MATHEMATICS OF OPERATIONS RESEARCH
Vol. 36, No. 3, August 2011, pp. 568–592
ISSN 0364-765X | EISSN 1526-5471 | 11 | 3603 | 0568

DOI 10.1287/moor.1110.0506

Convergence of Stationary Points of Sample Average Two-Stage Stochastic Programs: A Generalized Equation Approach

Daniel Ralph
Cambridge Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom,
d.ralph@jbs.cam.ac.uk, http://www.jbs.cam.ac.uk/research/faculty/ralphd.html

Huifu Xu
School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom,
h.xu@soton.ac.uk, http://www.personal.soton.ac.uk/hx/

This paper presents an asymptotic analysis of a Monte Carlo method, variously known as sample average approximation (SAA) or sample path optimization (SPO), for a general two-stage stochastic minimization problem. We study the case when the second-stage problem may have multiple local optima or stationary points that are not global solutions and SAA is implemented using a general nonlinear programming solver that is only guaranteed to find stationary points. New optimality conditions are developed for both the true problem and its SAA problem to accommodate Karush-Kuhn-Tucker points. Because the optimality conditions are essentially stochastic generalized equations, the asymptotic analysis is carried out for the generalized equations first and then applied to optimality conditions. For this purpose, we analyze piecewise continuous (PC^0) stochastic mappings to understand when their expectations are piecewise continuous and thereby derive exponential convergence of SAA. It is shown under moderate conditions that, with probability one, an accumulation point of the SAA stationary points satisfies a relaxed stationary condition for the true problem and further that, with probability approaching one exponentially fast with increasing sample size, a stationary point of SAA converges to the set of relaxed stationary points. These results strengthen or complement existing results where the second-stage problem is often assumed to have a unique solution and the exponential convergence is focused on how fast a solution of the true problem becomes an approximate solution of an SAA problem rather than the other way around.

Key words: stochastic generalized equation; sample average approximation; stochastic piecewise continuous (PC^0); set-valued mappings
MSC2000 subject classification: Primary: 90C15, 49J53; secondary: 90C31, 90C33
OR/MS subject classification: Primary: programming: stochastic, nondifferentiable; secondary: programming: algorithm, mathematics: sets
History: Received June 25, 2008; revised August 21, 2009, Sepember 1, 2010, and December 21, 2010.

1. Introduction.

In this paper, we study the following stochastic minimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^m,\, y(\cdot)\in Y} \quad & \mathbb{E}[f(x, y(\omega), \xi(\omega))] \\ \text{s.t.} \quad & x \in X, \\ & g(x, y(\omega), \xi(\omega)) \le 0, \quad \text{a.e. } \omega \in \Omega, \\ & h(x, y(\omega), \xi(\omega)) = 0, \quad \text{a.e. } \omega \in \Omega, \end{aligned} \tag{1}$$

where X is a nonempty, convex subset of $\mathbb{R}^m$; f, g, and h are continuously differentiable functions from $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^r$ to $\mathbb{R}$, $\mathbb{R}^s$, and $\mathbb{R}^t$, respectively; $\xi\colon \Omega \to \Xi \subset \mathbb{R}^r$ is a vector of random variables defined on probability space $(\Omega, \mathcal{F}, P)$; $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to (w.r.t.) the probability distribution P; and Y is a suitable space of functions $y(\cdot)\colon \Omega \to \mathbb{R}^n$ such that $\mathbb{E}[f(x, y(\omega), \xi(\omega))]$ is well-defined.

Problem (1) is a two-stage stochastic programming problem. At the first stage, a decision on x needs to be made, restricted to the feasible set X, *before* a realization of the random data $\xi(\omega)$. At the second stage, when x is fixed and a realization $\xi(\omega) = \xi$ is known, the following second-stage minimization problem is solved to calculate second-stage decision vector y:

$$\begin{aligned} \min_{y\in\mathbb{R}^n} \quad & f(x, y, \xi) \\ \text{s.t.} \quad & g(x, y, \xi) \le 0, \\ & h(x, y, \xi) = 0. \end{aligned} \tag{2}$$

When the second-stage problem is nonconvex, the meaning of "solving" may be to find a local minimum or even a stationary point to allow for a practical computational procedure. To explain this, consider a special case when

ξ has a finite distribution with N scenarios, $\xi^1, \ldots, \xi^N$, and corresponding probabilities $p_1, \ldots, p_N$. Then, the two-stage problem can be written as

$$\begin{aligned} \min_{x, y^1, \ldots, y^N} \quad & \sum_{i=1}^{N} p_i f(x, y^i, \xi^i) \\ \text{s.t.} \quad & x \in X, \\ & g(x, y^i, \xi^i) \le 0, \quad i = 1, \ldots, N, \\ & h(x, y^i, \xi^i) = 0, \quad i = 1, \ldots, N. \end{aligned} \tag{3}$$

Problem (3) is generally nonconvex, and therefore it is numerically difficult to obtain a global minimum.

Throughout this paper, we assume that the probability measure P of our considered space $(\Omega, \mathcal{F}, P)$ is nonatomic. To ease notation, we will use ξ to denote either the random vector $\xi(\omega)$ or an element of $\mathbb{R}^r$ depending on the context.

The stochastic program (1) has been well-studied and subsumes standard two-stage stochastic programs; see, for instance, Ruszczyński and Shapiro [40, Chapter 1] and Shapiro [42] and the references therein. Our interest here is with the sample average approximation (SAA) method, which is also known as the sample path optimization (SPO) method for solving stochastic optimization problems. Over the past few years, SAA has been increasingly investigated and recognized as one of the most effective methods for solving stochastic programs; see recent discussions in King and Wets [24], Plambeck et al. [31], Robinson [37], Shapiro and Homem-de-Mello [45], Shapiro [43], and references therein. The basic idea of SAA is to replace the true problem by approximating the expected value with the sample average.

Let $\xi^1, \ldots, \xi^N$ be an independent identically distribution (i.i.d.) sample of ξ. The SAA of (1) is

$$\begin{aligned} \min_{x, y^1, \ldots, y^N} \quad & \frac{1}{N} \sum_{i=1}^{N} f(x, y^i, \xi^i) \\ \text{s.t.} \quad & x \in X, \\ & g(x, y^i, \xi^i) \le 0, \quad i = 1, \ldots, N, \\ & h(x, y^i, \xi^i) = 0, \quad i = 1, \ldots, N. \end{aligned} \tag{4}$$

The SAA approach for two stage and multistage stochastic programs has been well-studied; see, for instance, Römisch and Schultz [39] and Shapiro [42]. In particular, Shapiro [42] has established the exponential convergence of optimal solutions of the two-stage SAA problem by studying an equivalent SAA of the optimal value function of the second-stage problem, roughly as follows: Let $v(x, \xi)$ denote the global optimal value of the second-stage problem (2). Assume, for $i = 1, \ldots, N$, that y^i is a global solution of the second-stage problem corresponding to x and ξ^i so that $v(x, \xi^i) = f(x, y^i, \xi^i)$. Then, (4) can be rewritten as

$$\begin{aligned} \min_{x} \quad & v_N(x) := \frac{1}{N} \sum_{i=1}^{N} v(x, \xi^i) \\ \text{s.t.} \quad & x \in X. \end{aligned} \tag{5}$$

This problem is the SAA for the following problem:

$$\begin{aligned} \min_{x} \quad & \vartheta(x) := \mathbb{E}[v(x, \xi)] \\ \text{s.t.} \quad & x \in X. \end{aligned} \tag{6}$$

Shapiro [42] and others study convergence of SAA for (6) rather than (1).

We refer to (5) and (6) as *implicit programming* approaches in that they do not explicitly involve the underlying functions in the second stage. Incidentally, the implicit programming approach has been used to study the convergence of stationary points of two-stage stochastic mathematical programs with equilibrium constraints (SMPEC) where the second-stage parametric MPEC problem has a unique solution (see Shapiro and Xu [46], Xu and Meng [49], and Meng and Xu [27]).

In this paper, we assume that (4) is *directly* solved by a nonlinear programming (NLP) code without necessarily decomposing this problem into two stages, and we study the convergence of its stationary points as the sample

size increases. The key difference between this work and the existing work is that here, at a stationary point $(x^N, y^1, \ldots, y^N)$ of (4), none of the points y^i need to be global or even local optimal solutions of (2) for $x = x^N$ and $\xi = \xi^i$. Consequently, we *cannot* use the implicit programming approach, which relies on the equivalence between (4) and (5). Our results are therefore of interest when the underlying functions are nonconvex and the second-stage problem has multiple solutions.

Our approach is to develop a relaxed optimality condition (see (41) in §6) by replacing the subdifferential of $v(x, \xi)$ with a set-valued mapping constructed from the gradients of the Lagrangian function of the second-stage problem at stationary points. This set-valued mapping is defined through an earlier sensitivity result established by Gauvin and Dubeau [14] on approximating the Clarke subdifferential (Clarke [8]) of the optimal value function of a parametric program. The relaxed optimality condition accommodates accumulation points of a sequence $\{x^N\}$ of stationary points of (4). In doing so, we explain what convergence of $\{x^N\}$ might mean when implementing SAA for (1) using standard NLP software.

Two particular cases explored in §6.3 are (1) when the second-stage problem (2) is unconstrained and nonconvex and has locally unique stationary points, and (2) when the second-stage problem is a nonconvex parametric quadratic program.

Observe that the first-order necessary conditions of (5) and (4) (w.r.t. x) are both sample averages of certain stochastic generalized equations. For simplicity of notation and the potential application of the convergence results to other areas, we first carry out convergence analysis of the SAA of a class of stochastic generalized equations and then apply these convergence results to the SAA applied to the true problem (1) in §6. Generalized equations have been extensively investigated by Robinson and many others for the sensitivity analysis of deterministic optimization problems and variational inequalities (see Robinson [36], Rockafellar and Wets [38], Facchinei and Pang [13], and references therein). In particular, King and Rockafellar [23] use generalized equations as a framework for asymptotic analysis of solutions in statistical estimation and stochastic programming. More recently, Shapiro [43] studies a class of stochastic generalized equations and establishes some useful convergence results for SAA problems as sample size increases. Note that the underlying functions (excluding the normal cone) in defining the generalized equations in all of the above-noted references are single valued.

In this paper, we study a class of stochastic generalized equations with the underlying function being random set valued rather than single valued. This is to accommodate the first-order optimality condition of (5) and its relaxation (41). Consequently we need the uniform strong law of large numbers for random set-valued mappings (Shapiro and Xu [47]) to deal with the sample average and metric regularity-type condition to obtain exponential convergence of $\{x^N\}$.

We now summarize the main contributions of this paper.

(a) We present a detailed discussion of random piecewise continuous set-valued mappings, including random polyhedral multifunctions,[1] and show, under some moderate conditions, that the expectation of such mappings is piecewise continuous. This result is used to show the uniform exponential convergence of the SAA solution of stochastic generalized equations. These results complement the asymptotic analysis on stochastic generalized equations by King and Rockafellar in [23], strengthen the recent convergence results for stochastic generalized equations by Shapiro [43, §7], and have applications, we believe, in stochastic equilibrium problems such as nonsmooth stochastic games.

(b) We establish almost sure convergence for the stationary points of a general stochastic minimization problem where the second-stage problem may have multiple optimal solutions and stationary points. This extends the recent literature that typically relies on the SAA finding a globally optimal solution of the second-stage problem (the second stage problem being convex or having a unique locally and globally optimal solution).

(c) Under some moderate conditions, we show that, with probability approaching one exponentially fast with the increase of sample size, a stationary point of the SAA problem (4) (w.r.t. x) converges to its true counterpart. This differs from the exponential convergence literature (e.g. Shapiro [43], Römisch and Schultz [39]) where the focus is on globally optimal solutions. We also allow the solution set of (4)—whether it is defined by stationary points, local optima, or global optima—to vary discontinuously in x, as suggested in (a).

The rest of the paper is organized as follows. In §2, we present some definitions and results on measurability and integrability of random sets and Clarke subdifferentials of random functions. We also review some preliminary results on set-valued mappings. In §3, we establish an exponential convergence result for sample averages of Hölder continuous set-valued mappings based on the work of Shapiro and Xu [46]. In §4, we investigate the expectation of random piecewise continuous set-valued mappings and show, under some moderate conditions,

[1] For consistency in terminology, we will use the term "polyhedral set-valued mapping" later on, although "polyhedral multifunction" is more frequently used in the literature.

that the expected value of such mappings are piecewise continuous. This allows us to establish exponential convergence of the SAA solution set of a stochastic generalized equation to the true solution set in the piecewise continuous case. Section 5 derives optimality conditions for the true problem (1). In §6, we apply the results of the preceding sections to study convergence of the first-stage decision vectors obtained from stationary points of the SAA problem (4).

2. Preliminaries. In this section, we present some notation and preliminary discussions about the measurability of a random set-valued mapping, implicit function theorem based on Clarke generalized Jacobians and generalized Karush-Kuhn-Tucker (KKT) conditions.

Throughout this paper, we use the following notations. $\mathbb{N}$ denotes the set of positive integers, $x^T y$ denotes the scalar products of two vectors x and y, and $\|\cdot\|$ denotes the Euclidean norm of either a vector and a nonempty set of vectors $\mathscr{C}$ where $\|\mathscr{C}\| := \sup_{x\in\mathscr{C}} \|x\|$. For a nonempty, closed set $\mathscr{C}$ in finite dimensional space, let $\|\mathscr{C}\| := \sup_{x\in\mathscr{C}} \|x\|$. Also, $d(x,\mathscr{C}) := \inf_{x'\in\mathscr{C}} \|x - x'\|$, which is the distance from point x to $\mathscr{C}$. For two nonempty, compact sets $\mathscr{C}$ and $\mathscr{D}$, $\mathbb{D}(\mathscr{C},\mathscr{D}) := \sup_{x\in\mathscr{C}} d(x,\mathscr{D})$ denotes the deviation from set $\mathscr{C}$ to set $\mathscr{D}$ (in some references (see Hess [17]), this is also called *excess* of $\mathscr{C}$ over $\mathscr{D}$), and $\mathbb{H}(\mathscr{C},\mathscr{D})$ denotes the Pompeiu-Hausdorff distance between the two sets, that is, $\mathbb{H}(\mathscr{C},\mathscr{D}) := \max(\mathbb{D}(\mathscr{C},\mathscr{D}), \mathbb{D}(\mathscr{D},\mathscr{C}))$. We use $\mathscr{C}+\mathscr{D}$ to denote the Minkowski addition of the two sets, that is, $\{x + x' \colon x \in \mathscr{C}, x' \in \mathscr{D}\}$. We use $B(x,\delta)$ to denote the closed ball with radius δ and center x, that is, $B(x,\delta) := \{x' \colon \|x' - x\| \le \delta\}$. When δ is dropped, $B(x)$ represents a neighborhood of point x. Finally, we use $\mathscr{B}$ to denote the unit ball in a finite dimensional space.

2.1. Set-valued mappings. Let X be a closed subset of $\mathbb{R}^n$. A set-valued mapping $F\colon X \to 2^{\mathbb{R}^m}$ is said to be *closed* at $\bar{x}$ if for $x^k \subset X$, $x^k \to \bar{x}$, $y^k \in F(x^k)$, and $y^k \to \bar{y}$ implies $\bar{y} \in F(\bar{x})$. F is said to be *uniformly compact* near $\bar{x} \in X$ if there is a neighborhood $B(\bar{x})$ of $\bar{x}$ such that the closure of $\bigcup_{x\in B(\bar{x})} F(x)$ is compact. F is said to be *outer semicontinuous* (osc for short)[2] at $\bar{x} \in X$ if, for every $\epsilon > 0$, $\rho > 0$, there exists a $\delta > 0$ such that

$$F(\bar{x} + \delta\mathscr{B}) \cap \rho\mathscr{B} \subset F(\bar{x}) + \epsilon\mathscr{B}.$$

The following result was established by Hogan [19].

Proposition 2.1. *Let $F\colon X \to 2^{\mathbb{R}^m}$ be uniformly compact near $\bar{x}$. Then, F is outer semicontinuous at $\bar{x}$ if and only if F is closed.*

From Proposition 2.1, we can see that the intersection with $\rho\mathscr{B}$ in the definition of osc of F can be removed when F is uniformly compact near $\bar{x}$.

2.2. Clarke subdifferentials and random sets. Let $\psi\colon \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function. Recall that the Clarke subdifferential (Clarke [8]) of ψ at $x \in \mathbb{R}^n$ is defined as $\partial\psi(x) := \operatorname{conv}\{\lim_{y\in D_\psi,\, y\to x} \nabla\psi(y)\}$, where D_ψ denotes the set of points at which ψ is Fréchet differentiable, $\nabla\psi(y)$ denotes the usual gradient of ψ, and "conv" denotes the convex hull of a set.

In this paper, for any fixed ξ, we will be mainly concerned with the Clarke subdifferential w.r.t x of the random function $\psi(x,\xi)$, denoted by $\partial_x\psi(x,\xi)$, and the subdifferential mapping $x \mapsto \partial_x\psi(x,\xi)$. For fixed x, the subdifferential is a random set. We need to deal with the expectation of subdifferentials, which is related to the measurability of random sets. In what follows, we make some preparations for this.

Let x be fixed, and consider the measurability of a general set-valued mapping $\mathscr{A}(x,\xi(\cdot))\colon \Omega \to 2^{\mathbb{R}^n}$. Here, ξ is a random vector defined on probability space $(\Omega,\mathscr{F},P)$. Let $\mathfrak{B}$ denote the space of nonempty, closed subsets of $\mathbb{R}^n$. Then, $\mathscr{A}(x,\xi(\cdot))$ can be viewed as a single valued mapping from Ω to $\mathfrak{B}$. Using Rockafellar and Wets [38, Theorem 14.4], we know that $\mathscr{A}(x,\xi(\cdot))$ is measurable if and only if for every $B \in \mathfrak{B}$, $\mathscr{A}(x,\xi(\cdot))^{-1}B$ is $\mathscr{F}$-measurable.

Proposition 2.2 (Xu and Zhang [50], Proposition 2.1). *Let $\psi(x,\xi)$ be a locally Lipschitz continuous function in both x and ξ. The Clarke subdifferential $\partial_x\psi(x,\xi)$ is measurable.*

[2] In some reference books, the outer semicontinuity of a set-valued mapping is defined through the outer limit of set convergence in the sense of Painlevé-Kuratowski (see, for instance, Rockafellar and Wets [38, Definition 5.4]). Here, the outer semicontinuity is in the sense of Berge [4, p. 109]. The relation between the two definitions are given in Rockafellar and Wets [38, Proposition 5.12]. Note, also, that in some references, outer (inner) semicontinuity is also called upper (lower) semicontinuity (see, e.g., Aubin and Frankowska [2]). In this paper, we follow the terminology of Rockafellar and Wets [38]. See commentary at the end of Chapter 4 in Rockafellar and Wets [38] about this, and also see Rockafellar and Wets [38, Chapters 4 and 5] for comprehensive treatment of set-valued convergence and set-valued mappings.

We now define the expected value of $\partial_x \psi(x, \xi)$. Because we will also need to consider the expected value of the set of Lagrange multipliers of the second-stage problem, we consider the definition of a general random set-valued mapping $\mathcal{A}(x, \xi(\omega))$: $X \times \Omega \to 2^{\mathbb{R}^n}$.

By a *selection* of the random set $\mathcal{A}(x, \xi(\omega))$, we refer to a random vector $A(x, \xi(\omega)) \in \mathcal{A}(x, \xi(\omega))$, which means $A(x, \xi(\omega))$ is measurable. Note that such selections exist (see Artstein and Vitale [1] and references therein). The *expectation of* $\mathcal{A}(x, \xi(\omega))$, denoted by $\mathbb{E}[\mathcal{A}, (x, \xi)]$, is defined as the collection of $\mathbb{E}[A(x, \xi(\omega))]$, where $A(x, \xi(\omega))$ is an integrable selection and $\mathbb{E}[\mathcal{A}(x, \xi)]$ is known as Aumann's integral (Aumann [3], Hess [17]) of the set-valued mapping $\mathcal{A}(x, \xi(\omega))$. We regard $\mathbb{E}[\mathcal{A}(x, \xi)]$ as well-defined if $\mathbb{E}[\mathcal{A}(x, \xi)] \in \mathfrak{B}$ is nonempty. A sufficient condition of the well-definedness of the expectation is $\mathbb{E}[\|\mathcal{A}(x, \xi)\|] := \mathbb{E}[\mathbb{H}(0, \mathcal{A}(x, \xi))] < \infty$; see Artstein and Vitale [1]. In such a case, $\mathcal{A}$ is sometimes called *integrably bounded* (see Aumann [3], Hess [17], and Papageorgiou [30]).

3. Sample average approximation of stochastic generalized equation. Generalized equations provide a unified framework for the first-order necessary condition in optimization and equilibrium conditions in game theory (Robinson [36], Rockafellar and Wets [38]). In this section, we consider a class of stochastic generalized equations and its SAA. This is to pave the way for studying the convergence of SAA stationary points of the true optimization problem (1) in §§5 and 6.

Let Γ: $\mathbb{R}^n \times \Xi \to 2^{\mathbb{R}^m}$ be a closed set-valued mapping, and let ξ: $\Omega \to \Xi \subset \mathbb{R}^r$ be a vector of random variables defined on probability space $(\Omega, \mathcal{F}, P)$. Through Rockafellar and Wets [38, Theorem 5.7(a) and Corollary 14.14], $\Gamma(x, \xi(\cdot))$: $\Omega \to 2^{\mathbb{R}^m}$ is measurable. We consider the following stochastic generalized equation:

$$0 \in \mathbb{E}[\Gamma(x, \xi)] + \mathcal{N}_X(x), \tag{7}$$

where X is a closed subset of $\mathbb{R}^n$ (here, we are slightly abusing the notation as X is used to denote the feasible set of problem (1)) and $\mathcal{N}_X(x)$ denotes a normal cone to X at x. We assume that $\mathcal{N}_X(x)$ is closed and $\mathcal{N}_X(\cdot)$ is outer semicontinuous, but we tacitly leave the normal cone to be unspecified in this section in order to allow the results to be used in various applications (although, in §§5 and 6, we will need it to be the Clarke normal cone).

When Γ is single valued, (7) reduces to a class of stochastic variational inequality (SVI) model (King and Rockafellar [23], Gürkan et al. [16]). Our interest here is in the case when Γ is set valued. This corresponds to first-order necessary conditions for stochastic nonsmooth optimization problems and equilibrium conditions for nonsmooth Nash games where a player's utility functions may be nonsmooth and, consequently, the first-order necessary conditions that characterize a player's optimal decision may involve generalized subdifferentials. Similar applications can also be found in stochastic equilibrium problems with equilibrium constraints (DeMiguel and Xu [11]). We will not discuss details of these applications as our main interest in this paper is to apply the framework of the stochastic generalized equation (7) to the first-order necessary optimality conditions of the stochastic optimization problem (1). Throughout this section, we make a blanket assumption that $\mathbb{E}[\Gamma(x, \xi)]$ is well-defined (a sufficient condition is that $\Gamma(x, \xi)$ is integrably bounded) and (7) has a nonempty solution set.

Let $\xi^1, \ldots, \xi^N$ be an i.i.d. sample of ξ and $\Gamma_N(x) := (1/N) \sum_{i=1}^N \Gamma(x, \xi^i)$. The SAA of (7) is defined as

$$0 \in \Gamma_N(x) + \mathcal{N}_X(x). \tag{8}$$

Shapiro [43, §7] presents an excellent analysis on the asymptotic behavior of SAA for this problem when Γ is single valued. Earlier analysis in this regard can also be tracked back to work by King and Rockafellar [23]. Here, we focus on the case when Γ is multivalued (set valued). It is convenient at this point to make the following assumption.

Assumption 3.1. *There exist positive constants N_0 and K (both independent of ω) such that for $N \geq N^0$, w.p.1 problem (8) has a solution x^N with $\|x^N\| \leq K$.*[3]

This assumption consists two parts: existence of a solution x^N to (8) and boundedness of $\{x^N\}$. Shapiro [43, §7] investigates this issue when Γ is *single* valued and $\mathbb{E}[\Gamma(x, \xi)]$ is continuously differentiable. Specifically, if (7) is strongly regular in the sense of Robinson [34] at a solution point x^* of (7) and $\Gamma_N(x)$ converges uniformly

[3] Our analysis in this section and §§5 and 6 can be carried out by replacing $K\mathcal{B}$ with some unspecified compact subset $\mathcal{X}$ of X that contains $\{x^N\}_{N \geq N^0}$. The latter might give a smaller set than $K\mathcal{B}$ when solution set of (7) is far away from the origin. We opt for $K\mathcal{B}$ only for the simplicity and clarity of presentation. Moreover, the reason we consider x^N for those $N \geq N^0$ is because here we consider the asymptotic analysis.

to $\mathbb{E}[\Gamma(x)]$ in a neighborhood of x^*, then w.p.1 (8) has a unique solution x^N in the neighborhood and $x^N \to x^*$ w.p.1 as $N \to \infty$; see Shapiro [43, Proposition 19] for details. When $\Gamma(x, \xi)$ is set valued, the issue becomes much more complicated. However, when applied to convergence analysis in §§5 and 6, this assumption is easily justified when the feasible set X is bounded in addition to being nonempty and closed. At that point, we may consider the generalized Equations (8) and (7) to be stationary conditions of optimization problems (5) and (6), respectively. In this situation, existence follows from existence of global minima of the optimization problems, which is standard given continuity of their respective objective functions. This does not require the algorithm for implementing SAA to find a global minimum of (5); any stationary point will do because boundedness follows from boundedness of X. When X is unbounded, sufficient conditions for the existence of a solution to (8) may be derived from the property of the underlying functions. We will not discuss this in detail as it is not our main focus. In the case when a solution x^N exists and with probability $\{x^N\}$ has a bounded subsequence, then we may carry out our convergence analysis by focusing on the subsequence. We omit this for simplicity and clarity of presentation.

Our first convergence result is concerned with almost sure convergence of $\{x^N\}$ but does not address the rate of convergence. The result follows easily from a similar argument in the proof of Shapiro and Xu [47, Theorem 7].

THEOREM 3.1. *Let X^* denote the solution set of the generalized equation (7). Let Assumption 3.1 hold. Suppose that* (a) *$\Gamma(x, \xi)$ is outer semicontinuous in x on $K\mathcal{B}$ for almost all (a.a.) ξ, and* (b) *there exists an integrable function $\kappa(\xi) > 0$ such that $\|\Gamma(x, \xi)\| \le \kappa(\xi)$ for all x in $K\mathcal{B}$ and a.a. ξ. Then, $d(x^N, X^*) \to 0$ w.p.1 as $N \to \infty$.*

3.1. Exponential convergence. From a numerical perspective, we cannot take sample size to infinity; therefore, it would be more interesting to know how close x^N is to the solution set X^* of the true problem when N is large. To do so, we need to relate the distance from x^N to X^* to that between $\Gamma_N(x)$ and $\mathbb{E}[\Gamma(x, \xi)]$.

ASSUMPTION 3.2. *There exist constants K, ν, α, and integer N^* (independent of ω) such that, w.p.1 $x^N \in K\mathcal{B}$, and*

$$d(x^N, X^*) \le \sup_{x \in K\mathcal{B}} \alpha \mathbb{D}(\Gamma_N(x), \mathbb{E}[\Gamma(x, \xi)])^\nu \tag{9}$$

for $N \ge N^*$.

In the following proposition, we give sufficient conditions for this assumption.

PROPOSITION 3.1. *If* (a) *X^* is nonempty and Assumption 3.1 holds with K and N^0 given there, and* (b) *there exist positive constants ν and α such that, for all $x \in K\mathcal{B}$,*

$$d(x, X^*) \le \alpha d(0, \mathbb{E}[\Gamma(x, \xi)] + \mathcal{N}_X(x))^\nu, \tag{10}$$

then (9) holds for $N \ge N^0$.

PROOF. Under Assumption 3.1, $x^N \in K\mathcal{B}$ w.p.1 for any $N \ge N^* = N^0$. Applying inequality (10) to x^N and because $0 \in \Gamma_N(x^N) + \mathcal{N}_X(x^N)$,

$$\begin{aligned} d(x^N, X^*) &\le \alpha d(0, \mathbb{E}[\Gamma(x^N, \xi)] + \mathcal{N}_X(x^N))^\nu \\ &\le \alpha \mathbb{D}(\Gamma_N(x^N), \mathbb{E}[\Gamma(x^N, \xi)])^\nu \le \alpha \sup_{x \in K\mathcal{B}} (\mathbb{D}(\Gamma_N(x), \mathbb{E}[\Gamma(x, \xi)])^\nu, \end{aligned}$$

which gives (9). □

The condition (10) provides an error bound for the set of solutions of the stochastic generalized equation (7). This condition, when $\nu = 1$, is a kind of global *calmness* of the inverse set-valued mapping $(\mathbb{E}[\Gamma(x, \xi)] + \mathcal{N}_X(x))^{-1}$ at zero. Calmness, in turn, is implied by the *metric regularity* of $\mathbb{E}[\Gamma(x, \xi)] + \mathcal{N}_X(x)$, which is equivalent to the *Aubin property* or *pseudo-Lipschitz continuity* (in Aubin's own terminology; see, for example, [2]) of the inverse set-valued mapping; see Dontchev et al. [12], Ioffe [22], and references therein for discussions in this regard.

Our focus here is on the rate of convergence of $\{x^N\}$ to X^* under Assumptions 3.1 and 3.2 using the constants K and N^* given there. That is, we will show $d(x^N, X^*) \to 0$ with probability approaching one exponentially fast. To do so, it is sufficient to show exponential convergence of $\sup_{x \in K\mathcal{B}} \mathbb{D}(\Gamma_N(x), \mathbb{E}[\Gamma(x, \xi)])$, namely, for any $\epsilon > 0$, there exist constants $c(\epsilon) > 0$ and $\beta(\epsilon) > 0$ independent of N such that

$$\text{Prob}\left\{ \sup_{x \in K\mathcal{B}} \mathbb{D}(\Gamma_N(x), \mathbb{E}[\Gamma(x, \xi)]) \ge \epsilon \right\} \le c(\epsilon) e^{-\beta(\epsilon) N} \tag{11}$$

for $N \geq N^*$. Equation (11) then gives, through (9),

$$\text{Prob}\{d(x^N, X^*) \geq \epsilon\} \leq \bar{c}(\epsilon)e^{-\bar{\beta}(\epsilon)N}, \tag{12}$$

where $\bar{c}(\epsilon) = c((\epsilon/\alpha)^{1/\nu})$ and $\bar{\beta}(\epsilon) = \beta((\epsilon/\alpha)^{1/\nu})$.

We need a number of intermediate results. The *support function* of a nonempty set A in $\mathbb{R}^m$ is the function that maps $u \in \mathbb{R}^m$ to $\sigma(u, A) = \sup_{a\in A} u^T a$.

Lemma 3.1 (Castaing and Valadier [7], Theorem II-18). *Let A, B be nonempty, compact, and convex subsets of $\mathbb{R}^m$ with support functions $\sigma(u, A)$ and $\sigma(u, B)$. Then,*

$$\mathbb{D}(A, B) = \max_{\|u\|\leq 1}(\sigma(u, A) - \sigma(u, B))$$

and

$$\mathbb{H}(A, B) = \max_{\|u\|\leq 1}|\sigma(u, A) - \sigma(u, B)|.$$

Assume from here to the end of this subsection that $\Gamma(x, \xi)$ is a nonempty, closed, and convex set for every $\xi \in \Xi$ and $x \in K\mathcal{B}$.

By Lemma 3.1, (11) is implied by

$$\text{Prob}\left\{\sup_{x\in K\mathcal{B},\, \|u\|\leq 1} |\sigma(u, \Gamma_N(x)) - \sigma(u, \mathbb{E}[\Gamma(x, \xi)])| \geq \epsilon\right\} \leq c(\epsilon)e^{-\beta(\epsilon)N}.$$

In what follows, under some moderate conditions, we will show the inequality above. For this purpose, we need some preparation.

Let $\phi(x, \xi)$: $\mathbb{R}^n \times \Xi \to \mathbb{R}$ be a real-valued function where ξ is a vector of random variables with support set Ξ. Let $\xi^1, \ldots, \xi^N$ be an i.i.d. sample of the random vector $\xi(\omega)$, and consider the corresponding sample average function $\phi_N(x) := 1/N \sum_{j=1}^N \phi(x, \xi^j)$. Let $\psi(x) = \mathbb{E}[\phi(x, \xi)]$. We will use a uniform exponential rate of convergence result of Shapiro and Xu [46, Theorem 5.1] to describe the convergence of $\phi_N(x)$ to $\psi(x)$ over a compact set $\mathcal{X} \subset \mathbb{R}^n$. We denote by $M_x(t) := \mathbb{E}\{e^{t[\phi(x,\xi)-\psi(x)]}\}$ the moment-generating function of the random variable $\phi(x, \xi(\omega)) - \psi(x)$. We assume:

Condition 1 (C1). For every $x \in \mathcal{X}$, the moment-generating function $M_x(t)$ is finite valued for t near 0.

Condition 2 (C2). $\phi(\cdot, \xi)$ is Hölder continuous on $\mathcal{X}$, that is, there exists a (measurable) function κ: $\Xi \to \mathbb{R}_+$ and constant $\gamma > 0$ such that

$$|\phi(x', \xi) - \phi(x, \xi)| \leq \kappa(\xi)\|x' - x\|^\gamma \tag{13}$$

for all $\xi \in \Xi$ and $x \in \mathcal{X}$.

Condition 3 (C3). For every $x \in \mathcal{X}$, the moment-generating function $M_{\kappa_x}(t)$ of $\kappa_x(\xi)$ is finite valued for t near 0.

It is known (Shapiro and Xu [46]) that the (C1) and (C3) hold when ξ has a finite support set.

Lemma 3.2 (Shapiro and Xu [46], Theorem 5.1). *Suppose that* (C1)–(C3) *hold and the set $\mathcal{X}$ is compact. Then, for any $\epsilon > 0$, there exist positive constants $c(\epsilon)$ and $\beta(\epsilon)$, independent of N, such that*

$$\text{Prob}\left\{\sup_{x\in\mathcal{X}} |\phi_N(x) - \psi(x)| \geq \epsilon\right\} \leq c(\epsilon)e^{-N\beta(\epsilon)}. \tag{14}$$

We are now ready to present the main result of this section. We give a kind of uniform exponential outer semiconvergence of $\Gamma_N(x)$ to $\mathbb{E}[\Gamma(x, \xi)]$ over a compact set $K\mathcal{B}$ and show the exponential convergence of x^N to X^* where x^N solves $0 \in \Gamma_N(x) + \mathcal{N}_X(x)$. To this end, we apply Lemma 3.2 to the support function $\sigma(u, \Gamma(x, \xi))$. Recall Pompeiu-Hausdorff distance $\mathbb{H}(\mathcal{C}, \mathcal{D})$ between the two sets.

Theorem 3.2 (Exponential Convergence). *Let $\Gamma(x, \xi)$: $X \times \Xi \to 2^{\mathbb{R}^m}$ be defined as in* (7). *Suppose that* (a) *Assumption 3.1 holds*; (b) *$\Gamma(\cdot, \xi)$ is Hölder continuous on $K\mathcal{B}$, that is, there exist a constant $\gamma > 0$ and integrable function κ_1: $\Xi \to \mathbb{R}_+$ such that for every $x \in K\mathcal{B}$,*

$$\mathbb{H}(\Gamma(x', \xi), \Gamma(x, \xi)) \leq \kappa_1(\xi)\|x' - x\|^\gamma \tag{15}$$

for all $x' \in K\mathcal{B}$ all $\xi \in \Xi$; and (c) *there exists an integrable function $\kappa_2(\xi) > 0$ such that $\|\Gamma(x, \xi)\| \leq \kappa_2(\xi)$ for $x \in X$ and almost everywhere (a.e.) ξ, and for $\kappa(\xi) := \kappa_1(\xi) + \kappa_2(\xi)$, the moment-generating function of $\kappa(\xi)$, $\mathbb{E}[e^{t\kappa(\xi)}]$, is finite valued for t near 0. Then,*

(i) *for sufficiently small* $\epsilon > 0$, *there exist positive constants* $c(\epsilon)$ *and* $\beta(\epsilon)$, *independent of* N, *and* $N^* > 0$ *such that* (11) *holds for* $N \geq N^*$;

(ii) *if, in addition, Assumption* 3.2 *holds, then* $d(x^N, X^*) \to 0$ *exponentially quickly*; *there are positive constants* $\bar{c}(\epsilon)$ *and* $\bar{\beta}(\epsilon)$, *independent of* N, *and* $N^* > 0$ *such that* (12) *holds w.p.*1 *for* $N \geq N^*$.

Proof. The thrust of the proof of Theorem 3.2, part (i) is to apply Lemma 3.2 to the function $\phi(u, x, \xi) := \sigma(u, \Gamma(x, \xi))$ with sample average $\phi_N(u, x)$, expected value $\psi(u, x)$, variables (u, x) on the compact set $\mathscr{X} = \mathscr{B} \times K\mathscr{B}$, and N^* taken to be the constant N^0 given in Assumption 3.1. To this end, we verify the hypotheses of Lemma 3.2. Observe first that $\phi(u, x, \xi)$ is Lipschitz continuous w.r.t. u with Lipschitz modulus $\|\Gamma(x, \xi)\|$. The latter is bounded by $\kappa_2(\xi)$ under condition (c) of Theorem 3.2. On the other hand, for fixed ξ, (15) implies

$$|\phi(u, x', \xi) - \phi(u, x, \xi)| \leq \kappa_1(\xi)\|x' - x\|^\gamma$$

for $\|u\| \leq 1$. This shows

$$|\phi(u', x', \xi) - \phi(u, x, \xi)| \leq \kappa(\xi)(\|u' - u\| + \|x' - x\|^\gamma)$$

for all $\xi \in \tilde{\Xi}$, where $\kappa(\xi) = \kappa_1(\xi) + \kappa_2(\xi)$ for $(u, x) \in \mathscr{B} \times K\mathscr{B}$, (u', x') close to (u, x), which verifies (C2).

Note that (C3) follows from the second part of (c). To verify (C1), we note that $\phi(u, x', \xi) - \mathbb{E}[\phi(u, x, \xi)]$ is bounded by $\kappa(\xi) + \mathbb{E}[\kappa(\xi)]$. Therefore, (c) implies (C1). By Lemma 3.2, for any $\epsilon > 0$, there exist positive constants $c(\epsilon)$ and $\beta(\epsilon)$, independent of N, such that

$$\text{Prob}\left\{\sup_{(u, x) \in \mathscr{B} \times K\mathscr{B}} |\phi_N(u, x) - \psi(u, x)| \geq \epsilon\right\} \leq c(\epsilon)e^{-N\beta(\epsilon)}. \tag{16}$$

Observe that $\mathbb{E}[\sigma(u, \Gamma(x, \xi))] = \sigma(u, \mathbb{E}[\Gamma(x, \xi)])$. This is proved by Papageorgiou [30, Proposition 3.4]. Using this relationship, we have $\psi(u, x) = \sigma(u, \mathbb{E}[\Gamma(x, \xi)])$. Note that $\phi_N(u, x) = \sigma(u, \Gamma_N(x))$ (see, e.g., Hess [17, p. 621]). Then, by Lemma 3.1,

$$\begin{aligned}\sup_{(u, x) \in \mathscr{B} \times K\mathscr{B}} |\phi_N(u, x) - \psi(u, x)| &\geq \sup_{x \in K\mathscr{B}} \sup_{\|u\| \leq 1} |\sigma(u, \Gamma_N(x)) - \sigma(u, \mathbb{E}[\Gamma(x, \xi)])| \\ &= \sup_{x \in K\mathscr{B}} \mathbb{H}(\Gamma_N(x), \mathbb{E}[\Gamma(x, \xi)]).\end{aligned}$$

Combining this with (16), we obtain (11).

The proof of Theorem 3.2, part (ii) now follows from the bounds (9) and (11) by taking N^* as the maximum of N^0 and the constant N^* appearing in Assumption 3.2. □

Theorem 3.2 studies the case when the set-valued mapping $\Gamma(x, \xi)$ is continuous in x for every ξ and hence its expectation F is continuous. In practical applications such as first-order optimality conditions for nonsmooth stochastic programming, $\Gamma(x, \xi)$ is often not continuous. Taking this even further, our interest is in the situation when F is not continuous. To illustrate this, let us consider the following examples. (Further examples of continuous and discontinous stochastic set-valued mappings relating to the stochastic program (1) will appear in §§6.3.1 and 6.3.2.)

Example 3.1. Let $\Gamma(x, \xi)$ be the subdifferential mapping of function $|x - \xi|$. That is,

$$\Gamma(x, \xi) = \Gamma_1(x, \xi) := \begin{cases} \{1\} & \text{if } x > \xi, \\ \{-1\} & \text{if } x < \xi, \\ [-1, 1] & \text{if } x = \xi, \end{cases}$$

where $x \in \mathbb{R}$ and ξ is a random variable satisfying uniform distribution on $[-1, 1]$. Let $X = \mathbb{R}$ and $K\mathscr{B} = [-1, 1]$. It is easy to observe that, for every fixed ξ, $\Gamma_1(\cdot, \xi)$ is continuous on $[-1, 1]$ except at point $x = \xi$ at which $\Gamma_1(\cdot, \xi)$ is osc.

Let us now consider the case when Γ is the Clarke subdifferential of function $|x|(\xi - x)$, that is,

$$\Gamma(x, \xi) = \Gamma_2(x, \xi) := \begin{cases} \{\xi - 2x\} & \text{if } x > 0, \\ \{-\xi + 2x\} & \text{if } x < 0, \\ \xi[-1, 1] & \text{if } x = 0, \end{cases}$$

where $x \in \mathbb{R}$ and ξ is a random variable satisfying uniform distribution on $[-1, 1]$. Let $K\mathscr{B} = [-1, 1]$. It is easy to observe that, for every $\xi \neq 0$, $\Gamma(\cdot, \xi)$ is continuous on $K\mathscr{B}$ except at $x = 0$.

If we look at the expected values of the above two random set-valued mappings, it is not difficult to calculate that

$$\mathbb{E}[\Gamma_1(x, \xi)] = \int_{-1}^{x} (-1)\frac{1}{2}\, d\xi + \int_{x}^{1} \frac{1}{2}\, d\xi = \{-x\},$$

which is continuous, and

$$\mathbb{E}[\Gamma_2(x, \xi)] = \begin{cases} \{-2x\} & \text{if } [-1, 0), \\ \{2x\} & \text{if } x \in (0, 1], \\ [-1/2, 1/2] & \text{if } x = 0, \end{cases}$$

which is only outer semicontinuous.

The next section will provide tools for tackling a stochastic generalized equation $0 \in F(x) + \mathcal{N}_X(x)$, where $F(\cdot) = \mathbb{E}[\Gamma(\cdot, \xi)]$ and Γ is discontinuous but has a piecewise continuous structure. The results of §4.3, in particular, will apply to both examples Γ_1 and Γ_2.

4. Piecewise continuity of expectations and related exponential convergence of SAA. In this section, we aim to extend the exponential convergence result Theorem 3.2 by relaxing the key assumption on the continuity of $F(x) = \mathbb{E}[\Gamma(x, \xi)]$ to piecewise continuity. Stochastic polyhedral set-valued mappings yield many of the properties needed for this extension to hold as shown in Proposition 4.2. Although polyhedral set-valued mappings are piecewise continuous, the fact that the expectation of a stochastic polyhedral set-valued mapping need not be polyhdedral (Remark 4.1) may help to explain why the development of this section is required.

The extension of Theorem 3.2 to piecewise continuous set-valued mappings requires a number of technical preparations. These are given in §§4.1 and 4.2 and the main (extended) convergence results are described in §4.3 (Theorem 4.1).

Throughout this section, we consider the expectation $F(x) = \mathbb{E}[\Gamma(x, \xi)]$ of

$$\Gamma(x, \xi) = \bigcup_{j=1}^{\hat{\jmath}} \Gamma_j(x, \xi), \tag{17}$$

which is the union of finitely many continuous set-valued mappings Γ_j. Other conditions on Γ are detailed in Definition 4.1.

Recall that $\xi\colon \Omega \to \Xi \subset \mathbb{R}^r$ is a vector of random variables defined on probability space $(\Omega, \mathcal{F}, P)$, where the probability measure P is nonatomic. The expectation $\mathbb{E}[\phi(\xi)]$ of a measurable function $\phi\colon \mathbb{R}^r \to \mathbb{R}^s$ will be written as $\int \phi(\xi)\, d\mu_P(\xi)$, where the measure induced by P on $\mathbb{R}^r$ is defined by $\mu_P(\{\xi \in T\}) := P(\{\omega\colon \xi(\omega) \in T\})$ for any Lebesgue measurable set T in $\mathbb{R}^r$. Here and throughout this section, unless otherwise specified, the support set of ξ is Ξ; a.e. means almost everywhere in Ξ with respect μ_P or, equivalently, with probability 1; for a.a. ξ means for almost all $\xi \in \Xi$ (w.r.t. μ_P); integrals are taken over Ξ; and full measure describes a subset Ξ' of Ξ such that $\xi(\omega)$ lies in Ξ' w.p.1 or $\mu_P(\{\xi \in \Xi'\}) = 1$.

4.1. Stochastic piecewise continuous set-valued mappings. Stochastic piecewise continuous set-valued mappings Γ are, roughly speaking, those that are defined as the union of finitely many continuous set-valued mappings as described in §4 (and part i of Definition 4.1). We further require a kind of regularity of the domains of each of the continuous set-valued mappings (part ii of the definition).

Stochastic piecewise continuity is relative to a fixed compact set of $\mathbb{R}^m$, denoted by $\mathscr{X}$, which is intended to be the set of interest with regard to the evaluation of $F(x) = \mathbb{E}[\Gamma(x, \xi)]$. For instance, in §4.3 (where we will extend our main convergence result Theorem 3.2 to Theorem 4.2), we take $\mathscr{X} := K\mathscr{B}$ (here, K is given in Assumption 3.1) because this is the compact set where we carry out our convergence analysis and where we require piecewise continuity of $F(x)$. Another possibility occurs when the solution set X^* of the stochastic generalized equation $0 \in F(x) + \mathcal{N}_X(x)$ is compact. Then, we may choose $\mathscr{X}$ to be a compact neighborhood of X^* relative to X because, under Assumption 3.1, we can easily use Xu and Zhang [51, Theorem 4.1] to show that $x^N \in \mathscr{X}$ w.p.1 for N sufficiently large.

Recall that the domain of a set-valued mapping, denoted by the prefix dom, is the set of points for which it takes nonempty values.

Definition 4.1. Let $\mathscr{X}$ be a nonempty, compact set in $\mathbb{R}^m$. We say Γ is a *stochastic piecewise continuous* or *stochastic* PC^0 set-valued mapping on $\mathscr{X}$ if it has the form (17) for some positive integer $\hat{\jmath}$ and if the following

conditions hold:

(i) (a) for each j, $\Gamma_j\colon \mathbb{T}_j \to 2^{\mathbb{R}^s}$ is a set-valued mapping that is (Pompeiu-Hausdorff) continuous;
(b) there is an integrable function $\kappa\colon \Xi \to \mathbb{R}$ such that $\|\Gamma(x,\xi)\| \le \kappa(\xi)$ for each x near $\mathscr{X}$ and a.a. ξ; and

(ii) (a) each $\mathbb{T}_j$ is the closure of its interior $\mathbb{S}_j$ such that $\bigcup_{j=1}^{\hat{j}} \mathbb{T}_j$ is a neighborhood of $\mathscr{X} \times \Xi$ and $\mathbb{S}_1, \ldots, \mathbb{S}_{\hat{j}}$ are pairwise disjoint.

The *sections* of $\mathbb{T}_j$ and $\mathbb{S}_j$ w.r.t. x are the following set-valued mappings:

$$T_j(x) = \{\xi\colon (x,\xi) \in \mathbb{T}_j\} \qquad \text{and} \qquad S_j(x) = \{\xi\colon (x,\xi) \in \mathbb{S}_j\}.$$

These sections satisfy additional conditions in (ii):

(b) the domain of T_j is closed; and

(c) if the sequence $\{x^k\} \subset \operatorname{dom} S_j$ converges to x, then, for a.a. $\xi \in T_j(x)$ and sufficiently large k (depending on ξ), we have that $\xi \in T_j(x^k)$.

Part ii(c) of the definition is a kind of lower semicontinuity or inner continuity of T_j. It holds most obviously if each Γ_j is a function whose domain $\mathbb{T}_j$ is sufficiently regular, e.g., as illustrated in §6.3.1. It also holds when the sets $\mathbb{T}_j$ are simple such as the polyhedral case that we will develop in §4.3 and in §6.3.2.

Now, $F(x) = \mathbb{E}[\Gamma(x,\xi)]$ is well-defined by the stochastic PC^0 property i and the measurability of each (closed) set $\mathbb{T}_j$ (property ii(a)). Disjointness of the sets $\mathbb{S}_j$ gives the identity

$$\Gamma(x,\xi) = \Gamma_j(x,\xi), \quad \forall\, (x,\xi) \in \mathbb{S}_j,$$

as we now show: Any element (x,ξ) of $\mathbb{S}_j$ has a neighborhood W contained in $\mathbb{S}_j$. Thus, for $k \ne j$, W does not intersect $\mathbb{S}_k = \operatorname{int} \mathbb{T}_k$, which implies $(x,\xi) \notin \operatorname{cl} \mathbb{S}_k = \mathbb{T}_k$. We will use this identity later without reference.

We develop notation relating to these "active pieces" of Γ. As each $\mathbb{S}_j$ is open, the Lebesgue measure $\mathbb{T}_j \backslash \mathbb{S}_j$ is zero. By Fubini's theorem, for a.a. $x \in \operatorname{dom} T_j$, $\mu_P(T_j(x) \backslash S_j(x)) = 0$ and indeed a.a. ξ $(\in \Xi)$ belong to the open set $\bigcup_j \mathbb{S}_j(x)$. Let

$$\mathscr{D}_\Gamma = \{x\colon \mu_P(T_j(x) \backslash S_j(x)) = 0 \text{ for each } j\}.$$

As there are only finitely many indices j, a.a. $x \in \operatorname{dom} \Gamma$ lie in $\mathscr{D}_\Gamma$. We also mention that $\operatorname{cl} D_\Gamma$ is a neighborhood of X.

Associated with $x \in \mathscr{D}_\Gamma$ is an index set $J = \{j\colon S_j(x) \ne \varnothing\}$ such that the family $S_j(x)$, $j \in J$ consists of open, nonempty, and pairwise disjoint sets whose union has full measure (in Ξ). The definition of $\mathscr{D}_\Gamma$ therefore yields, for any $x \in \mathscr{D}_\Gamma$ and associated J, that

(a) $\mu_P(T_j(x)) > 0$ for every $j \in J$;
(b) $\bigcup_{j \in J} T_j(x)$ has full measure in Ξ; and
(c) for distinct $j, k \in J$, $T_j(x) \cap T_k(x)$ has measure zero.

This leads to our next point of notation. For any nonempty index set $J \subset \{1, \ldots, \hat{j}\}$, let

$$X_J := \{x \in \operatorname{dom} \Gamma\colon \text{(b) and (c) above hold}\},$$

$$\Gamma^J(x,\xi) := \begin{cases} \Gamma_j(x,\xi) & \text{if } x \in X_J,\ j \in J, \text{ and } \xi \in T_j(x), \\ \varnothing & \text{if } x \notin X_J. \end{cases}$$

For each $x \in X_J$, it is clear that $\Gamma^J(x,\xi)$ is (well-defined, nonempty, and) continuous for a.a. ξ. Its expectation is the set-valued mapping

$$F^J(x) := \begin{cases} \mathbb{E}[\Gamma^J(x,\xi)] & \text{if } x \in X_J, \\ \varnothing & \text{otherwise.} \end{cases}$$

These definitions attempt to identify a family of set-valued mappings—$F^J(x)$ for certain index sets J—from which $F(x)$ is selected and which are continuous (and, in particular, nonempty for each $x \in X$). These properties of the mappings F^J are shown next.

4.2. The expectation of a stochastic PC^0 mapping is piecewise continuous. Throughout this subsection, Γ and $\mathscr{X}$ are assumed to be as in Definition 4.1, and $\operatorname{cl} S$ and $\operatorname{int} S$ denote the closure and interior of a set S, respectively.

Observe that $\mathscr{D}_\Gamma \subset \bigcup_{J\subset\{1,\dots,\hat{\jmath}\}} X_J$; hence, $\operatorname{dom}\Gamma = \bigcup_{J\subset\{1,\dots,\hat{\jmath}\}} \operatorname{cl} X_J$. Because there are only finitely many possible index sets J, there is a family $\mathscr{J}$ of these such that X_J has interior for each $J \in \mathscr{J}$ and $\bigcup_{J\in\mathscr{J}} \operatorname{int} X_J$ has full measure in a neighborhood of $\mathscr{X}$. Hence, we define

$$\begin{aligned} \mathscr{J} &:= \{J \subset \{1,\dots,\hat{\jmath}\}\colon \operatorname{int} X_J \neq \varnothing\}, \\ \bar{X}_J &:= \operatorname{cl}\operatorname{int} X_J, \quad \forall J \in \mathscr{J} \end{aligned} \tag{18}$$

and observe that $\operatorname{dom}\Gamma = \bigcup_{J\in\mathscr{J}} \bar{X}_J$.

Proposition 4.1. *For $J \in \mathscr{J}$,*
(i) *X_J contains $\bar{X}_J$; and*
(ii) *$F^J = \mathbb{E}[\Gamma^J(\cdot,\xi)]$ is a continuous set-valued mapping on $\bar{X}_J$.*

Proof. Much of the proof is based on elementary details that we provide in the appendix. We sketch the main argument: If $j \in J$, $x \in \bar{X}_J$, and $\{x^k\}$ is a sequence in $(\operatorname{int} X_J)\cap\mathscr{D}_\Gamma$ that converges to x, then $F_j(x^k) \to F_j(x)$. This uses the indicator function $\mathbb{I}(T_j(x),\xi)$, defined to be 1 if $\xi \in T_j(x)$ and 0 otherwise. Property ii(c) of Definition 4.1 (stochastic piecewise continuity) can be used to show that $\mathbb{I}(T_j(x^k),\xi)$ converges to $\mathbb{I}(T_j(x),\xi)$ for a.a. ξ. Together with continuity of Γ_j, we get, via Lebesgue's dominated convergence theorem,

$$F_j(x^k) = \int \Gamma_J(x^k,\xi)\mathbb{I}(T_j(x^k),\xi)\,d\mu_P(\xi) \to \int \Gamma_J(x,\xi)\mathbb{I}(T_j(x),\xi)\,d\mu_P(\xi) = F_j(x). \quad \square$$

We are now ready to link the mappings $F^J(x) = \mathbb{E}[\Gamma^J(x,\xi)]$ to $F(x) = \mathbb{E}[\Gamma(x,\xi)]$ using the following family of index sets:

$$\mathscr{J}(x) = \{J \in \mathscr{J}\colon x \in \bar{X}_J\}. \tag{19}$$

Our focus is on x such that $x \in \bar{X}_J \cap \bar{X}_K$ for at least two index sets J and K in $\mathscr{J}$.

The first main technical result of §4—Theorem 4.1—says that F is piecewise continuous. Recall that Γ is the stochastic set-valued mapping satisfying piecewise properties i and ii (Definition 4.1), and F is its expectation. Also recall the notations $\mathscr{J}$ and $\bar{X}_J$ from (18) and $\mathscr{J}(x)$ from (19). The next result says that F is *piecewise continuous* or PC^0, i.e., outer semicontinuous such that, for each x, its set value $F(x)$ coincides with that of one of finitely many continuous set-valued mappings at x. This reduces to the usual definition of piecewise continuity when F is a function rather than a set-valued mapping.

Theorem 4.1 (Expectation of Stochastic PC^0 Mapping Is PC^0). (i) *For each $x \in \operatorname{dom}\Gamma$ (a closed neighborhood of $\mathscr{X}$), $\mathscr{J}(x)$ is a nonempty subset of $\mathscr{J}$ such that*

$$F(x) = \mathbb{E}\left[\bigcup_{J\in\mathscr{J}(x)} \Gamma^J(x,\xi)\right];$$

(ii) *Given $\varnothing \neq \bar{\mathscr{J}} \subset \mathscr{J}$, the set-valued mapping $\mathbb{E}[\bigcup_{J\in\bar{\mathscr{J}}}\Gamma^J(x,\xi)]$ is continuous on $\bigcap_{J\in\bar{\mathscr{J}}}\bar{X}_J$; and*
(iii) *F is outer semicontinuous.*

Proof. *Part* (i). Let $x \in \operatorname{dom}\Gamma$. It suffices to show that

$$\bigcup_j \Gamma_j(x,\xi) = \bigcup_{J\in\mathscr{J}(x)} \Gamma^J(x,\xi) \text{ a.e.},$$

where the first set defines $\Gamma(x,\xi)$. Clearly $\bigcup_j \Gamma_j(x,\xi)$ contains $\bigcup_{J\in\mathscr{J}(x)}\Gamma^J(x,\xi)$ a.e. by definition of $\mathscr{J}(x)$. The converse requires several facts relating to an arbitrary index j and point $x \in \operatorname{dom} T_j$. First, we have seen from the proof of Proposition 4.1 that x is the limit of a sequence $\{x^k\} \subset \operatorname{dom} S_j$. Indeed, we can assume $\{x^k\} \subset (\operatorname{dom} S_j)\cap\mathscr{D}_\Gamma$ because $\mathbb{S}_j$ and hence $\operatorname{dom} S_j$ has interior. Assume further that, by taking a subsequence if necessary, $J = \{l\colon S_l(x^k) \neq \varnothing\}$ is unchanged for all k. Thus, J lies in $\mathscr{J}(x)$ by definition of the latter and also $j \in J$. By construction, $\Gamma_j(x,\xi) = \Gamma^J(x,\xi)$ for a.a. $\xi \in T_j(x)$ as needed.

Part (ii). By definition, $\Gamma^J(\cdot,\xi)$, $J \in \bar{\mathscr{J}}$ is continuous on $\bigcap_{J\in\bar{\mathscr{J}}}\bar{X}_J$ for a.a. $\xi \in \Xi$ and continuity is preserved for the union of finitely many set-valued mappings $\bigcup_{J\in\bar{\mathscr{J}}}\Gamma^J(\cdot,\xi)$. Because $\Gamma(x,\xi)$ is integrably bounded by $\kappa(\xi)$, then $\|\Gamma^J(x,\xi)\| \le \kappa(\xi)$. By Aumann [3, Corollary 5.2], $\mathbb{E}[\bigcup_{J\in\bar{\mathscr{J}}}\Gamma^J(x,\xi)]$ is continuous on the same set.

Part (iii). The outer semicontinuity of $F(x)$ follows from this property of $\Gamma(x,\xi)$ and Aumann [3, Corollary 5.2]. The proof is complete. $\square$

In hindsight, we see that the underlying idea of Theorem 4.1 is to partition $\mathscr{X}$ into a finite number of sets where, in each set the "active set" of pieces F^J that define F remains constant. Continuity of F^J yields continuity of F in each of these sets and extends to continuity on the closure of these sets, giving the piecewise continuity property of F.

REMARK 4.1. It is no surprise that, if F is a polyhedral set-valued mapping, i.e., its graph is the union of finitely many closed and convex polyhedra, then whatever the structure of $\Gamma = \mathbb{E}[F(\cdot, \xi)]$, F is piecewise continuous. This can be shown formally by decomposing F into the union of polyhedral convex set-valued mappings, each of which is continuous by Rockafellar and Wets [38, Example 9.35]. (A similar argument appears in our proof of Proposition 4.2 in the next section.) Note that polyhedrality of Γ is not generally implied by polyhedrality of F (a difficulty that arises because of nonlinearity of the probability density function).

Let $p(\xi) = 1/\xi^2$ for $\Xi = [1, \infty)$ so that $\int_1^\infty p(\xi)\, d\xi = 1$, i.e., p is a probability distribution on Ξ. Now, take $f(x, \xi) = \min\{x, \xi\}$, a piecewise affine function. For $x \geq 1$, the expectation of $f(x, \xi)$ is

$$\int_1^x (\xi/\xi^2)\, d\xi + x \int_x^\infty (1/\xi^2)\, d\xi = [\ln(x) - \ln(1)] + x[0 - (-1/x)] = \ln(x) + 1,$$

which is not piecewise affine.

4.3. Fast convergence of SAA for piecewise continuous generalized equations. We are now ready to present the main result of §4. We extend the exponentially fast convergence of Theorem 3.2 for SAA applied to the generalized equation (7)—here, $0 \in F(x) + \mathcal{N}_X(x)$ with $F = \mathbb{E}[\Gamma(\cdot, \xi)]$—to piecewise continuous F as described by Theorem 4.1. As in §3, X^* denotes the solution set of $0 \in F(x) + \mathcal{N}_X(x)$ while x^N denotes any solution of the SAA (8), i.e., $0 \in \Gamma_N(x^N) + \mathcal{N}_X(x^N)$. The case when Γ is a polyhedral set-valued mapping is of special interest (see Proposition 4.2).

THEOREM 4.2. *Let X be a closed, nonempty set in $\mathbb{R}^m$. Suppose that* (a) *Assumptions 3.1 and 3.2 hold*; (b) *Γ is a stochastic* PC^0 *set-valued mapping on $K\mathscr{B}$*; (c) *$\Gamma(\cdot, \xi)$ is piecewise Hölder continuous on $K\mathscr{B}$, i.e., there exist $\gamma > 0$ and a function $\kappa_1 \colon \Xi \to \mathbb{R}_+$ such that, for each j,*

$$\mathbb{H}(\Gamma_j(x', \xi),\, \Gamma_j(x, \xi)) \leq \kappa_1(\xi)\|x' - x\|^\gamma$$

for all $x, x' \in \operatorname{dom}(\mathbb{T}_j) \cap K\mathscr{B}$, and $\xi \in \Xi$; and (d) *κ_1 is integrable, there is an integrable function $\kappa_2(\xi)$ such that, for each j, $\|\Gamma_j(x, \xi)\| \leq \kappa_2(\xi)$ for $(x, \xi) \in \mathbb{T}_j$, and the moment-generating function of $\kappa_1(\xi) + \kappa_2(\xi)$ is finite valued near zero. Then, $\{x^N\}$ converges exponentially quickly to X^* as described in Theorem 3.2*(ii).

PROOF. We outline the proof, details of which are found in the appendix. Let $\mathscr{X}$ be the set $K\mathscr{B}$. First, from Theorem 4.1, $\mathscr{X}$ can be covered by nonempty sets of the form $\bigcap_{J \in \bar{\mathscr{J}}} \bar{X}^J$ for some nonempty $J \subset \bar{\mathscr{J}}$. On each set, the expectation of $\Gamma(x, \xi)$ is continuous. Next, we can apply Theorem 3.2 to show an exponential rate of convergence of SAA applied to $\Gamma(x, \xi)$ on each of these sets. Because there are only finitely many of these, the result follows. □

We finish this section by studying *polyhedral set-valued mappings* (also known as polyhedral multifunctions), namely, set-valued mappings whose graphs are the union of finitely many closed, convex polyhedra. The following result shows that a polyhedral convex set-valued mapping is stochastic piecewise continuous and also Hölder continuous, the latter from Rockafellar and Wets [38, Example 9.57], as needed in Theorem 4.2. This will be useful in applications in §6.

PROPOSITION 4.2. *Let $\mathscr{X}$ be a nonempty, compact set in $\mathbb{R}^m$ and $\Gamma \colon \mathbb{R}^{m+r} \to 2^{\mathbb{R}^s}$ be a polyhedral set-valued mapping. If* (a) *$\operatorname{dom}\Gamma$ is both the closure of its interior and a neighborhood of $\mathscr{X} \times \Xi$, and* (b) *there is an integrable function $\kappa \colon \Xi \to \mathbb{R}$ such that $\|\Gamma(x, \xi)\| \leq \kappa(\xi)$ for x near $\mathscr{X}$ and a.a. ξ, then Γ is stochastic* PC^0 *and piecewise Lipschitz* (*take $\gamma = 1$ in part* (c) *of Theorem 4.2*) *on $\mathscr{X}$.*

PROOF. We start by writing $\Gamma(x, \xi) = \bigcup_{j=1}^{\hat{j}} \Gamma_j(x, \xi)$, where the graph of each Γ_j is convex in addition to being nonempty, closed, and polyhedral. For brevity, we omit details of two claims that can be shown by elementary arguments. We may assume without loss of generality that, first, $\operatorname{int}\operatorname{dom}\Gamma_j \neq \varnothing$ for each $j = 1, \ldots, \hat{j}$ and, second, the convex polyhedral set-valued mappings Γ_j may be chosen such that their domains $\mathbb{T}_j = \operatorname{dom}\Gamma_j$ have nonempty but pairwise disjoint interiors $\mathbb{S}_j$. Thus, stochastic PC^0 property ii(a) (i.e., property ii(a) in Definition 4.1) holds. Details appear in the appendix.

Looking at stochastic PC^0 property i, it is only left to observe that the submappings of each Γ_j are continuous, which is a consequence of being polyhedral convex set-valued mappings (see Rockafellar and Wets [38, Example 9.35]). In fact, we have Lipschitz continuity of each Γ_j (see Robinson [35] and Rockafellar and Wets [38, Example 9.57]). Polyhedrality of each $\mathbb{T}_j$ also gives stochastic PC^0 property ii(b).

For stochastic PC^0 property ii(c), fix $\bar{x}$. Because $T_j(\bar{x})$ is convex, its measure equals the measure of its interior (which is zero if the interior is empty). Thus, instead of "a.a. $\xi \in T_j(\bar{x})$," we consider "$\xi \in \operatorname{int} T_j(\bar{x})$." If $x^k \to \bar{x}$ and $S_j(x^k) \neq \varnothing$, then, because $\mathbb{T}_j$ is polyhedral, $T_j(x^k) \to T_j(x)$. It follows for $\xi \in \operatorname{int} T_j(\bar{x})$ and large k that $\xi \in \operatorname{int} T_j(x^k)$. □

Proposition 4.2 applies direction to the two examples Γ_1 and Γ_2 at the end of §3, each of which is a polyhedral set-valued mapping.

5. Optimality conditions. This section lays the groundwork for convergence analysis of stationary points of the SAA problem (4) in the §6. We do so in two parts: In the first part, we derive the stationary conditions for the true implicit program (6) in terms of $\partial_x v(x, \xi)$, the Clarke generalized gradient w.r.t. x of the second-stage global optimal value function; see (27) and Proposition 5.1. In the second part, we apply classical sensitivity analysis to $v(x, \xi)$ to approximate $\partial_x v(x, \xi)$ by a set-valued mapping constructed from the gradients w.r.t. x of the Lagrangian function at the global minima of the second stage problem; see Lemma 5.2. This paves the way for §6.2, where we introduce relaxed optimality conditions (Equation (41)) by enlarging the set of global optimal solutions to include all stationary points.

5.1. First-order necessary conditions for the true implicit program (6). The integrand function in the true implicit program (6) is the optimal value function of the second-stage problem (2). We therefore start our discussion from the latter.

Observe that the second-stage problem is a parametric minimization problem, where x and ξ can be treated as parameters. Gauvin and Dubeau [14] present a thorough discussion on optimal value function (called marginal function) of a general class of parametric maximization problems. In particular, they show that the Clarke subdifferential of the optimal value function is contained by the set of gradients of the Lagrangian function at solution points. Bonnans and Shapiro [6] take this further to present a comprehensive sensitivity analysis on a general class of nonsmooth parametric optimization problems in Banach space in terms of directional derivatives of Lagrangian function. More recently, Hu and Ralph [21] use the approach for the sensitivity analysis of MPECs. Here, we apply the results of Gauvin and Dubeau [14] results to our second-stage problem (2). First, we define the Lagrangian of (2) by

$$L(x, y, \xi, \lambda, \mu) := f(x, y, \xi) + \lambda^T h(x, y, \xi) + \mu^T g(x, y, \xi),$$

where $\lambda \in \mathbb{R}^t$, $\mu \in \mathbb{R}^s_+$. The first-order necessary optimality conditions of the second-stage problem (2) (also known as the KKT conditions) can be written as

$$0 = \nabla_y L(x, y, \xi, \lambda, \mu), \tag{20}$$

$$0 \leq \mu \perp -g(x, y, \xi) \geq 0, \tag{21}$$

$$0 = h(x, y, \xi). \tag{22}$$

A point y^* satisfying (20)–(22) is called a *KKT point* and the corresponding vectors λ^* and μ^* are called *Lagrange multipliers*. Following the notation of Gauvin and Dubeau [14], we use $S(x, \xi)$ to denote the feasible solution set of (2) as follows:

$$S(x, \xi) := \{y \in \mathbb{R}^n\colon\ g(x, y, \xi) \leq 0,\ \ h(x, y, \xi) = 0\}; \tag{23}$$

$P(x, \xi)$ to denote the set of global optimal solutions

$$P(x, \xi) := \{y \in S(x, \xi)\colon\ v(x, \xi) = f(x, y, \xi)\}; \tag{24}$$

and $W(x, y, \xi)$ to denote the set of Lagrange multipliers (λ, μ) corresponding to a stationary point y of the second-stage problem

$$W(x, y, \xi) := \{(\lambda, \mu)\colon \text{(20)–(22) hold}\}. \tag{25}$$

For fixed x, ξ, let $Q(x, \xi)$ denote the set of the stationary points of the second-stage problem (2):

$$Q(x, \xi) := \{y\colon \text{there exist multipliers such that (20)–(22) hold}\}. \tag{26}$$

To simplify discussion, we subsequently assume that the feasible set of the second-stage problem $S(x, \xi)$ is nonempty for any $x \in X$ and $\xi \in \Xi$.

DEFINITION 5.1. A feasible point $y \in S(x, \xi)$ is said to be *Mangasarian-Fromovitz (M-F) regular* if (a) there exists a vector $d \in \mathbb{R}^n$ such that $\nabla_y g_i(x, y, \xi)^T d < 0$ for $i \in I(x, y, \xi) := \{i: g_i(x, y, \xi) = 0\}$, and (b) the Jacobian matrix $\nabla_y h(x, y, \xi)$ has full row rank.

By Gauvin and Tolle [15, Theorem 2.2], if y is a local minimizer of second-stage problem (2), then this regularity condition is known to be necessary and sufficient to ensure that $W(x, y, \xi)$ is nonempty and compact. If y is a stationary point, i.e., $W(x, y, \xi)$ is nonempty, then M-F regularity is equivalent to boundedness of $W(x, y, \xi)$; see Nocedal and Wright [28, p. 353]. Lemma 5.1 next is an application of Gauvin and Dubeau [14, Theorem 5.1] to our second-stage problem (2).

LEMMA 5.1. *Suppose that, for almost every $\xi \in \Xi$, $S(x, \xi)$ is nonempty and uniformly compact for x near x^*. Suppose further that M-F regularity holds at $y^* \in P(x^*, \xi)$ for almost every $\xi \in \Xi$. Then, the optimal value function $v(x, \xi)$ is Lipschitz continuous in x near x^* for almost every ξ.*

Note that the uniform nonemptiness assumption on $S(x, \xi)$ implies a relatively complete recourse of the second-stage problem, that is, $\mathbb{E}[v(x, \xi)] < \infty$ for all $x \in X$. The compactness assumption implies that the set $\{y: h(x, y, \xi) = 0,\ g(x, y, \xi) \le 0\}$ is compact.

We now consider the first-order necessary optimality condition of the first-stage problem (6). Here, we assume that the expected value of $v(x, \xi)$ is well-defined. A sufficient condition for this is that $|v(x, \xi)|$ is bounded by an integrable function. Lemma 5.1 implies that the Clarke subdifferential $\partial_x v(x, \xi)$ is well-defined for almost every $\xi \in \Xi$. Moreover, if the Lipschitz module of $v(x, \xi)$ in x is bounded by an integrable function, then the above result further implies that $\vartheta(x)$ is locally Lipschitz continuous and, hence, $\partial\vartheta(x)$ is well-defined. Using the Clarke subdifferentials, we can write down the first-order necessary condition of the first-stage problem as

$$0 \in \partial\vartheta(x) + \mathcal{N}_X(x) \tag{27}$$

here and, later on, $\mathcal{N}_X(x)$ denotes Clarke's normal cone to X at x (see Clarke [8, p. 11] for the definition). It is well-known that Clarke's normal cone is a closed convex set and $\mathcal{N}_X(\cdot)$ is outer semicontinuous. Because the feasible set X here is convex, the normal cone coincides with usual normal cone in convex analysis. A feasible solution $x \in X$ satisfying (27) is known as the *Clarke stationary point*. From a numerical perspective, the optimality condition (27) is not very useful because the subdifferential of the expected value function is usually difficult to obtain. A more numerically motivated condition will be in terms of $\mathbb{E}[\partial_x v(x, \xi)]$. We need the following result:

PROPOSITION 5.1. *Let the conditions of Lemma 5.1 hold. If the Lipschitz modulus of $v(x, \xi)$ in x is bounded by an integrable function, then $\partial\vartheta(x) \subset \mathbb{E}[\partial_x v(x, \xi)]$. Equality holds when v is Clarke regular (Clarke [8, Definition 2.3.4]) at x for almost every $\xi \in \Xi$.*

The result is essentially Clarke's Theorem 2.7.2 in Clarke [8], given that the probability space here is assumed to be nonatomic. Note that by Clarke [8, Proposition 2.3.1 and Remark 2.3.5], the equality also holds when $-v$ is Clarke regular at x for almost every ξ. Using Proposition 5.1, we may consider a necessary optimality condition for the true implicit problem (6) as follows:

$$0 \in \mathbb{E}[\partial_x v(x, \xi)] + \mathcal{N}_X(x). \tag{28}$$

The two optimality conditions coincide under Clarke regularity as described in Proposition 5.1, for example, when $v(x, \xi)$ is continuously differentiable in x for a.a. ξ.

5.2. Weak first-order necessary conditions for the true implicit program. To analyze convergence of SAA, the stationary condition (28) may only be useful if the second-stage global optimal value function $v(x, \xi)$ can be computed (making the SAA (4) equivalent to the implicit SAA (5)). We propose an alternative that may be useful in the absence of global second-stage optimality, namely, a relaxed optimality condition that is specified via second-stage stationary points. In such a case, $\partial_x v(x, \xi)$ is not useful because $v(x, \xi)$ is not supposed to be calculated. In Lemma 5.2 next, we will construct an approximation $\Psi(x, \xi)$ to the set $\partial_x v(x, \xi)$.

LEMMA 5.2. *Let $x \in X$. Suppose that, for every $\xi \in \Xi$, $S(x, \xi)$ is nonempty and uniformly compact for x' near x and M-F regularity holds at $y \in P(x, \xi)$. Then,*

$$\partial_x v(x, \xi) \subset \Psi(x, \xi), \tag{29}$$

where

$$\Psi(x,\xi) := \operatorname{conv}\left\{ \bigcup_{y\in P(x,\xi)} \bigcup_{(\lambda,\mu)\in W(x,y,\xi)} \nabla_x L(x,y,\xi;\mu,\lambda) \right\} \tag{30}$$

and conv *denotes the convex hull. If M-F regularity is replaced by a linear independence constraint qualification, then* $\partial_x v(x,\xi) = \Psi(x,\xi)$.

Note that the condition in Lemma 5.2 is a bit stronger than those in Lemma 5.1 because we require (29) to hold for every ξ rather than almost every ξ. The result follows from an argument parallel to the proofs of Gauvin and Dubeau [14, Theorem 5.2 and Corollary 5.4] in that the upper Dini directional derivative of $v(x,\xi)$ in x is bounded (for all unit directions) by the support function of $\Psi(x',\xi)$ for all x' close to x. Therefore, the Clarke generalized directional derivative of v at x is bounded by the support function of $\Psi(x,\xi)$. We omit details of the proof.

Using Lemma 5.2, we may consider a weak optimality condition obtained by replacing $\partial_x v(x,\xi)$ with $\Psi(x,\xi)$ in (28), that is,

$$0 \in \mathbb{E}[\Psi(x,\xi)] + \mathcal{N}_X(x). \tag{31}$$

Because $\partial_x v(x,\xi) \subset \Psi(x,\xi)$, the optimality condition (31) is generally weaker than its counterpart (28). This can be illustrated by thinking of a second-stage problem with many local minima and few global minima, where the SAA iterates may converge (statistically) to a set that is much larger than the set of "true" stationary points of (6). However, a clear benefit in adopting the former is that it does not require the derivative information of $v(x,\xi)$ as in (28).

Assumption 5.1. *Let X^* be the solution set of (31) and X^* be bounded. For a compact neighborhood $B(X^*)$ of X^* relative to X, there exists an integrable function $\kappa_2(\xi) > 0$ (depending on $B(X^*)$) such that $\mathbb{E}[\kappa_2(\xi)^2] < \infty$ and*

$$\max\{\|\nabla_x f(x,\xi)\|, \|\nabla_x g(x,y,\xi)\|, \|\nabla_x h(x,y,\xi)\|, \|W(x,y,\xi)\|\} \le \kappa_2(\xi) \tag{32}$$

for all $x \in B(X^*)$, $\xi \in \Xi$, and $y \in P(x,\xi)$.

Note that the assumption on the boundedness of $W(x,y,\xi)$ by $\kappa_2(\xi)$ may be satisfied under some conditions on the derivatives of the underlying functions. We give a specific condition that is a kind of linear independence constraint qualification, uniform w.r.t. x: Let $\varnothing \ne I \subset \{1,\dots,r\}$, $(x,\xi) \in \mathcal{X} \times \Xi$, and $y \in P(x,\xi)$ be such that $I = \{i\colon g_i(x,y,\xi) = 0\}$, the index set of active constraints of $g(x,\xi)$. Denote (x,y,ξ) by z and consider the matrix $M(z) = [\nabla_y g_I(z) \;\; \nabla_y h(z)]^\top$. Let J be an index set such that the square submatrix $M(z)_{JJ}$ has maximum rank. If this rank equals the number of rows of $M(z)$, i.e., $M(z)$ has full rank, and if the inverse of $M(z)_{JJ}$ is norm bounded by $c\kappa_2(\xi)$ for some constant $c > 0$ (that is, independent of x and I), then Assumption 5.1 holds. Note also that, under Assumption 5.1, $\mathbb{E}[W(x,y,\xi)]$ is well-defined.

To establish the following necessary optimality condition, we assume uniform compactness of the set of second-stage global solutions $S(x,\xi)$ and M-F regularity at these solutions to ensure that Ψ has good local properties. We use Assumption 5.1 to ensure its integrability.

Theorem 5.1 (Weak First-Order Necessary Conditions). *Let x^* be an optimal solution to the true implicit programming problem (1). Assume that* (a) *for every $\xi \in \Xi$, $S(x,\xi)$ is nonempty, $S(x,\xi)$ is uniformly compact for x near x^*, M-F regularity holds at $y \in P(x^*,\xi)$; and* (b) *for every $\xi \in \Xi$, $S(x^*,\xi')$ is uniformly compact for ξ' near ξ. Under Assumption 5.1,*

$$0 \in \mathbb{E}[\Psi(x^*,\xi)] + \mathcal{N}_X(x^*). \tag{33}$$

We attach a proof in the appendix. It is essentially a verification of the well-definedness of Aumann's [3] integral $\mathbb{E}[\Psi(x^*,\xi)]$ through measurability and the bounded integrability of the set-valued mapping $\Psi(x,\xi)$.

During the revision of this paper, we noticed that Outrata and Römisch [29] recently derive some first-order optimality conditions similar to Theorem 5.1 for a class of two-stage nonsmooth stochastic minimization problems. The feasible set of their second-stage problem is an abstract measurable set and their optimality conditions are represented in terms of Mordukhovich's subdifferentials and coderivatives (of set-valued mappings); see Outrata and Römisch [29, Theorem 3.5] for details. Despite the differences in the problem settings, we believe that, under the nonatomic probability conditions, there is no essential difference between Theorem 5.1 and Outrata and Römisch [29, Theorem 3.5].

To conclude this section, let us comment that the aim of Theorem 5.1 is to derive the weak first-order necessary conditions for (6) in terms of the derivative information of the underlying function at the *global optimal solutions*

to the second-stage problem (2). This is sensible in the cases when the second-stage problem is convex (w.r.t. y) or the second-stage problem has a unique optimal solution. In the next section, we will consider the case that $P(x,\xi)$ is replaced by the set of stationary points $Q(x,\xi)$ and, consequently, the weak optimality condition is replaced by the relaxed optimality condition (41).

6. Convergence of stationary points. In this section, we analyze the convergence of stationary points of the SAA problem (4) by applying the results of §§3–5. Specifically, we relax the requirement of global optimality by enlarging the second-stage global solution set $P(x,\xi)$ (Equation (24)) to the corresponding set of stationary points $Q(x,\xi)$ (Equation (26)). By extension, we relax the stationary condition (33) to (41) by enlarging $\Psi(x,\xi)$ from (30) to the set-valued mapping $\Phi(x,\xi)$ defined in (39)–(40). Our main result is Theorem 6.2 on the exponential rate of convergence of $\{x^N\}$ (in §6.3).

Before proceeding, we briefly discuss the tractability of each SAA subproblem (4). One way is to solve it through (5), that is, for given x, we solve N second-stage problems for y^i, $i=1,\dots,N$. This is known as a two-stage decomposition method, which is popular in dealing with large-scale stochastic programs. Higle and Sen [18] and Rusczyński and Shapiro [40, Chapter 3] present comprehensive reviews of various decomposition methods for solving two-stage stochastic programs.

Another way to solve the SAA problem is to treat it as a standard nonlinear program and solve for $x, y^1,\dots,y^N$ simultaneously. A concern is that, when the sample size N is large, the SAA problem size may be too large to be solved effectively in computation. However, SAA methods have been shown to converge quickly, i.e., for N "not too large" in many cases (see Linderoth et al. [26], where sample sizes of 50 to 5,000 are used for various two-stage problems problems with 1 to 121 first-stage variables and 5 to 1,259 second stage variables). This underlies the hope that SAA will be practical for many problems when implemented by large-scale NLP codes.

6.1. KKT conditions of SAA problem (4). Our interest is in solving the SAA problem directly by an NLP code that may find a stationary point that is not a local or global solution. Thus, we cannot use the implicit program (5) for our convergence analysis. We formulate the KKT condition of the SAA problem (4) next, multiplying by $1/N$ on both sides in order to display sample averages of various functions:

$$0 \in \frac{1}{N}\sum_{i=1}^{N}\left[\nabla_x f(x,y^i,\xi^i)+\nabla_x h(x,y^i,\xi^i)^T\lambda^i+\nabla_x g(x,y^i,\xi^i)^T\mu^i\right]+\mathcal{N}_X(x) \tag{34}$$

and, for $i=1,\dots,N$,

$$\begin{aligned}
0 &= \frac{1}{N}\nabla_y f(x,y^i,\xi^i)+\frac{1}{N}\nabla_y h(x,y^i,\xi^i)^T\lambda^i+\frac{1}{N}\nabla_y g(x,y^i,\xi^i)^T\mu^i,\\
0 &\le \frac{1}{N}g(x,y^i,\xi^i)\perp\mu^i\ge 0,\\
0 &= \frac{1}{N}h(x,y^i,\xi^i).
\end{aligned}$$

Removing the factor $1/N$ in last three equations, we have (34) and, for $i=1,\dots,N$,

$$0 = \nabla_y f(x,y^i,\xi^i)+\nabla_y h(x,y^i,\xi^i)^T\lambda^i+\nabla_y g(x,y^i,\xi^i)^T\mu^i, \tag{35}$$

$$0 \le g(x,y^i,\xi^i)\perp\mu^i\ge 0, \tag{36}$$

$$0 = h(x,y^i,\xi^i). \tag{37}$$

An $(N+1)$-tuple $(x^N,y^1,\dots,y^N)$ is said to be a *stationary point* of the SAA problem (4) if it satisfies (34)–(37) for some multipliers $(\lambda^1,\mu^1),\dots,(\lambda^N,\mu^N)$.

A minor point on notation is that $(y^1,\lambda^1,\mu^1),\dots,(y^N,\lambda^N,\mu^N)$ will change as N changes, that is, it would be more accurate to denote each y^i by $y^{i,N}$ (and do so similarly with the other vectors). To keep the notation simple, we will take this point as understood.

6.2. Reduced (implicit) KKT conditions of SAA problem (4). We now begin to study the convergence of x^N as N tends to infinity. Specifically, we show that accumulation points are stationary for the true problem (1), where the meaning of "stationary" is developed in (41) (to follow). The difficulty for the convergence analysis here is that when N tends to infinity, so does the number of second-stage systems (see (34)–(37)). This motivates us to consider eliminating y^i, $i = 1, \ldots, N$ from the KKT conditions and focus on the convergence of x^N. In other words, we consider a reduced or implicit KKT condition for the SAA problem (4).

Observe that a point y^i satisfying (35)–(37) is a stationary point of the second-stage problem

$$\begin{aligned} \min_{y\in\mathbb{R}^n} \quad & f(x^N, y, \xi^i) \\ \text{s.t.} \quad & g(x^N, y, \xi^i) \le 0, \\ & h(x^N, y, \xi^i) = 0. \end{aligned} \tag{38}$$

Note that if (38) is a convex program or has a unique optimal solution, then y^i is the optimal solution to (38). Here, we consider the general case. Let

$$\hat{\Phi}(x, \xi) := \left\{ \bigcup_{y\in Q(x,\xi)} \bigcup_{(\lambda,\mu)\in W(x,y,\xi)} \nabla_x L(x, y, \xi; \lambda, \mu) \right\}, \tag{39}$$

where $Q(x, \xi)$ is the set of stationary points defined by (26), $W(x, y, \xi)$ is the set of corresponding multipliers defined by (25), and

$$\Phi(x, \xi) := \operatorname{conv} \hat{\Phi}(x, \xi), \tag{40}$$

that is, $\Phi(x, \xi)$ is the convex hull of the collection of gradient vectors $\nabla_x L(x, y, \xi; \lambda, \mu)$ at KKT points of the second-stage problem (2). By definition, it is easy to observe that

$$\partial_x v(x, \xi) \subset \Psi(x, \xi) \subset \Phi(x, \xi).$$

Consequently, we may weaken the optimality condition (31) to provide a *relaxed* optimality condition for the true implicit program (6):

$$0 \in \mathbb{E}[\Phi(x, \xi)] + \mathcal{N}_X(x). \tag{41}$$

It is relaxed in that the set of stationary points used to define the expectation is a relaxation of the set of globally optimal points used in (31). In fact, the optimality condition is both reduced (to x only) and relaxed.

Suppose $(x^N, y^1, \ldots, y^N)$ is a stationary point of SAA problem (4). Then, following from the definition of mapping Φ and KKT conditions (34)–(37), we can easily show that $x = x^N$ solves the following generalized equation:

$$0 \in \frac{1}{N} \sum_{i=1}^{N} \Phi(x, \xi^i) + \mathcal{N}_X(x). \tag{42}$$

This is the (reduced and) relaxed optimality condition for the SAA (4). In §6.4 (Theorem 6.2), we will show that w.p.1 every accumulation point of the sequence $\{x^N\}$ satisfies (41). Lemma 6.1 prepares us for this. Although (31) is sharper than the optimality condition (41), we cannot use it unless the SAA iterate $(x^N, y^1, \ldots, y^N)$ is sure to be globally optimal.

Assumption 6.1. *Let $\Phi(x, \xi)$ be defined as in (40).*

(a) *For any $x \in X$ and a.a. $\xi \in \Xi$, $S(\cdot, \cdot)$ is nonempty and uniformly compact near (x, ξ); M-F regularity holds at each stationary point $y \in Q(x, \xi)$ for a.a. ξ; and $Q(\cdot, \cdot)$ is nonempty and uniformly compact near (x, ξ).*

(b) *For every $K > 0$, there exists an integrable function $\kappa_3(\xi) > 0$ (depending on K) such that $\mathbb{E}[\kappa_3(\xi)^2] < \infty$ and*

$$\max\{\|\nabla_x f(x, \xi)\|, \|\nabla_x g(x, y, \xi)\|, \|\nabla_x h(x, y, \xi)\|, \|W(x, y, \xi)\|\} \le \kappa_3(\xi) \tag{43}$$

for all $x \in K\mathcal{B}$, $\xi \in \Xi$, and $y \in Q(x, \xi)$.

Assumption 6.1(b) covers Assumption 5.1 as $P(x, \xi) \subset Q(x, \xi)$. It holds trivially if Ξ is bounded; for each $K > 0$, the union of $Q(x, \xi)$ for $(x, \xi) \in K\mathcal{B} \times \Xi$ is bounded and the functions and gradients appearing in (43) are continuous.

LEMMA 6.1. *Let Assumption 6.1 hold. Then,*
(i) $\Phi(x, \xi(\cdot))$: $\Omega \to 2^{\mathbb{R}^m}$ *is a random set-valued mapping*; *and*
(ii) *there exists an integrable function* $\eta(\xi)$ *such that* $\|\Phi(x, \xi)\| \le \eta(\xi)$, *and* $\Phi(x, \xi)$ *is outer semicontinuous in* x *on* $K\mathcal{B}$.

PROOF. *Part* (i). The proof is similar to that of Theorem 5.1.
Part (ii). It is easy to demonstrate that there exists a constant $C > 0$ such that $\|\nabla_x L(x, y, \xi; \lambda, \mu)\| \le C(\kappa_3(\xi) + \kappa_3(\xi)^2)$. Obviously, $\eta(\xi) := C(\kappa_3(\xi) + \kappa_3(\xi)^2)$ is integrable under the condition that $\mathbb{E}[\kappa_3(\xi)^2] < \infty$. The outer semicontinuity comes from Proposition 2.1 and uniform local compactness of $\Phi(x, \xi)$ at every point in $B(X^*)$. □

Next, we investigate the convergence of the stationary sequence $\{x^N\}$ as N tends to infinity. We do so by investigating how the relaxed SAA optimality condition (42) approximates (41).

REMARK 6.1. The introduction of the mapping $\Phi(x, \xi)$ and the optimality conditions (42) and (41) are motivated by the necessity to accommodate the stationary points of the SAA problem (4). If we are guaranteed that each y^i is locally optimal for the second-stage problem, then clearly the first-order optimality condition (41) can be tightened. Furthermore, if a global solution $y(x, \xi)$ of the second-stage problem can always be found (for example, (38) has a nonempty, compact feasible set and only one stationary point), then the tighter optimality condition of §5 applies.

6.3. Convergence of sequence $\{x^N\}$. We proceed to the main result of §6. Let $X^\#$ denote its solution set of (41), that is,

$$X^\# := \{x \in X\colon 0 \in \mathbb{E}[\Phi(x, \xi)] + \mathcal{N}_X(x)\}. \tag{44}$$

We replicate Assumption 3.1 and Theorems 3.1 and 3.2 in the context here.

ASSUMPTION 6.2. *There are positive constants* N^0 *and* K (*independent of* ω) *such that for* $N \ge N^0$, *w.p.*1 *the SAA method finds a stationary point* $(x^N, y^1, \ldots, y^N)$ *of* (4) *such that* $x^N \in K\mathcal{B}$.

As discussed after Assumption 3.1, though the assumption holds almost automatically if X is bounded, sufficient conditions for the unbounded case is an open question. Next, we have an application of Theorem 3.1 where Assumption 6.2 replaces Assumption 3.1 and Lemma 6.1 is used to provide the necessary properties of $\Phi(x, \xi)$.

THEOREM 6.1. *Let Assumptions 6.1 and 6.2 hold and let* $X^\#$ *be defined by* (44). *Then, w.p.*1, $d(x^N, X^\#) \to 0$ *as* $N \to \infty$.

Note that the above almost sure convergence result follows essentially from uniform semiconvergence of $\Phi_N(x)$ to $\mathbb{E}[\Phi(x, \xi)]$ (see Theorem 3.1). Therefore, we mention a few references regarding uniform convergence of sample average random set-valued mappings. Shapiro [41] considers uniform convergence of sample average of the Clarke subdifferential of a random function. By assuming a.e. differentiability of the random function at every point, Shapiro [41] shows that the expected value function is continuously differentiable and that the sample average of the Clarke subgradients converge uniformly to the gradient of its true counterpart. Shapiro and Xu [46] use this result to establish the convergence of the SAA stationary sequence for an SMPEC. See Propositions 4.1(c) and 4.2 in Shapiro and Xu [46]. In the case when the expected value function is *not* continuously differentiable, a uniform strong law of large numbers for random set-valued mappings has recently been developed (Xu and Meng [49], Shapiro and Xu [47]) and applied to analyze the convergence of stationary points of SAA problems for a two-stage stochastic minimization problem with nonsmooth equality constraints (Xu and Meng [49]). In all of these papers, the second-stage problem is assumed to have a unique feasible solution for every x and ξ. Meng and Xu extend Xu and Meng [49] by considering a case when the second-stage problem has multiple solutions but they alter their model slightly by replacing the optimal solution at the second stage with an arbitrary feasible point.

We now move on to discuss exponential convergence.

ASSUMPTION 6.3. *Let* $K > 0$ *be given and* $X^\#$ *be defined by* (44). *Let* $X^\#$ *be nonempty, and let there exist positive constants* ν, α, *and* N^* *such that for* $N \ge N^*$, *w.p.*1 *the SAA solution* $(x^N, y^1, \ldots, y^N)$ *satisfies*

$$d(x^N, X^\#) \le \alpha \sup_{x \in K\mathcal{B}} \mathbb{D}(\Phi_N(x), \mathbb{E}[\Phi(x, \xi)])^\nu. \tag{45}$$

Note that calmness of $\mathbb{E}[\Phi(x, \xi)] + \mathcal{N}_X(x)$ is sufficient for (45); see Proposition 3.1 and subsequent discussions. The main result follows; it is an application of Theorem 4.2, i.e., an extension of Theorem 3.2.

THEOREM 6.2. *Let Assumptions 6.1, 6.2, and 6.3 hold. Suppose also that* (a) $\hat{\Phi}$ *is stochastic* PC^0 *on* $K\mathcal{B}$ *(see Definition 4.1);* (b) $\hat{\Phi}$ *is piecewise Hölder continuous on* $K\mathcal{B}$ *w.r.t. the bound* $\kappa_1(\xi)$ *(condition* (c) *of Theorem 4.2);* (c) κ_1 *is integrable and the integrable bound* $\eta(\xi)$ *on* $\|\hat{\Phi}(x,\xi)\|$ *given by Lemma 6.1*(ii) *is such that the moment-generating function of* $\kappa_1(\xi)+\eta(\xi)$ *is finite valued near zero. Then, for any small positive number* ϵ*, there exist positive constants* $c(\epsilon)$ *and* $\beta(\epsilon)$*, independent of* N*, and* $N^*>0$ *such that the following holds w.p.*1 *for* $N \geq N^*$:

$$\mathrm{Prob}\{d(x^N, X^\#) \geq \epsilon\} \leq c(\epsilon)e^{-\beta(\epsilon)N}. \tag{46}$$

Theorem 6.2 presents an exponential convergence if $\Phi(x,\xi)$ is stochastic PC^0 and piecewise Hölder continuous in a neighborhood of x^*. Over the past few years, various exponential convergence results have been established for SAA but the focuses are mainly on globally optimal solutions and globally optimal values. (See, for instance, Kaniovski et al. [25], Römisch and Schultz [39], Shapiro and Homem-de-Mello [45], Shapiro [43, 42], Shapiro and Xu [46], and the references therein). The *novelty* of Theorem 6.2 as far as we are concerned is that it is the first time that exponential convergence is established for *stationary points* of SAA problems.

In the next two subsections, we contrast the domain of application of Theorem 3.2 with that of Theorem 6.2 by looking at some special subclasses of the stochastic program (1).

6.3.1. Application when the second stage is unconstrained. We take an extremely simple version of (1) in which the second-stage problem is unconstrained:

$$\begin{aligned} \min_{x\in\mathbb{R}^m,\, y(\cdot)\in Y} \quad & \mathbb{E}[f(x, y(\omega), \xi(\omega))] \\ \text{s.t.} \quad & x \in X, \end{aligned} \tag{47}$$

where Y is a suitable space of functions $y(\cdot)\colon \Omega \to \mathbb{R}^n$ such that $\mathbb{E}[f(x, y(\omega), \xi(\omega))]$ is well-defined. Specifically, we will allow $f(x,\cdot,\xi)$ to have multiple but locally unique stationary points (hence, to be nonconvex). We will show that Theorem 3.2, which relies on continuity of the various set-valued mappings, can be applied under various assumptions to follow.

The set of second-stage stationary points for each first-stage decision x and second-stage scenario ξ is $Q(x,\xi) = \{y\colon \nabla_y f(x,y,\xi) = 0\}$. The mapping Φ from (40) used to define the relaxed optimality condition $0 \in \mathbb{E}[\Phi(x,\xi)] + \mathcal{N}_X(x)$ is

$$\Phi(x,\xi) := \mathrm{conv}\{\nabla_x f(x,y,\xi)\colon y \in Q(x,\xi)\}.$$

Let $X^\#$ be the solution set of the relaxed optimality condition as in (44).

The SAA subproblem, given samples $\xi_1, \ldots, \xi_N$, is to find x^N solving (42), that is,

$$0 \in \mathrm{conv}\{\nabla_x f(x^N, y_i^N, \xi^i)\colon i = 1, \ldots, N\} + N_X(x^N),$$

where $y_i^N \in Q(x^N, \xi^i)$ for each i.

Given (x,y,ξ), denote the Hessian matrix of f w.r.t. y by $H(x,\xi,y)$:

$$H(x,\xi,y) := \nabla^2_{yy} f(x,y,\xi).$$

We assume that:

(a) X and Ξ are nonempty and compact sets in $\mathbb{R}^m$ and $\mathbb{R}^r$, respectively.

(b) $f(x,y,\xi)$ is twice differentiable w.r.t. (x,y) such that its second derivatives $\nabla^2_{xy} f(x,y,\xi)$ and $H(x,y,\xi)$ are continuous in a neighborhood of $X \times \mathbb{R}^n \times \Xi$.

(c) $f(x,\cdot,\xi)$ and $\|\nabla_y f(x,\cdot,\xi)\|$ are coercive uniformly w.r.t. $(x,\xi) \in X \times \Xi$ and that, for each $\Delta > 0$, there is $R > 0$ such that $f(x,y,\xi) \geq \Delta$ and $\|\nabla_y f(x,y,\xi)\| \geq \delta$ for all $(x,\xi) \in X \times \Xi$ and y with $\|y\| \geq R$.

(d) For each $(x,\xi) \in X \times \Xi$, $H(x,y,\xi)$ is invertible for all[4] $y \in Q(x,\xi)$ such that $H(x,y,\xi)^{-1}$ is bounded in norm by a integrable function $\kappa(\xi)$ that is independent of $x \in X$.

[4] It would be simpler to require invertibility of $H(x,y,\xi)$ for a.a. ξ and all y (even nonstationary points), but with condition (c), it can be shown that this would imply convexity w.r.t. y.

For $(x, \xi) \in X \times \Xi$, coercivity (condition (c)) yields a global minimum of $f(x, \cdot, \xi)$ and hence nonemptiness of $Q(x, \xi)$. Uniform coercivity of $\nabla_y f(x, \cdot, \xi)$ (condition (c)) together with compactness of $X \times \Xi$ (condition (a)) yields boundedness of the graph of $Q(\cdot)$ restricted to $X \times \Xi$, denoted $G = \{(x, \xi, y)\colon (x, \xi) \in X \times \Xi,\ y \in Q(x, \xi)\}$. G is also closed and hence compact because if $\{(x^k, \xi^k, y^k)\}$ is a sequence in G that converges to $(\bar{x}, \bar{\xi}, \bar{y})$, then $\bar{y} \in Q(\bar{x}, \bar{y})$ by continuity of $\nabla_y f(\cdot)$ (condition (b)).

We further claim that $Q(\cdot, \xi)$ is a Lipschitz continuous set-valued mapping on X uniformly w.r.t. $\xi \in \Xi$, and we show this in five steps.

Step 1. Fix $(\bar{x}, \bar{\xi}) \in X \times \Xi$ and $\bar{y} \in Q(\bar{x}, \bar{\xi})$. By invertibility of $H(\bar{x}, \bar{y}, \bar{\xi})$ (condition (d)), the implicit function theorem says there are bounded neighborhoods U of $(\bar{x}, \bar{\xi})$ and V of $\bar{y}$, and a function $y(\cdot)\colon U \to V$ with the following two properties: (i) $y(\bar{x}, \bar{\xi}) = \bar{y}$ and $Q(x, \xi) \cap V = \{y(x, \xi)\}$ for each $(x, \xi) \in U$, and (ii) $y(\cdot)$ is continuous in (x, ξ) and partially differentiable w.r.t. x such that $\nabla_x y(\cdot)$ is continuous. For some integrable $\hat{\kappa}(\xi)$, $\|\nabla_x y(x, \xi)\| \le \hat{\kappa}(\xi)$ for $(x, \xi) \in U$; hence, $y(\cdot, \xi)$ has Lipschitz continuity given by $\|y(x_1, \xi) - y(x_2, \xi)\| \le \hat{\kappa}(\xi)\|x_1 - x_2\|$ for x_1, x_2 with (x_1, ξ), $(x_2, \xi) \in U$. We refer to $y(\cdot)$ as a stationary point function because $y(x, \xi) \in Q(x, \xi)$ for $(x, \xi) \in U$.

Step 2. Because each $(\bar{x}, \bar{\xi}, \bar{y}) \in G$ is associated with an open set $U \times V$ and function $y(\cdot)\colon U \to V$ as described in Step 1 and G is also compact, we may cover G with a finite number $\hat{\jmath}$ of these neighborhoods. For indices $j = 1, \ldots, \hat{\jmath}$, we label these neighborhoods as $U_j \times V_j$, and their associated functions as $y_j(\cdot)$. Note that, for any indices j_1, j_2, if there exists $(x, \xi) \in U_{j_1} \cap U_{j_2}$ with $y_{j_1}(x, \xi) \in V_{j_2}$, then these stationary point functions coincide on $U_{j_1} \cap U_{j_2}$. Thus, we can combine them: $Q(x, \xi) \cap (V_{j_1} \cup V_{j_2})$ is a function on $U_{j_1} \cup U_{j_2}$ that coincides with $y_{j_1}(\cdot)$ on U_{j_1} and $y_{j_2}(\cdot)$ on U_{j_2} and is differentiable w.r.t. x with the associated derivative being continuous and bounded on $U_{j_1} \cup U_{j_2}$. We can inductively combine such "overlapping" neighborhoods until there is no further overlapping. Thus, without loss of generality, we may assume that if j_1 and j_2 are distinct indices, then $y_{j_1}(x, \xi) \notin V_{j_2}$ for each $(x, \xi) \in U_{j_1} \cap U_{j_2}$.

Step 3. Next, we claim that each U_j is closed *relative* to $X \times \Xi$.[5] To see closedness, take a limit point $(\bar{x}, \bar{\xi})$ of a sequence $\{(x^k, \xi^k)\}$ in $U_J \cap (X \times \Xi)$. Let $y^k = y_j(x^k, \xi^k)$ so that the sequence $\{(x^k, \xi^k, y^k)\}$ lies in G. As noted earlier, there is a limit point $\bar{y}$ of $\{y^k\}$ with $(\bar{x}, \bar{\xi}, \bar{y}) \in G$. Thus, by Step 2, $\bar{y} = y_{j_2}(\bar{x}, \bar{\xi})$ for some index j_2. Hence, on an appropriate subsequence, each (x^k, ξ^k) lies in U_{j_2} and each y^k lies in V_{j_2}. Step 2, therefore, gives $j_2 = j$. Thus, $(\bar{x}, \bar{\xi}) \in \hat{U}_j$ and $\hat{U}_j$ is closed.

Step 4. For a given $(\bar{x}, \bar{\xi}) \in X \times \Xi$, we show that the list of stationary functions with $(\bar{x}, \bar{\xi})$ in their domains is locally constant. Take indices j_1 and j_2 such that $(\bar{x}, \bar{\xi}) \in U_{j_1} \setminus U_{j_2}$. Openness of U_{j_1} implies that $(x, \xi) \in U_{j_1}$ for all (x, ξ) in $X \times \Xi$ near $(\bar{x}, \bar{\xi})$. Closedness of U_{j_2} relative to $X \times \Xi$ implies that $(x, \xi) \notin U_{j_2}$ for all (x, ξ) in $X \times \Xi$ near $(\bar{x}, \bar{\xi})$. In other words, any $(\bar{x}, \bar{\xi}) \in X \times \Xi$ has a neighborhood U such that the set of indices $J = \{j\colon (\bar{x}, \bar{y}) \in U_j\}$ coincides with $\{j\colon (x, \xi) \in U_j\}$ for all $(x, \xi) \in U$. Thus, for $(x, \xi) \in U$, $Q(x, \xi) = \{y_j(x, \xi)\colon j \in J\}$.

Step 5. Take $(\bar{x}, \bar{\xi}) \in X \times \Xi$, the neighborhood U, and index set J as in Step 4. We may assume without loss of generality that U is a Cartesian product of $U^X \subset X$ and $U^\Xi \subset \Xi$. The Lipschitz property of $y(\cdot)$ from earlier gives a constant $\kappa > 0$, which is independent of $(\bar{x}, \bar{\xi})$, U, and $j \in J$ such that, for any $x_1, x_2 \in U^X$ and $\xi \in U^\Xi$, $\|y_j(x_1, \xi) - y_j(x_2, \xi)\| \le \hat{\kappa}(\xi)\|x_1 - x_2\|$. Obviously, in light of Step 4, this implies the same Lipschitz property for $Q(\cdot)$ on U:

$$\mathbb{D}_H(Q(x_1, \xi), Q(x_2, \xi)) \le \hat{\kappa}(\xi)\|x_1 - x_2\| \quad \text{for all } x_1, x_2 \in U^X,\ \xi \in U^\Xi.$$

We leave it to the reader to see that a set-valued mapping that is Lipschitz on each neighborhood of an open cover of a compact set is Lipschitz on that compact set. The same argument yields Lipschitz continuity of $Q(\cdot, \xi)$ on X such that $\Gamma(x, \xi) = Q(x, \xi)$ fulfills the Hölder criterion of Theorem 3.2 with $\kappa_1(\xi) = \hat{\kappa}(\xi)$ and $\gamma = 1$.

We conclude that Theorem 3.2 delivers exponential convergence of SAA solutions $\{x^N\}$ to the relaxed solution set $X^\#$ if, in addition to (a)–(d) earlier, Assumption 3.1 holds. (Of course, we can obtain a similar outcome in the more general setting of Theorem 6.2.)

6.3.2. Application when the second stage is a nonconvex quadratic program. Let us now look at a class of stochastic programs with fixed recourse

$$\begin{aligned} \min_{x \in X,\, y(\cdot)} \quad & \mathbb{E}[p(x, \xi(\omega))] + \mathbb{E}[q(x, y(\omega), \xi(\omega))] \\ \text{s.t.} \quad & y(\omega) \ge 0, \quad H^1 x + H^2 \xi(\omega) + M y(\omega) = h, \ \text{a.e. } \omega \in \Omega, \end{aligned} \tag{48}$$

[5] That is, U_j is open and $U_j \cap X \times \Xi$ is closed and thus U_j would contain $X \times \Xi$ if the latter set were connected. In this case, $Q(x, \xi)$ would be a list of $\hat{\jmath}$ functions, each with a Lipschitz continuity property in x. Hence, Lipschitz continuity of $Q(\cdot, \xi)$ would follow. The remaining Steps 4 and 5 deal with the possibility that $X \times \Xi$ is unconnected.

where $p\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ is continuously differentiable, $q\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ is a (possibly nonconvex) quadratic function, $\xi\colon \Omega \to \Xi$ is a random vector as above, $\Xi \subset \mathbb{R}^r$, $X \subset \mathbb{R}^n$, $H^1 \in \mathbb{R}^{q\times n}$, $H^2 \in \mathbb{R}^{q\times r}$ are constant matrices, $h \in \mathbb{R}^q$ is a constant vector, and $M \in \mathbb{R}^{q\times m}$ is a constant matrix.

This is a two-stage stochastic program with fixed recourse: At the first stage, a decision on x needs to be made before realization of ξ. In making this decision, it is assumed that, at the second stage, if x is given and ξ is realized, then $y(\xi) := y(x, \xi)$ will be determined by finding a stationary point of a possibly nonconvex quadratic program (QP):

$$\begin{aligned} \min_y \quad & q(x, y, \xi) \\ \text{s.t.} \quad & y \geq 0, \\ & H^1 x + H^2 \xi + My = h. \end{aligned} \tag{49}$$

In order to satisfy many of the hypotheses entailed in Theorem 6.2, especially Assumptions 6.1 and 6.2, we introduce further conditions:[6]

(a) X and Ξ are compact, convex polyhedra with nonempty interiors.

(b) M has full row rank and there exists $y > 0$ with $My = 0$.

(c) The constant matrix $B = \nabla_y^2 q(x, y, \xi)$, which defines the pure quadratic part of q w.r.t. y, is such that $0 \neq y \geq 0$ and $My = 0$ implies $y^\top B y > 0$.

We sketch the implications of these conditions and leave the details to readers. Part (b) is sufficient, for any (x, ξ), for the Mangasarian-Fromowitz constraint qualification (MFCQ) to hold at each feasible point y of the second-stage problem (49) and also for complete recourse (feasibility of the second-stage problem). It implies a perhaps more commonly known sufficient condition for completeness from Birge and Louveaux [5, p. 92]: there exists $y \geq 0$ such that $My = t$ for any $t \in \mathbb{R}^q$. Part (c) implies boundedness of the set of second-stage stationary points $Q(x, \xi)$, uniformly for (x, ξ) in any bounded set such as $X \times \Xi$. Thus, (a)–(c) give Assumption 6.1, where uniform boundedness of the second-stage feasible set $S(x, \xi)$ can be assured by introducing a ball constraint $\|y\| \leq \rho$ of sufficiently large radius ρ that it is never binding at stationary points.

Existence of stationary points of both (49) and the SAA of the underlying problem (48) follows from Wolfe's theorem for QPs because, in both cases, the QPs have nonempty feasible sets and are bounded below (using part (c)). Thus, Assumption 6.2 also holds and Theorem 6.1 says that the set of solutions $X^\#$ of the relaxed optimality conditions is nonempty and bounded.

With the introduction of multipliers $\lambda \in \mathbb{R}^q$ and $\mu \in \mathbb{R}^m$, the Lagrangian function of the second-stage problem is

$$L(x, y, \xi; \lambda, \mu) = q(x, y, \xi) + (H^1 x + H^2 \xi + My - h)^\top \lambda - y^\top \mu,$$

and the associated KKT conditions are

$$0 = \nabla_y L(x, y, \xi; \lambda, \mu), \tag{50}$$

$$0 = H^1 x + H^2 \xi + My - h, \tag{51}$$

$$0 \leq \mu \perp y \geq 0. \tag{52}$$

This system is piecewise linear because of linearity of $\nabla_y L$; hence, the set of its solutions $(x, y, \xi, \lambda, \mu)$ is polyhedral. Likewise, linearity of $\nabla_x L$ give polyhedrality of

$$\hat{\Phi}(x, \xi) = \{\nabla_x L(x, y, \xi; \lambda, \mu)\colon\ x \geq 0,\ Ax = b,\ \xi \in \Xi,\ (x, y, \xi, \lambda, \mu)\},$$

which satisfies (50)–(52).

From previous discussions, $\hat{\Phi}(x, \xi)$ is nonempty for any $(x, \xi) \in \mathbb{R}^n \times \mathbb{R}^r$, i.e., $\operatorname{dom} \hat{\Phi} = \mathbb{R}^{n+r}$. Proposition 4.2 and Theorem 4.1 yield that $\hat{\Phi}$ is piecewise continuous. Thus, the exponential convergence result Theorem 6.2 can be applied if, in addition, (45) is valid for constants ν, α, and large enough N. (Theorem 3.2, however, cannot be used here because of the discontinuity of $\hat{\Phi}$.)

Suppose, in addition to the earlier assumptions, $p(x, \xi)$ is convex in x and q is independent of x and is convex in y. Let $v(x, \xi) = p(x, \xi) + q(x, y(x, \xi), \xi)$ and $\vartheta(x) = \mathbb{E}[v(x, \xi)]$. Then, we can show that $\vartheta(x)$ is convex and

$$\partial \vartheta(x) = \mathbb{E}[\partial_x v(x, \xi)] = \mathbb{E}[\Psi(x, \xi)] = \mathbb{E}[\Phi(x, \xi)] = \mathbb{E}[\hat{\Phi}(x, \xi)].$$

Because a stationary point of $\vartheta(x)$ is a global minimizer, $X^\#$ is the set of global optimal solutions of the true problem. Note that, if $v(\cdot, \xi) - p(\cdot, \xi)$ is strongly convex on $X^\#$, then we can easily show that (45) holds and,

[6] Of course, these conditions can be significantly weakened. We consider them only for the simplicity and clarity of presentation.

hence, show the exponential convergence of x^N to $X^\#$ (a singleton). This recovers Römisch and Schultz [39, Proposition 3.2].

6.4. Further remarks. Note that, in some cases, NLP solvers can provide solutions that satisfy some kind of second-order necessary conditions in addition to first-order necessary (i.e., KKT) conditions. Consequently, one may obtain a local optimal solution in solving the SAA problems (4). In such a case, we may employ tighter optimality conditions than (41) and (42) by replacing $Q(x,\xi)$ with the set of points for which first- and second-order necessary conditions hold. This will allow us to strengthen the convergence result Theorem 6.2 with a correspondingly smaller set $X^\#$.

Another issue concerning practical implementation of (4) is the validation of an approximate solution x^N solved from (4), that is, to evaluate the quality of this candidate solution by looking at how close x^N is to its true counterpart with certain confidence. Let us use $\bar{x}$ instead of x^N to denote the current approximate solution. We would like to evaluate the quality of this candidate solution. Note that, if we are guaranteed to obtain global optimal optimum in solving (4), then validation analysis can be carried out through (5) and (6) (see the estimation of the optimality gap in Shapiro [43, §4.1]). Here, we treat the candidate solution as a stationary point obtained from solving (4). We give a brief discussion about this to distinguish the mathematical treatments of the cases when the mapping $\Phi(\bar{x},\xi)$ to be integrated is single or set valued. Assume that $\bar{x}$ is close to $X^\#$ and that $\Phi(\bar{x},\xi)$ is a singleton in a neighborhood of $\bar{x}$ for all $\xi\in\Xi$ (e.g., the second-stage problem has a unique stationary point that is the optimal solution). In this case, $v(\cdot,\xi)$ is differentiable near $\bar{x}$ and $\nabla\vartheta(\bar{x})=\mathbb{E}[\nabla_x v(\bar{x},\xi)]=\mathbb{E}[\Phi(\bar{x},\xi)]$ under some regularity conditions. Consequently, $\Phi^N(x)=(1/N)\sum_{i=1}^N \nabla_x v(x,\xi^i)$ is an unbiased estimator of $\mathbb{E}[\Phi(x,\xi)]$. Shapiro and Homem-de-Mello [44, §4] present a complete discussion on statistical validation of the sample average gradients and hence the KKT conditions. The approach can be applied to $\mathbb{D}(\Phi^N(\bar{x}),\mathbb{E}[\Phi(\bar{x},\xi)])$ in our case and, subsequently, through (45) to obtain an estimation of the error bound of the approximate stationary point. In the case when $\Phi(\bar{x},\xi)$ is set valued, it is difficult to validate $\Phi^N(\bar{x})$. Shapiro [43, §4.2] and Linderoth et al. [26, §§5.3–5.4] present some discussions on validation of subgradients of some random functions through sample average subgradients. Similar ideas may be used to build a subset of $\Phi^N(\bar{x})$ as in Linderoth et al. [26]. We omit the details as they are beyond the focus of this paper.

Acknowledgements. The authors thank Alexander Shapiro for providing the reference for Gauvin and Dubeau [14]. They also thank the two anonymous referees and the associate editor for insightful comments that significantly helped improve the presentation of this paper. Finally, they thank Fumio Hiai for valuable discussions on Aumann's integral of piecewise continuous random set-valued mappings.

Appendix

Proof of Proposition 4.1. We begin by showing a kind of continuity of the set-valued mapping $T_j(x)$ on $\operatorname{dom} T_j$, for an arbitrary index j. First, observe that, because $\mathbb{T}_j=\operatorname{cl}\mathbb{S}_j$, $\operatorname{cl}(\operatorname{dom} T_j)=\operatorname{cl}(\operatorname{dom} S_j)$; hence, $\operatorname{dom} T_j=\operatorname{cl}(\operatorname{dom} S_j)$ and $\operatorname{dom} T_j$ is closed (stochastic PC^0 property ii(b)). Define the indicator function:

$$\mathbb{I}(T_j(x),\xi):=\begin{cases}1 & \text{if } \xi\in T_j(x),\\ 0 & \text{otherwise.}\end{cases}$$

Now, take a convergent sequence $\{x^k\}\subset\operatorname{dom} S_j$ with limit x. Because $\operatorname{dom} T_j$ is closed, we have $T_j(x)\neq\varnothing$. For a.a. $\xi\in T_j(x)$, we have $\xi\in T_j(x^k)$ for large k from stochastic PC^0 property ii(c). If $\xi\notin T_j(x)$, then $(x,\xi)\notin\mathbb{T}_j$, which, by openness of the complement of $\mathbb{T}_j$, implies $\xi\notin T_j(x^k)$ for large k. This shows that $\mathbb{I}(T_j(x^k),\xi)$ converges to $\mathbb{I}(T_j(x),\xi)$ for a.a. ξ.

Proof of Proposition 4.1. Part (i). Let $x\in\bar{X}_J=\operatorname{cl}(\operatorname{int} X_J)$ and note that $(\operatorname{int} X_J)\cap\mathscr{D}_\Gamma$ is dense in $\bar{X}_J$ because a.a. points of $\mathscr{X}$ lie in $\mathscr{D}_\Gamma$. Hence, there exists $\{x^k\}\subset(\operatorname{int} X_J)\cap\mathscr{D}_\Gamma$ converging to x. Applying the above to each $j\in J$ gives a.e. convergence of $s_k(\xi)=\sum_{j\in J}\mathbb{I}(T_j(x^k),\xi)$ to $s(\xi)=\sum_{j\in J}\mathbb{I}(T_j(x),\xi)$. By hypothesis, each $s_k(\xi)=1$ a.e.; therefore, $s(\xi)$ has the same property. This implies both that $\bigcup_{j\in J}T_j(x)$ has full measure and that $\mu_P(T_j(x)\cap T_l(x))=0$ for distinct $j,l\in J$, i.e., $x\in X_J$.

Part (ii). The properties of J imply that $F^J(x)=\sum_{j\in J}F_j(x)$ for $x\in X_J$, where $F_j(x):=\int_{T_j(x)}\Gamma_j(x,\xi)\,d\mu_P(\xi)$. Because the closure of $(\operatorname{int} X_J)\cap\mathscr{D}_\Gamma$ is $\bar{X}_J$, it is sufficient to take $x\in\bar{X}_J$, $\{x^k\}\subset(\operatorname{int} X_J)\cap\mathscr{D}_\Gamma$ that converges to x, and take $j\in J$ and show that $F_j(x^k)\to F_j(x)$. We have $F_j(x^k)=\int\Gamma_J(x^k,\xi)\mathbb{I}(T_j(x^k),\xi)\,d\mu_P(\xi)$. Therefore, continuity of Γ_j on $\mathbb{T}_j$ together with a.e. convergence $\mathbb{I}(T_j(x^k),\xi)\to\mathbb{I}(T_j(x),\xi)$ gives, via Lebesgue's dominated convergence theorem,

$$F_j(x^k)\to\int\Gamma_J(x,\xi)\mathbb{I}(T_j(x),\xi)\,d\mu_P(\xi)=F_j(x).\quad\square$$

PROOF OF THEOREM 4.2. We apply Theorem 4.1 with $\mathscr{X} = KB$. For each $\bar{x}$ near $K\mathscr{B}$, there is a nonempty subset $\bar{\mathcal{J}} = \mathcal{J}(\bar{x})$ of $\mathcal{J}$ with the following properties: $\bar{x} \in \bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J$,

$$\Gamma(\bar{x}, \xi) = \bigcup_{J\in\bar{\mathcal{J}}} \Gamma^J(\bar{x}, \xi) \quad \text{a.e.} \tag{53}$$

(from the proof of Theorem 4.1), and $x \to \mathbb{E}[\bigcup_{J\in\bar{\mathcal{J}}} \Gamma^J(x, \xi)]$ continuous on $\bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J$. Theorem 3.2 can now be applied to $\bigcup_{J\in\bar{\mathcal{J}}} \Gamma^J(x, \xi)$ with $K\mathscr{B}$ replaced by $\bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J$. The SAA of $\mathbb{E}[\bigcup_{J\in\bar{\mathcal{J}}} \Gamma^J(x, \xi)]$ given N samples $\xi^1, \ldots, \xi^N$ is $\bigcup_{J\in\bar{\mathcal{J}}} \Gamma_N^J(x)$, where $\Gamma_N^J(x)$ is $(1/N)\sum_{i=1}^N \Gamma^J(x, \xi^i)$, the SAA of $\mathbb{E}[\Gamma^J(x, \xi)]$. By virtue of part (i) of Theorem 3.2, for small $\epsilon > 0$, there exist constants $c^{\bar{\mathcal{J}}}(\epsilon) > 0$ and $\beta^{\bar{\mathcal{J}}}(\epsilon) > 0$ independent of N such that, for N sufficiently large,

$$\text{Prob}\left\{ \sup_{x\in\bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J} \mathbb{D}\left(\bigcup_{J\in\bar{\mathcal{J}}} \Gamma_N^J(x),\ \mathbb{E}\left[\bigcup_{J\in\bar{\mathcal{J}}} \Gamma^J(x, \xi) \right] \right) \geq \epsilon \right\} \leq c^{\bar{\mathcal{J}}}(\epsilon) e^{-\beta^{\bar{\mathcal{J}}}(\epsilon)N}. \tag{54}$$

Let us call a subfamily $\bar{\mathcal{J}}$ of $\mathcal{J}$ *essential* if both it and $\bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J$ are nonempty. Then, because $\mathscr{X}$ is covered by nonempty sets of the type $\bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J$,

$$\sup_{x\in K\mathscr{B}} \mathbb{D}(\Gamma_N(x), F(x)) \leq \max_{\text{essential}\,\bar{\mathcal{J}}} \sup_{x\in\bigcap_{J\in\bar{\mathcal{J}}} \bar{X}^J} \mathbb{D}\left(\bigcup_{J\in\bar{\mathcal{J}}} \Gamma_N^J(x),\ \mathbb{E}\left[\bigcup_{J\in\bar{\mathcal{J}}} \Gamma^J(x, \xi) \right] \right)$$

w.p.1. This gives, via (54), that

$$\text{Prob}\left\{ \sup_{x\in K\mathscr{B}} \mathbb{D}(\Gamma_N(x), F(x)) \geq \epsilon \right\} \leq \sum_{\text{essential}\,\bar{\mathcal{J}}} c^{\bar{\mathcal{J}}}(\epsilon) e^{-\beta^{\bar{\mathcal{J}}}(\epsilon)N}.$$

Part (i) of Theorem 3.2, in our setting, follows. Part (ii) of Theorem 3.2 also holds under Assumption 3.2. □

Claims from proof of Proposition 4.2. Write $\Gamma(x, \xi) = \bigcup_{j=1}^{\hat{\jmath}} \Gamma_j(x, \xi)$, i.e., (17), where the graph of each Γ_j is convex in addition to being nonempty, closed, and polyhedral.

CLAIM 1. We may assume without loss of generality that $\operatorname{int}\operatorname{dom}\Gamma_j \neq \varnothing$ for each $j = 1, \ldots, \hat{\jmath}$. Clearly, as a polyhedral set, $\operatorname{dom}\Gamma$ satisfies $\operatorname{cl}\operatorname{int}\operatorname{dom}\Gamma = \bigcup_{j=1}^{\hat{\jmath}} \operatorname{cl}\operatorname{int}\operatorname{dom}\Gamma_j$, which means we can drop Γ_j if its domain does not have interior. For a.a. ξ, $\Gamma(x, \xi)$ is the union of $\Gamma_j(x, \xi)$ for j such that $\operatorname{int}\operatorname{dom}\Gamma_j \neq \varnothing$.

CLAIM 2. We assume without loss of generality that the convex polyhedral set-valued mappings Γ_j may be chosen such that their domains $\mathbb{T}_j = \operatorname{dom}\Gamma_j$ have nonempty but pairwise disjoint interiors $\mathbb{S}_j$, i.e., stochastic PC^0 property ii(a) holds. To show Claim 2, take any nonempty subset J of $\{1, \ldots, \hat{\jmath}\}$ and let J^c denote its complement and $\mathbb{S}_J := \bigcap_{j\in J} \mathbb{S}_j \setminus \bigcup_{j\in J^c} \mathbb{T}_j$. Let $\mathcal{J}_0$ be the family of such index sets J, where $\mathbb{S}_J \neq \varnothing$. The closed sets $\mathbb{T}_J = \operatorname{cl}\mathbb{S}_J$, $J \in \mathcal{J}_0$ have nonempty and disjoint interiors. We take for granted that such a set $\mathbb{T}_J$ is polyhedral; hence, even if nonconvex, it is the union of finitely many (closed) polyhedral convex sets called "subdomains" of $\mathbb{T}_J$ whose interiors are nonempty and disjoint. Denote by Γ_J the restriction of Γ to $\mathbb{T}_J$. We conclude the argument by observing that Γ_J can be decomposed into a family of polyhedral convex "submappings" by restricting it to each of those polyhedral convex subdomains of $\mathbb{T}_J$. □

PROOF OF THEOREM 5.1. Under condition (a), it follows from Lemma 5.2 that (29) holds at $x = x^*$. Hence, $0 \in \mathbb{E}[\Psi(x^*, \xi)] + \mathcal{N}_X(x^*)$ by the usual nonsmooth optimality conditions (27) and Proposition 5.1, and it suffices to show the well-definedness of $\mathbb{E}[\Psi(x^*, \xi)]$. Let

$$\hat{\Psi}(x, \xi) := \bigcup_{y\in P(x,\xi)} \bigcup_{(\lambda,\mu)\in W(x,y,\xi)} \nabla_x L(x, y, \xi; \mu, \lambda).$$

By the property of Aumann's integral (Hess [17, Theorem 5.4(d)]), it suffices to show that $\mathbb{E}[\hat{\Psi}(x^*, \xi)]$ is well-defined.

Let $\xi^k \in \Xi$ and $\xi^k \to \xi$. There exist bounded sequences defined by $y^k \in P(x^*, \xi^k)$, $(\lambda^k, \mu^k) \in W(x^*, y^k, \xi^k)$ and

$$\eta^k = \nabla_x L(x^*, y^k, \xi^k; \mu^k, \lambda^k) \in \hat{\Psi}(x^*, \xi^k),$$

where boundedness of multipliers follows from M-F regularity (see Gauvin and Tolle [15, Theorem 2.2]). Let $(\bar{y}, \bar{\lambda}, \bar{\mu}, \bar{\eta})$ be an accumulation point of $\{(y^k, \lambda^k, \mu^k, \eta^k)\}$. It is clear, by continuity of all functions, that $\bar{y}$ is a global minimizer of the lower-level problem at (x^*, ξ), i.e., $\bar{y}$ belongs to $P(x^*, \xi)$. On the other hand, it follows

from KKT conditions (20)–(22) that $(\bar{\lambda}, \bar{\mu})$ must satisfy the second-stage KKT conditions with $(x^*, \bar{y}, \bar{\xi})$. This shows the closedness of $\hat{\Psi}(x, \xi)$.

Next, we show the uniform compactness near ξ. Under Assumption 5.1, $\nabla_x L(x, y, \xi'; \lambda, \mu)$ is bounded by $\kappa_2(\xi') + c\kappa_2(\xi')^2$ for all $\xi' \in \Xi$, where c is some positive constant. Therefore, $\hat{\Psi}(x^*, \xi')$ is bounded. By Proposition 2.1, the mapping $\hat{\Psi}(x^*, \cdot)$: $\Xi \to 2^{\mathbb{R}^{s+t}}$ is outer semicontinuous on Ξ. Hence, by Aubin and Frankowska [2, Proposition 8.2.1], $\hat{\Psi}(x^*, \xi(\cdot))$: $\Omega \to 2^{\mathbb{R}^{s+t}}$ is measurable. Moreover, because $\mathbb{E}[\kappa_2(\xi) + c\kappa_2(\xi)^2] < \infty$ under Assumption 5.1, then $\mathbb{E}[\Psi(x^*, \xi)]$ is well-defined. The proof is complete. □

References

[1] Artstein, Z., R. A. Vitale. 1975. A strong law of large numbers for random compact sets. *Ann. Probab.* **3**(5) 879–882.
[2] Aubin, J.-P., H. Frankowska. 1990. *Set-Valued Analysis*. Birkhauser, Boston.
[3] Aumann, R. J. 1965. Integrals of set-valued functions. *J. Math. Anal. Appl.* **12** 1–12.
[4] Berge, C. 1963. *Topological Spaces*. Macmillan, New York.
[5] Birge, J., F. Louveaux. 1997. *Introduction to Stochastic Programming*. Springer Series in Operations Research, Springer-Verlag, New York.
[6] Bonnans, J. F., A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research, Springer-Verlag, New York.
[7] Castaing, C., M. Valadier. 1977. *Convex Analysis and Measurable Multifunctions*. Lecture Notes in Mathematics, Vol. 580. Springer, Berlin.
[8] Clarke, F. H. 1983. *Optimization and Nonsmooth Analysis*. Wiley, New York.
[9] Dai, L., C. H. Chen, J. R. Birge. 2000. Convergence properties of two stage stochastic programming. *J. Optim. Theory Appl.* **106**(3) 489–509.
[10] Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*. Springer-Verlag, New York.
[11] DeMiguel, V., H. Xu. 2009. A stochastic multiple leader Stackelberg model: Analysis, computation, and application. *Oper. Res.* **57**(5) 1220–1235.
[12] Dontchev, A. L., A. S. Lewis, R. T. Rockafellar. 2004. The radius of metric regularity. *Trans. Amer. Math. Soc.* **355**(2) 493–517.
[13] Facchinei, F., J.-S. Pang. 2003. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, New York.
[14] Gauvin, J., F. Dubeau. 1982. Differential properties of the marginal functions in mathematical programming. *Math. Programming Stud.* **19** 101–119.
[15] Gauvin, J., J. W. Tolle. 1977. Differential stability in nonlinear programming. *SIAM J. Control Optim.* **15**(2) 294–311.
[16] Gürkan, G., A. Y. Özge, S. M. Robinson. 1999. Sample-path solution of stochastic variational inequalities. *Math. Programming* **84**(2) 313–333.
[17] Hess, C. 2002. Set-valued integration and set-valued probability theory: An overview. *Handbook of Measure Theory*, Vols. I and II. North-Holland, Amsterdam, 617–673.
[18] Higle, J. L., S. Sen. 1996. *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming*. Kluwer Academic Publishers, Dordrecht, The Netherlands.
[19] Hogan, W. W. 1973. Point-to-set maps in mathematical programming. *SIAM Rev.* **15** 591–603.
[20] Homem-de-Mello, T. 2008. On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling. *SIAM J. Optim.* **19**(2) 524–551.
[21] Hu, X., D. Ralph. 2002. A note on sensitivity of value functions of mathematical programs with complementarity constraints. *Math. Programming* **93**(2) 265–279.
[22] Ioffe, A. D. 2010. On regularity concepts in variational analysis. *J. Fixed Point Theory Appl.* **8**(2) 339–363.
[23] King, A. J., R. T. Rockafellar. 1993. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* **18**(1) 148–162.
[24] King, A. J., R. J.-B. Wets. 1991. Epi-consistency of convex stochastic programs. *Stochastics Stochastics Rep.* **34**(1-2) 83–92.
[25] Kaniovski, Y. M., A. J. King, R. J.-B. Wets. 1995. Probabilistic bounds (via large deviation) for solutions of stochastic programming problems. *Ann. Oper. Res.* **56**(1) 189–208.
[26] Linderoth, J., A. Shapiro, S. Wright. 2006. The empirical behavior of sampling methods for stochastic programming. *Ann. Oper. Res.* **142**(1) 215–241.
[27] Meng, F., H. Xu. 2006. A regularized sample average approximation method for stochastic mathematical programs with nonsmooth equality constraints. *SIAM J. Optim.* **17**(3) 891–919.
[28] Nocedal, J., J. J. Wright. 1999. *Numerical Optimization*. Springer, New York.
[29] Outrata, J., W. Römisch. 2005. On optimization conditions for some nonsmooth optimization problems over L^p spaces. *J. Optim. Theory Appl.* **126**(2) 411–438.
[30] Papageorgiou, N. 1985. On the theory of Banach space valued multifunctions 1. Integration and conditional expectation. *J. Multivariate Anal.* **17** 185–206.
[31] Plambeck, E. L., B. R. Fu, S. M. Robinson, R. Suri. 1996. Sample-path optimization of convex stochastic performances functions. *Math. Programming* **75**(2) 137–176.
[32] Ralph, D., S. Dempe. 1995. Directional derivatives of the solution of a parametric nonlinear program. *Math. Programming* **70**(2) 159–172.
[33] Robinson, S. M. 1979. Generalized equations and their solutions, Part I: Basic theory, point-to-set maps and mathematical programming. *Math. Programming Stud.* **10** 128–141.
[34] Robinson, S. M. 1980. Strongly regular generalized equations. *Math. Oper. Res.* **5**(1) 43–62.
[35] Robinson, S. M. 1981. Some continuity properties of polyhedral multifunctions. *Math. Programming Stud.* **14** 206–214.

[36] Robinson, S. M. 1983. Generalized equations. A. Bachem, M. Grötschel, B. Korte, eds. *Mathematical Programming: The State of the Art*. Springer-Verlag, Berlin, 346–367.
[37] Robinson, S. M. 1996. Analysis of sample-path optimization. *Math. Oper. Res.* **21**(3) 513–528.
[38] Rockafellar, R. T., R. J.-B. Wets. 1998. *Variational Analysis*. Springer-Verlag, Berlin.
[39] Römisch, W., R. Schultz. 1996. Lipschitz stability for stochastic programs with complete recourse. *SIAM J. Optim.* **6**(2) 531–547.
[40] Ruszczyński, A., A. Shapiro. 2003. Stochastic programming models. A. Rusczyński, A. Shapiro eds. *Stochastic Programming*: *Handbooks in OR & MS*, Vol. 10. North-Holland Publishing Company, Amsterdam.
[41] Shapiro, A. 1989. Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* **17**(2) 841–858.
[42] Shapiro, A. 2005. Complexity of two and multi-stage stochastic programming problems. http://www2.isye.gatech.edu/~ashapiro/publications.html.
[43] Shapiro, A. 2003. Monte Carlo sampling methods. A. Rusczyński, A. Shapiro, eds. *Stochastic Programming*: *Handbooks in OR & MS*, Vol. 10. North-Holland Publishing Company, Amsterdam.
[44] Shapiro, A., T. Homem-de-Mello. 1998. A simulation-based approach to two stage stochastic programming with recourse. *Math. Programming* **81**(3) 301–325.
[45] Shapiro, A., T. Homem-de-Mello. 2000. On rate of convergence of Monte Carlo approximations of stochastic programs. *SIAM J. Optim.* **11**(1) 70–86.
[46] Shapiro, A., H. Xu. 2008. Stochastic mathematical programs with equilibrium constraints, modeling and sample average approximation. *Optimization* **57**(3) 395–418.
[47] Shapiro, A., H. Xu. 2007. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *J. Math. Anal. Appl.* **325**(2) 1390–1399.
[48] Wets, R. J.-B. 1989. G. L. Nemhauser et al., eds. *Stochastic Programming: Handbooks in OR & MS*, Vol. 1. 573–629.
[49] Xu, H., F. Meng. 2007. Convergence analysis of sample average approximation methods for a class of stochastic mathematical programs with equality constraints. *Math. Oper. Res.* **32**(3) 648–668.
[50] Xu, H., D. Zhang. 2009. Smooth sample average approximation of stationary points in nonsmooth stochastic problems and applications. *Math. Programming Ser. A* **119**(2) 371–401.
[51] Xu, H., D. Zhang. 2009. Stochastic Nash equilibrium problems: Sample average approximation and applications. *Optimization* online.