A Decomposition Algorithm for Two-Stage Stochastic Programs with Nonconvex Recourse

Hanyang Li Ying Cui∗
November 15, 2022

Abstract

In this paper, we have studied a decomposition method for solving a class of nonconvex two-stage stochastic programs, where both the objective and constraints of the second-stage problem are nonlinearly parameterized by the first-stage variables. Due to the failure of the Clarke regularity of the resulting nonconvex recourse function, classical decomposition approaches such as Benders decomposition and (augmented) Lagrangian-based algorithms cannot be directly generalized to solve such models. By exploring an implicitly convex-concave structure of the recourse function, we introduce a novel decomposition framework based on the so-called partial Moreau envelope. The algorithm successively generates strongly convex quadratic approximations of the recourse function based on the solutions of the second-stage convex subproblems and adds them to the first-stage master problem. Convergence under both fixed scenarios and interior samplings is established. Numerical experiments are conducted to demonstrate the effectiveness of the proposed algorithm.

Keywords: two-stage stochastic program, nonconvex recourse, decomposition

1 Introduction

Stochastic programming (SP) is a mathematical framework to model the decision making in the presence of uncertainty [6, 55]. Two-stage SPs constitute a special class of this paradigm where partial decisions have to be made before the observation of the entire information, while the rest decisions are determined after the full information is revealed. Most of the existing computational study of the continuous two-stage SPs is devoted to convex problems, especially linear problems [52, 6, 31, 55]. Although there is a significant amount of literature that tackles the nonconvexity of the SPs caused by the integrality of the decision variables [7, 8, 53, 13], the nonconvexity is usually abandoned and ignored in the algorithmic development of continuous SPs.

However, there are many emerging applications in operations research and machine learning that call for complex nonlinear two-stage SP models and computational methods. Let us first introduce the mathematical formulation of such problems before discussing the applications. The central optimization problem under consideration in this paper takes the following form:

\[
\begin{align*}
\min_{x \in X} \quad & \zeta(x) \triangleq \varphi(x) + \mathbb{E}_{\tilde{\xi}} \left[ \psi(x; \tilde{\xi}) \right],
\end{align*}
\]

∗Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455 (li002492@umn.edu yingcui@umn.edu). The authors are partially supported by NSF CCF-2153352.
where $\psi(x; \xi)$ is the second-stage recourse function that is given by:

$$
\psi(x; \xi) \triangleq \begin{cases} 
\min_y \{ f(x, y; \xi) \text{ subject to } G(x, y; \xi) \leq 0 \} & \text{if } x \in \mathcal{X} \\
+\infty & \text{if } x \notin \overline{\mathcal{X}} 
\end{cases} 
$$

(2)

In the above formulation, $\mathcal{X}$ and $\overline{\mathcal{X}}$ are nonempty convex compact subsets in $\mathbb{R}^{n_1}$ with $\overline{\mathcal{X}} \subseteq \text{int}(\mathcal{X})$, $\varphi : \mathbb{R}^{n_1} \to \mathbb{R}$ is a deterministic convex function that only depends on the first-stage decision $x$, $\xi : \Omega \to \Xi$ is a random vector on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Xi \subseteq \mathbb{R}^{m_1}$ being a measurable closed set; $\xi = \tilde{\xi}(\omega)$ for some $\omega \in \Omega$ represents a realization of the random vector $\tilde{\xi}$; and $f : \mathbb{R}^{n_1+n_2} \times \Xi \to \mathbb{R}$ and $G \triangleq (g_1, \ldots, g_\ell) : \mathbb{R}^{n_1+n_2} \times \Xi \to \mathbb{R}^\ell$ are two Carathéodory functions (i.e., $f(\bullet, \bullet; \xi)$ and $G(\bullet, \bullet; \xi)$ are continuous for almost any $\xi \in \Xi$; $f(x, y; \bullet)$ and $G(x, y; \bullet)$ are measurable for any $(x, y) \in \mathbb{R}^{n_1+n_2}$ that are jointly determined by the first-stage variable $x$ and the second-stage variable $y$). We assume that, for almost any $\xi \in \Xi$, the function $f(\bullet, \bullet; \xi)$ is concave-convex (i.e., $f(\bullet, y; \xi)$ is concave for $y \in \mathbb{R}^{n_2}$ and $f(x, \bullet; \xi)$ is convex for $x \in \mathbb{R}^{n_1}$), and $g_i(\bullet, \bullet; \xi)$ is jointly convex for each $i = 1, \ldots, \ell$.

An example of concave-convex $f(\bullet, \bullet; \xi)$ is a bilinear function $x^\top D(\xi)y$ for some random matrix $D(\xi) \in \mathbb{R}^{n_1 \times n_2}$. The above settings notably extend the classical paradigm for continuous two-stage SPs [55, Chapter 2.3] in the following directions:

(i) The first-stage variable $x$ not only appears in the constraints of the second-stage problem, but also in the objective $f$. The recourse function $\psi(\bullet; \xi)$ is nonconvex since $f(\bullet, \bullet; \xi)$ is not jointly convex. This is fundamentally different from the recent papers [28, 27] that have assumed the joint convexity of $f(\bullet, \bullet; \xi)$.

(ii) Both the objective function $f$ and the constraint map $G$ can be nonsmooth.

These two features together lead to a complex nonconvex and nonsmooth recourse function $\psi(\bullet; \xi)$ (see Figure 1 below), which constitutes the major challenge for designing rigorous and efficient numerical methods to solve problem (1).

![Figure 1: The nonconvex nonsmooth recourse functions for fixed $\xi$'s. Left: $x \in \mathbb{R}$; Right: $x \in \mathbb{R}^2$.](image)

Recourse functions in the form of (2) arise from many applications. One important source of the nonconvex recourse in (2) comes from the decision dependent/influenced uncertainty [30, 40, 34, 23], where the probability distribution $\mathbb{P}(x)$ of the random vector $\xi$ is dependent on the first-stage variable $x$. This is in contrast to the classical SP paradigm under exogenous uncertainty, where the distribution of $\xi$ is not affected by the first-stage decisions. There is growing interest in the
endogenous uncertainty in recent literature of stochastic and robust programs [24, 25, 57, 30]. A typical example where the random parameters can be altered by a decision is that the price (as a first-stage variable) may affect the distribution of the product demand. Assume that the probability distribution of \(\xi\) is given by \(P_x\) that depends on the first-stage variable \(x\). We consider the corresponding two-stage SP model:

\[
\begin{align*}
\text{minimize} \quad & \varphi(x) + \mathbb{E}_{\xi \sim P_x}[\psi(x; \xi)] = \varphi(x) + \int_{\Xi} \psi(x; \xi) dP_x \\
\text{subject to} \quad & \text{constraints}
\end{align*}
\]

If there exists a decision-independent distribution \(\hat{P}\) such that \(P_x\) is absolutely continuous with respect to \(\hat{P}\) for any \(x \in X\), we can reformulate the above problem by applying [22, Proposition 3.9] as:

\[
\begin{align*}
\text{minimize} \quad & \varphi(x) + \mathbb{E}_{\xi \sim \hat{P}}[\psi(x; \xi)], \quad \text{with} \quad \hat{\psi}(x; \xi) \triangleq \psi(x; \xi) \frac{dP_x}{d\hat{P}}, \\
\text{subject to} \quad & \text{constraints}
\end{align*}
\]

where \(dP_x/d\hat{P}\) is the Radon-Nikodym derivative of \(P_x\) with respect to \(\hat{P}\). Even if originally the first-stage decision \(x\) only appears in the constrains of the second-stage problem in \(\psi(x; \xi)\), the above transformation would make \(x\) also appear in the second-stage objective function through the multiplication of \(dP_x/d\hat{P}\).

A specific example of the decision-dependent uncertainty in SPs is a power system planning problem originated in [36] and expanded in [30]. Assume that \(\xi = (\{d_j\}_{j \in J}, \{\pi_j\}_{j \in J}, \{q_i\}_{i \in I})\) follows a discrete distribution with the support \(\{\xi^s\}_{s=1}^S\), where \(d_j\) and \(\pi_j\) represent the demand and the price of electricity in the location \(j \in J\), and \(q_i\) is the unit production cost of the power plant \(i \in I\). For each \(s = 1, \ldots, S\), the probability of \(\xi = \xi^s\) is the linear combination of \(|G|\) given different distributions (each with probability \(p_{sg}\)) whose weights \(\{x_g\}_{g \in G}\) are determined as parts of the first-stage decisions, i.e., \(P_x(\xi = \xi^s) = \sum_{g \in G} p_{sg} x_g\) for each \(s\). The capacity of each power plant \(\{x_i\}_{i \in I}\) also needs to be determined in the first stage. The second-stage decisions are the production \(y_{ij}\) from the power plant \(i\) to the location \(j\) for each \(s\). By letting \(\hat{P}(\xi = \xi^s) = 1/S\) for each \(s\) in [3], we can rewrite the recourse function as

\[
\hat{\psi}(\{x_i\}_{i \in I}, \{x_g\}_{g \in G}; \xi^s) = \min_{i \leq y \leq u_x} S \sum_{g \in G} p_{sg} x_g \sum_{i \in I, j \in J} (q_{is} - \pi_{js}) y_{ij} \\
\text{subject to} \quad & \sum_{j \in J} y_{ij} \leq x_i, \quad i \in I; \quad \sum_{i \in I} y_{ij} = d_{js}, \quad j \in J
\]

and obtain a decision-independent SP with the recourse \(\hat{\psi}\). Observe that both the objective and constraints depend on the first-stage variables. In particular, the objective function is convex in \(x_g\) and concave in \(y\), which fits our problem setting. Later, we will apply our proposed algorithms to solve a two-stage SP with the above recourse in Section [6].

The second example of the nonconvex recourse in [2] is the stochastic interdiction problem [18, 29], where the defender may want to maximize the second-stage objective function instead of minimizing it. Even for the simple linear second-stage problem with only \(x\) appearing in the constraints, the recourse function

\[
\hat{\psi}(x; \xi) \triangleq \max_y \quad c(\xi)^T y \\
\text{subject to} \quad T(\xi)x + W(\xi)y = h(\xi)
\]
is not convex in \( x \). One may take the dual of the second-stage maximization problem so that the recourse is a parametrized minimization problem

\[
\tilde{\psi}(x;\xi) = \min_{\lambda} \lambda^\top T(\xi)x - \lambda^\top h(\xi)
\]

subject to \( W(\xi)^\top \lambda + c(\xi) = 0 \).

However, this dualization would bring a bilinear term \( \lambda^\top T(\xi)x \) of the first-stage variable \( x \) and the second-stage variable \( \lambda \) to the objective function that necessitates the concave-convex structure of \( f(\bullet,\bullet;\xi) \). To the best of our knowledge, there is no known rigorous decomposition method to solve a general nonconvex two-stage min-max stochastic programs even under the linear setting.

When the distribution of \( \tilde{\xi} \) is taken as the empirical distribution of observed realizations \( \xi^1, \ldots, \xi^S \), the simplest way to tackle the problem (1) is to simultaneously solve the first-stage variable \( x \) and second-stage variables \( y^1, \ldots, y^S \) (each \( y^s \) is associated with one scenario \( \xi^s \)) via the sample average approximation [55]:

\[
\begin{align*}
\minimize_{x \in X, y^1, \ldots, y^S} & \varphi(x) + \frac{1}{S} \sum_{s=1}^{S} f(x, y^s; \xi^s) \\
\text{subject to} & \ G(x, y^s; \xi^s) \leq 0, \quad s = 1, \ldots, S.
\end{align*}
\]

However, this approach can be prohibitive when the number of scenarios \( S \) is large since the dimension of the unknown variables is \( n_1 + n_2 S \). Even if \( S \) is small or moderate, the above formulation may still be difficult to handle under our setting as the function \( f(\bullet,\bullet;\xi) \) is not jointly convex (for example when \( f(\bullet,\bullet;\xi) \) is bilinear). In fact, the nonconvexity of the recourse function also makes it challenging to apply the stochastic approximation method [44, 43, 39] to solve (1), since it is not clear how to obtain a (Clarke) subdifferential of the nonconvex recourse function in (4). Without strong assumptions like the uniqueness of the second-stage solutions, only a superset of the subdifferential \( \partial \tilde{\psi}(\bullet;\xi) \) at given \( x \) is computable [10, Chapter 4]. When \( f \) and \( G \) are twice continuously differentiable, the authors in [11] have adopted a smoothing method to deal with the possibly nonconvex recourse by adding the Tikhonov-regularized barrier of the inequality constraints to the second-stage objective function. For a special class of two-stage nonconvex quadratic SPs under the simplex constraint, the paper [9] has derived upper and lower approximations of the objective values via copositive programs.

Notice that the constraints in (5) are in fact block-wise separable in \( y^1, \ldots, y^S \) so that there is a block-angular structure between the first- and second-stage variables. Decomposition algorithms of two-stage SPs take advantage of this special structure to efficiently handle a large number of scenarios via solving \( S \) numbers of low-dimensional subproblems [52]. Two classical decomposition algorithms for two-stage SPs are (augmented) Lagrangian decomposition and Benders decomposition. (Augmented) Lagrangian decompositions (including the progressive hedging algorithm) copy the first-stage variable \( S \) times and attach one to each scenario [26, 19]. In order to force the non-anticipativity of the first-stage decision, one has to add equality constraints among all copies to ensure that \( x \) is the same across different realizations of the uncertainty. However, there are two major bottlenecks to apply such kind of dual-based algorithms to solve the problem (1). One, each subproblem pertaining to one pair of variables \( (x^s, y^s) \) is still nonconvex if \( f \) is not a jointly convex function, so that it is in general not easy to obtain its global optimal solution. Two, the convergence of these dual approaches is largely restricted to the convex problems or special integer problems.
Although there are some recent advances for the convergence study of the progressive hedging algorithm for solving nonconvex SPs under the local convexity conditions, it is not clear whether the problem satisfies those conditions without further assumptions on f and g. Benders decomposition (or L-shaped methods) alternatively updates the first-stage and second-stage variables, where the second-stage subproblem can be solved in parallel to save the computational time and reduce storage burden. In order to derive valid inequalities of x and add them to the first-stage master problem, one usually uses subgradient inequalities of the (convex) recourse function to generate a sequence of lower approximations. However, when the recourse function is associated with the complex nonconvex function in (1), it is challenging to derive its lower bounds based on the computed second-stage solutions. In fact, for the recourse functions in Figure 1, there seems not to exist a convex function that passes one of the downward cusps and at the same time approximates the original function from below.

In this paper, we tackle the two-stage stochastic programs by a novel lifting technique that transforms the complex nonconvex and nondifferentiable recourse function in the original space to a structured convex-concave function in a lifted space. The reveal of this latent structure enables us to construct convex surrogation of all recourse functions at the latest first-stage iterate, whose evaluations are decomposable across different scenarios. Such surrogate functions are then added to the master problem to generate the next first-stage iterate. We shall prove that repeating the above procedure, the sequence of the first-stage iterates converges to a properly defined stationary solution of (1). In order to further reduce the computational cost per step when the number of scenarios S is large as well as to handle the case where ξ is continuously distributed, we also propose a framework that incorporates sequential sampling into the surrogation algorithm. The sequential sampling method gradually adds scenarios and generates cuts along the iterations, which has the advantage that one may obtain satisfactory descent progress in the early iterations with relatively small sample sizes to accelerate the overall procedure.

The paper is organized as follows. Section 2 introduces notation and provides preliminary knowledge. In Section 3, we discuss the implicitly convex-concave structure of the recourse function and derive its computationally tractable approximations. A decomposition algorithm for solving problem with a fixed number of scenarios is proposed and analyzed in Section 4. To further handle the continuously distributed random vectors as well as to reduce the computational cost of the decomposition algorithm in its early stage, we provide an internal sampling version of the algorithm in Section 5 and show the almost surely convergence of the iterative sequence. In Section 6, we conduct extensive numerical experiments to show the effectiveness of our proposed frameworks. The paper ends with a concluding section.

2 Preliminaries

We first summarize the notation used throughout the paper. We write Z_+ as the set of all nonnegative integers, and R^n as the n-dimensional Euclidean space equipped with the inner product \langle x, y \rangle = x^\top y and the induced norm \|x\| \triangleq \sqrt{x^\top x}. The symbol B(x, δ) is used to denote the closed ball of radius δ > 0 centered at a vector x ∈ R^n. Let A and C be two nonempty subsets of R^n. The diameter of A is defined as R(A) \triangleq \sup_{x, y \in A} \|x - y\|, and the distance from a vector x ∈ R^n to A is defined as dist(x, A) \triangleq \inf_{y \in A} \|y - x\|. The one-sided deviation of A from C is defined as
\[ \mathbb{D}(A,C) \triangleq \sup_{x \in A} \text{dist}(x,C). \]

We next introduce the concepts of generalized derivatives and subdifferentials for nonsmooth functions. Interested readers are referred to the monographs \cite{Gfrerer2021,Clarke1990,Borwein1996} for thorough discussions on these subjects. Consider a function \( f : O \to \mathbb{R} \) defined on an open set \( O \subseteq \mathbb{R}^n \). The classical one-sided directional derivative of \( f \) at \( \bar{x} \in O \) along the direction \( d \in \mathbb{R}^n \) is defined as
\[
f'(\bar{x};d) \triangleq \lim_{t \downarrow 0} \frac{f(\bar{x}+td)-f(\bar{x})}{t} \quad \text{if this limit exists.}
\]
The function \( f \) is said to be directionally differentiable at \( \bar{x} \in O \) if it is directionally differentiable along any direction \( d \in \mathbb{R}^n \). In contrast, the Clarke directional derivative of \( f \) at \( \bar{x} \in O \) along the direction \( d \in \mathbb{R}^n \) is defined as
\[
f^0(\bar{x};d) \triangleq \limsup_{x \to \bar{x}, t \downarrow 0} \frac{f(x+td)-f(x)}{t},
\]
which is finite when \( f \) is Lipschitz continuous near \( \bar{x} \). The Clarke subdifferential of \( f \) at \( \bar{x} \) is the set \( \partial_C f(\bar{x}) \equiv \{ v \in \mathbb{R}^n \mid f^0(\bar{x};d) \geq v^\top d \text{ for all } d \in \mathbb{R}^n \} \), which coincides with the usual subdifferential in convex analysis for a convex function. If \( f \) is strictly differentiable at \( \bar{x} \), then \( \partial_C f(\bar{x}) = \{ \nabla f(\bar{x}) \} \). We say that \( f \) is Clarke regular at \( \bar{x} \in O \) if \( f \) is directionally differentiable at \( \bar{x} \) and \( f^0(\bar{x};d) = f'(\bar{x};d) \) for all \( d \in \mathbb{R}^n \). This Clarke regularity at \( \bar{x} \) is equivalent to have \( f(x) \geq f(\bar{x}) + v^\top (x-\bar{x}) + o(\|x-\bar{x}\|) \) for any \( v \in \partial_C f(\bar{x}) \). Therefore, if a function fails to satisfy the Clarke regularity at \( \bar{x} \) (for example at the downward cusp on the left panel of Figure \ref{fig:fig1}), there does not exist an approximate linear lower bound of the original function based on the Clarke subdifferentials with small \( o \) error locally.

Let \( X \subseteq \mathbb{R}^n \) be a nonempty closed convex set and \( f : \mathbb{R}^n \to \mathbb{R} \) be a locally Lipschitz continuous function that is directionally differentiable. We say \( \bar{x} \in X \) is a directional-stationary point of \( f \) on \( X \) if \( f'(\bar{x};x-\bar{x}) \geq 0 \) for all \( x \in X \), and a Clarke-stationary point if \( f^0(\bar{x};d) \geq 0 \) for all \( d \in \mathbb{R}^n \). The latter is equivalent to \( 0 \in \partial_C f(\bar{x}) + N_X(\bar{x}) \) with \( N_X(\bar{x}) \) being the normal cone of \( X \).

Let \( \mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m \) be a set-valued mapping. Its outer limit at \( x \in \mathbb{R}^n \) is defined as
\[
\limsup_{x \to \bar{x}} \mathcal{F}(x) \triangleq \bigcup_{x' \to \bar{x}} \limsup_{\nu \to \infty} \mathcal{F}(x') = \{ u \mid \exists x' \to \bar{x}, \exists u' \to u \text{ with } u' \in \mathcal{F}(x') \}.
\]
We say \( \mathcal{F} \) is outer semicontinuous (osc) at \( \bar{x} \in \mathbb{R}^n \) if \( \limsup_{x \to \bar{x}} \mathcal{F}(x) \subseteq \mathcal{F}(\bar{x}) \).

### 3 The implicit convexity-concavity of the recourse functions

A key ingredient to design a decomposition method for solving the two-stage SP \cite{Eckstein1992} is to derive a computational-friendly approximation of the nonconvex recourse function \cite{Kochettov2023} at any given \( x \in X \) and \( \xi \in \Xi \). This is the main content of the present section.

For simplicity, we omit \( \xi \) in \cite{Kochettov2023} throughout this section and rewrite the recourse function as, for \( x \in \mathbb{R}^{n_1} \),
\[
\psi(x) \triangleq \begin{cases}
\text{minimum } \{ f(x,y) \mid y \in \mathbb{R}, G(x,y) \leq 0 \} & \text{if } x \in \overline{X}, \\
+\infty & \text{if } x \notin \overline{X},
\end{cases}
\]
where \( f(\bullet,\bullet) \) is concave-convex and \( G(\bullet,\bullet) \) is jointly convex. We assume that for any \( x \in \overline{X} \), the minimization problem of \( y \) in \cite{Eckstein1992} has an optimal solution, which implies the finiteness of \( \psi(x) \) on \( \overline{X} \). In the following, we show that the above function, although generally being nonconvex and nondifferentiable in \( \mathbb{R}^{n_1} \), has a benign structure in a lifted space. Leveraging this structure,
we then derive an approximate difference-of-convex decomposition of the recourse function that is computationally tractable. Such an approximation is the cornerstone of the decomposition methods to be presented in the next two sections.

3.1 The implicit convexity-concavity of $\psi$

As mentioned in the first section, the difficulty to design a decomposition method for solving (1) is due to the lack of a valid inequality of the recourse function, which is partially because $x$ appears in both the objective and constraints of the parametric problem in (6). However, if either $x$ in the objective or in the constraints is fixed, the resulting functions are relatively easy to analyze. Specifically, for any fixed $\bar{x} \in \mathbb{R}^n$, consider the functions

$$\psi_{\text{cvx}}(x) \triangleq \min_{y} f(\bar{x}, y) \text{ subject to } G(x, y) \leq 0$$

and

$$\psi_{\text{cve}}(x) \triangleq \min_{y} f(x, y) \text{ subject to } G(\bar{x}, y) \leq 0.$$

The following structural properties of $\psi_{\text{cvx}}$ and $\psi_{\text{cve}}$ can be easily derived. We include the proof here for completeness.

**Lemma 1.** Let $\bar{x} \in \bar{X} \subseteq \mathbb{R}^n$ be fixed. Suppose that the minimization problems in defining $\psi_{\text{cvx}}$ and $\psi_{\text{cve}}$ both have nonempty solution sets for any $x \in \mathbb{X}$. Then the function $\psi_{\text{cvx}}$ is convex and $\psi_{\text{cve}}$ is concave on $\mathbb{X}$.

**Proof.** It is not difficult to see that $\psi_{\text{cvx}}$ and $\psi_{\text{cve}}$ is finite for any $x \in \mathbb{X}$ since corresponding minimization problems have nonempty solution sets. Consider any $\lambda \in [0, 1]$ and any $x^1, x^2 \in \mathbb{X}$. We first show that

$$\psi_{\text{cvx}}(\lambda x^1 + (1 - \lambda) x^2) \leq \lambda \psi_{\text{cvx}}(x^1) + (1 - \lambda) \psi_{\text{cvx}}(x^2). \quad (7)$$

Let $\bar{y}^1$ and $\bar{y}^2$ be one of the optimal solutions of the optimization problems in defining $\psi_{\text{cvx}}(x^1)$ and $\psi_{\text{cvx}}(x^2)$ respectively. The joint convexity of each $g_j$ for $j = 1, \ldots, \ell$ implies that

$$G(\lambda x^1 + (1 - \lambda)x^2, \lambda \bar{y}^1 + (1 - \lambda) \bar{y}^2) \leq 0, \quad \forall j = 1, \ldots, \ell,$$

which yields the feasibility of $\lambda \bar{y}^1 + (1 - \lambda) \bar{y}^2$ for the minimization problem in (8) when $x$ takes the value $\lambda x^1 + (1 - \lambda)x^2$. We can thus deduce that $\psi_{\text{cvx}}(\lambda x^1 + (1 - \lambda)x^2) \leq f(\bar{x}, \lambda \bar{y}^1 + (1 - \lambda) \bar{y}^2)$. The convexity of $f(\bar{x}, \bullet)$ further yields that

$$f(\bar{x}, \lambda \bar{y}^1 + (1 - \lambda) \bar{y}^2) \leq \lambda f(\bar{x}, \bar{y}^1) + (1 - \lambda)f(\bar{x}, \bar{y}^2) = \lambda \psi_{\text{cvx}}(x^1) + (1 - \lambda)\psi_{\text{cvx}}(x^2),$$

which proves the inequality (7). The concavity of $\psi_{\text{cve}}$ is due to the fact that it is the minimum of concave functions $\{f(\bullet, y)\}$ for $y$ feasible to the problem in (8). $\square$

Lemma 1 suggests that the recourse function (6) has a hidden convex-concave structure. Indeed, such a function $\psi$ belongs to a special class of nonconvex functions coined *implicitly convex-concave (icc)* functions that are formally defined below. For an extended-real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\}$, the effective domain of $f$ is defined as $\text{dom } f \triangleq \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$. 7
Figure 2: An illustration of the implicitly convex-concave structure of the nonconvex recourse function. **Left**: the original 1-dimensional recourse function that is neither convex nor concave; **Middle**: the lifted bivariate counterpart in $\mathbb{R}^2$ that is convex in $x$ and concave in $z$; **Right**: the exposure of the concave component in the $z$ coordinate and the convex component in the $x$ coordinate.

**Definition 1** ([19, Definition 4.4.4]). A function $\theta : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ with $\text{dom} \theta$ being a convex set is said to be implicitly convex-concave if there exists a function $\bar{\theta} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ satisfying:

(a) $\bar{\theta}(x, z) = +\infty$ if $x \notin \text{dom} \theta$, $z \in \mathbb{R}^n$, and $\bar{\theta}(x, z) = -\infty$ if $x \in \text{dom} \theta$, $z \notin \text{dom} \theta$;

(b) $\bar{\theta}(\bullet, z)$ is convex for any fixed $z \in \text{dom} \theta$;

(c) $\bar{\theta}(x, \bullet)$ is concave for any fixed $x \in \text{dom} \theta$;

(d) $\theta(x) = \bar{\theta}(x, x)$ for any $x \in \text{dom} \theta$.

The above concept is firstly introduced in [35] to analyze the convergence property of a difference-of-convex algorithm to solve two-stage convex bi-parametric quadratic SPs. More properties of icc functions are studied in the recent monograph [19]. In fact, the term “icc” suggests that this class of functions is a generalization of the difference-of-convex (dc) functions, as the latter is “explicitly convex-concave”, i.e., for any dc function $\theta(x) = \theta_1(x) - \theta_2(x)$ with both $\theta_1$ and $\theta_2$ convex, one can always associate it with the bivariate function $\bar{\theta}(x, y) = \theta_1(x) - \theta_2(y)$ to explicitly expose the convexity-concavity of $\theta$ in the lifted pair $(x, y)$. Back to the recourse function $\psi$ (6), we consider its lifted bivariate counterpart

$$
\overline{\psi}(x, z) \triangleq \begin{cases} 
\min \{ f(z, y) \mid G(x, y) \leq 0 \} & \text{if } x, z \in \mathbb{X} \\
+\infty & \text{if } x \notin \mathbb{X} \\
-\infty & \text{if } x \in \mathbb{X} \text{ and } z \notin \mathbb{X}
\end{cases}.
$$

(8)

If the minimization problem of $y$ in (8) has a nonempty solution set for any $(x, z) \in \mathbb{X} \times \mathbb{X}$, it is not difficult to see that the assumption in Lemma 1 holds. Henceforth, the following result is a direct consequence of Lemma 1. No proof is needed.

**Proposition 1.** Assume that for any $(x, z) \in \overline{\mathbb{X}} \times \overline{\mathbb{X}}$, the minimization problem of $y$ in (8) has a nonempty solution set. Then $\psi$ in (6) is an implicitly convex-concave function associated with the lifted function $\overline{\psi}$ in (8).

One shall see from the subsequent sections that the derived implicitly convex-concave property of the recourse function is critical to study the two-stage SP (1). On one hand, we can leverage
this structure to construct an approximation of the nonconvex recourse function $\psi$ at any given $x$; on the other hand, it enables us to define a stationary point of \((\ref{prob:5})\) that is provably computable by our later designed algorithms. To fulfill these tasks, we first derive a superset of the Clarke subdifferential of $\psi$. To proceed, we denote $\partial_1 \overline{\psi}(x, z)$ as the subdifferential of the convex function $\overline{\psi}(\cdot, z)$ at $x$ for any $z \in \overline{X}$, and $\partial_2(-\overline{\psi})(x, z)$ as the subdifferential of the convex function $(-\overline{\psi})(x, \cdot)$ at $z$ for any $x \in \overline{X}$. We also write $\overline{\nabla}(x, z)$ as the set of all optimal solutions of problem \((\ref{prob:6})\).

Lemma 2. Assume that for any $(x, z) \in \overline{X} \times \overline{X}$, the minimization problem of $y$ in \((\ref{prob:8})\) has a nonempty solution set. The following two statements hold for all $(x, z) \in \overline{X} \times \overline{X}$.

(a) $\{\partial_1(-f)(z, y) \mid y \in \overline{\nabla}(x, z)\} \subseteq \partial_2(-\overline{\psi})(x, z)$;
(b) $\partial C \psi(x) \subseteq \partial_1 \overline{\psi}(x, x) - \partial_2(-\overline{\psi})(x, x)$.

Proof. (a) For any $(x, z) \in \overline{X} \times \overline{X}$, we take any $y \in \overline{\nabla}(x, z)$ and any $c \in \partial_1(-f)(z, y)$ to obtain

$$
\overline{\psi}(x, z') \leq f(z', y) \leq f(z, y) + (c)^\top (z' - z) = \overline{\psi}(x, z) + c^\top (z - z'), \quad \forall z' \in \mathbb{R}^n,
$$

where the first inequality holds because $\overline{\psi}(x, z') = -\infty$ if $z' \notin \overline{X}$ and $y \in \overline{\nabla}(x, z)$ must be feasible to the constraint $G(x, y) \leq 0$ if $z' \in \overline{X}$; and the second inequality is due to the concavity of $f(\cdot, y)$. By applying \([46, Theorem 23.5]\), we have $c \in \partial_2(-\overline{\psi})(x, z)$ and part (a) is proved. Part (b) is a consequence of \([19, Proposition 4.4.26 (c)]\) on the relationship between the subdifferentials of an implicitly convex-concave function and its lifted counterpart.

Part (a) of the above lemma can viewed as a weaker version of Danskin’s theorem \([20, 5, 15]\). Instead of a complete characterization of the subdifferential of an optimal value function in the aforementioned papers, we only need to obtain one element from this subdifferential to design our algorithms later. Therefore, only a one-sided inclusion as in part (a) is needed, which holds without the compactness of the feasible set in terms of $y$.

3.2 The partial Moreau envelope

Equipped with the lifted function $\overline{\psi}$, one may be able to construct computationally friendly surrogates of the recourse function in \((\ref{func:6})\) via a modification of the usual Moreau envelope. Let us first recall the definition of the classical Moreau envelope. An extended-real-valued function $f$ is said to be proper if $f(x) < +\infty$ for some $x \in \mathbb{R}^n$, and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. Given a proper lower semicontinuous function $\theta : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ and a positive scalar $\gamma$, its Moreau envelope is

$$
e^{\text{ori}}_{\gamma} \theta(x) \triangleq \inf_{z \in \mathbb{R}^n} \left\{ \theta(z) + \frac{1}{2\gamma} \|x - z\|^2 \right\}, \quad x \in \mathbb{R}^n.
$$

We use the superscript “ori” to emphasize that this is the original definition of the Moreau envelope and is different from our later modification. The function $\theta$ is said to be prox-bounded if there exists $\gamma > 0$ such that $e^{\text{ori}}_{\gamma} \theta(x) > -\infty$ for some $x \in \mathbb{R}^n$. It is known that a convex function is always prox-bounded and its Moreau-envelope is continuously differentiable (c.f. \([50, Theorem 2.26]\)). In general, for any prox-bounded function $\theta$, the parametric functions $e^{\text{ori}}_{\gamma} \theta(x) \uparrow \theta(x)$ as $\gamma \downarrow 0$ for all $x \in \mathbb{R}^n$. Therefore, one can view the Moreau envelope as a lower approximation of the original
function. However, if $\theta$ is nonconvex and nonsmooth, the function $e^{\text{ori}}_\gamma \theta$ may be neither convex nor smooth. Nevertheless, for any $x \in \mathbb{R}^n$, it holds that (see, e.g., \cite{3,37})

$$e^{\text{ori}}_\gamma \theta(x) = \frac{1}{2\gamma} \|x\|^2 - \sup_{z \in \mathbb{R}^n} \left\{ -\theta(z) - \frac{1}{2\gamma} \|z\|^2 + \frac{1}{\gamma} z^\top x \right\},$$

(9)

which indicates that one can always obtain a dc decomposition of $e^{\text{ori}}_\gamma \theta$ no matter $\theta$ is convex or not. The only trouble brought by the nonconvexity of $\theta$ is that the inner sup problem for $y$ may not be concave (especially if $\theta$ is not weakly convex), thus one may not be able to evaluate the subgradient of the second term at a given $x$ when using the dc algorithm to minimize the function $e^{\text{ori}}_\gamma \theta$. Specifically, in the context of the recourse function (6), its associated Moreau envelope is

$$e^{\gamma}_\psi(x) = \inf_{x \in \mathbb{R}^n} \left\{ \psi(z) + \frac{1}{2\gamma} \|x - z\|^2 \right\} - \sup_{z \in \mathbb{R}^n} \left\{ -f(z, y) - \frac{1}{2\gamma} \|z\|^2 + \frac{1}{\gamma} z^\top y \bigg| G(z, y) \leq 0 \right\},$$

where the inner max problem is not jointly concave in $(z, y)$ since $f$ is not assumed to be jointly convex. This issue motivates us to introduce the following new type of envelopes tailored to icc functions that is more computationally tractable:

$$e^{\gamma}_\theta(z) \triangleq \inf_{x \in \mathbb{R}^n} \left\{ \theta(x, z) + \frac{1}{2\gamma} \|x - z\|^2 \right\},$$

(10)

where $\theta : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is any icc function and $\overline{\theta} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\}$ is its lifted counterpart as in Definition 1. When $\overline{\theta}(x, z)$ is independent of $z$ (so that this function only has the convex part), the above definition reduces to the usual Moreau envelope. Hence, we term the new regularization of $\theta$ in (10) its partial Moreau envelope. Similarly as in (9), the newly defined partial Moreau has the following explicit dc decomposition

$$e^{\gamma}_\theta(z) = \frac{1}{2\gamma} \|z\|^2 - \sup_{x \in \mathbb{R}^n} \left\{ -\overline{\theta}(x, z) - \frac{1}{2\gamma} \|x\|^2 + \frac{1}{\gamma} z^\top x \right\} \left\{ \text{strongly convex} \right\},$$

(11)

denoted as $g^\gamma(z)$, convex.

We denote the optimal solution mapping of the minimization problem in (10) as

$$P_{\gamma} \theta(z) \triangleq \arg\min_{x \in \mathbb{R}^n} \left\{ \overline{\theta}(x, z) + \frac{1}{2\gamma} \|x - z\|^2 \right\},$$

for any $z \in \mathbb{R}^n$, it holds that $\emptyset \neq P_{\gamma} \theta(z) \subseteq \text{dom} \, \theta$. When $z \in \text{dom} \, \theta$, the mapping is single-valued since the inner objective function is strongly convex in $x$; for this case, we follow the terminology in the literature to call $P_{\gamma} \theta(z)$ the proximal point of $\overline{\theta}$ at $z$. Similar to the classical Moreau envelope, the partial Moreau envelope approximates the original function from below. The following lemma provides the gap between the partial Moreau envelope and the original function under the Lipschitz continuity of $\overline{\theta}(\bullet, z)$. The proof is adapted from \cite[Proposition 3.4]{42} on a similar property regarding the Moreau envelope.
Lemma 3. Consider an icc function $\theta : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and its lifted counterpart $\overline{\theta} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Let $X \subseteq \text{dom } \theta$. Assume that $\overline{\theta}(\cdot, z)$ is Lipschitz continuous on $\text{dom } \theta$ with Lipschitz constant $\kappa(z)$ for every $z \in X$, i.e.,
\[ |\overline{\theta}(x_1, z) - \overline{\theta}(x_2, z)| \leq \kappa(z) \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \text{dom } \theta, \, z \in X, \]
then $0 \leq \theta(z) - e_\gamma \theta(z) \leq \gamma \kappa(z)^2 / 2$ for any $z \in X$.

Proof. For any $z \in X$, it holds that
\[ 0 \leq \theta(z) - e_\gamma \theta(z) = \overline{\theta}(z, z) - \overline{\theta}(P_\gamma \theta(z), z) - \|P_\gamma \theta(z) - z\|^2 / (2\gamma) \leq \kappa(z) \|P_\gamma \theta(z) - z\| - \|P_\gamma \theta(z) - z\|^2 / (2\gamma) \leq \gamma \kappa(z)^2 / 2, \]
where the second inequality follows from the Lipschitz continuity of $\overline{\theta}(\cdot, z)$ and the last inequality uses the fact that $\max_{t \geq 0} \left[ \kappa(z) t - t^2 / (2\gamma) \right] = \gamma \kappa(z)^2 / 2$. \hfill \Box

With $g_\gamma$ defined in (11), it follows from similar arguments in the proof of Lemma 2 that for any $z \in \text{dom } \theta$,
\[ 1 / \gamma P_\gamma \theta(z) + \partial_2(-\overline{\theta})(P_\gamma \theta(z), z) \subseteq \partial g_\gamma(z). \]

One can then obtain the following convex majorization of $e_\gamma \theta(z)$ at any given point $z' \in \text{dom } \theta$ based on the subgradient inequality of the convex function $g_\gamma$:
\[ e_\gamma \theta(z) \leq \overline{e}_\gamma \theta(z; z') \triangleq \frac{1}{2\gamma} \|z\|^2 - g_\gamma(z') - (P_\gamma \theta(z') / \gamma + c)^\top (z - z'), \quad \forall z \in \text{dom } \theta, \]
where $c \in \partial_2(-\overline{\theta})(P_\gamma \theta(z'), z')$.

4 The decomposition algorithm and its convergence

Based on the discussion in the last section, we are now ready to present the decomposition algorithm for solving the nonconvex two-stage SP (1) and analyze its convergence. In this section, we focus on the case where there are fixed scenarios $\{\xi^1, \ldots, \xi^S\}$, each realized with probability $1 / S$. The problem (1) then reduces to
\[ \min_{x \in X \subseteq \mathbb{R}^n} \left\{ \varphi(x) + \frac{1}{S} \sum_{s=1}^S \psi(x; \xi^s) \right\}, \]
where each $\psi(x; \xi^s)$ is given by (2). The above problem can be viewed as a sample average approximation of the two-stage SP (1) under a prescribed sample size $S$. All the discussions in this section can be easily adapted to the case where the distribution of $\tilde{\xi}$ has finite support (but unequal probability mass for different $\xi^s$). We will work on the internal sampling scheme for continuously distributed $\xi$ in the next section.
4.1 The algorithmic framework

Our goal is to solve the nonconvex problem (13) via a successive approximation scheme. For any \( \xi \in \Xi \) and \( z \in X \), the partial Moreau envelope of the recourse function (2) associated with the bivariate function (8) is

\[
e_{\gamma} \psi(z; \xi) \triangleq \min_{x \in \mathbb{R}^{n_1}} \left\{ \overline{\psi}(x, z; \xi) + \frac{1}{2\gamma} \|x - z\|^2 \right\} = \min_{x \in \bar{X}, y \in \mathbb{R}^{n_2}} \left\{ f(z, y; \xi) + \frac{1}{2\gamma} \|x - z\|^2 \right\} \quad \text{subject to} \quad G(x, y; \xi) \leq 0.
\]

We consider a double-loop algorithm where the outer loop updates the parameter \( \gamma \) in the partial Moreau envelope and the inner loop solves the nonconvex problem \( \min_{x \in X} (\phi(x) + \frac{1}{S} \sum_{s=1}^{S} e_{\gamma_s} \psi(x; \xi_s)) \) to stationarity for a fixed \( \gamma \). To solve the latter nonconvex problem during the \( \nu \)-th inner loop, we replace the Moreau-regularized recourse function \( e_{\gamma_s} \psi(x; \xi) \) with its upper approximation constructed at the latest iterate \( x_{\nu, i} \), where \( i \) denotes the inner iterate index. For each \( s \), let \( (x_{\nu, i}^s, y_{\nu, i}^s) \) be one of the optimal solutions of (14) at \( (z, \xi) = (x_{\nu, i}, \xi_s) \), which can be computed by solving separable convex optimization problems. Notice that \( x_{\nu, i}^s = P_{\gamma_s} \psi(x_{\nu, i}; \xi_s) \in \bar{X} \). One may then derive from (12) the following upper approximating function of \( e_{\gamma_s} \psi(x; \xi) \):

\[
\tilde{e}_{\gamma_s} \psi(x; \xi^s; x_{\nu, i}) \triangleq \frac{1}{2\gamma_s} \|x\|^2 - g_{\gamma_s}(x_{\nu, i}; \xi^s) - (x_{\nu, i}^s/\gamma_s + c_{\nu, i}^s)^\top (x - x_{\nu, i}),
\]

where \( g_{\gamma_s}(x_{\nu, i}; \xi^s) \triangleq \frac{1}{2\gamma_s} \|x_{\nu, i}\|^2 - e_{\gamma_s} \psi(x_{\nu, i}; \xi^s) \) and \( c_{\nu, i}^s \in \partial_2 (\overline{\psi})(x_{\nu, i}^s, x_{\nu, i}; \xi^s) \). Due to Lemma 2, a particular way to choose \( c_{\nu, i}^s \) is to take an element from \( \partial_1 (-f)(x_{\nu, i}^s, y_{\nu, i}^s; \xi^s) \). The resulting master problem to generate the next first-stage iterate \( x_{\nu, i+1} \) is

\[
\min_{x \in \mathcal{X}} \left\{ \phi(x) + \frac{1}{S} \sum_{s=1}^{S} \tilde{e}_{\gamma_s} \psi(x; \xi^s; x_{\nu, i}) \right\}.
\]

The inner iteration continues until the distance between two consecutive iterates \( x_{\nu, i} \) and \( x_{\nu, i+1} \) is sufficiently close. We summarize the procedure of the decomposition algorithm below. When \( S = 1 \), it reduces to the algorithm in [19, Algorithm 7.2.1] to minimize an implicitly convex-concave function (without decomposition).
Algorithm 1: A decomposition algorithm for the nonconvex two-stage SP (13)

**Input:** Initial point \( x_0 \in X \) and two scalar sequences \( \{\gamma_\nu\}_{\nu \geq 0} \downarrow 0 \) and \( \{\varepsilon_\nu\}_{\nu \geq 0} \downarrow 0 \).

**Outer loop:** Set \( \nu = 0 \).
1. Execute the inner loop with the initial point \( x_{\nu,0} \) and parameters \( \gamma_\nu \) and \( \varepsilon_\nu \).
2. Set \( \nu \leftarrow \nu + 1 \) and repeat step 1 until a prescribed stopping criterion is satisfied.

**Inner loop:** Set \( i = 0 \) and \( x_{\nu,0} = x_\nu \).
1. Solve the subproblem (14) at \((z, \xi) = (x_{\nu,i}, \xi^s)\) for all \( s \) to get solutions \( x_{\nu,i}^s, y_{\nu,i}^s \).
2. Select \( c_{\nu,i} \in \partial_1 \left(-f\right)(x_{\nu,i}, y_{\nu,i}; \xi^s) \subseteq \partial_2 \left(-\psi\right)(x_{\nu,i}, x_{\nu,i}; \xi^s) \) for each \( s \).
3. Solve the master problem (16) to obtain \( x_{\nu,i+1} \).
4. Set \( i \leftarrow i + 1 \) and repeat the above steps if \( \|x_{\nu,i+1} - x_{\nu,i}\| > \varepsilon_\nu \gamma_\nu \). Otherwise, break the inner loop with \( x_{\nu+1} \triangleq x_{\nu,i+1} \).

Notice that each \( \bar{\psi}_{\nu,\psi}(x^s; \xi^s; x_{\nu,i}) \) is a strongly convex quadratic function in \( x \). Therefore, the master problem (16) is a strongly convex optimization problem with \( n_1 \) number of variables, which is usually easy to solve.

### 4.2 Convergence Analysis

This subsection is about the global convergence of the sequence generated by Algorithm 1. We begin with several technical assumptions that will be used in our convergence analysis. Since our focus of this section is to solve problem (13) with fixed scenarios \( \{\xi^1, \ldots, \xi^S\} \), one should interpret Assumptions A-B as requirements on all realizations \( \{\xi^1, \ldots, \xi^S\} \). The “almost surely” part will be used in the next section where a generally distributed \( \tilde{\xi} \) is considered.

**Assumption A.** The minimization problem of \( y \) in defining \( \bar{\psi}(x, z; \xi) \) in (8) has an optimal solution for any \( (x, z) \in X \times X \) (almost surely).

**Assumption B.** There exists a measurable function \( \kappa_1 : \Xi \rightarrow \mathbb{R}_+ \) such that the following condition holds (almost surely):

\[
|\bar{\psi}(x_1, z; \xi) - \bar{\psi}(x_2, z; \xi)| \leq \kappa_1(\xi) \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \overline{X}, z \in X.
\]

Some remarks are in order. Assumption A guarantees that for any \( z \in \overline{X} \), the optimal solution of (14) in terms of \( y \) exists. This assumption also implies the relatively complete recourse of the original problem (1) that \( \psi(x; \xi) \) is finite for all feasible \( x \in X \) almost surely. Assumption B is a stochastic version of the Lipschitz condition in Lemma 3. Since \( X \) is a compact set, we assume without loss of generality that \( \kappa_1(\xi) \) is independent of \( z \).

For convenience, we denote

\[
\zeta_S(x) \triangleq \varphi(x) + \frac{1}{S} \sum_{s=1}^{S} \psi(x; \xi^s) \quad \text{and} \quad \zeta_{S,\gamma_\nu}(x) \triangleq \varphi(x) + \frac{1}{S} \sum_{s=1}^{S} \gamma_\nu \psi(x; \xi^s).
\]

In the proposition below, we show that for any prescribed positive scalar \( \varepsilon_\nu \), the \( \nu \)-th inner loop of Algorithm 1 terminates in finite steps.

---

13
Proposition 2. (Convergence of the inner loop for Algorithm 1) Suppose that Assumptions A-B hold. Then the following statements hold for any $\nu$-th inner loop.

(a) $\hat{\zeta}_{S,\gamma_{\nu}}(x_{\nu,i+1}) \leq \hat{\zeta}_{S,\gamma_{\nu}}(x_{\nu,i}) - \frac{1}{2\gamma_{\nu}}\|x_{\nu,i} - x_{\nu,i+1}\|^2$ for any $i \geq 0$.

(b) $\lim_{i \to 0} \|x_{\nu,i} - x_{\nu,i+1}\| = 0$ and the stopping criterion $\|x_{\nu,i+1} - x_{\nu,i}\| \leq \varepsilon_{\nu}\gamma_{\nu}$ is achievable in finite steps, i.e.,

\[ i_{\nu} \triangleq \min \{ i \in \mathbb{Z}_+ \mid \|x_{\nu,i+1} - x_{\nu,i}\| \leq \varepsilon_{\nu}\gamma_{\nu} \} < +\infty. \]

In addition, we have

\[
\text{dist} \left( 0, \frac{1}{S} \sum_{s=1}^{S} \left[ \partial_1 \psi(x_{\nu,i,s}; x_{\nu,i},\xi^s) - \partial_2 (\psi) (x_{\nu,i,s}; x_{\nu,i},\xi^s) + N_X(x_{\nu,i,s}) \right] \right) \leq \varepsilon_{\nu}.
\]

Proof. Consider the $\nu$-th inner loop of Algorithm 1. We first show that the sequence $\left\{ \hat{\zeta}_{S,\gamma_{\nu}}(x_{\nu,i}) \right\}_{i \geq 0}$ is nonincreasing. Since the function $\hat{\mathcal{S}}_{\gamma_{\nu}} \psi(\bullet; \xi^s; x_{\nu,i})$ defined in (15) is quadratic, we have, by writing $a_{\nu,i}^s \triangleq \nabla_x \hat{\mathcal{S}}_{\gamma_{\nu}} \psi(\bullet; \xi^s; x_{\nu,i})(x_{\nu,i+1})$,

\[
\hat{\zeta}_{S,\gamma_{\nu}}(x_{\nu,i}) = \varphi(x_{\nu,i}) + \frac{1}{S} \sum_{s=1}^{S} \hat{\mathcal{S}}_{\gamma_{\nu}} \psi(x_{\nu,i}; \xi^s; x_{\nu,i})
\]

\[
= \varphi(x_{\nu,i}) + \frac{1}{S} \sum_{s=1}^{S} \hat{\mathcal{S}}_{\gamma_{\nu}} \psi(x_{\nu,i+1}; \xi^s; x_{\nu,i})
\]

\[
+ \frac{1}{S} \sum_{s=1}^{S} (a_{\nu,i}^s)^T (x_{\nu,i} - x_{\nu,i+1}) + \frac{1}{2\gamma_{\nu}} \|x_{\nu,i} - x_{\nu,i+1}\|^2,
\]

where the first equality is because $\hat{\mathcal{S}}_{\gamma_{\nu}} \psi(x_{\nu,i}; \xi^s; x_{\nu,i}) = e_{\gamma_{\nu}} \psi(x_{\nu,i}; \xi^s)$. Since $x_{\nu,i+1}$ is the optimal solution of the master problem (16), one may obtain that

\[
\left( b_{\nu,i} + \frac{1}{S} \sum_{s=1}^{S} a_{\nu,i}^s \right)^T (x_{\nu,i} - x_{\nu,i+1}) \geq 0 \quad \text{for some } b_{\nu,i} \in \partial \varphi(x_{\nu,i+1}).
\]

The convexity of $\varphi$ and the above inequalities imply that

\[
\hat{\zeta}_{S,\gamma_{\nu}}(x_{\nu,i}) \geq \varphi(x_{\nu,i+1}) + \frac{1}{S} \sum_{s=1}^{S} e_{\gamma_{\nu}} \psi(x_{\nu,i+1}; \xi^s) + \frac{1}{2\gamma_{\nu}} \|x_{\nu,i} - x_{\nu,i+1}\|^2
\]

\[
\geq \varphi(x_{\nu,i+1}) + \frac{1}{S} \sum_{s=1}^{S} e_{\gamma_{\nu}} \psi(x_{\nu,i+1}; \xi^s) + \frac{1}{2\gamma_{\nu}} \|x_{\nu,i} - x_{\nu,i+1}\|^2
\]

\[
= \hat{\zeta}_{S,\gamma_{\nu}}(x_{\nu,i+1}) + \frac{1}{2\gamma_{\nu}} \|x_{\nu,i} - x_{\nu,i+1}\|^2,
\]

where the second inequality uses the fact that $\hat{\mathcal{S}}_{\gamma_{\nu}} \psi(x; \xi^s; \bar{x}) \geq e_{\gamma_{\nu}} \psi(x; \xi^s)$ for any $x, \bar{x}$ and $\xi^s$. We thus prove the part (a). Using the compactness of $X$ and Assumption B, one may further derive
from Lemma 3 that

$$\inf_{x \in X} \gamma \zeta_{S, \gamma}(x) \geq \left[ \inf_{x \in X} \varphi(x) \right] + \frac{1}{s} \sum_{s=1}^{S} \left[ \inf_{x \in X} \psi(x; \xi^s) - \frac{\gamma}{2} \cdot \kappa_1(\xi^s)^2 \right] > -\infty.$$ 

Hence, the sequence \( \{ \zeta_{S, \gamma}(x_{\nu, i}) \}_{i \geq 0} \) is bounded below, which further yields that \( \{ \zeta_{S, \gamma}(x_{\nu, i}) \}_{i \geq 0} \) converges and \( \|x_{\nu, i+1} - x_{\nu, i}\| \) converges to 0 as \( i \to \infty \). The latter convergence indicates that the inner iterations terminate after finite steps.

To show the rest statement of this proposition, we first obtain from the optimality condition of the subproblem (14) at \( x_{\nu, i+1} \) that

$$0 \in \partial \left( \varphi(x_{\nu, i+1}) + \frac{1}{s} \sum_{s=1}^{S} \widetilde{\psi}(x_{\nu, i+1}; \xi^s, x_{\nu, i}) + \mathcal{N}(x_{\nu, i+1}) \right)$$

where the equation is due to the sum rule of the subdifferentials for convex functions [46, Theorem 23.8]. From the optimality condition of the subproblem (14), we obtain

$$0 \in \partial \psi(x_{\nu, i+1}) + \frac{1}{s} \sum_{s=1}^{S} \left( x_{\nu, i+1} - x_{\nu, i} \right) / \gamma \nu + \mathcal{N}(x_{\nu, i+1}), \quad \forall s = 1, \ldots, S.$$ 

Taking the sum over (19) and the above inclusions from \( s = 1, \ldots, S \), we get

$$0 \in \frac{1}{s} \sum_{s=1}^{S} \partial \psi(x_{\nu, i+1}) + \left[ \partial \varphi(x_{\nu, i+1}) - \frac{1}{s} \sum_{s=1}^{S} e_{\nu, i}^s + \mathcal{N}(x_{\nu, i+1}) \right]$$

$$+ \left( x_{\nu, i+1} - x_{\nu, i} \right) / \gamma \nu + \frac{1}{s} \sum_{s=1}^{S} \mathcal{N}(x_{\nu, i})$$

$$\leq \frac{1}{s} \sum_{s=1}^{S} \left[ \partial \psi(x_{\nu, i+1}) - \partial \psi(x_{\nu, i+1}) \right] + \partial \varphi(x_{\nu, i+1})$$

$$+ \left( x_{\nu, i+1} - x_{\nu, i} \right) / \gamma \nu + \mathcal{N}(x_{\nu, i+1}) + \frac{1}{S} \sum_{s=1}^{S} \mathcal{N}(x_{\nu, i}),$$

where the last inclusion is due to the definition of \( c_{\nu, i}^s \). Consequently, we derive

$$\text{dist} \left( 0, \frac{1}{S} \sum_{s=1}^{S} [ \partial \psi(x_{\nu, i+1}) - \partial \psi(x_{\nu, i+1}) + \mathcal{N}(x_{\nu, i}) ] \right)$$

$$\leq \left\| x_{\nu, i+1} - x_{\nu, i} \right\| / \gamma \nu \leq \varepsilon_{\nu},$$

where the last inequality is due to the stopping rule of the inner loop.
Let \( \bar{x} \) be an accumulation point of the sequence \( \{x_\nu\}_{\nu \geq 0} \), which must exist due to the compactness of \( X \). We let \( \{x_{\nu+1}\}_{\nu \in \alpha} \) be a convergent subsequence to \( \bar{x} \), where \( \alpha \) is a subset of \( \mathbb{N} \). In the following lemma, we show the convergence of the sequence \( \{x_{\nu,i}\}_{\nu \in \alpha} \) for each \( s \) with \( x_{\nu,i}^s \) being the partial optimal solution of the subproblem \( (14) \) at \( (z, \xi) = (x_{\nu,i}, \xi^s) \).

**Lemma 4.** Let Assumptions A-B hold and \( \bar{x} \) be an accumulation point of the sequence \( \{x_\nu\}_{\nu \geq 0} \) generated by Algorithm 1. If the subsequence \( \{x_{\nu,i}\}_{\nu \in \alpha} \) converges to \( \bar{x} \) and, for each \( s = 1, \ldots, S \), the corresponding subsequence \( \{x_{\nu,i}^s\}_{\nu \in \alpha} \) also converges to \( \bar{x} \).

**Proof.** Since \( \lim_{\nu \to \infty} \|x_{\nu+1} - x_{\nu,i}\| = \lim_{\nu \to \infty} \|x_{\nu,i+1} - x_{\nu,i}\| = 0 \) from Theorem 2, we know that \( \{x_{\nu,i}\}_{\nu \in \alpha} \) also converges to \( \bar{x} \). For each \( s = 1, \ldots, S \) and \( \nu \in \alpha \), we have, by the definition of \( x_{\nu,i}^s = P_{\gamma_{\nu}}\psi(x_{\nu,i}; \xi^s) \in \overline{X} \),

\[
\frac{1}{2\gamma_{\nu}} \|x_{\nu,i}^s - x_{\nu,i}\|^2 + \inf_{x \in \overline{X}} \overline{\psi}(x, x_{\nu,i}; \xi^s) \leq \frac{1}{2\gamma_{\nu}} \|x_{\nu,i}^s - x_{\nu,i}\|^2 + \overline{\psi}(x_{\nu,i}^s, x_{\nu,i}; \xi^s) \\
= e_{\gamma_{\nu}} \psi(x_{\nu,i}^s; \xi^s) \leq \frac{1}{2\gamma_{\nu}} \|x - x_{\nu,i}\|^2 + \overline{\psi}(x, x_{\nu,i}; \xi^s).
\]

The above inequality yields that

\[
\|x_{\nu,i}^s - x_{\nu,i}\| \leq \sqrt{\|x - x_{\nu,i}\|^2 + 2\gamma_{\nu}\left(\overline{\psi}(x, x_{\nu,i}; \xi^s) - \inf_{x \in \overline{X}} \overline{\psi}(x, x_{\nu,i}; \xi^s)\right)} \\
\leq \|x - x_{\nu,i}\| + \sqrt{2\gamma_{\nu}\left(\overline{\psi}(x, x_{\nu,i}; \xi^s) - \inf_{x \in \overline{X}} \overline{\psi}(x, x_{\nu,i}; \xi^s)\right)} \\
\leq \|x - x_{\nu,i}\| + \sqrt{2\gamma_{\nu}\kappa_1(\xi^s) R(\overline{X})} \to 0 \quad \text{as} \quad \nu \to \infty,
\]

where \( R(\overline{X}) \) denotes the diameter of the compact set \( \overline{X} \), and the last inequality follows from the uniform Lipschitz continuity of \( \overline{\psi}(\bullet, z; \xi) \) on \( \overline{X} \) in Assumption B. Since \( \kappa_1(\xi^s) < \infty \) for each \( s \) and \( \gamma_{\nu} \to 0 \), we derive that

\[
\|x_{\nu,i}^s - \bar{x}\| \leq \|x_{\nu,i}^s - x_{\nu,i}\| + \|x_{\nu,i} - \bar{x}\| \to 0 \quad \text{as} \quad \nu \to \infty.
\]

The proof is completed. \( \square \)

For any \( r > 0 \), we denote

\[
\partial^+_1 \overline{\psi}(\bar{x}, \bar{x}) \equiv \bigcup_{x, z \in B(\bar{x}, r)} \partial^+_1 \overline{\psi}(x, z),
\]

and similarly for \( \partial^+_2(-\overline{\psi}) \). We summarize a few results of the subgradient and the normal cone mappings below that will be used in the subsequent analysis.

**Lemma 5.** Under Assumption A, the following statements hold.

\( (a) \) The mappings \( \partial^+_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_1} \), \( N_X : X \to \mathbb{R}^{n_1} \), \( \partial^+_1 \overline{\psi} : \text{int}(\overline{X}) \times \text{int}(\overline{X}) \to \mathbb{R}^{n_1} \) and \( \partial^+_2(-\overline{\psi}) : \text{int}(\overline{X}) \times \text{int}(\overline{X}) \to \mathbb{R}^{n_1} \) are osc;

\( (b) \) For any fixed \( \bar{x} \in X(\subseteq \text{int}(\overline{X})) \), there exists \( r > 0 \) such that \( \partial^+_1 \overline{\psi}(\bar{x}, \bar{x}) \) is bounded.
Proof. (a) The osc of $\partial \varphi$ and $\mathcal{N}_X$ are standard results in convex analysis (see [50, Propositions 6.6 and 8.7]). To prove that $\partial_1 \overline{\psi}$ is osc on $\text{int}(X) \times \text{int}(X)$, we first notice that for any fixed $d \in \mathbb{R}^{n_1}$, the directional derivative $\overline{\psi}'_1((\bullet, \bullet); d)$ is upper semicontinuous jointly at $(\overline{x}, \overline{z}) \in \text{int}(X) \times \text{int}(X)$ [19, Proposition 4.4.26 (a)]. Since $\overline{\psi}'_1((x, z); \bullet)$ is the support function of the partial subgradient $\partial_1 \overline{\psi}(x, z)$, we have
\[
\limsup_{(x, z) \to (\overline{x}, \overline{z})} \left( \sup_{a \in \partial_1 \overline{\psi}(x, z)} a^\top d \right) \leq \sup_{a' \in \partial_1 \overline{\psi}(\overline{x}, \overline{z})} (a')^\top d \quad \forall d \in \mathbb{R}^{n_1},
\]
which implies $\limsup_{(x, z) \to (\overline{x}, \overline{z})} \partial_1 \overline{\psi}(x, z) \subseteq \partial_1 \overline{\psi}(\overline{x}, \overline{z})$ according to [44, Corollary 13.1.1]. The osc of $\partial_2(-\overline{\psi})$ on $\text{int}(X) \times \text{int}(X)$ can be proved similarly.

(b) Suppose for the sake of contradiction that for any $r > 0$, the set $\partial_1 \overline{\psi}(\overline{x}, \overline{z})$ is unbounded. Then there exists a sequence of subgradients $c^k \in \partial_1 \overline{\psi}(x^k, z^k)$ with $(x^k, z^k) \to (\overline{x}, \overline{z})$, and $\|c^k\| \to +\infty$. By taking a subsequence if necessary, we assume that the normalized subgradient $d_k = c^k/\|c^k\|$ converges to some $d \in \mathbb{R}^{n_1}$. Since $x^k \to \overline{x} \in \text{int}(X)$, there exists a positive scalar $t > 0$ such that $B(\overline{x}, t) \subseteq \text{int}(X)$ and $B(x^k, t) \subseteq \text{int}(X)$ for all sufficiently large $k$. Using the convexity of $\overline{\psi}(\bullet, z^k)$, we obtain $\overline{\psi}(x^k + td_k, z^k) - \overline{\psi}(x^k, z^k) \geq (c^k)^\top td_k = t\|c^k\|$. Taking limits on both sides and using the continuity of $\overline{\psi}(\bullet, z)$, we have
\[
\limsup_{k \to \infty} \|c^k\| = 0 < \|c^k\| = \lim_{k \to \infty} \|c^k\| = t\|c^k\|,
\]
which is a contradiction. The proof is thus completed. \hfill \square

We are now ready to present the global convergence of the sequence generated by Theorem 1. We shall prove that every accumulation point $\overline{x}$ of $\{x_\nu\}_{\nu \geq 0}$ is a critical point of problem (13) satisfying
\[
0 \in \partial \varphi(\overline{x}) + \frac{1}{S} \sum_{s=1}^{S} \left[ \partial_1 \overline{\psi}(\overline{x}, \overline{x}; \xi^s) - \partial_2(-\overline{\psi})(\overline{x}, \overline{x}; \xi^s) \right] + \mathcal{N}_X(\overline{x}). \tag{20}
\]
It has been shown in Lemma 2 that under Assumption A, we have $\partial_C \psi(x; \xi) \subseteq \partial_1 \overline{\psi}(x, x; \xi) - \partial_2(-\overline{\psi})(x, x; \xi)$. Hence, the condition (20) is weaker than the Clarke stationarity of problem (13) pertaining to $0 \in \partial \varphi(\overline{x}) + \frac{1}{S} \sum_{s=1}^{S} \partial_C \psi(\overline{x}; \xi^s) + \mathcal{N}_X(\overline{x})$. The term “critical point” is adapted from the result of the difference-of-convex (dc) algorithm to solve a dc problem minimize $[\theta_1(x) - \theta_2(x)]$, where the accumulation point satisfies $0 \in \partial \theta_1(x) - \partial \theta_2(x) + \mathcal{N}_X(x)$ [11].

Theorem 1 (Subsequential convergence for Algorithm 1). Let Assumptions A-B hold. Then any accumulation point of the sequence $\{x_\nu\}_{\nu \geq 0}$ generated by Algorithm 1 is a critical point of (13) satisfying (20).

Proof. Let $\overline{x}$ be the limit of a convergent subsequence $\{x_{\nu+1}\}_{\nu \in \alpha}$. By Theorem 2 and Lemma 4, we know that the subsequences $\{x_{\nu,i}\}_{\nu \in \alpha}$ and $\{x^s_{\nu,i}\}_{\nu \in \alpha}$ for all $s$ converge to the same accumulation
point $\bar{x}$ as $\nu \to \infty$. Using the triangle inequality of the distance function, we have that

$$
\text{dist} \left( 0, \frac{1}{S} \sum_{s=1}^{S} \left[ \partial_1 \overline{\psi}(\bar{x}, \bar{x}; \xi^s) - \partial_2(-\overline{\psi})(\bar{x}, \bar{x}; \xi^s) \right] + \partial \varphi(\bar{x}) + N_X(\bar{x}) \right) \\
\leq \text{dist} \left( 0, \frac{1}{S} \sum_{s=1}^{S} \left[ \partial_1 \overline{\psi}(x_{\nu, i_s}^s, x_{\nu, i_s}; \xi^s) - \partial_2(-\overline{\psi})(x_{\nu, i_s}^s, x_{\nu, i_s}; \xi^s) \right] + \partial \varphi(x_{\nu, i_s}^s + w_{\nu}) \right) \\
+ \frac{1}{S} \sum_{s=1}^{S} \mathbb{D} \left( \left[ \frac{\partial_1 \overline{\psi}(x_{\nu, i_s}^s, x_{\nu, i_s}; \xi^s)}{- \partial_2(-\overline{\psi})(x_{\nu, i_s}^s, x_{\nu, i_s}; \xi^s)} \right] \right) + \text{dist} \left( w_{\nu}, N_X(\bar{x}) \right) 
$$

where $w_{\nu}$ can be any element in $N_X(x_{\nu, i_s}^s + 1)$. By Theorem 2 there is a sequence $\{w_{\nu}\}_{\nu \geq 0}$ with $w_{\nu} \in N_X(x_{\nu, i_s}^s + 1)$ such that (i) converges to 0. Since for all $s$, $\lim_{\nu(\in \alpha) \to \infty} x_{\nu, i_s}^s = \bar{x} \in X \subseteq \text{int}(X)$, we thus obtain $N_X(x_{\nu, i_s}^s) = \{0\}$ for sufficiently large $\nu$ and any $s$. Then (iii) $\to 0$ as $\nu(\in \alpha) \to \infty$. To see the convergence of terms (ii), (iv), and (v), based on the derived osc properties in Lemma 5 (a) and 50 Proposition 5.12, we only need to prove that for each $s$, the sequences

$$
\{\partial_1 \overline{\psi}(x_{\nu, i_s}^s, x_{\nu, i_s}; \xi^s)\}_{\nu(\in \alpha) \geq 0}, \{\partial \varphi(x_{\nu, i_s}^s + 1)\}_{\nu(\in \alpha) \geq 0}, \text{ and } \{w_{\nu}\}_{\nu(\in \alpha) \geq 0} 
$$

are uniformly bounded for sufficiently large $\nu_0$. Indeed, Lemma 5 (b) implies the existence of $r > 0$ and integer $\nu_0$ such that for any $\nu(\in \alpha) \geq \nu_0$ and any $s$,

$$
\partial_1 \overline{\psi}(x_{\nu, i_s}^s, x_{\nu, i_s}; \xi^s) \leq \partial_1 \overline{\psi}(\bar{x}, \bar{x}; \xi^s) \text{ is bounded,} 
$$

which gives us the uniform boundedness of the first term in (21). The uniform boundedness of the second term in (21) is a direct consequence of Theorem 24.7 since $\varphi$ is real-valued and convex, and $x_{\nu, i_s}^s + 1 \to \bar{x}$. Observe that $\{w_{\nu}\}_{\nu \geq 0}$ must be bounded because (i) converges to 0 and all set-valued mappings in (i) except $w_{\nu}$ have proven to be uniformly bounded. Henceforth, we have proved that any accumulation point $\bar{x}$ is a critical point of (13) satisfying (20). $\square$

In the following, we establish the convergence to a stronger type of stationarity under additional assumptions. Suppose that $f(\bullet, \bullet; \xi^s)$ and $G(\bullet, \bullet; \xi^s)$ are continuously differentiable for all $\xi^s$. At $x = \bar{x}$, we say $\bar{y}^s$ is an optimal solution of the convex second-stage problem with $\xi = \xi^s$ and $\lambda^s$ being the corresponding multiplier if the following Karush-Kuhn-Tucker condition is satisfied:

$$
0 \in \nabla_y f(\bar{x}, \bar{y}^s; \xi^s) + \sum_{j=1}^{\ell} \lambda_j^s \nabla_y g_j(\bar{x}, \bar{y}^s; \xi^s) \text{ and } \lambda^s \in N_{\mathbb{R}^\ell}(G(\bar{x}, \bar{y}^s; \xi^s)). 
$$

We use $Y(\bar{x}, \xi^s)$ and $M(\bar{x}, \xi^s)$ to denote the set of all optimal solutions and multipliers satisfying the above condition, respectively. When $M(\bar{x}, \xi^s)$ is nonempty, one may write the critical cone of
the second-stage problem at \( \bar{y}^s \in Y(\bar{x}; \xi^s) \) as

\[
C_2(\bar{y}^s, \xi^s) \triangleq \left\{ d \in \mathbb{R}^{n_2} \middle| \begin{align*}
\nabla_y f(\bar{x}, \bar{y}^s; \xi^s) \top d &= 0, \\
\nabla_y g_j(\bar{x}, \bar{y}^s; \xi^s) \top d \in T_{\mathbb{R}^n}(g_j(\bar{x}, y; \xi)), &j = 1, \ldots, \ell
\end{align*} \right\}.
\]

where \( T_D(x) \) denotes the tangent cone of a closed convex set \( D \). In the following, we show that if the second-stage solutions \( \{\bar{y}^s\}_{s=1}^S \) are unique at the accumulation point \( \bar{x} \) for each \( s \), then \( \bar{x} \) in fact satisfies a stronger condition.

**Corollary 1** (Convergence to a directional stationary point). Let \( \bar{x} \) be an accumulation point of the sequence \( \{x_\nu\}_{\nu \geq 0} \) generated by Algorithm 1. In addition to the assumptions in Theorem 1 if the following conditions are satisfied for each \( s \):

(a) The feasible set \( \{y \in \mathbb{R}^{n_2} \mid G(x,y;\xi^s) \leq 0\} \) is bounded, uniformly for \( x \in X \);

(b) \( f(\cdot, \cdot; \xi^s) \) and \( G(\cdot, \cdot; \xi^s) \) are twice continuously differentiable;

(c) the set of multipliers \( M(\bar{x}, \xi^s) \) is nonempty and there exists \( \bar{y}^s \in Y(\bar{x}, \xi^s) \) satisfying the second order sufficient condition that for all \( d \in C_2(\bar{y}^s, \xi^s) \setminus \{0\},
\]

\[
\sup_{(\mu, \lambda) \in M(\bar{y}^s, \xi^s)} d^\top \nabla^2_y f(\bar{x}, \bar{y}^s; \xi^s) + \sum_{j=1}^\ell \lambda_j g_j(\bar{x}, \bar{y}^s; \xi^s) d > 0;
\]

then \( \bar{x} \) is a directional stationary point of problem (13), i.e.,

\[
\varphi'(\bar{x}; d) + \frac{1}{S} \sum_{s=1}^S \psi'(\bar{x}; d) \geq 0, \quad \forall \ d \in T_X(\bar{x}).
\]

**Proof.** We first prove that \( \partial_2(-\psi)(\bar{x}, \xi^s) \) is a singleton under given assumptions. By condition (a) and the convexity of \( -(\cdot)(\cdot, z) \), we can apply Danskin Theorem [15, Theorem 2.1] to get

\[
\partial_2(-\psi)(\bar{x}, \xi^s) = \text{conv}\left\{ -\nabla_x f(\bar{x}, y; \xi^s) \mid y \in Y(\bar{x}; \xi^s) \right\}.
\]

Since the second order sufficient condition of the second-stage problem holds at \( \bar{x} \) for any \( \xi^s \), we have that \( Y(\bar{x}; \xi^s) \) is a singleton [50, Example 13.25], which further implies that \( \partial_2(-\psi)(\bar{x}, \bar{x}; \xi^s) \) is a singleton. The desired directional stationarity of \( \bar{x} \) then follows from [19, Proposition 6.1.11].

Theorem 1 and Corollary 1 pertain to the subsequential convergence of the iterative sequence generated by Algorithm 1. In the following, we show that the full sequence of the objective values along the iterations converges if the sequence of the Moreau parameters \( \{\gamma_\nu\}_{\nu \geq 0} \) is summable. This result particularly indicates that although the sequence \( \{x_\nu\}_{\nu \geq 0} \) may have multiple accumulation points, the objective values at the accumulation points are the same. To proceed, we remind the readers of the definition of \( \bar{\zeta}_S \) in (17).

**Theorem 2** (Convergence of objective values for Algorithm 1). Suppose that Assumptions A-B hold. Let \( \{x_\nu\}_{\nu \geq 0} \) be the sequence generated by Algorithm 1 under the additional condition that \( \sum_{\nu=0}^\infty \gamma_\nu < \infty \). Then \( \lim_{\nu \to \infty} \bar{\zeta}_S(x_\nu) = \bar{\zeta}_S(\bar{x}) \), where \( \bar{x} \) is any accumulation point of the iterative sequence \( \{x_\nu\}_{\nu \geq 0} \).
Proof. One may derive that
\[
\zeta_S(x_{\nu+1}) - \zeta_S(x_{\nu}) = \left[ \zeta_S(x_{\nu+1}) - \zeta_{S,\gamma_{\nu}}(x_{\nu+1}) \right] \\
+ \left[ \zeta_{S,\gamma_{\nu}}(x_{\nu+1}) - \zeta_{S,\gamma_{\nu}}(x_{\nu}) \right] + \left[ \zeta_{S,\gamma_{\nu}}(x_{\nu}) - \zeta_S(x_{\nu}) \right],
\]
(23)
where the first and last terms on the right side are gaps between the partial Moreau envelopes \(\zeta_{S,\gamma_{\nu}}\) and original functions \(\zeta_S\) at \(x_{\nu+1}\) and \(x_{\nu}\), respectively. By Lemma \[3\] and Assumption B, we may obtain that
\[
\zeta_S(x_{\nu+1}) - \zeta_{S,\gamma_{\nu}}(x_{\nu+1}) \leq \frac{1}{S} \sum_{s=1}^{S} \gamma_{\nu} \kappa_1(\xi^s)^2/2 \quad \text{and} \quad \zeta_{S,\gamma_{\nu}}(x_{\nu}) - \zeta_S(x_{\nu}) \leq 0.
\]
(24)
Recall that \(x_{\nu+1} = x_{\nu,i_{\nu}+1}\) and \(x_{\nu} = x_{\nu,0}\). Then the second term on the right side of (23) can be bounded above based on Theorem \[2\](a) that
\[
\zeta_{S,\gamma_{\nu}}(x_{\nu+1}) - \zeta_{S,\gamma_{\nu}}(x_{\nu}) \leq -\frac{1}{2\gamma_{\nu}} \sum_{i=0}^{i_{\nu}} \|x_{\nu,i+1} - x_{\nu,i}\|^2.
\]
(25)
Substituting (24) and (25) into the inequality (23), we have
\[
\zeta_S(x_{\nu+1}) - \zeta_S(x_{\nu}) \leq -\frac{1}{2\gamma_{\nu}} \sum_{i=0}^{i_{\nu}} \|x_{\nu,i+1} - x_{\nu,i}\|^2 + \frac{\gamma_{\nu}}{2} \left[ \frac{1}{S} \sum_{s=1}^{S} \kappa_1(\xi^s)^2 \right].
\]
Since the sequence \(\{\zeta_S(x_{\nu})\}_{\nu \geq 0}\) must be bounded below due to Assumption A and \(\sum_{\nu=1}^{\infty} \gamma_{\nu} < \infty\), one may easily obtain the convergence of \(\{\zeta_S(x_{\nu})\}_{\nu \geq 0}\) that is a so-called quasi-Fejér monotone sequence; see, e.g., [17, Lemma 3.1]. For any convergent subsequence \(\{x_{\nu+1}\}_{\nu \in \alpha}\) and its limit \(\bar{x}\), we have \(\lim_{\nu(\in \alpha) \to \infty} \zeta_S(x_{\nu+1}) = \zeta_S(\bar{x})\) by the continuity of \(\psi(x;\xi^s)\) on \(X\) for each \(s\). Therefore, the full sequence \(\{\zeta_S(x_{\nu})\}_{\nu \geq 0}\) converges to \(\zeta_S(\bar{x})\) for any accumulation point \(\bar{x}\). 

5 A sampling-based decomposition algorithm

In this section, we consider a generally distributed random vector \(\xi\) with a known distribution. Instead of the approach in the previous section that deals with a fixed batch of samples throughout the algorithm, we incorporate the sampling strategy into the outer loop to progressively enlarge the problem size. In general, there are two ways to do the sampling for solving SPs. One is to use the sample average approximation to select a subset of data before the execution of the numerical algorithm [54, 55, 33]. The other is to adopt a sequential sampling technique [31, 32, 41, 51] where scenarios are gradually added along the iterations. Our method falls into the latter category.
Algorithm 2  A sampling-based decomposition algorithm for the SP  (1)

**Input:** Initial point  \( x_0 \in X \), two positive scalar sequences \( \{ \gamma_\nu \}_\nu \downarrow 0 \), \( \{ \varepsilon_\nu \}_\nu \downarrow 0 \), and a sequence of incremental sample size \( \{ S_\nu \}_\nu \).

**Outer loop:** Set  \( S_{-1} = 0 \) and  \( \nu = 0 \).

1: Generate i.i.d. samples \( \{ (\xi^{S_{\nu-1}+\Delta})_{S_{\nu-1}} \}_{\Delta=1} \) from the distribution of  \( \xi \) that are independent of previous samples.

2: Execute the inner loop of Algorithm 1 with the initial point  \( x_\nu \), samples \( \{ (\xi^s)_{S_{\nu}} \} \triangleq \{ (\xi^{S_{\nu-1}+\Delta})_{S_{\nu-1}} \} \), and parameters  \( \gamma_\nu \) and  \( \varepsilon_\nu \).

3: Set  \( \nu \leftarrow \nu + 1 \) and repeat step 1 until a prescribed stopping criterion is satisfied.

We rely on the LLN for convex subdifferentials to establish the almost surely convergence of the iterative sequence generated by Algorithm 2. To facilitate this tool, Lipschitz continuity of the original function is needed. In our setting, this requires that the lifted function \( \psi^{\bullet} \) is Lipschitz continuous. In addition to Assumption B on the Lipschitz continuity of \( \psi^{\bullet} \), we further pose the following assumption on the Lipschitz continuity of \( \psi(x, x; \xi) \).

**Assumption C.** There exists a measurable function  \( \kappa_2 : \Xi \to \mathbb{R}_+ \) such that  \( \mathbb{E} [\kappa_2(\xi)] < \infty \) and the following inequality holds for almost any  \( \xi \in \Xi \):

\[
|\psi(x, x_1; \xi) - \psi(x, x_2; \xi)| \leq \kappa_2(\xi) \|x_1 - x_2\|, \quad \forall x_1, x_2 \in X.
\]

We have the following result on the LLN for the subdifferentials of icc functions, which is a consequence of the LLN for random set-valued mappings [2]. In fact, the result can be viewed as a pointwise version of [56, Theorem 2].

**Lemma 6.** Suppose that Assumptions A-C hold. For any  \( x \in X \) and any  \( r > r' \geq 0 \) such that  \( B(x, r) \subseteq \text{int}(X) \), the following limit holds almost surely:

\[
\lim_{\nu \to \infty} \mathbb{D} \left( \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \left[ \frac{\partial_1 \psi(\xi^s; x, x; \xi)}{-\partial_2 \psi(\xi^s; x, x; \xi)} \right] - \mathbb{E}_\xi \left[ \frac{\partial_1 \psi(\xi; x, x; \xi)}{-\partial_2 \psi(\xi; x, x; \xi)} \right] \right) = 0,
\]

where the expectation of a random set-valued mapping  \( \mathbb{E}_\xi \left[ a(x; \tilde{\xi}) \right] \) is defined as the set of  \( \mathbb{E}_\xi \left[ a(x; \tilde{\xi}) \right] \) for all measurable selections  \( a(x; \xi) \in A(x; \xi) \).

A noteworthy remark about the preceding lemma is that we can interchange the partial subdifferential and expectation in the right-hand side of the distance, i.e.,

\[
\mathbb{E}_\xi \left[ \partial_1 \psi(x, \tilde{\xi}; x) \right] = \partial_1 \mathbb{E}_\xi \left[ \psi(x, \tilde{\xi}; x) \right].
\]

This is because a convex function is Clarke regular and the Clarke regularity ensures the interchangeability of the subdifferential and the expectation [16, Proposition 2.3.6 and Theorem 2.7.2]. Next we prove a technical lemma that is similar to Lemma 4.
Lemma 7. Suppose Assumptions A-C hold and \( \kappa_1(\xi) \) in Assumption B is essentially bounded, i.e.,
\[ \inf \{ t \mid \mathbb{P}\{ \xi \in \Xi \mid |\kappa_1(\xi)| > t \} = 0 \} \]
is finite. If a subsequence \( \{x_{\nu+1}\}_{\nu \in \alpha} \) generated by Algorithm 2 converges to a point \( \bar{x} \),
then the subsequence \( \{x_{\nu,i}\}_{\nu \in \alpha} \) also converges to \( \bar{x} \),
and for any \( r > 0 \), the subsequence \( \{x_{\nu,i}^s\}_{s=1}^{S_{\nu}} \subset \mathcal{B}(\bar{x}, r) \)
almost surely for all \( \nu \in \alpha \) sufficiently large.

Proof. Following similar derivation in the proof of Lemma 4, we obtain the convergence of \( \{x_{\nu,i}\}_{\nu \in \alpha} \)
to \( \bar{x} \) and the following inequalities
\[
\|x_{\nu,i}^s - \bar{x}\| \leq \|x_{\nu,i}^s - x_{\nu,i}\| + \|x_{\nu,i} - \bar{x}\|
\leq 2\|x_{\nu,i} - \bar{x}\| + \sqrt{2\gamma_{\nu}\left(\bar{\psi}(\bar{x}, x_{\nu,i}; \xi^*) - \inf_{x \in \mathcal{X}} \bar{\psi}(x, x_{\nu,i}; \xi^*)\right)}
\leq 2\|x_{\nu,i} - \bar{x}\| + \sqrt{2\gamma_{\nu} \kappa_1(\xi^*) R(\mathcal{X})}.
\]
Notice that \( \gamma_{\nu} \to 0 \) and \( \kappa_1(\xi^*) \) is almost surely bounded by a constant independent of \( s \).
Thus, for any given \( r > 0 \), there is a positive integer \( N \) such that \( \|x_{\nu,i}^s - \bar{x}\| \leq r \) holds
for any \( s = 1, 2, \ldots, S_{\nu} \).

Below is the main theorem of this section on the almost surely subsequential convergence of
the iterative sequence generated by Algorithm 2.

Theorem 3 (Subsequential convergence of Algorithm 2). Suppose that Assumptions A-C hold and
\( \kappa_1(\xi) \) in Assumption B is essentially bounded. Let \( \{x_{\nu}\}_{\nu \geq 0} \) be the sequence generated by Algorithm 2
and \( \bar{x} \) be any accumulation point. For any \( r > 0 \) such that \( \mathcal{B}(\bar{x}, r) \subseteq \text{int}(\mathcal{X}) \),
the following inclusion holds almost surely:
\[
0 \in \partial \varphi(\bar{x}) + \partial_1^r \mathbb{E}_\xi \left[ \bar{\psi}(\bar{x}, \bar{x}; \xi) \right] - \partial_2^r \mathbb{E}_\xi \left[ (-\bar{\psi})(\bar{x}, \bar{x}; \xi) \right] + \mathcal{N}_X(\bar{x}).
\]
In addition, if the set-valued mapping \( \partial_{1} \mathbb{E}_\xi \left[ \bar{\psi}(\cdot, \cdot; \xi) \right] - \partial_{2} \mathbb{E}_\xi \left[ (-\bar{\psi})(\cdot, \cdot; \xi) \right] \)
is continuous at \( (\bar{x}, \bar{x}) \),
then almost surely
\[
0 \in \partial \varphi(\bar{x}) + \partial_{1} \mathbb{E}_\xi \left[ \bar{\psi}(\bar{x}, \bar{x}; \xi) \right] - \partial_{2} \mathbb{E}_\xi \left[ (-\bar{\psi})(\bar{x}, \bar{x}; \xi) \right] + \mathcal{N}_X(\bar{x}),
\]
i.e., every accumulation point \( \bar{x} \) is a critical point of problem 1 almost surely.

Proof. Consider any subsequence \( \{x_{\nu+1}\}_{\nu \in \alpha} \) that converges to \( \bar{x} \).
It follows from Lemma 7 that for any \( r' > 0 \), the ball \( \mathcal{B}(\bar{x}, r') \) almost surely contains
\( x_{\nu,i}, x_{\nu,i+1} \), and the proximal points \( \{x_{\nu,i}^s\}_{s=1}^{S_{\nu}} \)
for all \( \nu \in \alpha \) sufficiently large. Consequently,
\[
\partial_1 \bar{\psi}(x_{\nu,i}, x_{\nu,i}; \xi^*) \subseteq \partial_{1}^r \bar{\psi}(\bar{x}, \bar{x}; \xi^*) \text{ almost surely for all } \nu \in \alpha \text{ sufficiently large.}
\]
We then obtain from Lemma 6 that, for any \( r > 0 \) such that \( \mathcal{B}(\bar{x}, r) \subseteq \text{int}(\mathcal{X}) \),
the following limit holds almost surely:
\[
\lim_{\nu \to \infty} \mathbb{D} \left( \frac{1}{S_{\nu}} \sum_{s=1}^{S_{\nu}} \left[ \partial_1 \bar{\psi}(x_{\nu,i}^s, x_{\nu,i}; \xi^*) - \partial_2 (-\bar{\psi})(x_{\nu,i}^s, x_{\nu,i}; \xi^*) \right] \right) = 0. \tag{26}
\]
Thus, the following estimation follows almost surely:

\[
\begin{align*}
\text{dist} \left(0, \partial \varphi(\bar{x}) + \partial \varphi(\bar{x}) \right)
&\leq \text{dist} \left(0, \frac{1}{S'} \sum_{s=1}^{S'} \left[ \partial \omega(x^s_{\nu,i}, x_{\nu,i}; \xi^s) - \partial \omega(-\bar{x}) (x^s_{\nu,i}, x_{\nu,i}; \xi^s) + N_X(x^s_{\nu,i}) \right] + \partial \varphi(x_{\nu,i}+1) + w_{\nu} \right) \\
&\quad + \mathbb{D} \left( \frac{1}{S'} \sum_{s=1}^{S'} \left[ \partial \omega(x^s_{\nu,i}, x_{\nu,i}; \xi^s) - \partial \omega(-\bar{x}) (x^s_{\nu,i}, x_{\nu,i}; \xi^s) \right] + \varphi(x_{\nu,i}+1) \right) \\
&\quad + \mathbb{D} \left( \varphi(x_{\nu,i}+1), \varphi(x_{\nu,i}) \right) + \text{dist} (w_{\nu}, N_X(\bar{x})),
\end{align*}
\]

where \(w_{\nu}\) can be any element in \(N_X(x_{\nu,i}+1)\). By Theorem \(2\) (with sample size \(S'\) instead of \(S\)), there is a sequence \(\{w_{\nu}\}_{\nu \geq 0}\) with \(w_{\nu} \in N_X(x_{\nu,i}+1)\) such that \((i')\) converges to 0. As shown in \(26\), we have \((ii') \to 0\). The term \((iii') \to 0\) because \(\{x^s_{\nu,i}\}_{s=1}^{S'} \subseteq \mathbb{B}(\bar{x}, r) \subseteq \text{int}(X)\) holds almost surely for sufficiently large \(\nu\). The convergence of the last two terms \((iv')\) and \((v')\) to 0 can be derived based on similar arguments to their counterparts in the proof of Theorem \(4\). Finally, if the set-valued mapping \(\partial \omega \in \mathbb{E}_{\xi} - \partial \omega (-\bar{x}; \bar{\xi})\) is continuous at \((\bar{x}, \bar{x})\), one may adopt similar arguments as the proof of \(56\) Theorem 3 to derive the almost surely convergence to a critical point.

One can further derive an analogous result of Corollary \(1\) for the sequence generated by Algorithm \(2\) by strengthening the conditions \((a), (b), (c)\) in the former corollary to almost any \(\xi \in \Xi\) so that \(\partial \omega_{\mu,\nu} (-\bar{x}; \bar{\xi})\) is a singleton. We omit the details here for brevity. The last result of this section is the almost surely convergence of the objective values of \(\tilde{\zeta}_{S_{\nu}}(x_{\nu,i})\) under proper assumptions on the sample sizes \(S_{\nu}\) and Moreau parameters \(\gamma_{\nu}\). To proceed, we first present a lemma on the convergence rate of the SAA in expectation. This result is obtained by using the Rademacher average of the random function \(\psi(x; \bar{\xi})\), which has its source in \(21\) Corollary 3.2; see also \(19\) Theorem 10.1.5.

Lemma 8. Let \(X\) be a compact set in \(\mathbb{R}^n\) and \(\tilde{\zeta}: \Omega \to \Xi \subseteq \mathbb{R}^m\) be a random vector defined on a probability space \((\Omega, F, \mathbb{P})\). Let \(\psi: X \times \Xi \to \mathbb{R}\) be a Carathéodory function. Suppose that \(\psi\) is uniformly bounded on \(X \times \Xi\) and Lipschitz continuous in \(x\) with modulus independent of \(\xi\). Let \(\{\xi^s\}_{s=1}^{S}\) be independent and identically distributed random vectors following the distribution of \(\tilde{\zeta}\). Then there exists a constant \(C\) such that for any \(\eta \in (0, 1/2)\), we have

\[
\mathbb{E} \left[ \sup_{x \in X} \left| \frac{1}{S} \sum_{s=1}^{S} \psi(x; \xi^s) - \mathbb{E}_{\xi} \left[ \psi(x; \bar{\xi}) \right] \right| \right] \leq \frac{C \sqrt{1-2\eta} \nu}{S \eta}, \quad \forall S > 0.
\]
We make a remark about the above lemma. For an icc function $\psi(x; \xi)$ associated with the lifted counterpart $\overline{\psi}(x, z; \xi)$, the uniform Lipschitz continuity of $\overline{\psi}(\cdot; \xi)$ holds on $X$ when $\overline{\psi}(\cdot, \cdot; \xi)$ is uniformly Lipschitz continuous over $(z; \xi) \in X \times \Xi$ and $\overline{\psi}(x, \cdot; \xi)$ is uniformly Lipschitz continuous over $(x; \xi) \in X \times \Xi$. Indeed, one can deduce this uniform Lipschitz continuity from Assumptions B and C with $\sup_{\xi \in \Xi} \left[ \max(\kappa_1(\xi), \kappa_2(\xi)) \right] < \infty$, which also implies the essential boundedness of $\kappa_1(\xi)$ assumed in Lemma 4 and Theorem 3. We are now ready to present the almost surely sequential convergence of the objective values generated by the internal sampling scheme.

**Theorem 4 (Sequential convergence of objective values for Algorithm 2).** Suppose that Assumptions A-C and conditions for the function $\psi$ in Lemma 8 hold. Let $\{x_\nu\}_{\nu \geq 0}$ be the sequence generated by Algorithm 2. Assume that the parameter of the partial Moreau envelope $\gamma_\nu$ and the sample size $S_\nu$ satisfy

$$\sum_{\nu=1}^{\infty} \gamma_\nu < \infty, \quad \sum_{\nu=1}^{\infty} \frac{S_{\nu+1} - S_{\nu}}{S_{\nu+1}(S_{\nu})^\eta} < \infty \quad \text{for some } \eta \in (0, 1/2).$$

Then $\lim_{\nu \to \infty} \zeta_{S_\nu}(x_\nu) = \zeta(\bar{x})$ almost surely, where $\bar{x}$ is any accumulation point of the iterative sequence $\{x_\nu\}_{\nu \geq 0}$.

**Proof.** We first prove the almost sure convergence of $\{\zeta_{S_\nu}(x_\nu)\}_{\nu \geq 0}$. We have

$$\zeta_{S_{\nu+1}}(x_{\nu+1}) - \zeta_{S_\nu}(x_\nu) = R_{\nu, 1} + R_{\nu, 2} + R_{\nu, 3} + R_{\nu, 4},$$

where

$$R_{\nu, 1} = \sum_{s=1}^{S_\nu} \frac{\gamma_\nu \kappa_1(\xi_s)^2}{2}, \quad R_{\nu, 3} = -\frac{1}{2\gamma_\nu} \sum_{i=0}^{i_\nu} \|x_{\nu,i+1} - x_{\nu,i}\|^2 \quad \text{and} \quad R_{\nu, 4} \leq 0.$$

Next we compute $R_{\nu, 1}$ that is the error of the sample-augmentation. It holds that

$$R_{\nu, 1} = \frac{1}{S_{\nu+1}} \left[ \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi_s) + \sum_{s=S_{\nu+1}}^{S_\nu+1} \psi(x_{\nu+1}; \xi_s) - \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi_s) \right] = \frac{1}{S_{\nu+1}} \sum_{s=1}^{S_\nu} \left[ \frac{S_\nu}{S_{\nu+1}} \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi_s) - \frac{1}{S_{\nu+1}} \sum_{s=S_{\nu+1}}^{S_{\nu+1}} \psi(x_{\nu+1}; \xi_s) \right].$$

Let $F_\nu \triangleq \sigma(\xi^1, \xi^2, \ldots, \xi^{S_\nu})$ be a filtration, i.e., an increasing sequence of $\sigma$-fields generated by samples used in outer iterations. Obviously $x_{\nu+1}$ is adapted to $F_\nu$ and $\{\xi_s\}_{s=S_{\nu+1}}^{S_{\nu+1}}$ are independent
of $F_\nu$. Therefore, by taking conditional expectation of $R_{\nu,1}$ given $F_\nu$, we obtain

$$
E[R_{\nu,1} \mid F_\nu] = \left( \frac{S_\nu}{S_{\nu+1}} - 1 \right) \left[ \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi^s) - E_\tilde{\xi} \left[ \psi(x_{\nu+1}; \tilde{\xi}) \right] \right],
$$

where $\{\xi^s\}_{s=1}^{S_\nu}$ and $\tilde{\xi}$ are independent and identically distributed. Based on the estimations of the terms $R_{\nu,1}, R_{\nu,2}, R_{\nu,3}$ and $R_{\nu,4}$, we have, by taking conditional expectation of (27) given $F_\nu$,

$$
E \left[ \tilde{\zeta}_{S_{\nu+1}}(x_{\nu+1}) \mid F_\nu \right] - \tilde{\zeta}_{S_\nu}(x_\nu) - \frac{1}{2\gamma_\nu} \sum_{i=0}^{i_\nu} \|x_{\nu,i+1} - x_{\nu,i}\|^2 
\leq \frac{S_{\nu+1} - S_\nu}{S_{\nu+1}} \left| \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi^s) - E_\tilde{\xi} \left[ \psi(x_{\nu+1}; \tilde{\xi}) \right] \right| + \frac{1}{2\gamma_\nu} \sum_{s=1}^{S_\nu} \gamma_\nu \kappa_1(\xi^s)^2.
$$

(28)

In order to show the almost sure convergence of $\{\tilde{\zeta}_{S_\nu}(x_\nu)\}_{\nu \geq 0}$, we need to verify that the right side of the preceding inequality is summable over $\nu$ almost surely and the sequence $\{\tilde{\zeta}_{S_\nu}(x_\nu)\}_{\nu \geq 0}$ is bounded below almost surely. We have

$$
E \left[ \sum_{\nu=1}^{\infty} \left( \frac{S_{\nu+1} - S_\nu}{S_{\nu+1}} \right) \left| \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi^s) - E_\tilde{\xi} \left[ \psi(x_{\nu+1}; \tilde{\xi}) \right] \right| \right] 
\leq \sum_{\nu=1}^{\infty} \frac{S_{\nu+1} - S_\nu}{S_{\nu+1}} E \left[ \sup_{x \in X} \left| \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \psi(x; \xi^s) - E_\tilde{\xi} \left[ \psi(x; \tilde{\xi}) \right] \right| \right] 
\leq \sum_{\nu=1}^{\infty} \frac{(S_{\nu+1} - S_\nu) C \sqrt{1 - 2\eta}}{S_{\nu+1} (S_\nu)^\eta} < \infty, \text{ for some } \eta \in (0, 1/2), \text{ by Lemma 8.}
$$

Hence, we derive that

$$
\sum_{\nu=1}^{\infty} \left( \frac{S_{\nu+1} - S_\nu}{S_{\nu+1}} \right) \left| \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \psi(x_{\nu+1}; \xi^s) - E_\tilde{\xi} \left[ \psi(x_{\nu+1}; \tilde{\xi}) \right] \right| < \infty \text{ almost surely.}
$$

Since $\{\gamma_\nu\}_{\nu=0}^{\infty}$ is assumed to be summable, we can obtain

$$
E \left[ \sum_{\nu=0}^{\infty} \left( \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \gamma_\nu \kappa_1(\xi^s)^2 \right) \right] = \sum_{\nu=0}^{\infty} E \left[ \frac{\gamma_\nu}{S_\nu} \sum_{s=1}^{S_\nu} \kappa_1(\xi^s)^2 \right] = \sum_{\nu=0}^{\infty} \gamma_\nu E_\tilde{\xi} \left[ \kappa_1(\tilde{\xi})^2 \right] < \infty.
$$

Consequently, $\sum_{\nu=1}^{\infty} \left( \frac{1}{S_\nu} \sum_{s=1}^{S_\nu} \gamma_\nu \kappa_1(\xi^s)^2 \right) < \infty$ almost surely. We have thus proved that the right side of (28) is summable over $\nu$ almost surely. Next, we show that $\{\tilde{\zeta}_{S_\nu}(x_\nu)\}_{\nu \geq 0}$ is bounded below almost surely. To see this, note that

$$
\sup_{x \in X} |\tilde{\zeta}_{S_\nu}(x)| \leq \sup_{x \in X} |\tilde{\zeta}_{S_\nu}(x) - \zeta(x)| + \sup_{x \in X} |\zeta(x)|,
$$

25
where the first term converges to 0 almost surely by the uniform LLN (c.f. [55] Theorem 9.60) and the second one is bounded due to the continuity of $\zeta(x) = \varphi(x) + \mathbb{E}_\xi [\psi(x; \xi)]$ on the compact set $X$. Therefore, there exists a constant $M$ such that $\zeta_{S_\nu}(x_\nu)$ is bounded below by $M$ almost surely for any $\nu$. Applying Robbins-Siegmund nonnegative almost supermartingale convergence lemma (c.f. [45] Theorem 1), we have $\sum_{\nu=1}^{\infty} \frac{1}{2^{\gamma_\nu}} \sum_{i=0}^{c_{\nu,i}} \|x_{\nu,i+1} - x_{\nu,i}\|^2 < \infty$ almost surely and the sequence $\{\zeta_{S_\nu}(x_\nu)\}_{\nu \geq 0}$ converges almost surely. Finally, let $\bar{x}$ be the limit of a convergent subsequence $\{x_\nu\}_{\nu \in \alpha}$. Using the uniform convergence of $\zeta_{S_\nu}$ to $\zeta$ and the continuity of $\zeta$ on the compact set $X$, it follows from [55] Proposition 5.1 that $\zeta_{S_\nu}(x_\nu)$ converges to $\zeta(\bar{x})$ almost surely as $\nu(\in \alpha) \to \infty$. This argument, together with the convergence of the full sequence $\{\zeta_{S_\nu}(x_\nu)\}_{\nu \geq 0}$, completes the proof of this theorem.

6 Numerical experiments

In this section, we present numerical results for a power system planning problem with the recourse in [4] and a linear first-stage objective of $x = \{\{x_i\}_{i \in I}, \{x_g\}_{g \in G}\}$. The overall deterministic equivalent formulation to minimize the total cost is

$$
\begin{align*}
\text{minimize} & \quad \sum_{i \in I} c_i x_i + \sum_{g \in G} c_g x_g + \sum_{s=1}^S \left( \sum_{g \in G} p_{sg} x_g \right) \left( \sum_{i \in I} \sum_{j \in J} \left( q_{is} - \pi_{js} \right) y_{ij} \right) \\
\text{subject to} & \quad \sum_{i \in I} c_i x_i + \sum_{g \in G} c_g x_g \leq B \text{ (budget constraint)}, \quad \sum_{g \in G} x_g = 1, \\
& \quad \sum_{j \in J} y_{ij} \leq x_i, \quad i \in I, \ s = 1, \ldots, S \text{ (capacity constraints),} \\
& \quad \sum_{i \in I} y_{ij} = d_{js}, \quad j \in J, \ s = 1, \ldots, S \text{ (demand constraints).}
\end{align*}
$$

(29)

In our experiments, we set $|I| = |G| = 5$ and $|J| = 8$. The box constraints of $x$ and $z_s$ are $[8, 15]^5 \times [0, 1]^5$ and $[0, 5]^{5 \times 8}$ for each $s = 1, 2, \ldots, S$. The unit costs in the first-stage $\{c_i\}_{i \in I} \cup \{c_g\}_{g \in G}$ are independently generated from a uniform distribution on $[0, 5]$. For each scenario, $\{q_{is}\}_{i \in I}, \{\pi_{js}\}_{j \in J}$ and $\{d_{js}\}_{j \in J}$ are generated from truncated normal distributions $\mathcal{N}(1, 5^2)$ on $[2, 4], [3, 5]$ and $[2, 5]$, respectively. To construct a set of probabilities $\cup_{s=1}^{S} \{p_{sg}\}_{g \in G}$ satisfying $\sum_{s=1}^{S} p_{sg} = 1$, we first randomly generate $S \times |G|$ values from uniform distributions on $[0, 1]$, and then group every $|G|$ values and normalizing them such that the sum of values in each group is 1. All the experiments are conducted in MATLAB 2021b on a Intel Xeon workstation with sixteen 3.70 GHz processors and 128 GB of RAM.

6.1 Fixed scenarios

Since the text example with fixed scenarios is in fact a large-scale nonconvex quadratic problem, it can also be directly solved by off-the-shelf nonlinear programming solvers. We compare the performance of our proposed decomposition algorithm based on the partial Moreau envelope (DPME) with the interior-point-based solvers Knitro [12] and IPOPT [59], both of which run with linear

26
solver MUMPS 5.4.1. The absolute and relative feasibility and optimality errors are computed according to the termination criteria of Knitro. The quantities $KKT_{\text{abs}}$ and $KKT_{\text{rel}}$ are defined as the max of absolute and relative feasibility and optimality errors, respectively. The initial points are chosen to be the same for all algorithms. Although this may not necessarily force all algorithms to converge to the same objective values, we do observe such a phenomenon in the experiments. Further implementation details of these algorithms are provided below.

**Knitro** (version 13.0.0): “knitro_qp” function is called in our numerical experiments to solve non-convex quadratic programs from the MATLAB environment. We set “hessopt = 0” to compute the exact Hessian in the interior point method instead of using the (L)BFGS approximations, as we have observed that the former choice is faster for all the problems tested here. We directly set “convex = 0” to declare our problems are nonconvex so that the solver does not need to spend time on checking the convexity of the problems. We report the results based on three different settings:

1. Knitro-direct: Set “algorithm = 1” so that the direct solver is used to solve linear equations. For the termination options, the KKT relative and absolute tolerance is set to be $10^{-4}$ and $10^{-2}$, respectively, i.e., “feastol = opttol = $10^{-4}$” and “feastol,abs = opttol,abs = $10^{-2}$”.

2. Knitro-CG-1: Set “algorithm = 2” so that the KKT system is solved using a projected conjugate gradient method. Stopping criteria are the same as above.

3. Knitro-CG-2: All are the same as Knitro-CG-1 except that the KKT relative and absolute tolerance is set to be $10^{-6}$ and $10^{-3}$, respectively.

**IPOPT** (version 3.14.4): Due to different scaling strategies and reformulations, the termination criteria of Knitro and IPOPT are not directly comparable. We set “ipopt.tol = $5 \times 10^{-2}$” in our experiments as we find the computed solutions based on this tolerance are about the same quality as those provided by Knitro. We also set “ipopt.hessian_constant = ‘yes’” to use exact Hessian in the interior point method, and have not adopted the (L)BFGS method for the same reason as mentioned above.

**DPME**: Each master problem for the first stage and the subproblem of the second stage are convex quadratic programs, which we have called Gurobi to solve. We compute the absolute (denoted as $\text{Feas}_{\text{abs},\nu}$) and relative (denoted as $\text{Feas}_{\text{rel},\nu}$) feasibility errors using the same way as Knitro. Let the overall objective value at the $\nu$-th outer loop be $\text{Obj}_{\nu}$. We terminate our algorithm if all of the conditions below are satisfied and then compute the absolute and relative KKT errors (as defined in Knitro) at the last iteration point:

$$\text{Feas}_{\text{abs},\nu} \leq 10^{-2}, \quad \text{Feas}_{\text{rel},\nu} \leq 10^{-4} \quad \text{and} \quad \frac{|\text{Obj}_{\nu-1} - \text{Obj}_{\nu}|}{\max\{1, |\text{Obj}_{\nu-1}|\}} \leq 10^{-4}. \quad (30)$$

Table 2 and Figure 3 summarize the performance of different algorithms when the number of scenarios $S$ varies from 1,000 to 120,000 over 100 independent replications (the sizes of the deterministic equivalent problems are listed in Table 1). For each algorithm, we report the mean and the standard deviation of the total iteration numbers, the absolute and relative KKT errors, objective values, and the wall-clock time. We also conduct experiments for $S = 500,000$ over 10

---

1Knitro user guide: https://www.artelys.com/docs/knitro/2_userGuide/termination.html
independent replications to demonstrate the scalability of our decomposition algorithm and put the results in Table 2. One may find that for small-sized problems (such as when \( S < 10,000 \)), the interior point method that is implemented by both Knitro and IPOPT can solve the problem faster than our DPME, which may be due to two reasons: one is that the gain of the decomposition cannot compensate for the overhead of the communication between the master problem and the subproblems; the other is that we have not used the second order information as in the interior point method. However, for the cases where \( S \) is large (such as \( S \geq 10,000 \)), the DPME is the fastest method and attains the smallest average objective values. While the computational time of IPOPT and Knitro-direct quickly exceeds the preset limit, the computational time of the DPME scales approximately linearly in terms of the number of scenarios. Although the Knitro-CG-1 can solve all the problems within the time limit, it cannot produce solutions that have similar objective values with other methods. When we switch the solver to the Knitro-CG-2 setting, it could provide solutions with similar objective values as DPME but needs significantly longer computational time. Therefore, we conclude that the DPME for fixed scenarios can significantly reduce the computational time for solving large-scale nonconvex two-stage SPs.

### 6.2 Sampling-based decomposition

We test the sampling-based DPME proposed in Algorithm 2 for the same test problem with the total number of scenarios \( S = 50,000 \). Instead of using all scenarios at each iteration, we gradually add them to reduce the computational cost especially at the early stage. In our experiments, the sample size \( S_\nu \) is taken as \( \eta \nu \) for different positive parameter \( \eta \). In order to understand how the growth rate \( \eta \) depends on the distributions of the random scenarios, we generate \( \{q_{ia}\}_{i \in I}, \{\pi_{ja}\}_{j \in J} \) and \( \{d_{ja}\}_{j \in J} \) from truncated normal distributions \( \mathcal{N}(1, \sigma^2) \) on \([2, 13],[3, 10]\) and \([2, 5]\) with \( \sigma \in \{0.5, 2, 5\} \), and varies the values of \( \eta \).

The stopping criteria for the sampling-based DPME is the same as (30), where we check the violation of the KKT system for the deterministic equivalent problem formulated by all scenarios \( S = 50,000 \). However, unlike the case for fixed scenarios, we do not have all the second-stage solutions \( \{y_s\}_{s=1}^S \) to compute the KKT residual since some samples may not have been used yet. To resolve this issue, we compute all \( \{y_s\}_{s=1}^S \) at every \( \nu \)-th outer iteration, and then estimate the multipliers corresponding to the first-stage budget and box constraints by minimizing the current KKT residual.
| S  | iterations | $\eta$ | $\sigma$ | $\kappa$ | $\kappa_T$ | $\kappa_T$ | $\eta$ | $\sigma$ |
|---|---|---|---|---|---|---|---|---|
| 1,000 | 8 | 17 | 28 | 29 | 12 | 1.6-3 | 3.0-4 | 4.0-6 |
| 5,000 | 9 | 16 | 42 | 31 | 11 | 1.6-3 | 1.1-3 | 1.5-6 |
| 10,000 | 10 | 16 | 43 | 32 | 11 | 1.6-3 | 2.4-6 | 3.0-6 |
| 30,000 | 10 | 16 | 47 | 35 | 11 | 1.6-3 | 3.0-6 | 4.0-6 |
| 80,000 | 11 | 14 | 46 | 35 | 11 | 1.6-3 | 3.8-1 | 5.0-6 |
| 120,000 | 11 | 11 | 14 | 12 | 8.3-1 | 5.0-6 | 4.0-4 |
| 500,000 | 14 | 12 | 11 | 15 | 2.4-2 | 1.5-3 | 5.0-6 |

Table 3: The performance of Knitro-direct, Knitro-CG, IPOPT, and DPME. In the table, “S” is the number of scenarios; the numbers without parentheses are the means over 100 replications and the numbers in parentheses are the standard deviations; “t” means the method exceeds the preset time limit, which is 1,800s for $S \leq 120,000$ and 18,000s for $S = 500,000$; “a” stands for Knitro-direct; “b” stands for Knitro-CG-1; “c” stands for Knitro-CG-2; “d” stands for IPOPT and “e” stands for DPME.
In Table 3, we summarize the performance of Algorithm 2 for different combinations of \((\sigma, \eta)\), where we also provide the results obtained from Algorithm 1 without sampling for benchmarks. It can be observed from the table that problems with larger variability may need faster growth rate of the batch size to retain the same level of solution quality. If the growth rate is properly chosen, the sampling-based DPME can outperform the fixed-scenario version in the computational time with comparable solution qualities.

7 Conclusion

Compared with the extensive literature on the algorithms for convex (especially linear) two-stage SPs, efficient computational algorithms for solving continuous nonconvex two-stage SPs have been much less explored. In this paper, we have made a first attempt on developing the decomposition scheme for a special class of latter problems. The key of the proposed algorithm is the derivation of successive strongly convex approximations of the nonconvex recourse functions. We hope the work done in paper could stimulate researchers’ interests in a broader paradigm of SPs that goes beyond the classical convex settings. There are a lot of open questions that deserve future investigations, such as how to combine the stochastic dual dynamic programming approach with the tools developed in the current paper to solve nonconvex multistage SPs, as well as how to design rigorous stopping criteria for the general nonconvex SPs with continuous distributions.

References

[1] Le Thi Hoai An and Pham Dinh Tao. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46, 2005.

[2] Zvi Artstein and Richard A Vitale. A strong law of large numbers for random compact sets. *The Annals of Probability*, pages 879–882, 1975.

[3] Edgar Asplund. Differentiability of the metric projection in finite-dimensional Euclidean space. *Proceedings of the American Mathematical Society*, 38(1):218–219, 1973.

[4] Jacques F Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4(1):238–252, 1962.

[5] Dimitri P Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty*. PhD thesis, Massachusetts Institute of Technology, 1971.

[6] John R Birge and Francois Louveaux. *Introduction to Stochastic Programming*. Springer Science & Business Media, 2011.

[7] Merve Bodur, Sanjeeb Dash, Oktay Günlük, and James Luedtke. Strengthened Benders cuts for stochastic integer programs with continuous recourse. *INFORMS Journal on Computing*, 29(1):77–91, 2017.

[8] Natasha Boland, Jeffrey Christiansen, Brian Dandurand, Andrew Eberhard, Jeff Linderoth, James Luedtke, and Fabricio Oliveira. Combining progressive hedging with a Frank–Wolfe
method to compute Lagrangian dual bounds in stochastic mixed-integer programming. *SIAM Journal on Optimization*, 28(2):1312–1336, 2018.

[9] Immanuel M Bomze, Markus Gabl, Francesca Maggioni, and Georg Ch Pflug. Two-stage stochastic standard quadratic optimization. *European Journal of Operational Research*, 299(1):21–34, 2022.

[10] J Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media, 2013.

[11] Pedro Borges, Claudia Sagastizábal, and Mikhail Solodov. A regularized smoothing method for fully parameterized convex problems with applications to convex and nonconvex two-stage stochastic programming. *Mathematical Programming*, pages 1–33, 2020.

[12] Richard H Byrd, Jorge Nocedal, and Richard A Waltz. Knitro: An integrated package for nonlinear optimization. In *Large-scale nonlinear optimization*, pages 35–59. Springer, 2006.

[13] Claus C Carøe and Rüdiger Schultz. Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24(1-2):37–45, 1999.

[14] Bock Jin Chun and Stephen M Robinson. Scenario analysis via bundle decomposition. *Annals of Operations Research*, 56(1):39–63, 1995.

[15] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[16] Frank H Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[17] Patrick L Combettes. Quasi-Fejérian analysis of some optimization algorithms. In *Studies in Computational Mathematics*, volume 8, pages 115–152. Elsevier, 2001.

[18] Kelly J Cormican, David P Morton, and R Kevin Wood. Stochastic network interdiction. *Operations Research*, 46(2):184–197, 1998.

[19] Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. SIAM, 2021.

[20] John M Danskin. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 2012.

[21] Yuri M Ermoliev and Vladimir I Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM Journal on Optimization*, 23(4):2231–2263, 2013.

[22] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[23] Vikas Goel and Ignacio E Grossmann. A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers & Chemical Engineering*, 28(8):1409–1429, 2004.

[24] Vikas Goel and Ignacio E Grossmann. A Lagrangian duality based branch and bound for solving linear stochastic programs with decision dependent uncertainty. In *Computer Aided Chemical Engineering*, volume 20, pages 55–60. Elsevier, 2005.
[25] Vikas Goel and Ignacio E Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical Programming*, 108(2):355–394, 2006.

[26] Monique Guignard. Lagrangean relaxation. *Top*, 11(2):151–200, 2003.

[27] Vincent Guigues. Inexact stochastic mirror descent for two-stage nonlinear stochastic programs. *Mathematical Programming*, 187(1):533–577, 2021.

[28] Vincent Guigues, Renato Monteiro, and Benar Svaiter. Inexact cuts in stochastic dual dynamic programming applied to multistage stochastic nondifferentiable problems. *SIAM Journal on Optimization*, 31(3):2084–2110, 2021.

[29] Tianyu Hao and Jong-Shi Pang. Piecewise affine parameterized value-function based bilevel non-cooperative games. *Mathematical Programming*, 180(1):33–73, 2020.

[30] Lars Hellemo, Paul I Barton, and Asgeir Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018.

[31] Julia L Higle and Suvrajeet Sen. *Stochastic decomposition: a statistical method for large scale stochastic linear programming*, volume 8. Springer Science & Business Media, 2013.

[32] Tito Homem-de Mello. Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation*, 13(2):10–133, 2003.

[33] Tito Homem-de Mello and Güzin Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

[34] Tore W Jonsbråten, Roger J.-B. Wets, and David L Woodruff. A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82:83–106, 1998.

[35] Junyi Liu, Ying Cui, Jong-Shi Pang, and Suvrajeet Sen. Two-stage stochastic programming with linearly bi-parameterized quadratic recourse. *SIAM Journal on Optimization*, 30(3):2530–2558, 2020.

[36] François V Louveaux. Optimal investments for electricity generation: A stochastic model and a test problem. *Numerical Techniques for Stochastic Optimization*, 1988.

[37] Yves Lucet. Fast Moreau envelope computation I: Numerical algorithms. *Numerical Algorithms*, 43(3):235–249, 2006.

[38] Boris S Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330. Springer Science & Business Media, 2006.

[39] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[40] Omid Nohadani and Kartikey Sharma. Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization*, 28(2):1773–1795, 2018.
[41] Raghu Pasupathy and Yongjia Song. Adaptive sequential sample average approximation for solving two-stage stochastic linear programs. *SIAM Journal on Optimization*, 31(1):1017–1048, 2021.

[42] Chayne Planiden and Xianfu Wang. Proximal mappings and Moreau envelopes of single-variable convex piecewise cubic functions and multivariable gauge functions. In *Nonsmooth optimization and its applications*, pages 89–130. Springer, 2019.

[43] Boris T Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

[44] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[45] Herbert Robbins and David Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

[46] R Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.

[47] R Tyrrell Rockafellar. Progressive decoupling of linkages in optimization and variational inequalities with elicitable convexity or monotonicity. *Set-Valued and Variational Analysis*, 27(4):863–893, 2019.

[48] R Tyrrell Rockafellar. Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality. *Mathematical Programming*, 192, 2022.

[49] R Tyrrell Rockafellar and Roger J-B Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147, 1991.

[50] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[51] Johannes Royset and Elijah Polak. Extensions of stochastic optimization results to problems with system failure probability functions. *Journal of Optimization Theory Applications*, 133:1–18, 2007.

[52] Andrzej Ruszczyński. Decomposition methods in stochastic programming. *Mathematical Programming*, 79(1-3):333–353, 1997.

[53] Rüdiger Schultz. Stochastic programming with integer variables. *Mathematical Programming*, 97(1):285–309, 2003.

[54] Alexander Shapiro. Monte Carlo sampling methods. *Handbooks in Operations Research and Management Science*, 10:353–425, 2003.

[55] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.
[56] Alexander Shapiro and Huifu Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *Journal of Mathematical Analysis and Applications*, 325(2):1390–1399, 2007.

[57] Bora Tarhan, Ignacio E Grossmann, and Vikas Goel. Stochastic programming approach for the planning of offshore oil or gas field infrastructure under decision-dependent uncertainty. *Industrial & Engineering Chemistry Research*, 48(6):3078–3097, 2009.

[58] Richard M Van Slyke and R. J.-B. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17(4):638–663, 1969.

[59] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[60] R. J.-B. Wets. Large scale linear programming techniques in stochastic programming. *IIASA Working Paper*, 1984.