEC$^{\mathbf{imp,2s}}$). In Sect. 4.1, we discuss an implicit framework for two-stage SMPECs and present inexact and exact schemes and an accelerated counterpart in Sects. 4.2 and 4.3. We conclude with a discussion of addressing nonconvexity in the implicit problem in Sect. 4.4.

4.1 An implicit framework

Consider the implicit problem (SMPEC$^{\mathbf{imp,2s}}$). Given the defined function $f^{\mathbf{imp}}$ and a scalar η, we consider a spherical smoothing $f_\eta^{\mathbf{imp}}$ as follows:

$$f_\eta^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}_{u\in\mathbb{B}}[f^{\mathbf{imp}}(\mathbf{x}+\eta u)] = \mathbb{E}_{u\in\mathbb{B}}[\mathbb{E}[\tilde{f}(\mathbf{x}+\eta u, \mathbf{y}(\mathbf{x}+\eta u, \omega), \omega)]]. \tag{G-Smooth$^{\mathbf{2s}}$}$$

Similar to the single-stage case discussed in Sect. 3.1, the zeroth-order approximation of the gradient is given by (13). An unbiased estimate of $g_\eta(\mathbf{x})$ is defined as

$$g_\eta(\mathbf{x}, v, \omega) \triangleq \left(\frac{n}{\eta}\right)\left[\frac{\left(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v, \omega), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)\right) v}{\|v\|}\right]. \tag{37}$$

Given a vector $\mathbf{x}_0 \in \mathcal{X}$, we may employ (37) in constructing a sequence $\{\mathbf{x}_k\}$ where $\mathbf{x}_k$ satisfies the following projected stochastic gradient update.

$$\mathbf{x}_{k+1} := \Pi_{\mathcal{X}} \left[\mathbf{x}_k - \gamma_k g_\eta(\mathbf{x}_k, v_k, \omega_k) \right]. \tag{38}$$

Lemma 9 (Properties of the two-stage exact zeroth-order gradient) *Suppose Assumption* 1(b) *holds. Consider* (SMPEC$^{\mathbf{imp},2\mathbf{s}}$). *Given* $\mathbf{x} \in \mathcal{X}$ *and* $\eta > 0$, *consider the stochastic zeroth-order mapping* $g_\eta(\mathbf{x}, v, \omega)$ *defined by* (37) *for* $v \in \eta\mathbb{S}$ *and* $k \geq 0$, *where* v *and* ω *are independent. Then,* $\nabla f_\eta^{\mathbf{imp}}(\mathbf{x}) = \mathbb{E}[g_\eta(\mathbf{x}, v, \omega) \mid \mathbf{x}]$ *and* $\mathbb{E}[\|g_\eta(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}] \leq L_0^2 n^2$ *almost surely for all* $k \geq 0$.

Proof The proof is similar to the proof of Lemma 2. We provide the details for the sake of completeness. From (37) and that $f^{\mathbf{imp}}(\mathbf{x}) \triangleq \mathbb{E}[\tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)]$ we can write

$$\begin{aligned}
\mathbb{E}[g_\eta(\mathbf{x}, v, \omega) \mid \mathbf{x}] &= \mathbb{E}_{v \in \eta\mathbb{S}} \left[\left(\frac{n}{\eta}\right) \frac{\left(f^{\mathbf{imp}}(\mathbf{x}+v) - f^{\mathbf{imp}}(\mathbf{x})\right) v}{\|v\|} \mid \mathbf{x} \right] \\
&= \left(\frac{n}{\eta}\right) \mathbb{E}_{v \in \eta\mathbb{S}} \left[f^{\mathbf{imp}}(\mathbf{x}+v) \frac{v}{\|v\|} \mid \mathbf{x} \right] \overset{\text{Lemma 1}(i)}{=} \nabla f_\eta^{\mathbf{imp}}(\mathbf{x}).
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}[\|g_\eta(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}, \omega] &= \left(\frac{n}{\eta}\right)^2 \mathbb{E} \left[\left\| \frac{\left(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v, \omega), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)\right) v}{\|v\|} \right\|^2 \mid \mathbf{x}, \omega \right] \\
&= \left(\frac{n}{\eta}\right)^2 \int_{\eta\mathbb{S}} \frac{\left\| \left(\tilde{f}(\mathbf{x}+v, \mathbf{y}(\mathbf{x}+v, \omega), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}(\mathbf{x}, \omega), \omega)\right) v \right\|^2}{\|v\|^2} p_v(v) dv \\
&\overset{\text{Assumption 1}(b.i)}{\leq} \frac{n^2}{\eta^2} \int_{\eta\mathbb{S}} L_0^2(\omega) \|v\|^2 p_v(v) dv \\
&\leq n^2 L_0^2(\omega) \int_{\eta\mathbb{S}} p_v(v) dv = n^2 L_0^2(\omega).
\end{aligned}$$

Taking the expectation with respect to ω from the both sides of the preceding inequality and invoking $L_0^2 \triangleq \mathbb{E}[L_0^2(\omega)] < \infty$, we obtain the desired bound. □

4.2 Inexact and exact schemes for convex regime

Consider the implicit form of (SMPEC$^{\mathbf{imp},2\mathbf{s}}$) where $\mathbf{y}(\mathbf{x}, \omega)$ solves VI($\mathcal{Y}$, $G(\mathbf{x}, \bullet, \omega)$). Computing such a $\mathbf{y}(\mathbf{x}, \omega)$ is often challenging, in particular, when $\mathcal{Y}$ is high-dimensional. To contend with this challenge, we employ gradient-like methods for computing inexact solutions to the lower-level ω-specific VI parametrized by $\mathbf{x}$, denoted by VI($\mathcal{Y}$, $G(\mathbf{x}, \bullet, \bullet)$). We consider the case where we have access to an

approximate solution $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega)$ such that

$$\|\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega) - \mathbf{y}(\mathbf{x}_k, \omega)\|^2 \leq \tilde{\epsilon}_k, \quad \text{where} \quad \mathbf{y}(\mathbf{x}_k, \omega) \in \text{SOL}(\mathcal{Y}, G(\mathbf{x}_k, \bullet, \omega)). \tag{39}$$

Similar to the single-stage case, we may define an inexact zeroth-order gradient mapping $g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)$ as follows.

$$g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) \triangleq \frac{n(\tilde{f}(\mathbf{x} + v, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x} + v, \omega), \omega) - \tilde{f}(\mathbf{x}, \mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}, \omega), \omega))v}{\|v\|\eta}, \tag{40}$$

where $v \in \eta\mathbb{S}$ and $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega)$ is an output of a gradient-like scheme. The outline of the proposed zeroth-order solver is presented in Algorithm 5 while an inexact approximation of $\mathbf{y}(\mathbf{x}, \omega)$ is computed by Algorithm 6. In the following, we extend Lemma 2 to the two-stage regime.

Remark 9 Throughout the algorithms in this section, in evaluation of the exact and inexact solution to the lower level problem, denoted by $\mathbf{y}(\bullet, \omega)$ and $\mathbf{y}_{\tilde{\epsilon}}(\bullet, \omega)$, respectively, we assume that we have access to an oracle that returns random replicates of ω.

Lemma 10 (Properties of the two-stage inexact zeroth-order gradient) *Suppose Assumption* 1(b) *holds. Consider* (SMPEC$^{\text{imp,2s}}$). *Let* $g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)$ *be defined as* (40) *for* $\omega \in \Omega$ *and* $v \in \eta\mathbb{S}$ *for* $\eta, \tilde{\epsilon} > 0$. *Suppose* $\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}, \omega) - \mathbf{y}(\mathbf{x}, \omega)\|^2 \leq \tilde{\epsilon}$ *almost surely for any* $\omega \in \Omega$ *and all* $\mathbf{x} \in \mathcal{X}$. *Then, the following hold for any* $\mathbf{x} \in \mathcal{X}$.

(a) $\mathbb{E}[\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}] \leq 3n^2 \left(\frac{2\tilde{L}_0^2\tilde{\epsilon}}{\eta^2} + L_0^2 \right)$, *almost surely.*

(b) $\mathbb{E}\left[\left\| g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega) \right\|^2 \mid \mathbf{x} \right] \leq \frac{4\tilde{L}_0^2 n^2 \tilde{\epsilon}}{\eta^2}$, *almost surely.*

Proof **(a)** In a similar fashion to the proof of Lemma 3 (a), we can show that

$$\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\| \leq \frac{\tilde{L}_0(\omega)\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x} + v, \omega) - \mathbf{y}(\mathbf{x} + v, \omega)\|n}{\eta} + \left\| g_\eta(\mathbf{x}, v, \omega) \right\| + \frac{\tilde{L}_0(\omega)\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}, \omega) - \mathbf{y}(\mathbf{x}, \omega)\|n}{\eta}.$$

Invoking Lemma 2, we may then bound the second moment of $\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\|$ in an almost sure sense as follows.

$$\begin{aligned}
\mathbb{E}[\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega)\|^2 \mid \mathbf{x}] &\leq 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2(\omega)n^2\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v,\omega)-\mathbf{y}(\mathbf{x}+v,\omega)\|^2}{\eta^2}\right) \mid \mathbf{x}\right] \\
&\quad + 3\mathbb{E}\left[\left\|g_\eta(\mathbf{x}, v, \omega)\right\|^2 \mid \mathbf{x}\right] \\
&\quad + 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2(\omega)n^2\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v,\omega)-\mathbf{y}(\mathbf{x}+v,\omega)\|^2}{\eta^2}\right) \mid \mathbf{x}\right] \\
&\leq 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2(\omega)n^2\tilde{\epsilon}^2}{\eta^2}\right) \mid \mathbf{x}\right] + 3L_0^2n^2 \\
&\quad + 3\mathbb{E}\left[\left(\frac{\tilde{L}_0^2(\omega)n^2\tilde{\epsilon}^2}{\eta^2}\right) \mid \mathbf{x}\right] \\
&\leq 3n^2\left(\tfrac{2\tilde{L}_0^2\tilde{\epsilon}}{\eta^2} + L_0^2\right).
\end{aligned}$$

(b) In a similar fashion to the proof of Lemma 3 (b), we can show that

$$\begin{aligned}
\left\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega)\right\| &\leq \frac{\tilde{L}_0(\omega)n\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v,\omega)-\mathbf{y}(\mathbf{x}+v,\omega)\|}{\eta} \\
&\quad + \frac{\tilde{L}_0(\omega)n\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x},\omega)-\mathbf{y}(\mathbf{x},\omega)\|}{\eta}.
\end{aligned}$$

Consequently, the following holds almost surely,

$$\begin{aligned}
&\mathbb{E}\left[\left\|g_{\eta,\tilde{\epsilon}}(\mathbf{x}, v, \omega) - g_\eta(\mathbf{x}, v, \omega)\right\|^2 \mid \mathbf{x}\right] \\
&\quad\leq \frac{2\mathbb{E}[\tilde{L}_0^2(\omega)n^2\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x}+v,\omega)-\mathbf{y}(\mathbf{x}+v,\omega)\|^2 \mid \mathbf{x}]}{\eta^2} \\
&\qquad + \frac{2\mathbb{E}[\tilde{L}_0^2(\omega)n^2\|\mathbf{y}_{\tilde{\epsilon}}(\mathbf{x},\omega)-\mathbf{y}(\mathbf{x},\omega)\|^2 \mid \mathbf{x}]}{\eta^2} \\
&\quad\leq \frac{2\mathbb{E}[\tilde{L}_0^2(\omega)n^2\tilde{\epsilon}^2 \mid \mathbf{x}]}{\eta^2} + \frac{2\mathbb{E}[\tilde{L}_0^2(\omega)n^2\tilde{\epsilon}^2 \mid \mathbf{x}]}{\eta^2} \leq \frac{4\tilde{L}_0^2n^2\tilde{\epsilon}}{\eta^2}.
\end{aligned}$$

□

Algorithm 5 $\texttt{ZSOL}_{\text{cnvx}}^{\mathbf{2s}}$: Zeroth-order method for convex (SMPEC$^{\mathbf{2s}}$)

1: **input:** Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize sequence $\{\gamma_k\}$, smoothing parameter sequence $\{\eta_k\}$, inexactness sequence $\{\tilde{\epsilon}_k\}$, $r \in [0, 1)$, and $S_0 := \gamma_0^r$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Generate $v_k \in \eta_k \mathbb{S}$
4: Do one of the following, depending on the type of the scheme.

- Inexact scheme: Call Alg. 6 twice to obtain $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega_k)$ and $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k + v_k, \omega_k)$
- Exact scheme: Evaluate $\mathbf{y}(\mathbf{x}_k, \omega_k)$ and $\mathbf{y}(\mathbf{x}_k + v_k, \omega_k)$

5: Evaluate the inexact or exact zeroth-order gradient approximation as follows.

$$g_{\eta_k, \tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k) := \frac{n\left(\tilde{f}(\mathbf{x}_k + v_k, \mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k + v_k, \omega_k), \omega_k) - \tilde{f}(\mathbf{x}_k, \mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega_k), \omega_k)\right) v_k}{\|v_k\| \eta_k} \qquad \text{(Inexact)}$$

$$g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k) := \frac{n\left(\tilde{f}(\mathbf{x}_k + v_k, \mathbf{y}(\mathbf{x}_k + v_k, \omega_k), \omega_k) - \tilde{f}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k, \omega_k), \omega_k)\right) v_k}{\|v_k\| \eta_k} \qquad \text{(exact)}$$

6: Update $\mathbf{x}_k$ as follows.

$$\mathbf{x}_{k+1} := \begin{cases} \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k, \tilde{\epsilon}_k}(\mathbf{x}_k, v_k, \omega_k)\right] & \text{(Inexact)} \\ \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k}(\mathbf{x}_k, v_k, \omega_k)\right] & \text{(Exact)} \end{cases}$$

7: Update the averaged iterate as follows. $S_{k+1} := S_k + \gamma_{k+1}^r$ and $\bar{\mathbf{x}}_{k+1} := \frac{S_k \bar{\mathbf{x}}_k + \gamma_{k+1}^r \mathbf{x}_{k+1}}{S_{k+1}}$
8: **end for**

Algorithm 6 Projection method for the VI in the lower-level of (SMPEC$^{\mathbf{2s}}$)

1: **input:** An arbitrary $\mathbf{y}_0 \in \mathcal{Y}$, vectors $\hat{\mathbf{x}}_k$ and ω, scalar $\rho \in (0, 1)$, stepsize $\alpha > 0$, integer k, and scalar $\tau > 0$
2: Compute $t_k := \lceil \tau \ln(k+1) \rceil$
3: **for** $t = 0, 1, \ldots, t_k - 1$ **do**
4: Evaluate the mapping $G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega)$
5: Update $\mathbf{y}_t$ as follows. $\mathbf{y}_{t+1} := \Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega)\right]$
6: **end for**
7: Return $\mathbf{y}_{t_k}$

Next we develop rate and complexity statements for Algorithm 5. The algorithm parameters for both inexact and exact schemes are defined next.

Definition 4 (*Parameters for Algorithms 5–6*) Let the stepsize and smoothing sequences in Algorithm 5 be given by $\gamma_k := \frac{\gamma_0}{(k+1)^a}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$, respectively for all $k \geq 0$ where γ_0, η_0, a, and b are strictly positive. In Algorithm 6, suppose $\alpha \leq \frac{\mu_F}{L_F^2}$. Let $t_k := \lceil \tau \ln(k+1) \rceil$ where $\tau \geq \frac{-2(a+b)}{\ln(1-\mu_F \alpha)}$. Finally, suppose $r \in [0, 1)$ is an arbitrary scalar.

Theorem 3 (Rate and complexity statements and a.s. convergence for inexact ($\texttt{ZSOL}_{\text{cnvx}}^{\mathbf{2s}}$)) *Consider the sequence* $\{\bar{\mathbf{x}}_k\}$ *generated by applying Algorithm 5 on*

(SMPEC$^{\mathbf{imp,2s}}$). *Suppose Assumptions 1–3 hold and algorithm parameters are defined by Definition 4.*
(a) *Suppose* $\hat{\mathbf{x}}_k \in \mathcal{X} + \eta_k \mathbb{S}$ *and let* $\{\mathbf{y}_{t_k}\}$ *be the sequence generated by Algorithm 6. Then for suitably defined scalars* $\tilde{d} < 1$ *and* $B > 0$*, the following holds for* $t_k \geq 1$.

$$\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 \leq \tilde{\epsilon}_k \triangleq B\tilde{d}^{t_k}.$$

(b) *Let* $a = 0.5$ *and* $b \in [0.5, 1)$ *and* $0 \leq r < 2(1-b)$*. Then, for all* $K \geq 2^{\frac{1}{1-r}} - 1$ *we have*

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq (2-r)\left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{2\theta_0\gamma_0}{1-r}\right)\frac{1}{\sqrt{K+1}} + (2-r)\left(\frac{\eta_0 L_0}{1-0.5r-b}\right)\frac{1}{(K+1)^b},$$

where $\theta_0 \triangleq D_{\mathcal{X}} + \frac{(2+3\gamma_0^2)n^2\tilde{L}_0^2 B}{\eta_0^2\gamma_0^2} + 1.5n^2L_0^2$*. In particular, when* $b := 1-\delta$ *and* $r = 0$*, where* $\delta > 0$ *is a small scalar, we have for all* $K \geq 1$

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq 2\left(\frac{D_{\mathcal{X}}}{\gamma_0} + 2\theta_0\gamma_0\right)\frac{1}{\sqrt{K+1}} + \left(\frac{2\eta_0 L_0}{\delta}\right)\frac{1}{(K+1)^{1-\delta}}.$$

(c) *Suppose* $\gamma_0 := \mathcal{O}(\frac{1}{L_0})$*,* $a := 0.5$*,* $b := 0.5$*, and* $r := 0$*. Let* $\epsilon > 0$ *be an arbitrary scalar and* K_ϵ *be such that* $\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_{K_\epsilon})\right] - f^* \leq \epsilon$*. Then,*
(c-1) *the total number of upper-level projection steps on* $\mathcal{X}$ *is* $K_\epsilon = \mathcal{O}\left(n^4L_0^2\tilde{L}_0^4\epsilon^{-2}\right)$.
(c-2) *the overall sample complexity of upper-level evaluations is* $\mathcal{O}\left(n^4L_0^2\tilde{L}_0^4\epsilon^{-2}\right)$.
(c-3) *the total number of lower-level projection steps on* $\mathcal{Y}$ *is* $\mathcal{O}\left(n^4L_0^2\tilde{L}_0^4\epsilon^{-2}\ln\left(n^2L_0\tilde{L}_0^2\epsilon^{-1}\right)\right)$.
(d) *For any* $a \in (0.5, 1]$ *and* $b > 1-a$*, there exists* $\mathbf{x}^* \in \mathcal{X}^*$ *such that* $\lim_{k\to\infty}\|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 = 0$ *almost surely.*

Proof **(a)** From $\mathbf{y}(\hat{\mathbf{x}}_k, \omega_k) \in \mathrm{SOL}(\mathcal{Y}, G(\hat{\mathbf{x}}_k, \bullet, \omega_k))$, we have that the following fixed-point relationship holds.

$$\mathbf{y}(\hat{\mathbf{x}}_k, \omega_k) = \Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k, \omega_k) - \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k), \omega_k)\right],$$

for any $\alpha > 0$. Thus, we can write

$$\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 &= \|\Pi_{\mathcal{Y}}\left[\mathbf{y}_t - \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_k)\right] \\
&\quad - \Pi_{\mathcal{Y}}\left[\mathbf{y}(\hat{\mathbf{x}}_k, \omega_k) - \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k), \omega_k)\right]\|^2 \\
&\leq \|\mathbf{y}_t - \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_k) - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k) + \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k), \omega_k)\|^2 \\
&= \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 + \|\alpha G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_k) - \alpha G(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k), \omega_k)\|^2 \\
&\quad - 2\alpha(\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k))^T(G(\hat{\mathbf{x}}_k, \mathbf{y}_t, \omega_k) - G(\hat{\mathbf{x}}_k, \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k), \omega_k)).
\end{aligned}$$

Invoking Assumption 1 (b) we obtain

$$\|\mathbf{y}_{t+1} - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 \leq \|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 + \alpha L_F(\omega)\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2$$

$$
\begin{aligned}
&- 2\alpha\mu_F(\omega)\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 \\
&\leq (1 + \alpha^2 L_F^2 - 2\alpha\mu_F)\|\mathbf{y}_t - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2.
\end{aligned}
$$

This implies that $\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_k)\|^2 \leq (1 + \alpha^2 L_F^2 - 2\alpha\mu_F)^{t_k}(\sup_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y} - \mathbf{y}_0\|^2)$. Note that $\alpha \leq \frac{\mu_F}{L_F^2}$ implies that $1 + \alpha^2 L_F^2 - 2\alpha\mu_F \leq 1 - \alpha\mu_F$. Defining $\tilde{d} \triangleq 1 - \alpha\mu_F$ and $B \triangleq \sup_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y} - \mathbf{y}_0\|^2$, we obtain the bound.

(b, d) Recall the properties of the exact and inexact zeroth-order gradient mappings in the two-stage model provided in Lemmas 9 and 10, respectively. Note that these results are identical to those of the single-stage model provided in Lemmas 2 and 3, respectively. For this reason, the proof of the remaining parts can be carried out in a similar fashion to the proofs in Theorem 1. As such, the proofs for (b) and (d) are omitted.

(c) Note that (c-1) and (c-2) follow directly from part (b) by substituting γ_0 and r. To show (c-3), note that the total projection steps in the lower-level is as follows.

$$
\begin{aligned}
2\sum_{k=0}^{K_\epsilon}\sum_{t=0}^{t_k} 1 &= 2(K_\epsilon + 1)(t_{K_\epsilon} + 1) = 2(K_\epsilon + 1)(\lceil \tau \ln(K_\epsilon + 1)\rceil + 1) \\
&= \mathcal{O}\left(n^4 L_0^2 \tilde{L}_0^4 \epsilon^{-2} \ln\left(n^2 L_0 \tilde{L}_0^2 \epsilon^{-1}\right)\right).
\end{aligned}
$$

□

Remark 10 The convergence rate in expectation in Theorem 1 (b) and Theorem 3 (b) can be extended to the case that $a \in [0.5, 1)$. However, the rate of convergence would be worse when $a \in (0.5, 1)$ compared to when $a = 0.5$. This is because employing Lemma 13, the rate of convergence is characterized as $\mathcal{O}\left(\frac{1}{k^{1-a}} + \frac{1}{k^a} + \frac{1}{k^b}\right)$. For this reason, we only present the rate analysis in those theorems for $a = 0.5$.

An exact zeroth-order scheme. Next, we address the two-stage model (SMPEC$^{\mathbf{imp,2s}}$) where we consider the case where an exact solution of the lower-level problem is available. In the following, we extend the convergence properties of the ZSOL scheme to the exact case.

Corollary 3 (Rate and complexity statements and almost sure convergence for exact (ZSOL$^{\mathbf{2s}}_{\mathrm{cnvx}}$)) *Consider the problem* (SMPEC$^{\mathbf{imp,1s}}$). *Suppose Assumptions 1–3 hold. Suppose* $\{\bar{\mathbf{x}}_k\}$ *denotes the sequence generated by Algorithm 5 (exact variant) in which the stepsize and smoothing sequences are defined as* $\gamma_k := \frac{\gamma_0}{(k+1)^a}$ *and* $\eta_k := \frac{\eta_0}{(k+1)^b}$, *respectively, for all* $k \geq 0$ *where* γ_0 *and* η_0 *are strictly positive. Then, the following statements hold.*

(a) *Let* $a = 0.5$ *and* $b \in [0.5, 1)$ *and* $0 \leq r < 2(1 - b)$. *Then, for all* $K \geq 2^{\frac{1}{1-r}} - 1$ *we have*

$$
\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq (2 - r)\left(\frac{D_{\mathcal{X}}}{\gamma_0} + \frac{L_0^2 n^2 \gamma_0}{1 - r}\right)\frac{1}{\sqrt{K+1}} + (2 - r)\left(\frac{\eta_0 L_0}{1 - 0.5r - b}\right)\frac{1}{(K+1)^b}.
$$

In particular, when $b := 1 - \delta$ and $r = 0$, where $\delta > 0$ is a small scalar, we have for all $K \geq 1$

$$\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_K)\right] - f^* \leq 2\left(\frac{D_{\mathcal{X}}}{\gamma_0} + L_0^2 n^2 \gamma_0\right)\frac{1}{\sqrt{K+1}} + \left(\frac{2\eta_0 L_0}{\delta}\right)\frac{1}{(K+1)^{1-\delta}}.$$

(b) *Let $a := 0.5$, $b = 0.5$, $r = 0$, $\gamma_0 := \frac{\sqrt{D_{\mathcal{X}}}}{nL_0}$, and $\eta_0 \leq \sqrt{D_{\mathcal{X}}}n$. Then, the iteration complexity in projection steps on $\mathcal{X}$ for achieving $\mathbb{E}\left[f^{\mathbf{imp}}(\bar{\mathbf{x}}_{K_\epsilon})\right] - f^* \leq \epsilon$ for some $\epsilon > 0$ is bounded as follows.*

$$K_\epsilon \geq \frac{64n^2 L_0^2 D_{\mathcal{X}}}{\epsilon^2}.$$

(c) *For any $a \in (0.5, 1]$ and $b > 1 - a$, there exists $\mathbf{x}^* \in \mathcal{X}^*$ such that $\lim_{k\to\infty} \|\bar{\mathbf{x}}_k - \mathbf{x}^*\|^2 = 0$ almost surely.*

Proof In view of the similarity between the results of Lemmas 9 and 10 with those of Lemmas 2 and 3, the proof can be done in a similar fashion to that of Corollary 1. □

4.3 Exact accelerated schemes for convex regime

In this subsection, we consider an accelerated scheme for resolving the problem (SMPEC$^{\mathbf{2s}}$), whose implicit form is defined as (SMPEC$^{\mathbf{imp,2s}}$) where $\mathbf{y}(\mathbf{x}, \omega)$ is the unique solution of an ω-specific strongly monotone variational inequality problem parametrized by $\mathbf{x}$. The deterministic counterpart of this problem is the standard MPEC in which the lower-level problem is a parametrized strongly monotone variational inequality problem. While the previous subsection has considered a standard gradient-based framework, we consider an accelerated counterpart motivated by Nesterov's celebrated accelerated gradient method [57] that produces a non-asymptotic rate of $\mathcal{O}(1/k^2)$ in terms of suboptimality for smooth convex optimization problems. In [59], Nesterov and Spokoiny develop an accelerated zeroth-order scheme for the unconstrained minimization of a smooth function. Instead, we present an accelerated gradient-free scheme for a nonsmooth function by leveraging the smoothing architecture. Notably, this scheme can contend with MPECs with convex implicit functions. In this subsection, we assume that $\mathbf{y}(\mathbf{x}, \omega)$ can be generated by invoking a suitable variational inequality problem solver.

We provide convergence theory for Algorithm 7 by appealing to related work on smoothed accelerated schemes for nonsmooth stochastic convex optimization [35]. There are two key differences between the framework presented here and that of our prior work.

(a) *Smoothing.* In [35], we employ a deterministic smoothing technique [6] while in this paper, we consider a locally randomized smoothing technique in a zeroth-order regime. Notably, the latter leads to similar (but not identical) smoothness properties with related relationships (but not identical) between the smoothed function and its original counterpart.

Algorithm 7 $\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx,acc}}$: Variance-reduced accelerated exact zeroth-order method for convex (SMPEC$^{\mathbf{2s}}$)

input: Given $\mathbf{x}_0 \in \mathcal{X}, \lambda_0 = 1$, stepsize sequence $\{\gamma_k\}$, smoothing parameter sequence $\{\eta_k\}$, sample-size $\{N_k\}$
for $k = 0, 1, \ldots, K-1$ **do**
  for $j = 1, \ldots, N_k$ **do**
    Generate $v_{j,k} \in \eta_k \mathbb{S}$
    Evaluate $\mathbf{y}(\mathbf{x}_k + v_{j,k}, \omega_{j,k})$
    Evaluate the exact zeroth-order gradient approximation as follows.

$$g_{\eta_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) := \frac{n\Big(\tilde{f}(\mathbf{x}_k + v_{j,k}, \mathbf{y}(\mathbf{x}_k + v_{j,k}, \omega_{j,k}), \omega_{j,k}) - \tilde{f}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k, \omega_{j,k}), \omega_{j,k})\Big) v_{j,k}}{\|v_{j,k}\| \eta_k}$$

  end for
  Evaluate the mini-batch exact zeroth-order gradient as $g_{\eta_k, N_k}(\mathbf{x}_k) = \frac{\sum_{j=1}^{N_k} g_{\eta_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k}$.
  Update $\mathbf{x}_k$ as follows.

$$\begin{aligned} \mathbf{z}_{k+1} &:= \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma_k g_{\eta_k, N_k}(\mathbf{x}_k, v_k)\right] \\ \lambda_{k+1} &:= \frac{1+\sqrt{1+4\lambda_k^2}}{2} \\ \mathbf{x}_{k+1} &= \mathbf{z}_{k+1} + \frac{(\lambda_k - 1)}{\lambda_{k+1}}\left(\mathbf{z}_{k+1} - \mathbf{z}_k\right). \end{aligned} \tag{41}$$

end for

(b) *Zeroth-order gradient approximation.* In [35], a sampled gradient of the smoothed function is available. However, faced by the need to resolve hierarchical problems, we do not have such access in this paper. Instead, we utilize an increasingly accurate zeroth-order approximation of the gradient by raising the sample-size N_k in constructing this approximation. We make the following assumption on the generated random samples in the proposed accelerated scheme in the upper-level.

Assumption 7 Given a mini-batch sequence $\{N_k\}$ and a smoothing sequence $\{\eta_k\}$, let $v_{j,k} \in \mathbb{R}^n$, for $j = 1, \ldots, N_k$ and $k \geq 0$ be generated randomly and independently, from $\eta_k \mathbb{S}$ for all $k \geq 0$. Also, let the random realizations $\{\omega_{j,k}\}$ be iid replicates.

We may define $\bar{w}_{k,N_k}$ as $\bar{w}_{k,N_k} \triangleq g_{\eta_k,N_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f^{\mathbf{imp}}_{\eta_k}(\mathbf{x}_k)$. The following claims can be made.

Lemma 11 *Consider $\bar{w}_{k,N_k}$ obtained by generating N_K independent realizations given by $\{v_{j,k}\}_{j=1}^{N_k}$ and $\{\omega_{j,k}\}_{j=1}^{N_k}$. Let Assumption 7 hold. Then the following hold almost surely for any $\mathbf{x}_k \in \mathcal{X}$.*

(a) $\mathbb{E}[\bar{w}_{k,N_k} \mid \mathbf{x}_k] = 0$.
(b) $\mathbb{E}[\|\bar{w}_{k,N_k}\|^2 \mid \mathbf{x}_k] \leq \frac{n^2 L_0^2}{N_k}$.

Proof Note that (a) holds in view of Lemma 9. Invoking Lemma 9, we may provide the following bound in an almost sure sense.

$$\begin{aligned}\mathbb{E}\left[\|\bar{w}_{k,N_k}\|^2 \mid \mathbf{x}_k\right] &= \mathbb{E}\left[\left\|g_{\eta_k,N_k}(\mathbf{x}_k) - \nabla_{\mathbf{x}} f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\ &= \mathbb{E}\left[\left\|\frac{\sum_{j=1}^{N_k} g_{\eta_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k} - \nabla_{\mathbf{x}} f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 \mid \mathbf{x}_k\right] \\ &\leq \frac{\sum_{j=1}^{N_k} \mathbb{E}\left[\left\|g_{\eta_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) - \nabla_{\mathbf{x}} f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2 | \mathbf{x}_k\right]}{N_k^2} \\ &\leq \frac{\sum_{j=1}^{N_k}\left(\mathbb{E}\left[\|g_{\eta_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})\|^2 | \mathbf{x}_k\right] - \left\|\nabla_{\mathbf{x}} f_{\eta_k}^{\mathbf{imp}}(\mathbf{x}_k)\right\|^2\right)}{N_k^2} \leq \frac{n^2 L_0^2}{N_k}.\end{aligned}$$

□

Lemma 12 [35, Lemma 4] *Consider the problem* (SMPEC$^{\mathbf{imp,2s}}$). *Suppose Assumptions 1–3, 7 hold. Suppose* $\{\mathbf{x}_k, \mathbf{z}_k\}$ *denote the sequence generated by Algorithm 7 in which the stepsize and smoothing sequences are defined as* $\eta_k = \frac{1}{k+1}$ *and* $\gamma_k = \frac{1}{2(k+1)}$, *and* $N_k = \lfloor (k+1)^a \rfloor$ *for* $k \geq 0$. *Suppose* $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq C$ *for some* $C > 0$. *Then the following holds.*

$$\mathbb{E}\left[f_{\eta_K}^{\mathbf{imp}}(\mathbf{z}_K) - f_{\eta_K}^{\mathbf{imp}}(\mathbf{x}^*)\right] \leq \frac{2}{\gamma_{K-1}(K-1)^2} \sum_{k=1}^{K-1} \frac{\gamma_k^2 k^2 n^2 L_0^2}{N_{k-1}} + \frac{2C^2}{\gamma_{K-1}(K-1)^2}. \tag{42}$$

We may now provide the main rate statement for the smoothed accelerated scheme by adapting [35, Thm. 5].

Proposition 5 (Rate statement for Algorithm 7) *Consider the problem* (SMPEC$^{\mathbf{imp,2s}}$). *Suppose Assumptions 1–3, 7 hold. Suppose* $\{\mathbf{x}_k, \mathbf{z}_k\}$ *denote the sequence generated by Algorithm 7 in which the stepsize and smoothing sequences are defined as* $\eta_k = \frac{1}{k+1}$ *and* $\gamma_k = \frac{1}{2(k+1)}$, *and* $N_k = \lfloor (k+1)^a \rfloor$ *for* $k \geq 0$. *Suppose* $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq C$ *for some* $C > 0$. *Then the following hold for* $a = 1 + \delta$ *where* $\delta > 0$. *Suppose* K_ϵ *is such that* $\mathbb{E}[f^{\mathbf{imp}}(\mathbf{z}_{K_\epsilon})] - f^* \leq \epsilon$. *Then the following holds.*

(a) *The iteration complexity in terms of zeroth-order gradient steps is* $\mathcal{O}(1/\epsilon)$.

(b) *We have* $\sum_{k=1}^{K_\epsilon} N_k \leq \mathcal{O}(1/\epsilon^{2+\delta})$ *implying that the sample complexity as well as the iteration complexity in terms of lower-level calls to the VI solver are both* $\mathcal{O}(1/\epsilon^{2+\delta})$.

Proof **(a)** From Lemma 12, we have that

$$\mathbb{E}\left[f_{\eta_K}^{\mathbf{imp}}(\mathbf{z}_K) - f_{\eta_K}^{\mathbf{imp}}(\mathbf{x}^*)\right] \leq \frac{2}{\gamma_{K-1}(K-1)^2} \sum_{k=1}^{K-1} \frac{\gamma_k^2 k^2 n^2 L_0^2}{N_{k-1}} + \frac{2C^2}{\gamma_{k-1}(K-1)^2}. \tag{43}$$

From Lemma 1 (v), we have that $f^{\mathbf{imp}}(\mathbf{x}) \leq f^{\mathbf{imp}}_{\eta_K}(\mathbf{x}) \leq f^{\mathbf{imp}}(\mathbf{x}) + \eta_K L_0$. Consequently, we have

$$\begin{aligned}\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{z}_K) - f^*\right] &\leq \mathbb{E}\left[f^{\mathbf{imp}}_{\eta_K}(\mathbf{z}_K) - f^{\mathbf{imp}}_{\eta_K}(\mathbf{x}^*)\right] + \eta_K L_0 \\ &\leq \frac{2}{\gamma_{K-1}(K-1)^2}\sum_{k=1}^{K-1}\frac{\gamma_k^2 k^2 n^2 L_0^2}{N_{k-1}} + \frac{2C^2}{\gamma_{K-1}(K-1)^2} \\ &\quad + \eta_K L_0 \leq \mathcal{O}\left(\frac{1}{K}\right),\end{aligned}$$

where we used $\eta_k = \frac{1}{k+1}$ and $\gamma_k = \frac{1}{2(k+1)}$, and $N_k = \lfloor (k+1)^a \rfloor$ where $a = 1+\delta$.

(b) The proof can be done in a similar vein to that of [35, Thm. 5] and thus, it is omitted.

□

Remark 11 Several points deserve emphasis. (i) The proposed scheme employs diminishing smoothing sequences rather than fixed, leading to asymptotic convergence guarantees, a key distinction from the scheme proposed in [59]. (ii) By adapting the framework employed for the inexact oracles, one may consider similar extensions to the accelerated framework. However, this would lead to bias in the gradient approximation and one would expect this to adversely affect the rate. This remains a goal of future study.

4.4 Nonconvex two-stage SMPEC

In this subsection, we address the two-stage model (SMPEC$^{\mathbf{imp,2s}}$) when the implicit function is nonconvex. The outline of the proposed zeroth-order scheme is given by Algorithm 8 in both inexact and exact variants. In the following we present the results for each of the two variants.

4.4.1 An inexact zeroth-order scheme

In the following, we present the rate and complexity result for the proposed inexact method for addressing the two-stage model in the nonconvex case.

Algorithm 8 $\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncnvx}}$: Variance-reduced zeroth-order method for nonconvex (SMPEC$^{\mathbf{2s}}$)

1: **input:** Given $\mathbf{x}_0 \in \mathcal{X}$, $\bar{\mathbf{x}}_0 := \mathbf{x}_0$, stepsize $\gamma > 0$, smoothing parameter $\eta > 0$, mini-batch sequence $\{N_k\}$ such that $N_k := k+1$, an integer K, a scalar $\lambda \in (0,1)$, and an integer R randomly selected from $\{\lceil \lambda K \rceil, \ldots, K\}$ using a uniform distribution
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: **for** $j = 1, \ldots, N_k$ **do**
4: Generate $v_{j,k} \in \eta\mathbb{S}$
5: Do one of the following.

- Inexact scheme: Call Alg. 6 twice to obtain $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega_{j,k})$ and $\mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k + v_{j,k}, \omega_{j,k})$
- Exact scheme: Evaluate $\mathbf{y}(\mathbf{x}_k, \omega_{j,k})$ and $\mathbf{y}(\mathbf{x}_k + v_{j,k}, \omega_{j,k})$

6: Evaluate the inexact or exact zeroth-order gradient approximation as follows.

$$g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) := \frac{n\Big(\tilde{f}(\mathbf{x}_k + v_{j,k}, \mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k + v_{j,k}, \omega_{j,k}), \omega_{j,k}) - \tilde{f}(\mathbf{x}_k, \mathbf{y}_{\tilde{\epsilon}_k}(\mathbf{x}_k, \omega_{j,k}), \omega_{j,k})\Big) v_{j,k}}{\|v_{j,k}\|\eta} \quad \text{(Inexact)}$$

$$g_{\eta}(\mathbf{x}_k, v_{j,k}, \omega_{j,k}) := \frac{n\Big(\tilde{f}(\mathbf{x}_k + v_{j,k}, \mathbf{y}(\mathbf{x}_k + v_{j,k}, \omega_{j,k}), \omega_{j,k}) - \tilde{f}(\mathbf{x}_k, \mathbf{y}(\mathbf{x}_k, \omega_{j,k}), \omega_{j,k})\Big) v_{j,k}}{\|v_{j,k}\|\eta} \quad \text{(Exact)}$$

7: **end for**
8: Evaluate the mini-batch zeroth-order gradient.

$$g_{\eta,N_k,\tilde{\epsilon}_k}(\mathbf{x}_k) := \frac{\sum_{j=1}^{N_k} g_{\eta,\tilde{\epsilon}_k}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k} \quad \text{(Inexact)}$$

$$g_{\eta,N_k}(\mathbf{x}_k) := \frac{\sum_{j=1}^{N_k} g_{\eta}(\mathbf{x}_k, v_{j,k}, \omega_{j,k})}{N_k} \quad \text{(exact)}$$

9: Update $\mathbf{x}_k$ as follows.

$$\mathbf{x}_{k+1} := \begin{cases} \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma g_{\eta,N_k,\tilde{\epsilon}_k}(\mathbf{x}_k)\right] & \text{(Inexact)} \\ \Pi_{\mathcal{X}}\left[\mathbf{x}_k - \gamma g_{\eta,N_k}(\mathbf{x}_k)\right] & \text{(Exact)} \end{cases}$$

10: **end for**
11: Return $\mathbf{x}_R$

Theorem 4 (Rate and complexity statements for inexact ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncnvx}}$)) *Consider Algorithms 8 and 6 for solving* (SMPEC$^{\mathbf{imp,2s}}$) *and suppose Assumptions 1 and 5 hold.* **(a)** *Given $\hat{\mathbf{x}}_k \in \mathcal{X}$, let $\mathbf{y}(\hat{\mathbf{x}}_k, \omega_{j,k})$ denote the unique solution of VI($\mathcal{Y}$, $G(\hat{\mathbf{x}}_k, \bullet, \omega_{j,k})$). Let $\mathbf{y}_{t_k}$ be generated by Algorithm 6. Then for suitably defined $\tilde{d} < 1$ and $B > 0$, the following holds for $t_k \geq 1$.*

$$\|\mathbf{y}_{t_k} - \mathbf{y}(\hat{\mathbf{x}}_k, \omega_{j,k})\|^2 \leq \tilde{\epsilon}_k \triangleq B\tilde{d}^{t_k}.$$

(b) *The following holds for any* $\gamma < \frac{\eta}{nL_0}$, $\ell \triangleq \lceil \lambda K \rceil$, *and all* $K > \frac{2}{1-\lambda}$.

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{n^2\gamma(1-2\ln(\lambda))\left(1-\frac{nL_0\gamma}{2\eta}\right)\left(\frac{4\tilde{L}_0^2 B}{\eta^2}+L_0^2\right)+\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right]-f^*+2L_0\eta}{\left(1-\frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1-\lambda)K}.$$

(c) *Suppose* $\gamma = \frac{\eta}{2nL_0}$ *and* $\eta = \frac{1}{L_0}$. *Let* $\epsilon > 0$ *be an arbitrary scalar and* K_ϵ *be such that* $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \epsilon$. *Then,*

(c-1) *the total number of upper-level projection steps on* $\mathcal{X}$ *is* $K_\epsilon = \mathcal{O}\left(n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1}\right)$.

(c-2) *the overall sample complexity of upper-level evaluations is* $\mathcal{O}\left(n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2}\right)$.

(c-3) *the total number of lower-level projection steps on* $\mathcal{Y}$ *is* $\mathcal{O}\left(\tau n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2} \ln(n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1})\right)$.

Proof **(a)** The proof of (a) is analogous to that of Theorem 3(a) and it is omitted.

(b) In view of the similarity between the results of Lemmas 9 and 10 with those of Lemmas 2 and 3, respectively, in a similar fashion to the proof of Theorem 3(b), we can obtain

$$\left(1-\frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(K-\ell)\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \left(1-\frac{nL_0\gamma}{2\eta}\right)\gamma\sum_{k=\ell}^{K-1}\left(\frac{8\tilde{L}_0^2 n^2\tilde{\epsilon}_k}{\eta^2}+\frac{2n^2L_0^2}{N_k}\right)+\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right]-f^*+2L_0\eta.$$

Next, we derive a bound on $\tilde{\epsilon}_k$. Note that from part (a), we have $\tilde{\epsilon}_k = B\tilde{d}^{t_k}$ where $t_k := \lceil \tau \ln(k+1) \rceil \geq \tau \ln(k+1)$. We have

$$(k+1)\tilde{\epsilon}_k \leq B\tilde{d}^{\tau\ln(k+1)}(k+1) = B\left(\tilde{d}^\tau \mathrm{e}\right)^{\ln(k+1)} \leq B,$$

where the last inequality is implied from $\tau \geq \frac{-1}{\ln(\tilde{d})}$ and $\tilde{d} < 1$. Thus, we have that $\tilde{\epsilon}_k \leq \frac{B}{k+1}$. Note that $K > \frac{2}{1-\lambda}$ implies $\ell \leq K-1$. From Lemma 13, using $\ell \geq 1$ we have $\sum_{k=\ell}^{K-1}\frac{1}{k+1} \leq \frac{1}{\ell+1}+\ln\left(\frac{K}{\ell+1}\right) \leq 0.5+\ln\left(\frac{N}{\lambda N+1}\right) \leq 0.5-\ln(\lambda)$. Also, $K-\ell \geq K-\lambda K = (1-\lambda)K$. Thus, we obtain

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \leq \frac{\left(1-\frac{nL_0\gamma}{2\eta}\right)2n^2\gamma\left(\frac{4\tilde{L}_0^2 B}{\eta^2}+L_0^2\right)(0.5-\ln(\lambda))+\mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right]-f^*+2L_0\eta}{\left(1-\frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1-\lambda)K}.$$

(c) The proofs of (c-1) and (c-2) are analogous to those of Theorem 2 (c-1) and (c-2), respectively. To show (c-3), note that the total number of lower-level projection steps is given by

$$\begin{aligned}\sum_{k=0}^{K_\epsilon} 2N_k t_k &= 2\sum_{k=0}^{K_\epsilon}(k+1)\lceil \tau \ln(k+1)\rceil \le 2\tau \int_1^{K_\epsilon}(x+1)(\ln(x+1)+1)dx \\ &= \mathcal{O}\left(\tau K_\epsilon^2 \ln(K_\epsilon)\right) \\ &= \mathcal{O}\left(\tau n^4 L_0^4 \tilde{L}_0^4 \epsilon^{-2} \ln(n^2 L_0^2 \tilde{L}_0^2 \epsilon^{-1})\right).\end{aligned}$$

□

4.4.2 An exact zeroth-order scheme

Here we present the rate and complexity results for the exact variant of Algorithm 8.

Corollary 4 (Rate and complexity statements for exact ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncnvx}}$)) *Consider Algorithms 8 (exact variant) for solving* (SMPEC$^{\mathbf{imp,2s}}$) *and suppose Assumptions 1 and 5 hold.*
(a) *The following holds for any* $\gamma < \frac{\eta}{nL_0}$, $\ell \triangleq \lceil \lambda K \rceil$, *and all* $K > \frac{2}{1-\lambda}$.

$$\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \le \frac{n^2 L_0^2 \gamma (0.5 - \ln(\lambda))\left(1 - \frac{nL_0\gamma}{2\eta}\right) + \mathbb{E}\left[f^{\mathbf{imp}}(\mathbf{x}_\ell)\right] - f^* + 2L_0\eta}{\left(1 - \frac{nL_0\gamma}{\eta}\right)\frac{\gamma}{4}(1-\lambda)K}.$$

(b) *Suppose* $\gamma = \frac{\eta}{2nL_0}$ *and* $\eta = \frac{1}{L_0}$. *Let* $\epsilon > 0$ *be an arbitrary scalar and* K_ϵ *be such that* $\mathbb{E}\left[\|G_{\eta,1/\gamma}(\mathbf{x}_R)\|^2\right] \le \epsilon$. *Then,*
(b-1) *The total number of upper-level projection steps on* $\mathcal{X}$ *is* $K_\epsilon = \mathcal{O}\left(n^2 L_0^2 \epsilon^{-1}\right)$.
(b-2) *The total sample complexity of upper-level is* $\mathcal{O}\left(n^4 L_0^4 \epsilon^{-2}\right)$.

Proof The proof can be done in a similar vein to that of Theorem 4 by noting that $\tilde{\epsilon}_k := 0$ in the exact variant. □

5 Numerical results

In this section, we demonstrate the proposed methodology by comparing the performance of the proposed scheme with sample-average approximation (SAA) schemes on a breadth of two-stage and single-stage SMPECs of varying structure and scale in Sects. 5.1 and 5.2, respectively. In Sect. 5.2, we also provide comparisons with the solvers `NLPEC` and `BARON`. We then provide confidence intervals in large-scale settings in Sect. 5.3 and conclude with a study of how the schemes perform on a set of test problems from the literature (Sect. 5.4). Implementations were developed in `MATLAB` on a PC with 16GB RAM and 6-Core Intel Core i7 processor (2.6 GHz).

5.1 Two-stage SMPECs

In this section, we apply the schemes on a stochastic Stackelberg–Nash–Cournot equilibrium problem which leads to a two-stage SMPEC. The deterministic setting of the problem is derived from [74]. Consider a market with N profit-maximizing firms by competing in Cournot (quantities) under the (Cournot) assumption that the remaining firms will hold their outputs at existing levels. In addition, there exists a leader, supplying the same product, that sets production levels by explicitly considering the reaction of the other N firms to its output variations. We assume that the ith Cournot firm (follower) supplies q_i units of the product while $f_i(q_i)$ denotes the cost of producing q_i units. In a similar fashion, suppose x denotes the output of the leader and let $f(x)$ denote the total cost. Next, let $p(\cdot, \omega)$ represent the random inverse demand curve. The N Cournot firms have sufficient capacity installed and can therefore wait to observe the quantities supplied by the leader as well as the realized demand function before making a decision on their supply quantities. For a given $x \geq 0$, let $\{q_1(x, \omega), \ldots, q_N(x, \omega)\}$ be the set of quantities for every $\omega \in \Omega$ where each $q_i(x, \omega)$ solve the following profit maximization problem assuming that $q_j(x, \omega)$, $j \neq i$ are fixed:

$$\max_{q_i \geq 0} \; q_i p\left(q_i + x + \sum_{j=1, j\neq i}^{N} q_j(x, \omega), \omega\right) - f_i(q_i). \tag{44}$$

Accordingly, let $Q(x, \omega) \triangleq \sum_{i=1}^{N} q_i(x, \omega)$. In addition, we assume there exists a capacity limit x^u for x. Then x^* is said to be a Stackelberg–Nash–Cournot equilibrium solution if x^* solves

$$\max_{0 \leq x \leq x^u} \mathbb{E}[xp(x + Q(x, \omega), \omega)] - f(x). \tag{45}$$

We consider the case of a linear demand curve with convex quadratic cost functions. Specifically, let $p(u, \omega) = a(\omega) - bu$ and let $f_i(q) = \frac{1}{2}cq^2$ for $i = 1, \ldots, N$, and $f(x) = \frac{1}{2}dx^2$. Under this condition, the follower's objective can be shown to be strictly concave in q^i [78]. Consequently, the concatenated necessary and sufficient equilibrium conditions of the follower-level game are given by the following conditions.

$$0 \leq q \perp F(q) - p(x + Q(x, \omega), \omega)\mathbf{1} - p'(x + Q(x, \omega), \omega)q \geq 0, \tag{46}$$

where $F(q) = \left(f_1'(q_1); \ldots; f_N'(q_N)\right)$. We observe that (46) is a strongly monotone linear complementarity problem for $x \geq 0$ and for every $\omega \in \Omega$. Consequently, $q : \mathbb{R}_+ \times \Omega \to \mathbb{R}_+^N$ is a single-valued map and is convex in its first argument for every ω if c_j is quadratic and convex [16, Prop. 4.2]. In fact, it can be claimed that $q(\cdot, \omega)$ is a piecewise C^2 and non-increasing function with $\partial_x q(x, \omega) \subset (-1, 0]$ for $X \geq 0$.

Table 3 Errors and time comparison of the three schemes with different parameters

			($\texttt{ZSOL}^{\mathbf{2s}}_{\texttt{cnvx}}$)		($\texttt{ZSOL}^{\mathbf{2s}}_{\texttt{acc,cnvx}}$)		SAA	
			$f^* - f(\bar{x}_K)$	Time	$f^* - f(x_K)$	Time	$f^* - f(\hat{x})$	Time
$N = 10$	$b = 1$	$c = 0.05$	1.2e−3	0.1	6.6e−5	1.4	5.4e−4	130.2
		$c = 0.1$	8.2e−4	0.1	4.8e−5	1.4	4.2e−4	109.2
	$b = 0.5$	$c = 0.05$	1.7e−3	0.1	7.0e−5	1.3	3.8e−4	122.5
		$c = 0.1$	1.2e−3	0.1	6.3e−5	1.4	2.2e−4	116.8
$N = 20$	$b = 1$	$c = 0.05$	4.5e−4	0.1	2.6e−5	1.5	2.6e−4	426.7
		$c = 0.1$	4.0e−4	0.1	1.3e−5	1.4	5.7e−4	443.1
	$b = 0.5$	$c = 0.05$	6.3e−4	0.1	2.3e−5	1.4	4.8e−4	419.1
		$c = 0.1$	4.2e−4	0.1	2.9e−5	1.5	3.1e−4	450.0
$N = 100$	$b = 1$	$c = 0.05$	9.9e−5	0.2	3.2e−6	4.3	–	–
		$c = 0.1$	2.3e−5	0.2	1.3e−6	4.4	–	–
	$b = 0.5$	$c = 0.05$	2.6e−4	0.2	4.7e−6	4.2	–	–
		$c = 0.1$	2.5e−5	0.2	1.4e−6	4.5	–	–
$N = 1000$	$b = 1$	$c = 0.05$	2.2e−5	0.6	3.6e−7	27.9	–	–
		$c = 0.1$	1.7e−6	0.6	8.3e−8	28.8	–	–
	$b = 0.5$	$c = 0.05$	2.5e−5	0.6	3.1e−7	29.1	–	–
		$c = 0.1$	1.4e−6	0.6	8.9e−8	28.4	–	–
$N = 10000$	$b = 1$	$c = 0.05$	1.0e−5	4.6	5.2e−7	403.5	–	–
		$c = 0.1$	6.0e−6	4.5	3.8e−8	392.4	–	–
	$b = 0.5$	$c = 0.05$	1.1e−5	4.7	5.6e−8	334.2	–	–
		$c = 0.1$	7.1e−6	4.6	2.7e−8	399.7	–	–

The errors and time in the table are based on averaging over 20 runs ('–' implies runtime > 3600 s)

Consider the leader's problem (45). Consequently, we have that

$$\mathbb{R}_+ \ni x \perp \mathbb{E}\left[-p(x + Q(x,\omega),\omega) + (1 + \partial_x Q(x,\omega))bx - a(\omega)\right] + \nabla_x f(x) \in \mathbb{R}_+.$$

This may be viewed as the following inclusion which has been shown to be monotone [16, Thm. 4.4].

$$\begin{aligned} 0 &\in \mathbb{E}[T(x,\omega)] + \mathcal{N}_{\mathbb{R}_+}, \\ \text{where } T(x,\omega) &\triangleq [-p(x + Q(x,\omega),\omega)\mathbf{1} - a(\omega)\mathbf{1}] + \nabla_x f(x) \\ &\quad + \{[(1 + \partial_x Q(x,\omega))bx]\}. \end{aligned}$$

Problem and algorithm parameters. Suppose there are $N = 10$ Cournot firms and $c = d = 0.1$. Furthermore, $b = 1$ and $a(\omega) \sim \mathcal{U}(7.5, 12.5)$ where $\mathcal{U}(l,u)$ denotes the uniform distribution on $[l,u]$. We choose $\gamma_k = \frac{1}{\sqrt{k+1}}$ and $\eta_k = \frac{1}{\sqrt{k+1}}$, $\forall k \geq 1$ in ($\texttt{ZSOL}^{\mathbf{2s}}_{\texttt{cnvx}}$) and $\gamma_k = \frac{1}{2(k+1)}$ and $\eta_k = \frac{1}{k+1}$, $\forall k \geq 1$ in ($\texttt{ZSOL}^{\mathbf{2s}}_{\texttt{acc,cnvx}}$). In addition, we choose sample size $N_k = \lfloor k^{1.01} \rfloor$.

Description of testing. We compare the performance of (ZSOL) and (acc-ZSOL) with Nesterov's fixed smoothing scheme under the same number of iterations in Fig. 1. Next we change the size and parameters of the original game to ascertain parametric sensitivity. In Table 3, we consider a set of 12 problems where the settings, the empirical errors, and elapsed time are shown in Table 3. Note that we have access to the true solution from [74] and this is employed for computing the sub-optimality metrics. In addition, to show the performance of our proposed schemes, we consider the (SAA) scheme (utilizing the average of 1000 samples) used in [16]. Let $(\omega_k)_{k=1}^K$ denote independent identically distributed (i.i.d.) samples. Then, with (SAA) we solve the following formulation of problem:

$$\max_{0\leq x\leq x^u} \frac{1}{K}\sum_{k=1}^{K}\left[x\cdot(a(\omega_k)-b\cdot(x+Q(x,\omega_k)))\right]-\tfrac{1}{2}dx^2$$
$$\text{subject to } 0\leq q_{i,k}\perp (c+2b)q_{i,k}-a(w_k)+b\cdot\left(x+\textstyle\sum_{j=1,j\neq i}^{N}q_{j,k}(x,\omega_k)\right)\geq 0,\ \forall i,k.$$

This problem allows for utilizing NLPEC [23] in GAMS to compute a solution. For comparison, we employ an alternative method to solve (SAA). (SAA) can be equivalently formulated as

$$\max_{0\leq x\leq x^u} \frac{1}{K}\sum_{k=1}^{K}[x\cdot(a(\omega_k)-b\cdot(x+Q(x,\omega_k)))]-\tfrac{1}{2}dx^2,$$

where $Q(x,\omega_k)\triangleq\sum_{i=1}^{N}q_i(x,\omega_k)$ and $q_i(x,\omega_k)$ is the solution to the following optimization problem:

$$\max_{q_i\geq 0}\ q_i\,p\left(q_i+x+\sum_{j=1,j\neq i}^{N}q_j(x,\omega_k),\omega_k\right)-f_i(q_i).$$

This problem allows for utilizing gradient based methods to compute a solution. The results are shown in 4. Next, we provide some key insights from our testing.

Insights.

(i) *Scalability.* Both ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx}}$) and ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{acc,cnvx}}$) show far better scalability in terms of N with modest impact on accuracy and run-time. (SAA) schemes on the other hand grow by a factor of 10 when number of firms double. In fact, for $N=20$, the (SAA) framework requires CPU time which is between 50 and 100 times greater than that required by the zeroth-order schemes. (SAA) schemes could not produce solutions for $N\geq 100$ in our tests while our proposed schemes can contend with problems with $N=10,000$ within 5s in the unaccelerated regime. The lack of scalability tends to be less surprising since the sample-average subproblems require solving MPECs with $\mathcal{O}(N)$ constraints and as N becomes large, direct solutions become challenging, as reflected by the computational times. We observe that the gradient based approach that uses sample-averages appears to scale better than NLPEC. However, we still see

Table 4 Errors and time comparison of (SAA) with different solution methods

			SAA(NLPEC)		SAA(Gradient)	
			$f^* - f(\hat{x})$	Time	$f^* - f(\hat{x})$	Time
$N = 10$	$b = 1$	$c = 0.05$	5.4e−4	130.2	4.6e−4	1.0
		$c = 0.1$	4.2e−4	109.2	4.5e−4	1.0
	$b = 0.5$	$c = 0.05$	3.8e−4	122.5	3.3e−4	1.0
		$c = 0.1$	2.2e−4	116.8	2.4e−4	1.0
$N = 20$	$b = 1$	$c = 0.05$	2.6e−4	426.7	3.1e−4	1.1
		$c = 0.1$	5.7e−4	443.1	4.2e−4	1.1
	$b = 0.5$	$c = 0.05$	4.8e−4	419.1	5.6e−4	1.1
		$c = 0.1$	3.1e−4	450.0	3.8e−4	1.1
$N = 100$	$b = 1$	$c = 0.05$	–	–	1.1e−4	5.5
		$c = 0.1$	–	–	2.8e−5	5.5
	$b = 0.5$	$c = 0.05$	–	–	3.0e−4	5.5
		$c = 0.1$	–	–	3.2e−5	5.6
$N = 1000$	$b = 1$	$c = 0.05$	–	–	2.3e−5	324.7
		$c = 0.1$	–	–	1.9e−6	312.8
	$b = 0.5$	$c = 0.05$	–	–	2.6e−5	306.2
		$c = 0.1$	–	–	2.1e−6	316.5

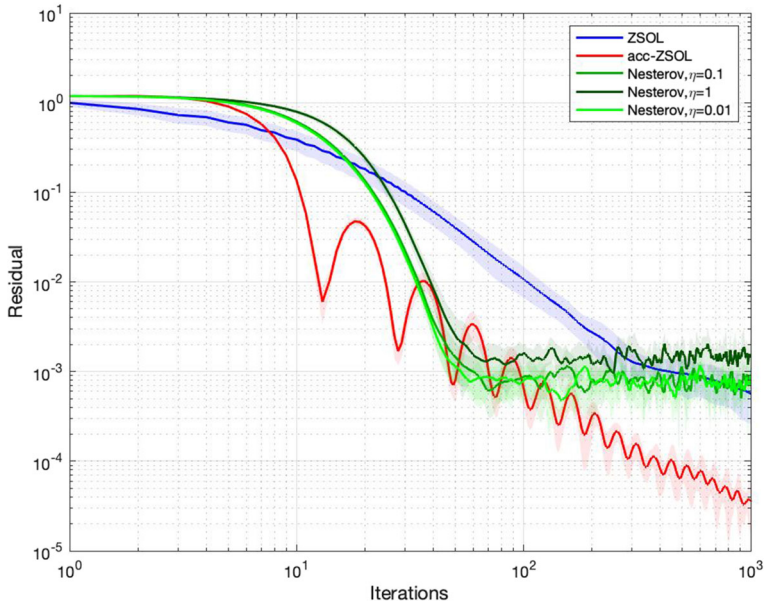

Fig. 1 Comparision of ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx}}$) and ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{acc,cnvx}}$) with acceleration with fixed smoothing (Nesterov) on convex (SMPEC$^{\mathbf{2s}}$)

Table 5 Errors of ($\texttt{ZSOL}^{\textbf{2s}}_{\text{cnvx}}$) with various γ_k and η_k

	(a, b)	(0.5, 0.5)	(0.5, 0.7)	(0.5, 0.9)	(0.7, 0.4)	(0.9, 0.2)
$f^* - f(\bar{x}_K)$	$N = 10$	1.2e−3	1.7e−3	1.5e−3	1.9e−3	7.7e−2
	$N = 100$	2.5e−5	3.0e−5	2.6e−5	1.1e−3	1.6e−2
	$N = 1000$	1.4e−6	4.8e−7	4.4e−7	2.9e−4	7.1e−4

a difference in performance and quality between the gradient-enabled SAA scheme and the proposed implicit SA framework.
(ii) *Accuracy.* The accelerated scheme provides nearly 10 times more accurate solutions than the unaccelerated scheme at a modest computational cost. This is aligned with the superior error bounds of such schemes compared to their unaccelerated counterparts.
(iii) *Comparison of accelerated schemes.* Figure 1 demonstrates the benefits of diminishing smoothing sequences as the scheme suggested in [59] degenerates for different values of the fixed smoothing parameter. Notably, ($\texttt{ZSOL}^{\textbf{2s}}_{\text{acc,cnvx}}$) shows no such degeneration and progressively improves in function value. We notice in Table 3, ($\texttt{ZSOL}^{\textbf{2s}}_{\text{acc,cnvx}}$) takes longer than ($\texttt{ZSOL}^{\textbf{2s}}_{\text{cnvx}}$) with the same number iterations, arising from the fact that ($\texttt{ZSOL}^{\textbf{2s}}_{\text{acc,cnvx}}$) utilizes an increasing sample size and solves more lower-level problems than ($\texttt{ZSOL}^{\textbf{2s}}_{\text{cnvx}}$).
(iv) *Performance of* ($\texttt{ZSOL}^{\textbf{2s}}_{\text{cnvx}}$) *with various* γ_k *and* η_k. As shown in Table 5, we compare the results generated by ($\texttt{ZSOL}^{\textbf{2s}}_{\text{cnvx}}$) with various values of (a, b) used in $\gamma_k := \frac{\gamma_0}{(k+1)^a}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$. As shown in the table, for this particular problem, we find that smaller a ($a = 0.5$) generates better results in ($\texttt{ZSOL}^{\textbf{2s}}_{\text{cnvx}}$). When the size of problem is large ($N = 1000$), fixing $a = 0.5$, larger values of b lead to smaller residuals.

5.2 Single-stage SMPECs

We consider both the convex and the nonconvex regimes next.

5.2.1 A convex implicit function

First, we consider a single-stage SMPEC where the the lower level is a parametrized stochastic variational inequality, i.e., given x, the lower-level problem is a noncooperative game in which the ith player solves the following problem.

$$\max_{q_i \geq 0} \mathbb{E}\left[q_i(a(\omega) - b(q_i + x + \textstyle\sum_{j \neq i} q_j(x))\right] - \frac{1}{2} c q_i^2,$$

Accordingly, the upper-level problem in x is defined as follows

$$\max_{0 \leq x \leq x^u} \mathbb{E}\left[x(a(\xi) - b(x + \textstyle\sum_{i=1}^N q_i(x)))\right] - \frac{1}{2} d x^2.$$

Table 6 Comparison of ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) and (SAA) (Convex implicit function)

			($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$)		SAA	
			$f^* - f(\bar{x}_K)$	Time	$f^* - f(\hat{x})$	Time
$N = 10^2$	$b = 0.01$	$c = 3$	6.9e−4	0.1	2.2e−4	0.05
		$c = 5$	3.7e−4	0.1	2.4e−4	0.05
	$b = 0.02$	$c = 3$	8.1e−4	0.1	7.3e−4	0.05
		$c = 5$	3.5e−4	0.1	4.0e−4	0.05
$N = 10^3$	$b = 0.01$	$c = 3$	7.0e−4	0.4	7.0e−4	1.2
		$c = 5$	4.3e−4	0.4	5.0e−4	1.1
	$b = 0.02$	$c = 3$	8.0e−4	0.4	6.8e−4	1.2
		$c = 5$	4.7e−4	0.4	4.2e−4	1.2
$N = 10^4$	$b = 0.01$	$c = 3$	5.1e−4	5.8	7.3e−4	88.6
		$c = 5$	2.5e−4	5.2	5.4e−4	85.7
	$b = 0.02$	$c = 3$	6.4e−4	5.6	4.3e−4	93.5
		$c = 5$	3.1e−4	5.3	4.7e−4	87.3
$N = 10^5$	$b = 0.01$	$c = 3$	8.7e−4	45.6	–	–
		$c = 5$	6.5e−4	47.1	–	–
	$b = 0.02$	$c = 3$	9.7e−4	46.3	–	–
		$c = 5$	7.5e−4	46.7	–	–

The errors and time in the table are based on averaging over 20 runs ('–' implies runtime > 3600 s)

Since the lower-level equilibrium problem has a unique solution (since it is characterized by a strongly monotone map), the resulting implicit function can be shown to be convex.

Algorithm and Problem parameters. We assume $b = 0.01$ and $c = 3$ here, other parameters are the same as in the previous section. It can be shown that $\mu_F = 3.01$ and $L_F = 3.11$. We assume that $\gamma_k = \frac{1}{\sqrt{k+1}}$ and $\eta_k = \frac{1}{\sqrt{k+1}}$ for ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$). In ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$), we run 10^3 iterations. In the lower-level's variance-reduced stochastic approximation scheme, we choose steplength $\alpha = 0.15$, sampling rate $\rho = \frac{1}{1.5}$ and the sample size $M_t = \lceil 10^{-4} \cdot 1.5^t \rceil$. Thus we may calculate that $\tau \geq 4.9$ and then we choose $t_k = \lceil 5 \ln(k+1) \rceil$. In Fig. 2, we show the trajectories for ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) under various algorithm parameters.

Again, we compare the errors and time between ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) and (SAA) in Table 6. Here, with (SAA) we solve the following optimization problem

$$\underset{0 \leq x \leq x^u}{\text{maximize}} \ \frac{1}{K} \sum_{k=1}^{K} \left[x(a(\omega_k) - b(x + Q(x))) \right] - \frac{1}{2} d x^2$$

$$\text{subject to } 0 \leq q_i \perp \frac{1}{L} \sum_{\ell=1}^{L} \left[(c+2b) q_i - a(w_\ell) + b \left(x + \sum_{j=1, j \neq i}^{N} q_j(x) \right) \right] \geq 0, \ \forall i.$$

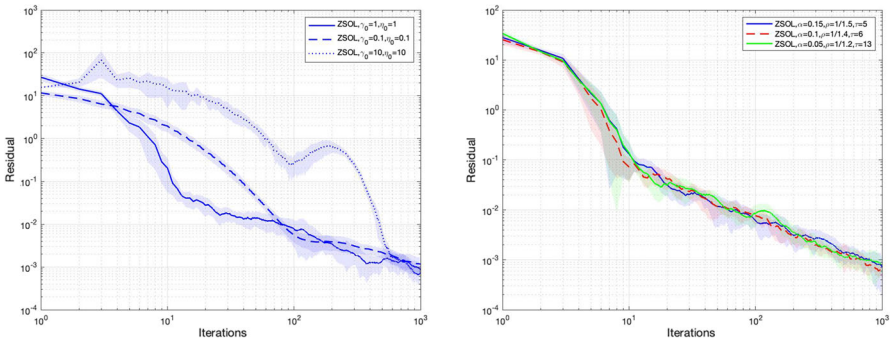

Fig. 2 Trajectories for ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) on the convex SMPEC$^{\mathbf{1s}}$

In (SAA), we use 10^3 samples in both the upper and lower-level problems. We also employ a gradient based method (Fig. 7) to solve the following equivalent (SAA) model:

$$\max_{0\le x\le x^u} \frac{1}{K}\sum_{k=1}^{K}\Big[x(a(\omega_k)-b(x+Q(x)))\Big]-\tfrac{1}{2}dx^2,$$

where $Q(x)\triangleq\sum_{i=1}^{N}q_i(x)$ and $q_i(x)$ is the solution to the following optimization problem:

$$\max_{q_i\ge 0}\mathbb{E}\Big[q_i(a(\omega))-b\Big(q_i+x+\sum_{j\ne i}q_j(x)\Big)\Big]-\tfrac{1}{2}cq_i^2.$$

Insights.
(i) *Scalability.* We observe that the CPU times for ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) grow by a factor of approximately 450 when N grows by a factor of 1000 (from 10^2 to 10^5); however (SAA) schemes show a growth in CPU time of 1770 when N grows by a factor of 100 (from 10^2 to 10^4). In fact, (SAA) schemes cannot process problems for $N=10^5$ in the prescribed time.
(ii) *Accuracy.* Both approaches provide similar accuracy but zeroth-order schemes require less than 6s in CPU time when $N=10^4$ while the (SAA) framework requires approximately 85s. The accuracy of ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) is relatively robust to changing steplength and sampling rates at the lower-level but does tend to be sensitive to changing the initial steplength at the upper-level; however, as the scheme progresses, the impact of initial steplengths tends to be muted.

5.2.2 A nonconvex implicit function

The second example, inspired from [3], is a bilevel problem with a strongly monotone mapping in the lower-level. We add a stochastic component in the lower-level to make

Table 7 Comparison of (SAA) with different solution methods

			SAA(NLPEC)		SAA(Gradient)	
			$f^* - f(\hat{x})$	Time	$f^* - f(\hat{x})$	Time
$N = 10^2$	$b = 0.01$	$c = 3$	2.2e−4	0.05	3.9e−4	0.4
		$c = 5$	2.4e−4	0.05	2.6e−4	0.4
	$b = 0.02$	$c = 3$	7.3e−4	0.05	5.9e−4	0.4
		$c = 5$	4.0e−4	0.05	3.7e−4	0.4
$N = 10^3$	$b = 0.01$	$c = 3$	7.0e−4	1.2	6.0e−4	2.5
		$c = 5$	5.0e−4	1.1	4.4e−4	2.5
	$b = 0.02$	$c = 3$	6.8e−4	1.2	5.9e−4	2.6
		$c = 5$	4.2e−4	1.2	3.8e−4	2.6
$N = 10^4$	$b = 0.01$	$c = 3$	7.3e−4	88.6	5.9e−4	25.3
		$c = 5$	5.4e−4	85.7	4.5e−4	25.3
	$b = 0.02$	$c = 3$	4.3e−4	93.5	5.2e−4	25.2
		$c = 5$	4.7e−4	87.3	4.2e−4	25.9
$N = 10^5$	$b = 0.01$	$c = 3$	–	–	6.7e−4	94.7
		$c = 5$	–	–	5.4e−4	95.0
	$b = 0.02$	$c = 3$	–	–	8.1e−4	96.3
		$c = 5$	–	–	6.0e−4	95.2

The errors and time in the table are based on averaging over 20 runs ('–' implies runtime > 3600 s)

Table 8 Errors comparison of the three schemes with different parameters

		$\texttt{ZSOL}^{\mathbf{1s}}_{\text{ncvx}}$	NLPEC	BARON
		$f(x_K)$	Stationary point	Global optimum
$(a, b) = (1, 0)$	$(c, d) = (1, 1)$	− 7.50	− 7.20	− 7.50
	$(c, d) = (2, 2)$	− 9.23	− 9.04	− 9.23
	$(c, d) = (3, 3)$	− 9.25	− 9.10	− 9.25
$(a, b) = (5, 0)$	$(c, d) = (1, 1)$	− 11.50	− 7.20	− 11.50
	$(c, d) = (2, 2)$	− 13.23	− 9.04	− 13.23
	$(c, d) = (3, 3)$	− 13.25	− 9.10	− 13.25
$(a, b) = (10, 0)$	$(c, d) = (1, 1)$	− 16.48	− 7.20	− 16.50
	$(c, d) = (2, 2)$	− 18.20	− 9.04	− 18.23
	$(c, d) = (3, 3)$	− 18.23	− 9.10	− 18.25

The errors of ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{ncvx}}$) are based on averaging over 20 runs

the mapping expectation-valued. Formally, this problem is defined as follows.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & -x_1^2 - 3x_2 - 4y_1(x) + (y_2(x))^2 \\ \text{subject to} \quad & x_1^2 + 2x_2 \leq 4, \quad 0 \leq x_1 \leq 1, \quad 0 \leq x_2 \leq 2, \end{aligned}$$

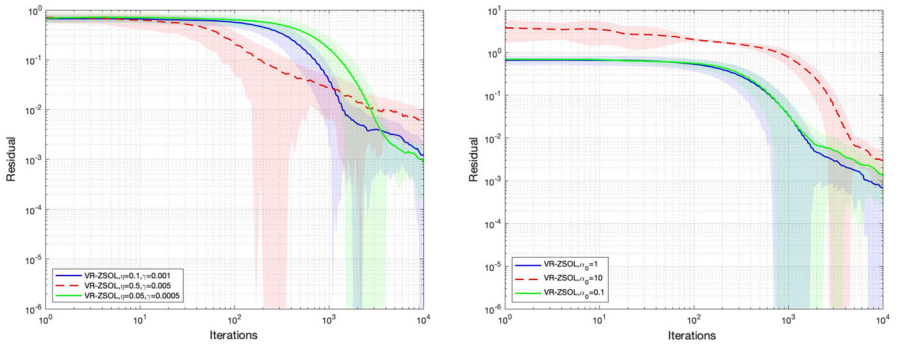

Fig. 3 Trajectories for ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$) on the nonconvex (SMPEC[1s])

where $y(x)$ is a solution to the following parametrized optimization problem.

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & \mathbb{E}\left[2x_1^2 + y_1^2 + y_2^2 - \xi(\omega)y_2\right] \\ \text{subject to} \quad & x_1^2 - 2x_1 + x_2^2 - 2y_1 + y_2 \geq -3,\ x_2 + 3y_1 - y_2 \geq 4,\ y_1 \geq 0, y_2 \geq 0. \end{aligned}$$

Problem and algorithm parameters. We assume $\xi(\omega) \sim \mathcal{U}(4, 6)$ and run ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$) for 10^4 iterations, choosing $\eta = 10^{-2}$ and $\gamma = 10^{-3}$ in ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$). In addition, we choose $\alpha_0 = 1$ and $\alpha_t = \frac{\alpha_0}{t+0.01}$ for $t = 0, 1, \ldots, t_k - 1$ in the stochastic approximation method applied to the lower-level. We compare the performance of ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$) on this problem in Fig. 3 for varying algorithm parameters, all of which suggest that the resulting sequences steadily converge to the global minimizer. To test the power of ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$) on different problems, we change the objective function of upper-level and lower-level to $-ax_1^2 - bx_2^2 - 3x_2 - 4y_1 + y_2^2$ and $\mathbb{E}[2x_1^2 + cy_1^2 + dy_2^2 - \xi(\omega)y_2]$, respectively. Then we vary the values of a, b, c and d. For comparison, we also run each problem using solvers `NLPEC` and `BARON` [69, 77] on the NEOS Server [14, 17, 29]. We record the empirical errors of each scheme for 9 different settings, as shown in Table 8. In ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$), we use 10^4 samples in each test problem.

Insights.

Global minimizers. From Fig. 3, we observe that while all of the implementations perform well, large initial steplengths at the lower-level tend to lead to a relatively worse behavior compared to more modest steplengths. Table 8 is instructive in that it shows that ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{ncvx}}$) produces values close to the global minimum as obtained by `BARON` for all nine problem instances. Notably, solvers such as `NLPEC` are equipped with convergence guarantees to stationary points and provide somewhat poorer values upon termination.

5.3 Confidence intervals for high-dimensional problems

To validate the effectiveness of solutions generated by ($\texttt{ZSOL}^{\mathbf{1s}}_{\texttt{cnvx}}$) and ($\texttt{ZSOL}^{\mathbf{2s}}_{\texttt{cnvx}}$), we construct 95% confidence intervals for large-scale test problems from Tables 3

Table 9 Errors and confidence intervals for high dimensional problems from Tables 3 and 6

			($\texttt{ZSOL}^{\mathbf{2s}}_{\text{cnvx}}$) [Table 3], ($\texttt{ZSOL}^{\mathbf{1s}}_{\text{cnvx}}$) [Table 6]		($\texttt{ZSOL}^{\mathbf{2s}}_{\text{acc,cnvx}}$)	
			$f^* - f(\bar{x}_K)$	CI	$f^* - f(x_K)$	CI
Table 3	$b = 1$	$c = 0.05$	1.0e−5	[0.9e−5, 1.1e−5]	5.2e−7	[5.0e−7, 5.4e−7]
		$c = 0.1$	6.0e−6	[5.9e−6, 6.1e−6]	3.8e−8	[3.4e−8, 4.2e−8]
$N = 10^4$	$b = 0.5$	$c = 0.05$	1.1e−5	[1.0e−5, 1.2e−5]	5.6e−8	[5.2e−8, 6.0e−8]
		$c = 0.1$	7.1e−6	[7.0e−6,7.2e−6]	2.7e−8	[2.4e−8, 3.0e−8]
Table 6	$b = 0.01$	$c = 3$	8.7e−4	[7.5e−4, 9.9e−4]	n/a	n/a
		$c = 5$	6.5e−4	[5.9e−4, 7.1e−4]	n/a	n/a
$N = 10^5$	$b = 0.02$	$c = 3$	9.7e−4	[8.0e−4, 1.1e−3]	n/a	n/a
		$c = 5$	7.5e−4	[6.4e−4, 8.6e−4]	n/a	n/a

Table 10 Results comparison with solutions from the literature

Problem		($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncvx}}$)		Literature	
		f^*	x^*	f^*	x^*
Problem 1	$L = 150, \gamma = 1.0$	− 343.35	55.57	− 343.35	55.55
	$L = 150, \gamma = 1.1$	− 203.15	42.57	− 203.15	42.54
	$L = 150, \gamma = 1.3$	− 68.14	24.19	− 68.14	24.14
Problem 2		− 1.00	(0.50,0.50)	− 1.00	(0.50,0.50)
Problem 3		0.01	(0.00,0.00)	0.01	(0.00,0.00)
Problem 4		0.00	(5.00,8.99)	0.00	(5.00,9.00)
Problem 5	$0.5((y_1 - 3)^2 + (y_2 - 4)^2)$	3.20	4.06	3.20	4.06
	$0.5((y_1 - 3)^2 + (y_2 - 4)^2 + (y_3 - 1)^2)$	3.45	5.13	3.45	5.15
	$0.5((y_1 - 3)^2 + (y_2 - 4)^2 + 10y_4^2)$	4.60	2.39	4.60	2.39

and 6. The results are shown in Table 9. Note that ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{acc,cnvx}}$) can process two-stage SMPECs. All confidence intervals presented are relatively narrow, validating the quality of corresponding solutions.

5.4 Additional tests on deterministic and two-stage stochastic MPECs

We test our schemes on test problems from the literature. In all of the test problems, the lower-level parametrized VI is strongly monotone, implying that the lower-level decision is uniquely determined by a $\mathbf{x} \in \mathcal{X}$.

Problem and algorithm parameters. The problems and their parameters are described in Appendix. We use the same algorithm parameters as those in 5.2.2(II). In Table 10, we compare the results generated by ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncvx}}$) and those from the literature, while in Table 11, we extend some of the existing problems to their stochastic counterparts with larger dimensions.

Insights.

Table 11 Results of high-dimensional counterparts

Problem	N	($\texttt{ZSOL}^{\mathbf{2s}}_{\texttt{ncvx}}$)					SAA		
		$\hat{f}(x_K)$	CI	Time	$\underline{lb}$	CI	$\hat{f}(\hat{x})$	CI	Time
Problem 1	5	−462.6	[−463.1, −462.1]	0.8	−462.8	[−464.0, −461.5]	−461.9	[−463.1, −460.7]	5.3
	10	−174.4	[−174.6, −174.2]	0.9	−174.7	[−175.2, −174.2]	−174.2	[−174.8, −173.6]	23.3
	100	−5.101	[−5.105, −5.097]	1.3	–	–	–	–	–
	1000	−0.071	[−0.072, −0.071]	5.2	–	–	–	–	–
Problem 2	2	−0.882	[−0.883, −0.881]	0.6	−0.883	[−0.886, −0.880]	−0.882	[−0.886, −0.878]	4.2
	10	−4.408	[−4.410, −4.406]	0.9	−4.408	[−4.414, −4.402]	−4.406	[−4.414, −4.398]	29.6
	100	−44.07	[−44.08, −44.07]	5.5	–	–	–	–	–
	1000	−439.7	[−439.7, −439.7]	98.1	–	–	–	–	–

(i) *Scalability.* Again, ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncvx}}$) shows far better scalability in terms of N with modest impact on accuracy and run-time. For both problems in Table 11, (SAA) schemes take around 5–20 times more time on small scale problems while when $N \geq 100$ on the other hand, no solutions are produced within the imposed time limit.
(ii) *Accuracy.* For deterministic MPECs, ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncvx}}$) provides almost the same solutions as the globally optimal solutions in all problems from the literature, which shows both efficacy and wide applicability of ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncvx}}$). In high-dimensional SMPECs, ($\texttt{ZSOL}^{\mathbf{2s}}_{\text{ncvx}}$) provides similar accuracy as (SAA) but takes far less computational time.

6 Concluding remarks

Motivated by the apparent lacuna in non-asymptotic rate guarantees and efficient first/zeroth-order schemes for MPECs, we consider a subclass of stochastic MPECs where the parametrized lower-level equilibrium problem is given by a deterministic/stochastic variational inequality (VI) problem whose mapping is strongly monotone, uniformly in upper-level decisions. Under suitable assumptions, the implicit objective is Lipschitz continuous over a compact and convex feasibility set, paving the way for developing a gradient-free locally randomized smoothing framework applied to the implicit form the SMPEC. This avenue allows for developing complexity guarantees in settings where the implicit objective is either convex or nonconvex, the lower-level oracle is exact (allowing for accelerated schemes in convex regimes) or inexact (requiring the use of stochastic approximation to compute an inexact lower-level decisions). We believe that this is but the first step in developing a comprehensive zeroth-order foundation for contending with SMPECs under far weaker assumptions. Possible extensions include settings where the lower-level map is merely monotone or possibly non-monotone.

7 Appendix

Lemma 13 (cf. Lemma 10 in [82] and Lemma 2.14 in [40]) *Let ℓ and N be arbitrary integers where $0 \leq \ell \leq N-1$. The following hold.*

(a) $\ln\left(\frac{N+1}{\ell+1}\right) \leq \sum_{k=\ell}^{N-1} \frac{1}{k+1} \leq \frac{1}{\ell+1} + \ln\left(\frac{N}{\ell+1}\right)$.

(b) *If $0 \leq \alpha < 1$, then for any $N \geq 2^{\frac{1}{1-\alpha}} - 1$, we have* $\frac{(N+1)^{1-\alpha}}{2(1-\alpha)} \leq \sum_{k=0}^{N} \frac{1}{(k+1)^{\alpha}} \leq \frac{(N+1)^{1-\alpha}}{1-\alpha}$.

Lemma 14 (Theorem 6, p. 75 in [41]) *Let $\{u_t\} \subset \mathbb{R}^n$ denote a sequence of vectors where $\lim_{t\to\infty} u_t = \hat{u}$. Also, let $\{\alpha_k\}$ denote a sequence of strictly positive scalars such that $\sum_{k=0}^{\infty} \alpha_k = \infty$. Suppose $v_k \in \mathbb{R}^n$ is defined by $v_k \triangleq \frac{\sum_{t=0}^{k} \alpha_t u_t}{\sum_{t=0}^{k} \alpha_t}$ for all $k \geq 0$. Then, $\lim_{k\to\infty} v_k = \hat{u}$.*

Lemma 15 (cf. [65]) *Let v_k, u_k, α_k, and β_k be nonnegative random variables, and let the following relations hold almost surely:*

$$E\Big[v_{k+1} \mid \tilde{\mathcal{F}}_k\Big] \le (1+\alpha_k)v_k - u_k + \beta_k \quad \text{for all} k, \quad \sum_{k=0}^{\infty}\alpha_k < \infty, \quad \sum_{k=0}^{\infty}\beta_k < \infty,$$

where $\tilde{\mathcal{F}}_k$ denotes the collection $v_0, \ldots, v_k, u_0, \ldots, u_k, \alpha_0, \ldots, \alpha_k, \beta_0, \ldots, \beta_k$. Then, we have almost surely $\lim_{k\to\infty} v_k = v$ and $\sum_{k=0}^{\infty} u_k < \infty$, where $v \ge 0$ is some random variable.

Proof of Lemma 8 We use induction on k for $k \ge 0$. We have $e_0 = \frac{\Gamma e_0}{0+\Gamma} \le \frac{\max\left\{\frac{\beta\gamma^2}{\alpha\gamma-1}, \Gamma e_0\right\}}{0+\Gamma}$ implying that the hypothesis statement holds for $k = 0$. Let us assume that $e_k \le \frac{\theta_0}{k+\Gamma}$ for some $k \ge 0$ where $\theta_0 \triangleq \max\left\{\frac{\beta\gamma^2}{\alpha\gamma-1}, \Gamma e_0\right\}$. Let the induction hypothesis hold for $k \ge 0$. We show that it holds for $k+1$ as well. We have

$$\begin{aligned}
&\theta_0 \ge \tfrac{\beta\gamma^2}{\alpha\gamma-1} \Rightarrow \theta_0 \le \gamma(\theta_0\alpha - \beta\gamma) \Rightarrow \tfrac{\theta_0}{k+\Gamma} \le \tfrac{\gamma(\theta_0\alpha-\beta\gamma)}{k+\Gamma} \Rightarrow \tfrac{\theta_0}{k+\Gamma+1} \le \tfrac{\gamma(\theta_0\alpha-\beta\gamma)}{k+\Gamma} \\
&\Rightarrow \tfrac{\theta_0}{(k+\Gamma+1)(k+\Gamma)} \le \tfrac{\gamma(\theta_0\alpha-\beta\gamma)}{(k+\Gamma)^2} \Rightarrow \theta_0\left(\tfrac{1}{k+\Gamma} - \tfrac{1}{k+\Gamma+1}\right) \\
&\le \tfrac{\gamma(\theta_0\alpha-\beta\gamma)}{(k+\Gamma)^2} \Rightarrow \tfrac{\theta_0}{k+\Gamma} - \tfrac{\gamma(\theta_0\alpha-\beta\gamma)}{(k+\Gamma)^2} \le \tfrac{\theta_0}{k+\Gamma+1} \\
&\Rightarrow \left(1 - \alpha\tfrac{\gamma}{k+\Gamma}\right)\tfrac{\theta_0}{k+\Gamma} + \tfrac{\beta\gamma^2}{(k+\Gamma)^2} \le \tfrac{\theta_0}{k+\Gamma+1} \Rightarrow (1-\alpha\gamma_k)\tfrac{\theta_0}{k+\Gamma} + \beta\gamma_k^2 \le \tfrac{\theta_0}{k+\Gamma+1} \\
&\Rightarrow (1-\alpha\gamma_k)\, e_k + \beta\gamma_k^2 \le \tfrac{\theta_0}{k+\Gamma+1} \Rightarrow e_{k+1} \le \tfrac{\theta_0}{k+\Gamma+1}.
\end{aligned}$$

□

Academic examples and their stochastic counterparts in Sect. 5.4

Problem 1. **This problem is described in** [61, **Definition 4.1**]

$$f(\mathbf{x}, \mathbf{y}) = r_1(x) - xp(x + y_1 + y_2 + y_3 + y_4),$$

where $r_i(v) = c_i v + \frac{\beta_i}{\beta_i+1} K_i^{1/\beta_i} v^{(1+\beta_i)/\beta_i}$, $p(Q) = 5000^{1/\gamma} Q^{-1/\gamma}$, c_i, β_i, K_i, $i = 1, \ldots, 5$ are given positive parameters in Table 12, γ is a positive parameter, $Q = x + y_1 + y_2 + y_3 + y_4$.

$$\begin{aligned}
&\mathcal{X} = \{0 \le x \le L\}. \\
&F(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \nabla r_2(y_1) - p(Q) - y_1\nabla p(Q) \\ \vdots \\ \nabla r_5(y_4) - p(Q) - y_4\nabla p(Q) \end{pmatrix}. \\
&\mathcal{Y} = \{0 \le y_j \le L, \quad j = 1, 2, 3, 4\}.
\end{aligned}$$

The following three examples were tested in [20, 61].
Problem 2.

$$
\begin{aligned}
f(\mathbf{x},\mathbf{y}) &= x_1^2 - 2x_1 + x_2^2 - 2x_2 + y_1^2 + y_2^2.\\
\mathcal{X} &= \{0 \le x_i \le 2, \quad i = 1,2\}.\\
F(\mathbf{x},\mathbf{y}) &= \begin{pmatrix} 2y_1 - 2x_1 \\ 2y_2 - 2x_2 \end{pmatrix}.\\
\mathcal{Y} &= \{(y_j - 1)^2 \le 0.25, \quad j = 1,2\}.
\end{aligned}
$$

Problem 3.

$$
\begin{aligned}
f(\mathbf{x},\mathbf{y}) &= 2x_1 + 2x_2 - 3y_1 - 3y_2 - 60 + R[\max\{0, x_1 + x_2 + y_1 - 2y_2 - 40\}]^2.\\
\mathcal{X} &= \{0 \le x_i \le 50, \quad i = 1,2\}.\\
F(\mathbf{x},\mathbf{y}) &= \begin{pmatrix} 2y_1 - 2x_1 + 40 \\ 2y_2 - 2x_2 + 40 \end{pmatrix}.\\
\mathcal{Y} &= \{-10 \le y_j \le 20,\ x_j - 2y_j - 10 \ge 0, \quad j = 1,2\}.
\end{aligned}
$$

Problem 4.

$$
\begin{aligned}
f(\mathbf{x},\mathbf{y}) &= \tfrac{1}{2}((x_1 - y_1)^2 + (x_2 - y_2)^2).\\
\mathcal{X} &= \{0 \le x_i \le 10, \quad i = 1,2\}.\\
F(\mathbf{x},\mathbf{y}) &= \begin{pmatrix} -34 + 2y_1 + \frac{8}{3}y_2 \\ -24.25 + 1.25y_1 + 2y_2 \end{pmatrix}.\\
\mathcal{Y} &= \{-x_{3-j} - y_j + 15 \ge 0, \quad j = 1,2\}.
\end{aligned}
$$

The next problem is taken from [20, 62]. In all tests, the only difference lies in the objective function.
Problem 5.

$$
\begin{aligned}
\mathcal{X} &= \{0 \le x \le 10\}.\\
F(\mathbf{x},\mathbf{y}) &= \begin{pmatrix} (1 + 0.2x)y_1 - (3 + 1.333x) - 0.333y_3 + 2y_1y_4 - y_5 \\ (1 + 0.1x)y_2 - x + y_3 + 2y_2y_4 - y_6 \\ 0.333y_1 - y_2 + 1 - 0.1x \\ 9 + 0.1x - y_1^2 - y_2^2 \\ y_1 \\ y_2 \end{pmatrix}.\\
\mathcal{Y} &= \{y_j \ge 0, \quad j = 3,4,5,6\}.
\end{aligned}
$$

High-dimensional stochastic counterparts.

Table 12 Parameter specification for Problem 1

i	1	2	3	4	5
c_i	10	8	6	4	2
K_i	5	5	5	5	5
β_i	1.2	1.1	1.0	0.9	0.8

Consider the stochastic N-dimensional counterpart of Problem 1, defined as follows.

$$f(\mathbf{x}, \mathbf{y}) = \mathbb{E}\left[r_1(x) - xp\left(x + \sum_{i=1}^{n} y_i, \omega \right) \right],$$

where $r_i(v) = c_i v + \frac{\beta_i}{\beta_i+1} K_i^{1/\beta_i} v^{(1+\beta_i)/\beta_i}$, $p(Q, \omega) = 5000^{1/\gamma(\omega)} Q^{-1/\gamma(\omega)}$, $c_i = 6$, $\beta_i = 1$, $K_i = 5$, $i = 1, \ldots, 5$, $\gamma(\omega) \in \mathcal{U}(0.9, 1.1)$ is a positive parameter, $Q = x + \sum_{i=1}^{N} y_i$.

$$\begin{aligned} \mathcal{X} &= \{0 \le x \le L\}. \\ F(\mathbf{x}, \mathbf{y}, \omega) &= \begin{pmatrix} \nabla r_2(y_1) - p(Q, \omega) - y_1 \nabla p(Q, \omega) \\ \vdots \\ \nabla r_n(y_n) - p(Q, \omega) - y_n \nabla p(Q, \omega) \end{pmatrix}. \\ \mathcal{Y} &= \{0 \le y_j \le L, \quad j = 1, \ldots, n\}. \end{aligned}$$

The stochastic N-dimensional counterpart of Problem 2.

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}, \mathbf{y}(\omega))], \text{ where } f(x, y(\omega)) &= \|x - \mathbf{1}\|^2 + \|y(\omega)\|^2. \\ \mathcal{X} &= \{0 \le x_i \le 2, \quad i = 1, \ldots, n\}. \\ F(\mathbf{x}, \mathbf{y}, \omega) &= \left(2y - 2x + \omega\right). \\ \mathcal{Y} &= \{\|y - \mathbf{1}\|^2 \le 0.25\}, \text{ where } \omega \in \mathcal{U}(-0.5, 0.5). \end{aligned}$$

References

1. Agdeppa, R.P., Yamashita, N., Fukushima, M.: An implicit programming approach for the road pricing problem with nonadditive route costs. J. Ind. Manag. Optim. **4**, 183–197 (2008)
2. Anitescu, M.: On solving mathematical programs with complementarity constraints as nonlinear programs. SIAM J. Optim. **15**(4), 1203–1236 (2005)
3. Bard, J.F.: Convex two-level optimization. Math. Program. **40**, 15–27 (1988)
4. Baringo, L., Conejo, A.J.: Strategic offering for a wind power producer. IEEE Trans. Power Syst. **28**, 4645–4654 (2013)
5. Beck, A.: Introduction to nonlinear optimization, vol. 19 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA (2014). Theory, algorithms, and applications with MATLAB
6. Beck, A.: First-Order Methods in Optimization. SIAM, Philadelphia (2017)

7. Beremlijski, P., Haslinger, J., Kočvara, M., Outrata, J.: Shape optimization in contact problems with Coulomb friction. SIAM J. Optim. **13**, 561–587 (2002)
8. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM J. Optim. **15**, 751–779 (2005)
9. Chen, X.: Smoothing methods for nonsmooth, nonconvex minimization. Math. Program. **134**, 71–99 (2012)
10. Chen, X., Sun, H., Wets, R.J.-B.: Regularized mathematical programs with stochastic equilibrium constraints: estimating structural demand models. SIAM J. Optim. **25**, 53–75 (2015)
11. Clarke, F.H., Ledyaev, Y.S., Stern, R.J., Wolenski, P.R.: Nonsmooth analysis and control theory. In: Graduate Texts in Mathematics, vol. 178. Springer, New York (1998)
12. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. SIAM, Philadelphia (2009)
13. Cui, S., Shanbhag, U.V.: On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. Set-Valued Var. Anal. **29**, 453–499 (2021)
14. Czyzyk, J., Mesnier, M.P., Moré, J.J.: The NEOS server. IEEE J. Comput. Sci. Eng. **5**, 68–75 (1998)
15. DeMiguel, V., Friedlander, M.P., Nogales, F.J., Scholtes, S.: A two-sided relaxation scheme for mathematical programs with equilibrium constraints. SIAM J. Optim. **16**, 587–609 (2005)
16. DeMiguel, V., Xu, H.: A stochastic multiple-leader Stackelberg model: analysis, computation, and application. Oper. Res. **57**, 1220–1235 (2009)
17. Dolan, E.D.: The NEOS Server 4.0 Administrative Guide. Technical Memorandum ANL/MCS-TM-250, Mathematics and Computer Science Division, Argonne National Laboratory (2001)
18. Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM J. Optim. (SIOPT) **22**, 674–701 (2012)
19. Evgrafov, A., Patriksson, M.: Stochastic structural topology optimization: discretization and penalty function approach. Struct. Multidiscip. Optim. **25**, 174–188 (2003)
20. Facchinei, F., Jiang, H., Qi, L.: A smoothing method for mathematical programs with equilibrium constraints. Math. Program. **85**, 107–134 (1999)
21. Facchinei, F., Pang, J.-S.: Finite-dimensional variational inequalities and complementarity problems. In: Springer Series in Operations Research, vol. I. Springer, New York, II (2003)
22. Fang, X., Hu, Q., Li, F., Wang, B., Li, Y.: Coupon-based demand response considering wind power uncertainty: a strategic bidding model for load serving entities. IEEE Trans. Power Syst. **31**, 1025–1037 (2015)
23. Dirkse, S. P., Ferris, M. C., Meeraus, A.: Mathematical programs with equilibrium constraints: automatic reformulation and solution via constraint optimization. Technical Report NA-02/11, Oxford University Computing Laboratory (2002)
24. Flaxman, A., Kalai, A. T., McMahan, B.: Online convex optimization in the bandit setting: Gradient descent without a gradient. In: SODA '05 Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 385–394 (January 2005)
25. Fletcher, R., Leyffer, S., Ralph, D., Scholtes, S.: Local convergence of SQP methods for mathematical programs with equilibrium constraints. SIAM J. Optim. **17**, 259–286 (2006)
26. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**, 2341–2368 (2013)
27. Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math. Program. **155**, 267–305 (2016)
28. Goldstein, A.A.: Optimization of Lipschitz continuous functions. Math. Program. **13**, 14–22 (1977)
29. Gropp, W., Moré, J.J.: Optimization environments and the NEOS server. In: Buhman, M.D., Iserles, A. (eds.) Approximation Theory and Optimization, p. 167. Cambridge University Press, Cambridge (1997)
30. Hintermüller, M., Surowiec, T.: A bundle-free implicit programming approach for a class of elliptic MPECs in function space. Math. Program. **160**, 271–305 (2016)
31. Hobbs, B.F., Metzler, C.B., Pang, J.-S.: Strategic gaming analysis for electric power systems: an MPEC approach. IEEE Trans. Power Syst. **15**, 638–645 (2000)
32. Hu, X., Ralph, D.: Convergence of a penalty method for mathematical programming with complementarity constraints. J. Optim. Theory Appl. **123**, 365–398 (2004)
33. Iusem, A.N., Jofré, A., Oliveira, R.I., Thompson, P.: Variance-based extragradient methods with line search for stochastic variational inequalities. SIAM J. Optim. **29**, 175–206 (2019)

34. Jalilzadeh, A., Shanbhag, U.V.: A proximal-point algorithm with variable sample-sizes (PPAWSS) for monotone stochastic variational inequality problems. In: Winter Simulation Conference, WSC 2019, National Harbor, MD, USA, December 8–11, 2019, vol. 2019, pp. 3551–3562. IEEE (2019)
35. Jalilzadeh, A., Shanbhag, U. V., Blanchet, J. H., Glynn, P. W.: Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs. arXiv:1803.00718 (2018)
36. Jiang, H., Ralph, D.: Smooth SQP methods for mathematical programs with nonlinear complementarity constraints. SIAM J. Optim. **10**(3), 779–808 (2000)
37. Jiang, H., Xu, H.: Stochastic approximation approaches to the stochastic variational inequality problem. IEEE Trans. Autom. Control **53**, 1462–1475 (2008)
38. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm. Stoch. Syst. **1**, 17–58 (2011)
39. Kanno, Y.: An implicit formulation of mathematical program with complementarity constraints for application to robust structural optimization. J. Oper. Res. Soc. Jpn. **54**, 65–85 (2011)
40. Kaushik, H. D., Yousefian, F.: A method with convergence rates for optimization problems with variational inequality constraints. arXiv:2007.15845v2 (2021)
41. Knopp, K.: Theory and Applications of Infinite Series. Blackie & Son Ltd, Bishopbriggs (1951)
42. Kočvara, M., Outrata, J.V.: Optimization problems with equilibrium constraints and their numerical solution. Math. Program. **101**, 119–149 (2004)
43. Kočvara, M., Outrata, J.V.: Inverse truss design as a conic mathematical program with equilibrium constraints. Discrete Contin. Dyn. Syst. Ser. S **10**, 1329–1350 (2017)
44. Lakshmanan, H., Farias, D.: Decentralized recourse allocation in dynamic networks of agents. SIAM J. Optim. **19**, 911–940 (2008)
45. Lawphongpanich, S., Hearn, D.W.: An MPEC approach to second-best toll pricing. Math. Program. **101**, 33–55 (2004)
46. Leyffer, S., López-Calva, G., Nocedal, J.: Interior methods for mathematical programs with complementarity constraints. SIAM J. Optim. **17**, 52–77 (2006)
47. Lin, G.-H., Chen, X., Fukushima, M.: Solving stochastic mathematical programs with equilibrium constraints via approximation and smoothing implicit programming with penalization. Math. Program. **116**, 343–368 (2009)
48. Liu, T., Pong, T.K., Takeda, A.: A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. Math. Program. **176**, 339–367 (2019)
49. Liu, Y., Lin, G.-H.: Convergence analysis of a regularized sample average approximation method for stochastic mathematical programs with complementarity constraints. Asia-Pac. J. Oper. Res. **28**, 755–771 (2011)
50. Luo, Z.-Q., Pang, J.-S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
51. Mayne, D.Q., Polak, E.: Nondifferential optimization via adaptive smoothing. J. Optim. Theory Appl. **43**, 601–613 (1984)
52. Migdalas, A., Pardalos, P.M., Värbrand, P.: Multilevel Optimization: Algorithms and Applications, vol. 20. Springer, Berlin (1998)
53. Mordukhovich, B.S.: Characterizations of linear suboptimality for mathematical programs with equilibrium constraints. Math. Program. **120**, 261–283 (2009)
54. Murphy, F.H., Sherali, H.D., Soyster, A.L.: A mathematical programming approach for determining oligopolistic market equilibrium. Math. Program. **24**, 92–106 (1982)
55. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**, 1574–1609 (2009)
56. Nemirovsky, A. S., Yudin, D. B.: Problem complexity and method efficiency in optimization, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, New York (1983)
57. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Doklady AN USSR **269**, 543–547 (1983)
58. Nesterov, Y.: Introductory Lectures on Convex Programming Volume I: Basic Course, Lecture Notes (1998)
59. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**, 527–566 (2017)
60. Outrata, J., Kočvara, M., Zowe, J.: Nonsmooth Approach to Optimization Problems with Equilibrium Constraints, vol. 28 of Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht (1998). Theory, Applications and Numerical Results

61. Outrata, J., Zowe, J.: A numerical approach to optimization problems with variational inequality constraints. Math. Program. **68**, 105–130 (1995)
62. Outrata, J.V.: On optimization problems with variational inequality constraints. SIAM J. Optim. **4**, 340–357 (1994)
63. Patriksson, M.: On the applicability and solution of bilevel optimization models in transportation science: a study on the existence, stability and computation of optimal solutions to stochastic mathematical programs with equilibrium constraints. Transp. Res. Part B Methodol. **42**, 843–860 (2008)
64. Patriksson, M., Wynter, L.: Stochastic mathematical programs with equilibrium constraints. Oper. Res. Lett. **25**, 159–167 (1999)
65. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc, New York (1987)
66. Raghunathan, A.U., Biegler, L.T.: An interior point method for mathematical programs with complementarity constraints (MPCCs). SIAM J. Optim. **15**, 720–750 (2005). (**electronic**)
67. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
68. Rockafellar, R.T., Wets, R.J.-B.: Stochastic variational inequalities: single-stage to multistage. Math. Program. **165**, 331–360 (2017)
69. Sahinidis, N. V.: BARON 21.1.13: Global Optimization of Mixed-Integer Nonlinear Programs, User's Manual (2017)
70. Scheel, H., Scholtes, S.: Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. Math. Oper. Res. **25**, 1–22 (2000)
71. Shapiro, A.: Stochastic programming with equilibrium constraints. J. Optim. Theory Appl. **128**, 223–243 (2006)
72. Shapiro, A., Xu, H.: Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. Optimization **57**, 395–418 (2008)
73. Sherali, H.D.: A multiple leader Stackelberg model and analysis. Oper. Res. **32**, 390–404 (1984)
74. Sherali, H.D., Soyster, A.L., Murphy, F.H.: Stackelberg–Nash–Cournot equilibria: characterizations and computations. Oper. Res. **31**, 253–276 (1983)
75. Steklov, V.A.: Sur les expressions asymptotiques decertaines fonctions définies par les équations différentielles du second ordre et leers applications au problème du dévelopement d'une fonction arbitraire en séries procédant suivant les diverses fonctions. Comm. Charkov Math. Soc. **2**, 97–199 (1907)
76. Su, C.-L.: Analysis on the forward market equilibrium model. Oper. Res. Lett. **35**, 74–82 (2007)
77. Tawarmalani, M., Sahinidis, N.V.: A polyhedral branch-and-cut approach to global optimization. Math. Program. **103**, 225–249 (2005)
78. Xu, H.: An implicit programming approach for a class of stochastic mathematical programs with complementarity constraints. SIAM J. Optim. **16**, 670–696 (2006)
79. Xu, H., Ye, J.: Approximating stationary points of stochastic mathematical programs with equilibrium constraints via sample averaging. Set-Valued Var. Anal. **128**, 283–309 (2011)
80. Xu, Y., Qi, Q., Lin, Q., Jin, R., Yang, T.: Stochastic optimization for dc functions and non-smooth non-convex regularizers with non-asymptotic convergence. In: International Conference on Machine Learning, PMLR, pp. 6942–6951 (2019)
81. Yousefian, F., Nedić, A., Shanbhag, U.V.: On stochastic gradient and subgradient methods with adaptive steplength sequences. Automatica **48**, 56–67 (2012)
82. Yousefian, F., Nedic, A., Shanbhag, U.V.: On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. Math. Program. **165**, 391–431 (2017)
83. Yousefian, F., Nedić, A., Shanbhag, U. V.: Convex nondifferentiable stochastic optimization: a local randomized smoothing technique. In: Proceedings of the 2010 American Control Conference, pp. 4875–4880 (2010)