Math. Program., Ser. A (2012) 135:293–321
DOI 10.1007/s10107-011-0453-3

FULL LENGTH PAPER

Optimality functions in stochastic programming

J. O. Royset

Received: 2 December 2009 / Accepted: 9 April 2011 / Published online: 11 June 2011

Abstract Optimality functions define stationarity in nonlinear programming, semi-infinite optimization, and optimal control in some sense. In this paper, we consider optimality functions for stochastic programs with nonlinear, possibly nonconvex, expected value objective and constraint functions. We show that an optimality function directly relates to the difference in function values at a candidate point and a local minimizer. We construct confidence intervals for the value of the optimality function at a candidate point and, hence, provide a quantitative measure of solution quality. Based on sample average approximations, we develop an algorithm for classes of stochastic programs that include CVaR-problems and utilize optimality functions to select sample sizes.

Keywords Stochastic programming · Optimality conditions · Validation analysis · Algorithms

Mathematics Subject Classification (2000) 90C15 · 90C46 · 90C26

1 Introduction

Stochastic optimization problems arise in numerous contexts where decisions must be made under uncertainty; see, e.g., [16,19,41,46] for algorithms, models, and applications. In this paper, we specifically deal with problems defined in terms of expected values of random functions. Let $F^j : \mathbb{R}^n \times \Omega \to \mathbb{R}$, $j = 0, 1, 2, \ldots, q$, be random functions defined on a common probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with $\Omega \subset \mathbb{R}^d$ and $\mathcal{F} \subset 2^\Omega$ being the Borel sigma algebra. Moreover, let the expected value functions

J. O. Royset (✉)
Operations Research Department, Naval Postgraduate School, Monterey, CA 93943, USA
e-mail: joroyset@nps.edu

$f^j : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, \infty\}$ be defined by

$$f^j(x) \triangleq E[F^j(x, w)]$$

for all $j \in \mathbf{q}_0 \triangleq \{0\} \cup \mathbf{q}$, with $\mathbf{q} \triangleq \{1, 2, \ldots, q\}$. Problems involving such expected value functions are generally challenging to solve due to the need for estimating expectations repeatedly. Even assessing how "close" a given candidate point $x \in \mathbb{R}^n$ is to optimality or stationarity may be nontrivial. We specifically consider the problem

$$P: \quad \min_{x \in \mathbb{R}^n} \{f^0(x) \mid f^j(x) \leq 0, \, j \in \mathbf{q}\}, \tag{1}$$

where we adopt assumptions as in Theorem 7.52 and p. 146 of [41] that ensure that expectation and gradient operators interchange and $f^j(\cdot)$ are continuously differentiable. However, $f^j(\cdot)$ may be nonconvex. We do allow certain classes of nonsmoothness in $F^j(\cdot, \omega)$, $j \in \mathbf{q}_0$, as described below, which may arise in two-stage stochastic programs [16], investment portfolio optimization [33], inventory control [48], and engineering design [32,35]. Inventory control and engineering design optimization as well as estimation of mixed logit models [3] may result in nonconvex models. Expected value constraints appear, for instance, in investment portfolio and engineering design optimization with restrictions on the Conditional Value-at-Risk (CVaR) (also called superquantile) [32,33]. Throughout the paper, we assume that an infeasible $x \in \mathbb{R}^n$ is meaningful, but undesirable, as often is the case for CVaR-constrained problems. If an infeasible point has little meaning and practical use, a chance-constrained model may be more suitable than P; see for example [20] and [41], Chap. 4. That topic, however, is outside the scope of the paper as in that case $F^j(\cdot, \cdot)$ is an indicator function, which is discontinuous and cannot easily be handled by our framework.

We consider two aspects of P. First, we focus on the assessment of the "quality" of a candidate point $x \in \mathbb{R}^n$, which we refer to as validation analysis. Second, we deal with algorithms that generates such candidate points. We then adopt a more specific assumption that requires $F^j(\cdot, \omega)$ to be given in terms of the maximum of a finite number of smooth random functions.

Stationary points of P are defined by the Karush-Kuhn-Tucker (KKT) or the Fritz-John (FJ) first-order necessary optimality conditions. (Recall that the conditions are equivalent for example under the Slater constraint qualification with convex inequality constraints.) However, the verification of these conditions at a given $x \in \mathbb{R}^n$ in the present context is challenging as it requires estimation of $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$.

Under the assumption of deterministic constraints, Shapiro and Homem-de-Mello [42] develops confidence regions for $\nabla f^0(x)$ as well as hypothesis tests for whether a point $x \in \mathbb{R}^n$ satisfies the KKT conditions; see also [11]. The results in [42] can be extended to constraints defined in terms of expectations [40]. The hypothesis tests require that the gradients of the active constraints are linearly independent, the strict complimentary condition holds at x, and that the inverse of an estimate of a variance-covariance matrix is nonsingular. For P, Bettonvil et al. [6] develops a series of hypothesis tests using bootstrapping for verification of KKT conditions that require

relatively small sample sizes. Other hypothesis tests for KKT conditions are found in [36,37], which also consider equality constraints.

Section 5.2 of [41] (see also [3,9,39]) uses stochastic variational inequalities to analyze optimality conditions for P. The results include conditions for almost sure convergence of stationary points of sample average problems (constructed by replacing expectations in P by their sample averages) to stationary points of P as the sample size grows. Extension of such results to second-order optimality conditions are found in [3]. A similar result for the case with a nonsmooth objective function and deterministic constraints is found in [48]. We find in Sect. 5.2 of [41] that under the linear independence constraint qualification and the strict complementarity condition, a stationary point of a sample average problem with sample size N is approximately normally distributed with mean equal to a stationary point of P and with standard deviation proportional to $N^{-1/2}$.

Another approach to validation analysis is based on estimating bounds on the optimal value of P; see [4,12,21,24]. Estimation of bounds in the case of constraints on expected value functions utilizes the Lagrangian function [41,47], p. 208. These bounding procedures are essentially limited to convex problems as they require global minima of sample average problems, or as they make use of strong duality. Even if global minima can be computed, nonconvex problems may have substantial duality gaps and bounds based on the Lagrangian function may be weak.

There are numerous algorithms for solving stochastic programs similar to P including decomposition algorithms in cases with special structure (see, e.g., [10]), stochastic approximations (see, e.g., [19,22]), other versions of stochastic search (see, e.g., [44]), and various algorithms based on sample average approximations (SAA) (see, e.g., [41]). Since P may involve constraints on nonconvex expected value functions, stochastic approximations may not be applicable and we focus on SAA. The SAA approach solves a sample average problem obtained from P by replacing $\mathcal{P}$ by an empirical distribution based on a sample from $\mathcal{P}$. Under mild assumptions, global minimizers and global minima of sample average problems converge to a global minimizer and a global minimum of P, respectively, as the sample size increases to infinity; see for example [41, Sect. 5.1]. The advantage of this approach is its simplicity and the fact that a large library of deterministic optimization algorithm may be applicable to solve the sample average problem. A more involved version of SAA approximately solves a sequence of sample average problems with gradually larger sample size [3,14,25,35]. This version may reduce the computational effort required to reach a near-optimal solution as early iterations can utilize small sample sizes, but it needs a rule for selecting the sequence of sample sizes [25,30].

In this paper, we propose an optimization algorithm and validation analysis techniques for P based on *optimality functions*. Optimality functions are optimal values of quadratic programs involving linearizations of objective and constraint functions and were introduced by E. Polak for use in nonlinear programming, semi-infinite optimization, and optimal control to characterize stationary points [27–29]. Optimality functions have not been applied previously for validation analysis and algorithm development in stochastic programming. As we see below, the use of optimality functions in the context of P appears promising for three reasons. First, they result in validation analysis procedures that appear more applicable than hypothesis test of

KKT conditions as they deal with the more general FJ conditions and do not require a constraint qualification. Second, they lead to bounds on the distance between the objective function value at a feasible point and a local minimum. Third, they result in sample-size adjustment rules that ensure convergence of algorithms for P based on approximately solving sequences of sample average problems.

The contributions of the paper are fourfold. (1) We introduce an optimality function to the area of stochastic programming and establish the properties of its estimator. (2) We derive bounds in terms of the optimality function on the distance between the objective function value at a feasible point and a local minimum of P. (3) We construct validation analysis techniques based on the optimality function and the FJ conditions. (4) We develop an implementable algorithm for P and prove its convergence to FJ points.

Section 2 defines optimality conditions in terms of an optimality function and show how that function relates to the distance to a local minimum of P. Section 3 constructs and analyzes an estimator for the optimality function. Section 4 develops procedures for validation analysis. Section 5 gives an algorithm for P. Section 6 presents numerical examples.

2 Optimality function

In this section, we introduce an optimality function and prove a relationship between the optimality function at a feasible point $x \in \mathbb{R}^n$ and the distance between $f^0(x)$ and a local minimum of P. We start by giving assumptions that ensure that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are finite valued and continuously differentiable and by stating optimality conditions.

Assumption 1 For a given set $S \subset \mathbb{R}^n$, the following hold for any nonempty compact set $X \subset S$ and for all $j \in \mathbf{q}_0$:

(i) There exists a measurable function $C : \Omega \to [0, \infty)$ such that $E[C(\omega)] < \infty$ and $|F^j(x, \omega)| \leq C(\omega)$ for all $x \in X$ and almost every $\omega \in \Omega$.

(ii) There exists a measurable function $L : \Omega \to [0, \infty)$ such that $E[L(\omega)] < \infty$ and

$$|F^j(x, \omega) - F^j(x', \omega)| \leq L(\omega)\|x - x'\|$$

for all $x, x' \in S$ and almost every $\omega \in \Omega$.

(iii) For every $x \in X$, $F^j(\cdot, \omega)$ is continuously differentiable at x for almost all $\omega \in \Omega$.

Assumption 1 is commonly made in the literature (see for example Theorem 7.52 in [41]) and allows for certain classes of nonsmoothness in $F^j(\cdot, \omega)$ that may arise in two-stage stochastic programs with recourse [16], CVaR problems [33], inventory control problems [48], and engineering design problems [35] when $\mathcal{P}$ has a continuous cumulative distribution function. Assumption 1(iii) excludes the possibility of atoms at a point $\omega \in \Omega$ for which $F^j(\cdot, \omega)$ is nonsmooth at some $x \in \mathbb{R}^n$. This occurs, for example, in the newsvendor problem with a discrete demand distribution.

If Assumption 1 holds on an open set S and $X \subset S$ is compact, then it follows from Theorem 7.52 in [41] that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are continuously differentiable on X and that $\nabla f^j(x) = E[\nabla_x F^j(x, \omega)]$ for all $x \in X$ and $j \in \mathbf{q}_0$.

We follow [29], see p. 190, and express the FJ conditions by means of a continuous optimality function $\theta : \mathbb{R}^n \to (-\infty, 0]$ defined by

$$\theta(x) \stackrel{\triangle}{=} \min_{h \in \mathbb{R}^n} \left\{ \max \left\{ -\psi^+(x) + \langle \nabla f^0(x), h \rangle, \max_{j \in \mathbf{q}} \{ f^j(x) - \psi^+(x) + \langle \nabla f^j(x), h \rangle \} \right\} + \tfrac{1}{2} \|h\|^2 \right\}, \tag{2}$$

where $\psi^+(x) \stackrel{\triangle}{=} \max\{0, \psi(x)\}$, with $\psi(x) \stackrel{\triangle}{=} \max_{j \in \mathbf{q}} f^j(x)$, is the constraint violation. We observe that $\theta(x)$ is the minimum value of a linear approximation of objective and constraint functions at x with a quadratic "regularizing" term. The dual problem of (2) takes the form:

$$\theta(x) = -\min_{\mu \in \mathcal{M}} \left\{ \mu^0 \psi^+(x) + \sum_{j \in \mathbf{q}} \mu^j [\psi^+(x) - f^j(x)] + \tfrac{1}{2} \left\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f^j(x) \right\|^2 \right\}, \tag{3}$$

where $\mathcal{M} \stackrel{\triangle}{=} \{\mu \in \mathbb{R}^{q+1} \mid \sum_{j \in \mathbf{q}_0} \mu^j = 1, \mu^j \geq 0, j \in \mathbf{q}_0\}$; see Theorem 2.2.8 in [29]. Here and below superscripts denote components of vectors. The optimality function equivalently expresses the FJ conditions in the sense stated next (see Theorem 2.2.8 in [29]), where $X_\psi \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid \psi(x) \leq 0\}$.

Proposition 1 *Suppose that $\hat{x} \in X_\psi$ and Assumption 1 holds on an open set $S \subset \mathbb{R}^n$ containing $\hat{x}$. Then, $\theta(\hat{x}) = 0$ if and only if $\hat{x}$ is a FJ point.* □

From Proposition 1 and the continuity of $\theta(\cdot)$, we see that an $x \in \mathbb{R}^n$ close to a feasible FJ point yields a near-zero value of $\theta(x)$. Under a positive definite assumption at a local minimizer $\hat{x}$ of P, $\theta(x)$ also gives a bound on the distance between $f^0(x)$ and $f^0(\hat{x})$ for $x \in X_\psi$ near $\hat{x}$ as the next result shows. We find related results for finite minimax problems in [29], p. 176, and for two-stage stochastic program with recourse in [11], but the present result is new.

Theorem 1 *Suppose that $\hat{x} \in \mathbb{R}^n$ is a local minimizer of P and $f^j(\cdot)$ is finite valued and twice continuously differentiable near $\hat{x}$ with $\nabla^2 f^j(\hat{x})$ being positive definite for all $j \in \mathbf{q}_0$. Then, there exist constants $\rho \in (0, \infty)$, $c \in (0, \infty)$, $m \in (0, 1]$, and $M \in [1, \infty)$ such that*

$$(\theta(x) - c\sqrt{-\theta(x)})/m \leq f^0(\hat{x}) - f^0(x) \leq \theta(x)/M \tag{4}$$

for any $x \in X_\psi$, with $\|x - \hat{x}\| \leq \rho$.

Proof Due to its length, we refer to the Appendix for the proof. □

The proof of Theorem 1 reveals that c is given by the size of $\|\nabla f^0(x)\|$ near $\hat{x}$. Moreover, if $f^j(\cdot)$, $j \in \mathbf{q}_0$, satisfy a strong convexity assumption, then (4) holds for all $x \in X_\psi$ with $\hat{x}$ being a *global* minimizer. In view of the above results, $\theta(\cdot)$ is a measure of quality of a candidate point. The computation of $\theta(x)$ for a given $x \in \mathbb{R}^n$ requires the solution of a convex quadratic program with linear constraints (see (3)), which can be achieved in finite time. However, the definition of $\theta(x)$ involves $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$, that, in general, cannot be computed in finite time. Consequently, we define an estimator for $\theta(x)$ using estimators for $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$.

3 Estimator of optimality function

Let $\omega_1, \omega_2, \ldots$ be a sequence of independent random vectors each with value in Ω and distributed as $\mathcal{P}$ and $f_N^j(x) \triangleq \frac{1}{N}\sum_{l=l}^N F^j(x,\omega_l)$, $\nabla f_N^j(x) \triangleq \frac{1}{N}\sum_{l=l}^N \nabla_x F^j(x,\omega_l))$, $\psi_N(x) \triangleq \max_{j\in\mathbf{q}} f_N^j(x)$, and $\psi_N^+(x) \triangleq \max\{0, \psi_N(x)\}$ be standard estimators of $f^j(x), \nabla f^j(x), \psi(x)$, and $\psi^+(x)$, respectively, for any $N \in \mathbb{N} \triangleq \{1,2,3,\ldots\}$, $j \in \mathbf{q}_0$, and $x \in \mathbb{R}^n$. Finally, we define the estimator of $\theta(x)$ by

$$\theta_N(x) \triangleq \min_{h\in\mathbb{R}^n} \left\{ \max\left\{ -\psi_N^+(x) + \langle \nabla f_N^0(x), h\rangle, \max_{j\in\mathbf{q}}\{f_N^j(x) - \psi_N^+(x) + \langle \nabla f_N^j(x), h\rangle\}\right\} + \tfrac{1}{2}\|h\|^2 \right\}.$$

Similar to (3), the dual problem of $\theta_N(x)$ takes the form:

$$\theta_N(x) = -\min_{\mu\in\mathcal{M}} \left\{ \mu^0\psi_N^+(x) + \sum_{j\in\mathbf{q}} \mu^j[\psi_N^+(x) - f_N^j(x)] + \tfrac{1}{2}\left\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f_N^j(x)\right\|^2 \right\}. \tag{5}$$

We next derive properties of $\theta_N(x)$ for a given $x \in \mathbb{R}^n$. We start by stating consistency, which follows from standard arguments (see for example the proof of Proposition 5.2 in [41]).

Proposition 2 *Suppose that Assumption 1 holds on an open set that contains a given $x \in \mathbb{R}^n$. Then, $\theta_N(x) \to \theta(x)$, as $N \to \infty$, almost surely.* □

We next examine the asymptotic distribution of $\theta_N(x)$ and adopt the following assumption.

Assumption 2 For a given $x \in \mathbb{R}^n$, $E[F^j(x,\omega)^2] < \infty$ for all $j \in \mathbf{q}$ and $E[(\partial F^j(x,\omega)/\partial x^i)^2] < \infty$ for all $j \in \mathbf{q}_0$ and $i = 1,2,\ldots,n$. □

In practice, one may need this assumption satisfied for all x in a region of interest as a specific candidate point is typically not known a priori. We denote the set of

minimizers in (3) by

$$\hat{\mathcal{M}}(x) \stackrel{\triangle}{=} \left\{ \mu \in \mathcal{M} \;\middle|\; -\theta(x) = \mu^0 \psi^+(x) + \sum_{j \in \mathbf{q}} \mu^j [\psi^+(x) - f^j(x)] + \tfrac{1}{2} \left\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f^j(x) \right\|^2 \right\}, \tag{6}$$

the set of active constraints by $\hat{\mathbf{q}}(x) \stackrel{\triangle}{=} \{j \in \mathbf{q} \mid \psi(x) = f^j(x)\}$, and the set of active functions in $\psi^+(x)$ by $\hat{\mathbf{q}}^+(x)$, which equals $\hat{\mathbf{q}}(x) \cup \{0\}$ if $\psi(x) = 0$, $\hat{\mathbf{q}}(x)$ if $\psi(x) > 0$, and $\{0\}$ otherwise. Moreover, for any $x \in \mathbb{R}^n$, we let $\overline{Y}(x)$ denote the $q + (q+1)n$-dimensional normal random vector with zero mean and variance-covariance matrix $\overline{V}(x)$, where $\overline{V}(x)$ is the variance-covariance matrix of the random vector $(F^1(x,\omega), F^2(x,\omega), \ldots, F^q(x,\omega), \nabla_x F^0(x,\omega)', \nabla_x F^1(x,\omega)', \ldots, \nabla_x F^q(x,\omega)')'$. Moreover, we define the q-dimensional random vector $Y_{-1}(x)$ and the n-dimensional random vectors $Y_j(x)$, $j \in \mathbf{q}_0$, such that $\overline{Y}(x) = (Y_{-1}(x)', Y_0(x)', Y_1(x)', \ldots, Y_q(x)')'$. The asymptotic distribution of $\theta_N(x)$ then takes the following form, where $\Rightarrow$ denotes convergence in distribution as $N \to \infty$.

Theorem 2 *Suppose that Assumption 2 holds at a given $x \in \mathbb{R}^n$ and that Assumption 1 is satisfied on an open set containing $x \in \mathbb{R}^n$. Then,*

$$N^{1/2}(\theta_N(x) - \theta(x)) \Rightarrow - \min_{\mu \in \hat{\mathcal{M}}(x)} \left\{ \mu^0 W(x) + \sum_{j \in \mathbf{q}} \mu^j [W(x) - Y_{-1}^j(x)] + \sum_{j \in \mathbf{q}_0} \mu^j \left\langle \sum_{k \in \mathbf{q}_0} \mu^k \nabla f^k(x), Y_j(x) \right\rangle \right\} \tag{7}$$

where $W(x) \stackrel{\triangle}{=} \max_{j \in \hat{\mathbf{q}}^+(x)} Y_{-1}^j(x)$, with $Y_{-1}^0(x) \stackrel{\triangle}{=} 0$.

Proof The proof is given in the Appendix. □

The following corollaries are of special interest.

Corollary 1 *Suppose that Assumption 2 holds at a given $x \in \mathbb{R}^n$ and that Assumption 1 is satisfied on an open set containing $x \in \mathbb{R}^n$. Then, the following statements hold:*

(i) *If the vectors* $\nabla f^j(x)$, $j \in \mathbf{q}_0$, *are linearly independent, then* $\hat{\mathcal{M}}(x) = \{\hat{\mu}(x)\}$ *is a singleton and*

$$\begin{aligned} & N^{1/2}(\theta_N(x) - \theta(x)) \\ & \Rightarrow -\hat{\mu}^0(x)W(x) - \sum_{j\in\mathbf{q}} \hat{\mu}^j(x)[W(x) - Y^j_{-1}(x)] \\ & \quad - \sum_{j\in\mathbf{q}_0} \hat{\mu}^j(x) \left\langle \sum_{k\in\mathbf{q}_0} \hat{\mu}^k \nabla f^k(x), Y_j(x) \right\rangle. \end{aligned} \tag{8}$$

(ii) *If* x *is a local minimizer of* P *and the vectors* $\nabla f^j(x)$, $j \in \hat{\mathbf{q}}(x)$, *are linearly independent, then* $\hat{\mathcal{M}}(x) = \{\hat{\mu}(x)\}$ *is a singleton and*

$$N^{1/2}\theta_N(x) \Rightarrow -W(x) + \sum_{j\in\hat{\mathbf{q}}^+(x)} \hat{\mu}^j(x) Y^j_{-1}(x). \tag{9}$$

Proof (i) If the vectors $\nabla f^j(x)$, $j \in \mathbf{q}_0$, are linearly independent, then the matrix $A(x) = (\nabla f^0(x), \nabla f^1(x), \ldots, \nabla f^q(x))$ has rank $q + 1$. Hence, $A(x)'A(x)$ is positive definite and the objective function in (3) is strictly convex. Consequently, $\hat{\mathcal{M}}(x)$ is a singleton and part (i) follows directly.

(ii) Since $x \in \mathbb{R}^n$ is a local minimizer of P, $\psi(x) \leq 0$ and, from Proposition 1, $\theta(x) = 0$. Hence, it follows from (3) that there exists a $\hat{\mu}(x) \in \hat{\mathcal{M}}(x)$ such that $\sum_{j\in\mathbf{q}_0} \hat{\mu}^j(x)\nabla f^j(x) = 0$ and $\sum_{j\in\mathbf{q}} \hat{\mu}^j(x)[\psi^+(x) - f^j(x)] = 0$. Consequently, $\hat{\mu}^j(x) = 0$ for all $j \in \mathbf{q}$ such that $j \notin \hat{\mathbf{q}}^+(x)$. We deduce from the KKT conditions for P that under the stated linear independence assumption, $\hat{\mathcal{M}}(x)$ is a singleton. Since $Y^0_{-1}(x) = 0$ by definition, (7) reduces to (9). □

Corollary 2 *Suppose that Assumption 2 holds at a given* $x \in \mathbb{R}^n$ *and that Assumption 1 holds on an open set containing* $x \in \mathbb{R}^n$. *If all constraints are deterministic, then*

$$N^{1/2}(\theta_N(x) - \theta(x)) \Rightarrow - \min_{\mu\in\hat{\mathcal{M}}(x)} \mu^0 \left\langle \sum_{k\in\mathbf{q}_0} \mu^k \nabla f^k(x), Y_0(x) \right\rangle. \tag{10}$$

Proof This result follows by similar argument as those leading to Theorem 2. □

The next corollary and (10) show that normality of $\theta_N(x)$ occurs when $\hat{\mathcal{M}}(x)$ is a singleton or no constraints exist. Let $\mathcal{N}(0, \sigma^2)$ denote a zero-mean normal random variable with variance σ^2.

Corollary 3 *Suppose that Assumption 2 holds at a given* $x \in \mathbb{R}^n$ *and that Assumption 1 holds on an open set containing* $x \in \mathbb{R}^n$. *If there are no constraints in* P, *then*

$$N^{1/2}(\theta_N(x) - \theta(x)) \Rightarrow \mathcal{N}(0, \nabla f^0(x)' V_0(x) \nabla f^0(x)),$$

where $V_0(x)$ *is the* n*-by-*n *variance-covariance matrix of* $Y_0(x)$.

Proof This result follows directly from Theorem 2. □

We next consider the bias $E[\theta_N(x)] - \theta(x)$. (We use E to denote the expectation with respect to any probability distribution. The meaning should be clear from the context.)

Proposition 3 *Suppose that Assumption 2 holds at a given $x \in \mathbb{R}^n$, Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$, and there exists an $\epsilon > 0$ such that $\sup_{N\in\mathbb{N}} E[|N^{1/2}(\theta_N(x) - \theta(x))|^{1+\epsilon}] < \infty$. Then,*

$$\begin{aligned} & E[\theta_N(x)] - \theta(x) \\ & = N^{-1/2} E\left[- \min_{\mu\in\hat{\mathcal{M}}(x)} \left\{ \mu^0 W(x) + \sum_{j\in\mathbf{q}} \mu^j [W(x) - Y^j_{-1}(x)] \right.\right. \\ & \quad \left.\left. + \sum_{j\in\mathbf{q}_0} \mu^j \left\langle \sum_{k\in\mathbf{q}_0} \mu^k \nabla f^k(x), Y_j(x) \right\rangle \right\} \right] \\ & \quad + o(N^{-1/2}). \end{aligned} \tag{11}$$

Moreover, if $\hat{\mathcal{M}}(x)$ is a singleton, then $E[\theta_N(x)] - \theta(x) = -N^{-1/2} E[W(x)] + o(N^{-1/2})$.

Proof From Theorem 25.12 in [7] and Theorem 2, we directly obtain (11). Since Y^j_{-1}, $j \in \mathbf{q}$, and $Y_j(x)$, $j \in \mathbf{q}_0$, have zero mean and $\sum_{j\in\mathbf{q}_0} \mu^j = 1$ for all $\mu \in \mathcal{M}$, the second part also holds. □

We observe that the bias identified above is similar to that of the optimal value of $\min_{x\in X_\psi} f^0_N(x)$ relative to the optimal value of $\min_{x\in X_\psi} f^0(x)$; see, for example p. 167 in [41]. In that case, the bias is always nonpositive. In the present case, $E[\theta_N(x)]$ may be larger than $\theta(x)$. However, in the absence of constraints in P, we find using Jensen's inequality that for any $N \in \mathbb{N}$, $E[\theta_N(x)] \le \theta(x)$.

4 Validation analysis

In this section, we develop confidence intervals of $\theta(x)$ and $\psi(x)$ for a candidate point $x \in \mathbb{R}^n$, which may be used to assess "near-optimality" and "near-feasibility" of x. In view of Corollary 3, confidence intervals can easily be obtained using standard techniques in the case of no constraint. Hence, we focus on situations with constraints.

4.1 Near feasibility in P

We adopt a simple batching approach to estimate the value of $\psi(x)$. (We refer to [17] and [37] for other approaches not pursued here.) By Jensen's inequality, we find that $\psi(x) \le E[\psi_N(x)]$. Hence, a confidence interval for $E[\psi_N(x)]$ provides a conservative confidence interval for $\psi(x)$, which we construct next. For given N and

M, let $\psi_{N,k}(x), k = 1, 2, \ldots, M$, be independent random variables distributed as $\psi_N(x)$. Then, $\overline{\psi}_{N,M}(x) \stackrel{\triangle}{=} \sum_{k=1}^{M} \psi_{N,k}(x)/M$ is an unbiased estimator of $E[\psi_N(x)]$. If $E[F^j(x,\omega)^2] < \infty$ for all $j \in \mathbf{q}$, then $\overline{\psi}_{N,M}(x)$ is approximately normal with mean $E[\psi_N(x)]$ and variance $Var[\psi_N(x)]/M$ for large M. Hence,

$$(-\infty, \overline{\psi}_{N,M}(x) + z_\alpha s_{\psi,N,M}(x)/\sqrt{M}] \tag{12}$$

is an approximate one-sided $100(1-\alpha)\%$-confidence interval for $E[\psi_N(x)]$ for large M and also a conservative $100(1-\alpha)\%$-confidence interval for $\psi(x)$, where z_α is the standard normal α-quantile and $s^2_{\psi,N,M}(x)$ is the standard unbiased estimator of $Var[\psi_N(x)]$.

4.2 Near optimality in P

We propose two approaches for obtaining confidence intervals for $\theta(x)$. We note that the optimality function synthesizes the lack of feasibility and optimality at a particular point into a real number. Hence, it is natural to supplement a confidence interval for $\theta(x)$ by one for $\psi(x)$ (see (12)), which assesses feasibility exclusively.

The first approach for obtaining confidence intervals for $\theta(x)$ makes use of the following result.

Proposition 4 *Suppose that Assumption 1 holds on an open set containing a given $x \in \mathbb{R}^n$. Then, for any $\mu \in \mathcal{M}$ and $N \in \mathbb{N}$,*

$$\theta(x) \geq E\left[-\mu^0 \psi_N^+(x) - \sum_{j\in\mathbf{q}} \mu^j (\psi_N^+(x) - f_N^j(x)) - \tfrac{1}{2} \left\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f_N^j(x) \right\|^2 \right]. \tag{13}$$

Proof For any $\mu \in \mathcal{M}$, let $\tilde{\eta} : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ be defined by $\tilde{\eta}(\overline{\zeta}) \stackrel{\triangle}{=} \max\{0, \max_{j\in\mathbf{q}} \zeta_{-1}^j\} - \sum_{j\in\mathbf{q}} \mu^j \zeta_{-1}^j + \frac{1}{2} \| \sum_{j\in\mathbf{q}_0} \mu^j \zeta_j \|^2$ for any $\overline{\zeta} = (\zeta_{-1}', \zeta_0', \zeta_1', \ldots \zeta_q')$ $\in \mathbb{R}^{q+(q+1)n}$, with $\zeta_{-1} \in \mathbb{R}^q$ and $\zeta_j \in \mathbb{R}^n$, $j \in \mathbf{q}_0$. Let $f_N(x) \stackrel{\triangle}{=} (f_N^1(x), f_N^2(x), \ldots, f_N^q(x))'$ and $\nabla \overline{f}_N(x) \stackrel{\triangle}{=} (\nabla f_N^0(x)', \nabla f_N^1(x)', \ldots, \nabla f_N^q(x)')'$. By the convexity of $\tilde{\eta}(\cdot)$, Jensen's inequality, (3), and the suboptimality of μ,

$$\begin{aligned} E[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')] &\geq \mu^0 \psi^+(x) + \sum_{j\in\mathbf{q}} \mu^j (\psi^+(x) - f^j(x)) \\ &\quad + \tfrac{1}{2} \left\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j(x) \right\|^2 \geq -\theta(x). \end{aligned}$$

The conclusion then follows from the definition of $\tilde{\eta}(\cdot)$. □

In view of Proposition 4, we construct a conservative confidence interval for $\theta(x)$ by computing a confidence interval for the right-hand side in (13). We adopt a batching approach and, for given N and M, let $\eta_{N,k}, k = 1, 2, \ldots, M$, be independent random variables distributed as $\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')$; see definition in the proof of Proposition 4. Then, $\overline{\eta}_{N,M} \stackrel{\triangle}{=} \frac{1}{M} \sum_{k=1}^{M} \eta_{N,k}$ is an unbiased estimator of $E[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ that is approximately normal with mean $E[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ and variance $Var[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]/M$ for large M under integrability assumptions on $(f_N(x)', \nabla \overline{f}_N(x)')$. Then, it follows that

$$[-\overline{\eta}_{N,M} - z_\alpha s_{\eta,N,M}(x)/\sqrt{M}, 0] \tag{14}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $E[-\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ for large M and also a conservative $100(1-\alpha)\%$-confidence interval for $\theta(x)$. Here, $s^2_{\eta,N,M}$ is the standard unbiased estimator of $Var[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$. To compute (14), it is necessary to select a $\mu \in \mathcal{M}$. In view of the proof of Proposition 4, we see that a tighter confidence interval can be expected when $\mu \in \hat{\mathcal{M}}(x)$. However, even when $\mu \in \hat{\mathcal{M}}(x)$, the inequality in (13) may be strict.

The second approach to constructing a confidence interval for $\theta(x)$ is motivated by a procedure for obtaining bounds on the optimal value of optimization problems with chance constraints [23]; see also Sect. 5.7.2 in [41]. The approach requires a slightly different sampling scheme. While we above use common random numbers, i.e., $f^j_N(x), \nabla f^j_N(x), j \in \mathbf{q}_0, \psi_N(x), \psi^+_N(x)$, and $\theta_N(x)$ are computed using the same sample, we now generate a sample of size N for each vector $(f^j_N(x), \nabla f^j_N(x)'), j \in \mathbf{q}_0$, independently, and also independently generate a sample of size N to compute $\psi_N(x)$. (Such independent sampling is for example discussed in [41], Chap. 5, Remark 9.) We refer to this modified scheme as the function-independent sampling scheme. Since the function-independent sampling scheme is only discussed in this subsection and used in numerical tests in Sect. 6, we slightly abuse notation by using the same notation for both sampling schemes.

From (2) we see that $\theta(x) = -\psi^+(x) + u(x)$, where

$$u(x) \stackrel{\triangle}{=} \min_{(h,z)\in\mathbb{R}^{n+1}} \{z + \tfrac{1}{2}\|h\|^2 \mid \langle \nabla f^0(x), h\rangle \le z,\ f^j(x) + \langle \nabla f^j(x), h\rangle \le z,\ j \in \mathbf{q}\}. \tag{15}$$

Here, $-\psi^+(x)$ is a measure of feasibility and $u(x)$ is a measure of optimality. Using the function-independent sampling scheme, we similarly let

$$u_N(x) \stackrel{\triangle}{=} \min_{(h,z)\in\mathbb{R}^{n+1}} \{z + \tfrac{1}{2}\|h\|^2 \mid \langle \nabla f^0_N(x), h\rangle \le z,\ f^j_N(x) + \langle \nabla f^j_N(x), h\rangle \le z,\ j \in \mathbf{q}\}. \tag{16}$$

The next lemma provides a useful relationship between $u(x)$ and $u_N(x)$.

Lemma 1 *Suppose that Assumption 2 holds at a given $x \in \mathbb{R}^n$, that Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$, and that the function-independent sampling scheme is used. Then,*

$$\liminf_{N\to\infty} \text{Prob}[u_N(x) \le u(x)] \ge \frac{1}{2^{q+1}}. \tag{17}$$

Proof Suppose that $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is a feasible point in (15). We want to determine the probability, denoted by $\hat{p}_N$, that $(\hat{h}, \hat{z})$ is feasible in (16). Since $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is feasible for (15),

$$\begin{aligned}\hat{p}_N &\triangleq \text{Prob}\left[\left\{\langle \nabla f_N^0(x), \hat{h}\rangle \le \hat{z}\right\} \bigcap \left(\bigcap_{j\in\mathbf{q}} \left\{f_N^j(x) + \langle \nabla f_N^j(x), \hat{h}\rangle \le \hat{z}\right\}\right)\right] \\ &\ge \text{Prob}\left[\left\{\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \le 0\right\} \bigcap \left(\bigcap_{j\in\mathbf{q}} \left\{f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) \right.\right.\right. \\ &\qquad \left.\left.\left. -\nabla f^j(x), \hat{h}\rangle \le 0\right\}\right)\right]. \end{aligned} \tag{18}$$

In view of the function-independent sampling scheme, it follows that

$$\begin{aligned}\hat{p}_N \ge\ & \text{Prob}\left[\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \le 0\right] \\ & \times \prod_{j\in\mathbf{q}} \text{Prob}\left[f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \le 0\right].\end{aligned}$$

By Assumption 2, a central limit theorem, and the continuous mapping theorem, $N^{1/2}\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle$ converges in distribution to a zero-mean normal random variable. Hence,

$$\lim_{N\to\infty} \text{Prob}\left[\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \le 0\right] \ge 1/2. \tag{19}$$

Similarly, for all $j \in \mathbf{q}$, $N^{1/2}(f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle)$ converges in distribution to a zero-mean normal random variable. Hence, for all $j \in \mathbf{q}$,

$$\lim_{N\to\infty} \text{Prob}\left[f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \le 0\right] \ge 1/2.$$

Consequently, $\liminf_{N\to\infty} \hat{p}_N \ge 1/2^{q+1}$. Since this result holds for any $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ that is feasible in (15), it also holds for the optimal solution in (15). If $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is optimal in (15) and it is also feasible in (16), then $u_N(x) \le \hat{z} + \frac{1}{2}\|\hat{h}\|^2 = u(x)$. This completes the proof. □

Lemma 1 provides the basis for the following procedure for obtaining a probabilistic lower bound on $u(x)$. This procedure is essentially identical to the one proposed in [23] in the context of chance constraints. Let $u_{N,k}(x), k = 1, 2, \ldots, K$, be independent random variables distributed as $u_N(x)$. After obtaining realizations of these random variables, we order them with respect to their values. Let $\tilde{u}_{N,1}, \tilde{u}_{N,2}, \ldots, \tilde{u}_{N,K}$, with $\tilde{u}_{N,k} \leq \tilde{u}_{N,k+1}$, be this ordered sequence. That is, $\tilde{u}_{N,1}$ is the smallest value of $u_{N,k}(x), k = 1, 2, \ldots, K$, $\tilde{u}_{N,2}$ is the second smallest, etc. Suppose that $\hat{\gamma}_N$ is a lower bound on $\text{Prob}[u_N(x) \leq u(x)]$ and suppose that for a given $\beta \in (0, 1)$, K and L satisfy

$$\sum_{k=0}^{L-1} \binom{K}{k} \hat{\gamma}_N^k (1 - \hat{\gamma}_N)^{K-k} \leq \beta. \tag{20}$$

Then, using the same arguments as in Sect. 5.7.2 of [41], we obtain that $\text{Prob}[\tilde{u}_{N,L} > u(x)] \leq \beta$. Hence, $[\tilde{u}_{N,L}, 0]$ is a $100(1 - \beta)\%$-confidence interval for $u(x)$. In view of Lemma 1, we recommend a number slightly smaller than $1/2^{q+1}$ as an estimate of $\hat{\gamma}_N$ when N is moderately large.

If $\overline{\psi}_{N,M}(x)$ of Sect. 4.1 is computed independently of the confidence interval for $u(x)$, then in view of the fact that $\theta_N(x) = -\psi_N^+(x) + u_N(x)$,

$$\left[-\max\{0, \overline{\psi}_{N,M}(x) + z_\alpha s_{\psi,N,M}(x)/\sqrt{M}\} + \tilde{u}_{N,L},\ 0 \right] \tag{21}$$

is an approximate $100(1 - \alpha)(1 - \beta)\%$-confidence interval for $\theta(x)$ for large M and N. The approach requires the solution of K quadratic programs. If $L = 1$, then $K \geq \log \beta / \log(1 - \hat{\gamma}_N)$. Hence, K is typically moderate. For example, if $\beta = 0.01$ and $\hat{\gamma}_N = 0.49$, then $K = 7$ suffices.

5 Algorithm and consistent approximations

In this section, we use the optimality function $\theta(\cdot)$ and an optimality function of an approximating problem to construct an implementable algorithm for P under the additional assumptions that $F^j(\cdot, \omega)$, $j \in \mathbf{q}_0$, are given by the maximum of continuously differentiable random functions. We therefore replace Assumption 1 by the following more specific assumption.

Assumption 3 The random functions $F^j : \mathbb{R}^n \times \Omega \to \mathbb{R}$, $j \in \mathbf{q}_0$, are given by

$$F^j(x, \omega) = \max_{k \in \mathbf{r}^j} g^{jk}(x, \omega),\ j \in \mathbf{q}_0, \tag{22}$$

where $\mathbf{r}^j = \{1, 2, \ldots, r^j\}$, $r^j \in \mathbb{N}$, and for a given set $S \subset \mathbb{R}^n$, the following hold for all $j \in \mathbf{q}_0$:

(i) For all $k \in \mathbf{r}^j$ and almost every $\omega \in \Omega$, $g^{jk}(\cdot, \omega)$ is continuously differentiable on S.

(ii) There exist a nonnegative-valued measurable function $C^j : \Omega \to [0, \infty)$ such that $E[C^j(\omega)] < \infty$, $|g^{jk}(x, \omega)| \leq C^j(\omega)$, and $\|\nabla_x g^{jk}(x, \omega)\| \leq C^j(\omega)$ for all $x \in S$ and $k \in \mathbf{r}^j$, and for almost every $\omega \in \Omega$.

(iii) For all $x \in S$, the set $\hat{\mathbf{r}}^j(x, \omega) \triangleq \{k \in \mathbf{r}^j \mid F^j(x, \omega) = g^{jk}(x, \omega)\}$ is a singleton for almost every $\omega \in \Omega$. □

Assumption 3(iii) excludes the possibility of atoms at a point $\omega \in \Omega$ for which there is more than one maximizer in (22) at a given x. If Assumption 3 holds on $S \subset \mathbb{R}^n$, then Assumption 1 also holds on S as the next result states.

Proposition 5 *Suppose that Assumption 3 holds on an open set $S \subset \mathbb{R}^n$. Then, (i) Assumption 1 holds on S and (ii) for any compact $X \subset S$, $f^j(\cdot)$, $j \in \mathbf{q}_0$, are finite valued and continuously differentiable on X with $\nabla f^j(x) = E[\nabla_x g^{\hat{k}^j(x,\omega)j}(x, \omega)]$, where $\hat{k}^j(x, \omega) \in \hat{\mathbf{r}}^j(x, \omega)$.*

Proof Assumption 1(i) holds directly from Assumption 3(i). For all $j \in \mathbf{q}_0$ and almost every $\omega \in \Omega$, $F^j(\cdot, \omega)$ is Lipschitz continuous on bounded sets and has a directional derivative at $x \in \mathbb{R}^n$ in direction $h \in \mathbb{R}^n$ given by $dF^j(x, \omega; h) = \max_{k \in \hat{\mathbf{r}}^j(x,\omega)} \langle \nabla_x g^{jk}(x, \omega), h \rangle$; see for example Theorem 5.4.5 in [29]. Hence, in view of Assumption 3(ii), $F^j(\cdot, \omega)$ is Lipschitz continuous on bounded sets with an integrable Lipschitz constant. Hence, Assumption 1(ii) holds. From Assumption 3(iii) we conclude that for all $x \in S$, $F^j(\cdot, \omega)$ is continuously differentiable at x and $\hat{\mathbf{r}}^j(x, \omega) = \{\hat{k}^j(x, \omega)\}$ for almost every $\omega \in \Omega$. Hence, $\nabla_x F^j(x, \omega) = \nabla_x g^{\hat{k}^j(x,\omega)j}(x, \omega)$ and Assumption 1(iii) holds. The conclusions then follows from Theorem 7.52 in [41]. □

If Assumption 1 holds on an open set $S \subset \mathbb{R}^n$ containing a compact set X, then $f_N^j(x)$ converges to $f^j(x)$ uniformly on X, as $N \to \infty$, almost surely for any $j \in \mathbf{q}_0$; see Theorem 7.48 in [41]. While this fact is useful, $f_N^j(\cdot)$ is nonsmooth and, hence, standard nonlinear programming algorithm may fail when applied to P with $f^j(\cdot)$ replaced by $f_N^j(\cdot)$ for a given realization of $\{\omega_l\}_{l=1}^N$. Consequently, we construct smooth approximations of $f_N^j(\cdot)$, $j \in \mathbf{q}_0$.

5.1 Sample average approximations and exponential smoothing

We adopt the exponential smoothing technique first proposed in [18]; see also [45,48] for recent applications. For any $\epsilon > 0$ and $j \in \mathbf{q}_0$, we define the smooth approximation $F_\epsilon^j : \mathbb{R}^n \times \Omega \to \mathbb{R}$ by

$$F_\epsilon^j(x, \omega) \triangleq \epsilon \log \sum_{k \in \mathbf{r}^j} \exp[g^{jk}(x, \omega)/\epsilon]. \tag{23}$$

Under Assumption 3, $F_\epsilon^j(\cdot, \omega)$, $j \in \mathbf{q}_0, \epsilon > 0$, are continuously differentiable for almost every $\omega \in \Omega$, with

$$\nabla F_\epsilon^j(x, \omega) = \sum_{k \in \mathbf{r}^j} \mu_\epsilon^{jk}(x, \omega) \nabla_x g^{jk}(x, \omega), \tag{24}$$

where

$$\mu_\epsilon^{jk}(x,\omega) \stackrel{\triangle}{=} \frac{\exp[g^{jk}(x,\omega)/\epsilon]}{\sum_{k'\in\mathbf{r}^j}\exp[g^{jk'}(x,\omega)/\epsilon]}, k \in \mathbf{r}^j. \tag{25}$$

Moreover, for any $j \in \mathbf{q}_0, \epsilon > 0, x \in \mathbb{R}^n$, and $\omega \in \Omega$,

$$0 \le F_\epsilon^j(x,\omega) - F^j(x,\omega) \le \epsilon \log r^j. \tag{26}$$

For any $j \in \mathbf{q}_0, \epsilon > 0$, and $N \in \mathbb{N}$, we define the smoothed sample average $f_{N\epsilon}^j : \mathbb{R}^n \to \mathbb{R}$ by

$$f_{N\epsilon}^j(x) \stackrel{\triangle}{=} \frac{1}{N}\sum_{l=1}^{N} F_\epsilon^j(x,\omega_l). \tag{27}$$

Finally, we define, for any $\epsilon > 0$ and $N \in \mathbb{N}$, the smoothed sample average problem

$$P_{N\epsilon}: \quad \min_{x\in\mathbb{R}^n}\{f_{N\epsilon}^0(x) \mid f_{N\epsilon}^j(x) \le 0, j \in \mathbf{q}\}. \tag{28}$$

For given $\epsilon > 0$, $N \in \mathbb{N}$, and realization of $\{\omega_l\}_{l=1}^N$, $P_{N\epsilon}$ is smooth and, hence, solvable by standard nonlinear programming algorithms. We note that if $\mathbf{r}^j$ is a singleton for all $j \in \mathbf{q}_0$, then smoothing is not required and the above expressions simplify.

A simple approach for solving P is to select a small ϵ and a large N to ensure small smoothing and sampling errors, respectively, and then to apply a standard algorithm to $P_{N\epsilon}$. In the case of deterministic constraints in P, the results of [48] provide theoretical backing for this approach by showing that every accumulation point of a sequence of stationary points of smoothed sample average problems of $P_{N\epsilon}$ (but with deterministic constraints) is a stationary point of P. In the next subsection, we extend the result of [48] in one direction by considering a sequence of near-stationary points of $P_{N\epsilon}$ (with expectation constraints) as expressed by optimality functions. We utilize this result to obtain an implementable algorithm that approximately solves sequences of smoothed sample average problems $P_{N\epsilon}$ for gradually smaller ϵ and larger N. There is evidence that such a gradual increase in precision tends to perform better numerically than the simple approach of solving a single approximating problem with high precision [2,5,14,15,25,26,30,35,42]. This effect is often caused by the fact that substantial objective function and constraint violation improvements can be achieved with low precision in the early stages of the calculations without paying the price associated with high precision. In the present context, a high precision requires a large N, which results in expensive function evaluations, and a small ϵ, which may cause ill-conditioning [26]. Hence, we proceed by considering a sequence of $P_{N\epsilon}$ with gradually higher precision.

5.2 Consistent approximations

We analyze $P_{N\epsilon}$ within the framework of consistent approximations (see [29, Sect. 3.3]), which allows us to related near-stationary points of $P_{N\epsilon}$ to stationary points of P through their respective optimality functions. We start by defining an optimality function for $P_{N\epsilon}$.

Let $\theta_{N\epsilon} : \mathbb{R}^n \to (-\infty, 0]$ denote an optimality function for $P_{N\epsilon}$ defined by

$$\theta_{N\epsilon}(x) \stackrel{\triangle}{=} -\min_{\mu\in\mathcal{M}} \left\{ \mu^0 \psi^+_{N\epsilon}(x) + \sum_{j\in\mathbf{q}} \mu^j [\psi^+_{N\epsilon}(x) - f^j_{N\epsilon}(x)] + \tfrac{1}{2} \left\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j_{N\epsilon}(x) \right\|^2 \right\}, \tag{29}$$

where $\psi^+_{N\epsilon}(x) \stackrel{\triangle}{=} \max\{\psi_{N\epsilon}(x), 0\}$, with $\psi_{N\epsilon}(x) = \max_{j\in\mathbf{q}} f^j_{N\epsilon}(x)$. Similar results as in Proposition 1 hold for $P_{N\epsilon}$ and $\theta_{N\epsilon}(\cdot)$, and hence if $x \in \mathbb{R}^n$ is feasible for $P_{N\epsilon}$, then x is a FJ point of $P_{N\epsilon}$ if and only if $\theta_{N\epsilon}(x) = 0$. To avoid dealing with N and ϵ individually, we let $\{\epsilon_N\}_{N=1}^\infty$ be such that $\epsilon_N > 0$ for all $N \in \mathbb{N}$ and $\epsilon_N \to 0$, as $N \to \infty$. We adopt the following definition of weakly consistent approximations from Sect. 3.3 in [29].

Definition 1 The elements of $\{(P_{N\epsilon_N}, \theta_{N\epsilon_N}(\cdot)\}_{N=1}^\infty$ are weakly consistent approximations of $(P, \theta(\cdot))$ if (i) $P_{N\epsilon_N}$ epi-converges to P, as $N \to \infty$, almost surely, and (ii) for any $x \in \mathbb{R}^n$ and $\{x_N\}_{N=1}^\infty \subset \mathbb{R}^n$ with $x_N \to x$, as $N \to \infty$, $\limsup_{N\to\infty} \theta_{N\epsilon_N}(x_N) \le \theta(x)$, almost surely. □

We proceed by showing that $\{(P_{N\epsilon_N}, \theta_{N\epsilon_N}(\cdot)\}_{N=1}^\infty$ indeed are weakly consistent approximations of $(P, \theta(\cdot))$. We need the following key result.

Proposition 6 *Suppose that Assumption 3 holds on an open set $S \subset \mathbb{R}^n$ and that $X \subset S$ is compact. Then, for all $j \in \mathbf{q}_0$,*

(i) *$f^j_{N\epsilon_N}(x)$ converges to $f^j(x)$ uniformly on X, as $N \to \infty$, almost surely, and*

(ii) *$\nabla f^j_{N\epsilon_N}(x)$ converges to $\nabla f^j(x)$ uniformly on X, as $N \to \infty$, almost surely.*

Proof See appendix. □

We need the following constraint qualification to ensure epi convergence.

Assumption 4 For a given set $S \subset \mathbb{R}^n$ the following holds almost surely. For every $x \in S \cap X_\psi$, there exists a sequence $\{x_N\}_{N=1}^\infty \subset S$, with $\psi_N(x_N) \le 0$, such that $x_N \to x$, as $N \to \infty$. □

Theorem 3 *Suppose that Assumptions 3 and 4 hold on an open set $S \subset \mathbb{R}^n$, $X \subset S$ is compact, and $X_\psi \subset X$. Then, $\{(P_{N\epsilon_N}, \theta_{N\epsilon_N}(\cdot)\}_{N=1}^\infty$ are weakly consistent approximations of $(P, \theta(\cdot))$.*

Proof Using Theorem 3.3.2 in [29], it follows directly from Proposition 6(i) and Assumption 4 that $P_{N\epsilon_N}$ epi-converges to P, as $N \to \infty$, almost surely. Next, we consider the optimality functions. Let $\eta : \mathcal{M} \times X \to \mathbb{R}$ and $\eta_N : \mathcal{M} \times X \to \mathbb{R}$ be defined by

$$\eta(\mu, x) \stackrel{\Delta}{=} \mu^0 \psi^+(x) + \sum_{j \in \mathbf{q}} \mu^j [\psi^+(x) - f^j(x)] + \tfrac{1}{2} \left\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f^j(x) \right\|^2$$

and $\eta_N(\mu, x)$ similarly with $\psi^+(x)$, $f^j(x)$, and $\nabla f^j(x)$ replaced by $\psi^+_{N\epsilon_N}(x)$, $f^j_{N\epsilon_N}(x)$, and $\nabla f^j_{N\epsilon_N}(x)$, respectively. In view of Proposition 6, $\eta_N(\cdot, \cdot)$ converges to $\eta(\cdot, \cdot)$ uniformly on $\mathcal{M} \times X$, as $N \to \infty$, almost surely. Since $\theta(x) = -\min_{\mu \in \mathcal{M}} \eta(\mu, x)$ and $\theta_{N\epsilon_N}(x) = -\min_{\mu \in \mathcal{M}} \eta_N(\mu, x)$, we conclude that $\theta_{N\epsilon_N}(\cdot)$ converges to $\theta(\cdot)$ uniformly on X, as $N \to \infty$, almost surely, which completes the proof. □

Consistent approximations lead to an algorithm for P, which approximately solves sequences of problems $\{P_{N\epsilon_N}\}_{N \in \mathcal{K}}$, where $\mathcal{K}$ is an order set of strictly increasing positive integers with infinite cardinality. As N increases, the precision with which $P_{N\epsilon_N}$ is solved increases too. We measure the precision of a candidate point of $P_{N\epsilon_N}$ by means of $\theta_{N\epsilon_N}(\cdot)$. When a point of sufficient precision is obtained for $P_{N\epsilon_N}$, the algorithm starts solving $P_{N'\epsilon_{N'}}$, where N' is the next integer in $\mathcal{K}$ after N. We allow flexibility in the choice of optimization algorithm for approximately solving $\{P_{N\epsilon_N}\}_{N \in \mathcal{K}}$. We only require that the optimization algorithm converges to a feasible FJ point of $P_{N\epsilon_N}$ as the next assumption formalizes. Here, we adopt the notation $A_{N\epsilon}(x)$ for the iterate obtained after a fixed number of iterations of the optimization algorithm, starting from $x \in \mathbb{R}^n$, when applied to $P_{N\epsilon}$.

Assumption 5 The following holds almost surely. For any $N \in \mathbb{N}$ and $\epsilon > 0$, every accumulation point $\hat{x} \in \mathbb{R}^n$ of $\{x_i\}_{i=0}^{\infty}$, with $x_{i+1} = A_{N\epsilon}(x_i)$, $i = 0, 1, 2, \ldots$, satisfies $\theta_{N\epsilon}(\hat{x}) = 0$ and $\psi_{N\epsilon}(\hat{x}) \leq 0$.

The algorithm for P, stated next, is a straightforward adaptation of Algorithm Model 3.3.14 in [29]. We use the notation $\mathcal{K}(N)$ to denote the smallest $N' \in \mathcal{K}$ strictly greater than N.

Algorithm 1 (Solves P under Assumptions 3, 4, and 5)

Input. Function $\Delta : \mathbb{N} \to (0, \infty)$ with $\Delta(N) \to 0$, as $N \to \infty$; ordered strictly increasing set $\mathcal{K} \subset \mathbb{N}$ with infinite cardinality; $\{\epsilon_N\}_{N \in \mathcal{K}} \subset (0, \infty)$, with $\epsilon_N \to^{\mathcal{K}} 0$, as $N \to \infty$; $\delta_1, \delta_2 > 0$; $N_0 \in \mathcal{K}$; $x_0 \in \mathbb{R}^n$; and realizations $\{\omega_l\}_{l=1}^{\infty}$ obtained by independent sampling from $\mathcal{P}$.

Step 0. Set $i = 0$, $x_0^* = x_0$, and $N = N_0$.

Step 1. Compute $x_{i+1} = A_{N\epsilon_N}(x_i)$.

Step 2. If $\theta_{N\epsilon_N}(x_{i+1}) \geq -\delta_1 \Delta(N)$ and $\psi_{N\epsilon_N}(x_{i+1}) \leq \delta_2 \Delta(N)$, then set $x_N^* = x_{i+1}$ and replace N by $\mathcal{K}(N)$.

Step 3. Replace i by $i + 1$, and go to Step 1. □

In view of Theorem 3, convergence of Algorithm 1 is deduced from Theorem 3.3.15 in [29]:

Theorem 4 *Suppose that Assumptions 3, 4, and 5 hold on a sufficiently large open subset of* $\mathbb{R}^n$. *Moreover, suppose that Algorithm 1 has generated the sequences* $\{x_N^*\}$ *and* $\{x_i\}_{i=0}^{\infty}$ *and they are bounded. Then,* $\{x_N^*\}$ *is an infinite sequence and every accumulation point* $\hat{x}$ *of* $\{x_N^*\}$ *satisfies* $\theta(\hat{x}) = 0$ *and* $\psi(\hat{x}) \leq 0$ *almost surely.* □

6 Numerical examples

In this section, we present numerical tests of Algorithm 1 and the validation analysis procedures in Sect. 4 as applied to five examples involving constraints. We also carried out validation analysis for an unconstrained example using Corollary 3. However, as the results are conceptually similar to those below, they are omitted. All calculations are performed in Matlab 7.4 on a 2.16 GHz laptop computer with 1 GB of RAM and Windows XP, unless stated otherwise.

6.1 Example 1: Validation analysis for deterministically constrained problem

This problem instance arises in search and detection applications where an area is divided into n cells, one of which contains a stationary target. Let $x \in \mathbb{R}^n$, with x^i representing the number of time units a searcher allocates to cell i. Then, the probability of not detecting the target is $f^0(x) = E[F^0(x, \omega)]$, where $F^0(x, \omega) = \sum_{i=1}^n p_i \exp(-\omega^i x^i)$, p_i is the prior probability that the target is located in cell i, and $\omega = (\omega^1, \omega^2, \ldots, \omega^n)'$ is an independent lognormally distributed random vector (with parameters[1] $\xi^i = 100u^i$ and $\lambda^i = 0$, where $u^i \in (0, 1)$ are given data generated by independent sampling from a uniform distribution) representing the random search effectiveness in the cells. The searcher is constrained by $\sum_{i=1}^n x^i \leq 1$ and $x \geq 0$, where we use $n = 100$. Assumption 3 holds for this problem instance. We consider three candidate solutions: $x_1 \in \mathbb{R}^{100}$, which is nearly optimal, $x_2 = (1/100, 1/100, \ldots, 1/100)' \in \mathbb{R}^{100}$, and $x_3 = (1/50, 1/50, \ldots, 1/50)' \in \mathbb{R}^{100}$, which is infeasible. Hence, $\psi(x_1) = \psi(x_2) = 0$ and $\psi(x_3) = 1$. We verify using long simulations (sample size 10^8) that $\theta(x_1) \approx 8 \cdot 10^{-7}$, $\theta(x_2) \approx -0.00736$, and $\theta(x_3) \approx -0.99318$; see the last row of Table 1.

We consider both confidence intervals (14) and (21). To compute (14), we first determine μ by solving (5) using sample size N. Second, we compute $\overline{\eta}_{N,M}$ using μ with M replications. In (21), we use $L = 1$ which leads to $K = 5$ when $\beta = 0.05$; see (20). Table 1 provides 95%-confidence intervals for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (14) (rows 3–6) and (21) (rows 7–10) with varying sample size N and replications M and K. We observe that the confidence intervals cover the exact value of the optimality function. When the value of the optimality function is some distance from zero, a tight confidence interval is obtained using a moderate sample size N. However, when the

[1] We note that λ^i and ξ^i are the mean and standard deviation, respectively, of the normal distribution from which the lognormal distribution is obtained.

Table 1 95%-Confidence intervals in Example 1 for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (14) (rows 3–6) and (21) (rows 7–10) with varying sample size N and replications M and K

Method	N	M	K	Confidence intervals		
				$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
(14)	10^2	30	–	$[-0.004254, 0]$	$[-0.008125, 0]$	$[-1.049167, 0]$
	10^3	30	–	$[-0.000630, 0]$	$[-0.007837, 0]$	$[-1.048609, 0]$
	10^4	30	–	$[-0.000050, 0]$	$[-0.007783, 0]$	$[-1.048554, 0]$
	10^5	100	–	$[-0.000006, 0]$	$[-0.007483, 0]$	$[-1.009602, 0]$
(21)	10^2	–	5	$[-0.001886, 0]$	$[-0.007628, 0]$	$[-0.994375, 0]$
	10^3	–	5	$[-0.000464, 0]$	$[-0.007497, 0]$	$[-0.993391, 0]$
	10^4	–	5	$[-0.000049, 0]$	$[-0.007359, 0]$	$[-0.993278, 0]$
	10^5	–	5	$[-0.000006, 0]$	$[-0.007365, 0]$	$[-0.993201, 0]$
"Exact"				$\approx 8 \times 10^{-7}$	≈ -0.00736	≈ -0.99318

The last row gives approximate but accurate values of $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$

optimality function is close to zero, a large sample size is required. While the confidence intervals reported are from a single generation, we also verify the coverage and variability of the confidence intervals across independent replications. Specifically, we confirm the confidence level in (21) by estimating coverage probabilities, i.e., the probability that the random confidence interval (21) includes $\theta(x)$. We find that 100, 99, 98 and 99% of 1,000 (200 in the case of $N = 10^5$) independent replications of (21) cover $\theta(x_1)$ for $N = 10^2, 10^3, 10^4$, and 10^5, respectively. Similar calculations for $\theta(x_2)$ and $\theta(x_3)$ result in coverage percentages of at least 97%. All these percentages are well above the stipulated 95%. We also compute the coefficients of variation across 20 replications of (14) and (21), and obtain at most 11%, 2%, and 0.01% coefficients of variation in confidence interval for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$, respectively, regardless of sample size or method used in Table 1. Hence, the variability of the confidence intervals is modest across independent replications.

We also apply the hypothesis test of [42] and find a p value of 0.65 for the case with x_1. Hence, we are unable to reject the null hypothesis that x_1 is a KKT point using any reasonable test size. In the case of x_2 and x_3, the p values are essentially zero and the null hypothesis is rejected even with a small test size. While these conclusions are reasonable, they do not directly provide information about how "close" a candidate solution is to a FJ point. In practice, we are rarely able to obtain a candidate solution that is a FJ point. Hence, the "distance" to such a point becomes important as measured by the optimality function.

6.2 Example 2: Validation analysis for problem with expectation constraint

We next consider an engineering design problem where the cost of a short structural column needs to be minimized subject to constraints on the failure probability and the aspect ratio; see [34]. The design variables are the width x^1 and depth x^2 of the

Table 2 95%-Confidence intervals in Example 2 for $\psi(x_1)$, $\psi(x_2)$, and $\psi(x_3)$ using (12) with varying sample size N and replications M

N	M	Confidence intervals		
		$\psi(x_1)$	$\psi(x_2)$	$\psi(x_3)$
10^2	30	$(-\infty, 0.1338]$	$(-\infty, 0.9153]$	$(-\infty, 10.1632]$
10^3	30	$(-\infty, 0.0079]$	$(-\infty, 1.0616]$	$(-\infty, 10.1894]$
10^4	30	$(-\infty, -0.0014]$	$(-\infty, 0.8175]$	$(-\infty, 10.2649]$
10^5	100	$(-\infty, -0.0067]$	$(-\infty, 0.7898]$	$(-\infty, 9.9154]$

column. In [35], we find that the failure probability for design $x = (x^1, x^2)$ can be approximated with high-precision by the expression $E[1 - \chi_4^2(r^2(x, \omega))]$, where ω is a four-dimensional standard normal random vector modeling random loads and material property, $\chi_4^2(\cdot)$ is the cumulative distribution function of a Chi-squared distributed random variable with four degrees of freedom, and $r(x, \omega)$ is the minimum distance from $0 \in \mathbb{R}^4$ to a limit-state surface describing the performance of the column given design x and realization ω; see [34,35]. The failure probability is constrained to be no greater than 0.00135. Hence, we set $f^1(x) = E[1 - \chi_4^2(r^2(x, \omega))]/0.00135 - 1$. As in [34], we adopt the objective function $f^0(x) = x^1x^2$ and the additional constraints $f^2(x) = -x^1$, $f^3(x) = -x^2$, $f^4(x) = x^1/x^2 - 2$, and $f^5(x) = 0.5 - x^2/x^1$. In view of results in [35], Assumption 3 holds for this problem instance.

We consider three designs: $x_1 = (0.334, 0.586)'$ is the best point reported in [34]; $x_2 = (0.346, 0.553)'$ is an infeasible solution reported in [34], and $x_3 = (0.586, 0.334)'$ is the "mirror image" of x_1. Table 2 gives 95%-confidence intervals for $\psi(x_1)$, $\psi(x_2)$, and $\psi(x_3)$ for various sample sizes and replications. Table 3 presents confidence intervals for $\theta(x_1), \theta(x_2)$, and $\theta(x_3)$, with $\alpha = 0.1$ in (14) and $\alpha = \beta = 0.05$ in (21). We see that (14) and (21) give comparable results and that a near-optimal solution may require a large sample size to ensure a tight confidence interval.

6.3 Example 3: Optimization and validation analysis for full problem

We illustrate Algorithm 1 by considering the following randomly generated problem instance. Let $n = 20$, $F^0(x, \omega) = \sum_{i=1}^{20} a^i(x^i - b^i\omega^i)^2$, where $a^i = i, b^i = 21 - i, i = 1, 2, \ldots, 20$, and $\omega = (\omega^1, \omega^2, \ldots, \omega^{20})'$ is a vector of independent and uniformly distributed random variables between 0 and 1. $F^1(\cdot, \cdot)$ and $F^2(\cdot, \cdot)$ are defined similarly, but with a^i and b^i being randomly and independently generated from a uniform distribution supported on [0, 10] and [0, 2], respectively. Moreover, we subtract 100 from these expressions to construct constraints of the form $E[\sum_{i=1}^{20} a^i(x^i - b^i\omega^i)^2 - 100] \leq 0$. The resulting instance of P involves 60 independent random variables, an expected value objective function, and two expected value constraint functions.

Table 3 90%-Confidence intervals in Example 2 for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (14) (rows 3–6) and (21) (rows 7–10) with varying sample size N and replications M and K

Method	N	M	K	Confidence intervals		
				$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
(14)	10^2	30	–	$[-0.2597, 0]$	$[-0.8055, 0]$	$[-10.2772, 0]$
	10^3	30	–	$[-0.0554, 0]$	$[-0.7856, 0]$	$[-10.0301, 0]$
	10^4	30	–	$[-0.0074, 0]$	$[-0.8179, 0]$	$[-10.1692, 0]$
	10^5	100	–	$[-0.0014, 0]$	$[-0.7816, 0]$	$[-9.8631, 0]$
(21)	10^2	30	5	$[-0.1540, 0]$	$[-0.9465, 0]$	$[-12.1029, 0]$
	10^3	30	5	$[-0.0595, 0]$	$[-0.8129, 0]$	$[-10.6630, 0]$
	10^4	30	5	$[-0.0031, 0]$	$[-0.8229, 0]$	$[-10.1777, 0]$
	10^5	30	5	$[-0.0003, 0]$	$[-0.8137, 0]$	$[-10.3143, 0]$

We apply Algorithm 1 to this problem instance using $x_0 = 0$, $N_0 = 100$, $\Delta(N) = 1/\sqrt{N}$, and $\delta_1 = \delta_2 = 1$. Moreover, we let $\mathcal{K}(N) = 2N$ and $A_{N\epsilon}(x)$ be the iterate obtained after one iteration of the Polak-He Phase 1-Phase 2 algorithm started from x; see Sect. 2.6 in [29]. We refer to the iterations of Algorithm 1 with the same sample size N as a stage. No smoothing is required as $F^j(\cdot, \omega)$, $j = 0, 1, 2$, are already smooth for all $\omega \in \Omega$. We run Algorithm 1 for ten stages and generate the candidate points $x_0^*, x_{100}^*, x_{200}^*, \ldots, x_{51200}^*$. For each candidate point x_N^*, we compute the confidence intervals (12) and (21) using sample size $10N$ (1,000 for x_0^*), replications $M = 30$ and $K = 23$, and $L = 1$; see Table 4. Columns 2 and 3 give the sample size and number of iterations used in each stage, respectively. Columns 4 and 5 give 95% confidence intervals for $\psi(x_N^*)$ and 90% confidence intervals for $\theta(x_N^*)$, respectively. We also compute two-sided 95% confidence intervals for $f^0(x_N^*)$ using the standard estimator; see column 6. The ten stages require 6,900 seconds of run time. The verification analysis needs 3,300 s.

6.4 Examples 4 and 5: Engineering design optimization

We consider two engineering design problems where the goal is to minimize the design cost subject to a buffered failure probability constraint and other constraints. Hence, the problem instances involve the expectation of a random function of the form (22) as constraint [32]. We note that a buffered failure probability constraint is essentially equivalent to a CVaR constraint [32]. Both examples involve seven design variables and seven random variables. The first design example, referred to as Example 4, is taken from [38] and use 10 random functions ($r^1 = 10$ in (22)). The second example, called Example 5, is taken from [31], pp. 472–473, and involves nine random functions ($r^1 = 9$); see [1] for details. We apply Algorithm 1, setting x_0 equals to the variables upper bounds, $N_0 = 1000$, $\epsilon_N = 1000/N$, and $\mathcal{K}(N) = N + \min\{10^4, \lfloor 0.5N \rfloor\}$. Instead of defining $\Delta(\cdot)$ for the test in Step 2 of Algorithm 1, we simply set $\Delta(N) = 1$ for all N and multiply the parameters δ_1 and δ_2 by a factor $\zeta \in (0, 1)$ after each time

Table 4 95%-Confidence intervals in Example 3 for $\psi(x_N^*)$ and $f^0(x_N^*)$, and 90%-confidence intervals for $\theta(x_N^*)$ for candidate points generated by Algorithm 1

Candidate point	N	#iter.	Confidence intervals		
			$\psi(x_N^*)$	$\theta(x_N^*)$	$f^0(x_N^*)$
x_0^*	100	–	$(-\infty, -48.1472]$	$[-431.1261, 0]$	[5296, 5447)
x_{100}^*	100	302	$(-\infty, -2.0657]$	$[-8.9403, 0]$	[3411, 3533]
x_{200}^*	200	106	$(-\infty, -0.4903]$	$[-3.5880, 0]$	[3439, 3521]
x_{400}^*	400	104	$(-\infty, 0.5280]$	$[-2.0762, 0]$	[3419, 3477]
x_{800}^*	800	149	$(-\infty, 0.0672]$	$[-1.4028, 0]$	[3458, 3498]
x_{1600}^*	1,600	66	$(-\infty, -0.0001]$	$[-0.7915, 0]$	[3453, 3482]
x_{3200}^*	3,200	60	$(-\infty, -0.0107]$	$[-0.4043, 0]$	[3462, 3482]
x_{6400}^*	6,400	75	$(-\infty, 0.0785]$	$[-0.2027, 0]$	[3466, 3481]
x_{12800}^*	12,800	129	$(-\infty, 0.0125]$	$[-0.1082, 0]$	[3470, 3480]
x_{25600}^*	25,600	79	$(-\infty, 0.0607]$	$[-0.1085, 0]$	[3467, 3474]
x_{51200}^*	51,200	99	$(-\infty, 0.0499]$	$[-0.0609, 0]$	[3467, 3472]

Table 5 Sample sizes after 1 h of calculations in Algorithm 1 when applied to Examples 4 and 5 as well as 95%- and 90%-confidence intervals for $\psi(x_N^*)$ and $\theta(x_N^*)$, respectively, at the corresponding solution

Example	Final sample size	Confidence intervals	
		$\psi(x_N^*)$	$\theta(x_N^*)$
4	65624	$(-\infty, 0.0066]$	$[-0.0154, 0]$
5	65624	$(-\infty, 0.2132]$	$[-0.2408, 0]$

both tests are satisfied. We use $\zeta = 0.1$ and 0.8 in Examples 4 and 5, respectively. Since Examples 4 and 5 are more complex than Example 3, we utilize a desktop computer at 3.16 GHz with 3GB of RAM and let $A_{N\epsilon}(x)$ denote the iterate obtained after 20 iterations of SNOPT [8] as implemented in TOMLAB [13], started at x.

Table 5 presents final sample sizes (column 2), 95%-confidence intervals for $\psi(x_N^*)$ (column 3), and 90%-confidence intervals for $\theta(x_N^*)$ (column 4) at the last point obtained by Algorithm 1 after one hour of computations. The confidence intervals are based on (12) and (21) using sample size 10^6, replications $M = 30$ and $K = 5$, and $L = 1$. In the case of Example 4, the obtained design appears feasible and nearly stationary. However, for Example 5, the one hour of calculation time is insufficient to achieve a near-feasible and near-stationary design.

In view of the numerical results, the proposed procedures for estimating the optimality function and constraint violation result in informative confidence intervals. The required sample size and number of replications are typically modest except when estimating $\theta(x)$ for a solution x close to a stationary point, where a large sample size is needed. We also see that the sample-size adjustment rule of Algorithm 1 based on the optimality function yields reasonable results.

7 Conclusions

We have proposed the use of optimality functions for validation analysis and algorithm development in nonlinear stochastic programs with expected value functions as both objective and constraint functions. The validation analysis assesses the quality of a candidate solution $x \in \mathbb{R}^n$ by its proximity to a Fritz-John stationary point as measured by the value of an optimality function at x or, in practice, by a confidence interval for that value. In algorithmic development, optimality functions determine the sample size in variable-sample size schemes. Preliminary numerical tests indicate that the approach is promising.

Acknowledgments The author acknowledges financial supported from Air Force Office of Scientific Research Young Investigator grant F1ATA08337G003 and numerical assistance from G.H. Basova, Turkish Army.

Appendix

Proof of Theorem 1 Let $\mathbb{B}(x, \rho) \stackrel{\triangle}{=} \{x' \in \mathbb{R}^n \mid \|x' - x\| \leq \rho\}$ for any $x \in \mathbb{R}^n$ and $\rho > 0$. Since $f^j(\cdot)$ is finite valued and twice continuously differentiable near $\hat{x}$ and $\nabla^2 f^j(\hat{x})$ is positive definite for all $j \in \mathbf{q}_0$, there exist constants $\hat{\rho} > 0$ and $0 < m \leq 1 \leq M < \infty$ such that $f^j(\cdot)$, $j \in \mathbf{q}_0$, are finite valued and twice continuously differentiable on $\mathbb{B}(\hat{x}, \hat{\rho})$ and that

$$m\|x' - x\|^2 \leq \langle x' - x, \nabla^2 f^j(x)(x' - x)\rangle \leq M\|x' - x\|^2, \tag{30}$$

for all $x \in \mathbb{B}(\hat{x}, \hat{\rho})$, $x' \in \mathbb{R}^n$, and $j \in \mathbf{q}_0$.

For a given $x \in \mathbb{R}^n$, we define $\tilde{\psi}(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ for any $x' \in \mathbb{R}^n$ by $\tilde{\psi}(x, x') \stackrel{\triangle}{=} \max\{f^0(x') - f^0(x), \psi(x')\}$. It follows by the mean value theorem and (30) that for any $x \in \mathbb{B}(\hat{x}, \hat{\rho}) \cap X_\psi$, $x' \in \mathbb{B}(\hat{x}, \hat{\rho})$, and some $s^j \in [0, 1]$, $j \in \mathbf{q}_0$,

$$\begin{aligned}
\tilde{\psi}(x, x') &= \max\Big\{\langle \nabla f^0(x), x' - x\rangle + \tfrac{1}{2}\langle x' - x, \nabla^2 f^0(x + s^0(x' - x))(x' - x)\rangle, \\
&\qquad \times \max_{j \in \mathbf{q}}\{f^j(x) + \langle \nabla f^j(x), x' - x\rangle + \tfrac{1}{2}\langle x' - x, \nabla^2 f^j(x + s^j(x' - x))(x' - x)\rangle\}\Big\} \\
&\leq \frac{1}{M}\max\Big\{\langle \nabla f^0(x), M(x' - x)\rangle + \tfrac{1}{2}\|M(x' - x)\|^2, \\
&\qquad \times \max_{j \in \mathbf{q}}\{f^j(x) + \langle \nabla f^j(x), M(x' - x)\rangle + \tfrac{1}{2}\|M(x' - x)\|^2\}\Big\},
\end{aligned} \tag{31}$$

where we use that $M \geq 1$ and $x \in X_\psi$, and therefore $Mf^j(x) \leq f^j(x)$ for all $j \in \mathbf{q}$.

Let $h(x)$ denote the optimal solution of (2), which according to Theorem 2.2.8 in [29] is unique and continuous as a function of x. Since $\hat{x}$ is a FJ point, $h(\hat{x}) = 0$. Hence, there exists a $\rho' > 0$ such that $\|h(x)\| \leq m\hat{\rho}/2$ for all $x \in \mathbb{B}(\hat{x}, \rho')$. Let $\rho = \min\{\hat{\rho}/2, \rho'\}$. For any $x \in \mathbb{B}(\hat{x}, \rho) \cap X_\psi$, in view of (2) and the property $\psi^+(x) = 0$, the minimization of the right-hand side in (31) with respect to x' yields

an optimal value $\theta(x)/M$. Let $\xi_x \in \mathbb{R}^n$ be the optimal solution of that minimization. Then, due to the equivalence between minimization of the right-hand side in (31) with respect to x' and the minimization in (2), we find that $M(\xi_x - x) = h(x)$. Hence, $\|\xi_x - x\| = \|h(x)\|/M \leq m\hat{\rho}/(2M)$. Moreover, $\|\xi_x - \hat{x}\| \leq \|\xi_x - x\| + \|x - \hat{x}\| \leq m\hat{\rho}/(2M) + \hat{\rho}/2 \leq \hat{\rho}$. We therefore obtain by minimizing the left-hand size of (31) with respect to x' over $\mathbb{B}(\hat{x}, \hat{\rho})$ that

$$\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x') \leq \theta(x)/M \tag{32}$$

for all $x \in \mathbb{B}(\hat{x}, \rho) \cap X_\psi$. Using similar arguments, we also obtain that

$$\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x') \geq \theta(x)/m \tag{33}$$

for all $x \in \mathbb{B}(\hat{x}, \hat{\rho}) \cap X_\psi$.

First, consider an $x \in \mathbb{B}(\hat{x}, \rho) \cap X_\psi$ and let $\hat{x}' \in \mathbb{R}^n$ be the unique optimal solution of $\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x')$. Since $\tilde{\psi}(x, x) = 0$, it follows that $\hat{x}' \in X_\psi$. From (32), we obtain that

$$\begin{aligned} f^0(\hat{x}) - f^0(x) &= \min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \{f^0(x') - f^0(x) \mid \psi(x') \leq 0\} \\ &\leq \min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \{\tilde{\psi}(x, x') \mid \psi(x') \leq 0\} = \tilde{\psi}(x, \hat{x}') \leq \theta(x)/M, \end{aligned}$$

which proves the right-most inequality in (4).

Second, we prove the left-most inequality and consider three cases. Let $x \in \mathbb{B}(\hat{x}, \hat{\rho}) \cap X_\psi$ and $\hat{x}'$ be as in the previous paragraph.

(i) Suppose that $\psi(\hat{x}') < \tilde{\psi}(x, \hat{x}')$ and $f^0(\hat{x}') - f^0(x) = \tilde{\psi}(x, \hat{x}')$. Then,

$$\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x') = \min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \{f^0(x') - f^0(x) \mid \psi(x') \leq 0\} = f^0(\hat{x}) - f^0(x).$$

Hence, by(33), $\theta(x)/m \leq f^0(\hat{x}) - f^0(x)$.

(ii) Suppose that $\psi(\hat{x}') = \tilde{\psi}(x, \hat{x}')$ and $f^0(\hat{x}') - f^0(x) = \tilde{\psi}(x, \hat{x}')$. If $\hat{x}' = \hat{x}$, then we find that $\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x') = \tilde{\psi}(x, \hat{x}) = f^0(\hat{x}) - f^0(x)$. Hence, in view of (33), $\theta(x)/m \leq f^0(\hat{x}) - f^0(x)$. We next consider the possibility $\hat{x} \neq \hat{x}'$ and define $\hat{h} = \hat{x} - \hat{x}'$. Since $\hat{x}'$ is the constrained minimizer of $\tilde{\psi}(x, \cdot)$ over $\mathbb{B}(\hat{x}, \hat{\rho})$, it follows that the directional derivative of $\tilde{\psi}(x, \cdot)$ at $\hat{x}'$ is nonnegative in all feasible directions, i.e., $d\tilde{\psi}(x, \hat{x}'; y - \hat{x}') = \max\{\langle \nabla f^0(\hat{x}'), y - \hat{x}' \rangle, d\psi(\hat{x}', y - \hat{x}')\} \geq 0$, for all $y \in \mathbb{B}(\hat{x}, \hat{\rho})$. By strong convexity of $f^0(\cdot)$ on $\mathbb{B}(\hat{x}, \hat{\rho})$,

$$\langle \nabla f^0(\hat{x}'), \hat{h} \rangle < (f^0(\hat{x}) - f^0(x)) - (f^0(\hat{x}') - f^0(x)) < 0. \tag{34}$$

Consequently,

$$d\psi(\hat{x}', \hat{h}) \geq 0. \tag{35}$$

Now, let $j' \in \hat{\mathbf{q}}(\hat{x}')(= \{j \in \mathbf{q} \mid \psi(\hat{x}') = f^j(\hat{x}')\})$ be such that $d\psi(\hat{x}'; \hat{h}) = \langle \nabla f^{j'}(\hat{x}'), \hat{h}\rangle$. Then, by the mean value theorem and (30), $f^{j'}(\hat{x}) \geq f^{j'}(\hat{x}') + \langle \nabla f^{j'}(\hat{x}'), \hat{h}\rangle + \frac{1}{2}m\|\hat{h}\|^2$. Hence, using (35) and (33), we obtain

$$\psi(\hat{x}) \geq f^{j'}(\hat{x}) \geq \psi(\hat{x}') + d\psi(\hat{x}'; \hat{h}) + \tfrac{1}{2}m\|\hat{h}\|^2 \geq \theta(x)/m + \tfrac{1}{2}m\|\hat{h}\|^2. \tag{36}$$

Since $\psi(\hat{x}) \leq 0$, we find that $\|\hat{h}\| \leq \sqrt{-2\theta(x)}/m$. There exists a constant $c \in (0, \infty)$ such that $\|\nabla f^0(x')\| \leq c/4$ for all $x' \in \mathbb{B}(\hat{x}, \hat{\rho})$. It now follows from (34) and (33) that

$$\begin{aligned} f^0(\hat{x}) - f^0(x) &> f^0(\hat{x}') - f^0(x) + \langle \nabla f^0(\hat{x}'), \hat{h}\rangle \\ &\geq \theta(x)/m - \|\nabla f^0(\hat{x}')\|\|\hat{h}\| \geq (\theta(x) - c\sqrt{-\theta(x)})/m. \end{aligned}$$

(iii) Suppose that $\psi(\hat{x}') = \tilde{\psi}(x, \hat{x}')$ and $f^0(\hat{x}') - f^0(x) < \tilde{\psi}(x, \hat{x}')$. Then, due to the optimality of $\hat{x}'$ for $\tilde{\psi}(x, \cdot)$, $d\psi(\hat{x}', x' - \hat{x}') \geq 0$ for all $x' \in \mathbb{B}(\hat{x}, \hat{\rho})$. Using similar arguments as in (36), we obtain that for any $x' \in \mathbb{B}(\hat{x}, \hat{\rho})$,

$$\begin{aligned} 0 \geq \psi(x') \geq \psi(\hat{x}') + d\psi(\hat{x}'; x' - \hat{x}') + \tfrac{1}{2}m\|x' - \hat{x}'\|^2 \geq \theta(x)/m \\ + \tfrac{1}{2}m\|x' - \hat{x}'\|^2 \end{aligned}$$

and $\|x' - \hat{x}'\| \leq \sqrt{-2\theta(x)}/m$. Hence, $\|\hat{x} - x\| \leq \|\hat{x} - \hat{x}'\| + \|x - \hat{x}'\| \leq 2\sqrt{-2\theta(x)}/m$. It now follows from strong convexity of $f^0(\cdot)$ on $\mathbb{B}(\hat{x}, \hat{\rho})$ and (33) that

$$f^0(\hat{x}) - f^0(x) > \langle \nabla f^0(x), \hat{x} - x\rangle \geq -\|\nabla f^0(x)\|\|\hat{x} - x\| \geq -\frac{c}{m}\sqrt{-\theta(x)}.$$

The left-most inequality (4) now follows as a consequence of these three cases. □

Proof of Theorem 2 The proof is based on the Delta Theorem 7.59 (see also Exercise 5.4, p. 249) in [41]. Let $g : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ be defined for any $\overline{\zeta} = (\zeta_{-1}, \zeta_0', \zeta_1', \ldots, \zeta_q') \in \mathbb{R}^{q+(q+1)n}$, with $\zeta_{-1} \in \mathbb{R}^q$, $\zeta_j \in \mathbb{R}^n$, $j \in \mathbf{q}_0$, by

$$g(\overline{\zeta}) \stackrel{\triangle}{=} -\min_{\mu \in \mathcal{M}} \left\{ \mu^0 w(\overline{\zeta}) + \sum_{j \in \mathbf{q}} \mu^j [w(\overline{\zeta}) - \zeta_{-1}^j] + \tfrac{1}{2} \left\| \sum_{j \in \mathbf{q}_0} \mu^j \zeta_j \right\|^2 \right\},$$

where $w : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ is defined by $w(\overline{\zeta}) \stackrel{\triangle}{=} \max\{0, \max_{j \in \mathbf{q}} \zeta_{-1}^j\}$. Since $\sum_{j \in \mathbf{q}_0} \mu^j = 1$ for all $\mu \in \mathcal{M}$, it follows that $g(\overline{\zeta}) = -w(\overline{\zeta}) - \phi(\overline{\zeta})$,

where $\phi : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ is defined by $\phi(\overline{\zeta}) \triangleq \min_{\mu\in\mathcal{M}}\{-\sum_{j\in\mathbf{q}} \mu^j \zeta_{-1}^j + \frac{1}{2}\|\sum_{j\in\mathbf{q}_0} \mu^j \zeta_j\|^2\}$. Let $\hat{\mathbf{q}}_w(\overline{\zeta}) \triangleq \{j \in \mathbf{q} \mid \max_{k\in\mathbf{q}} \zeta_{-1}^k = \zeta_{-1}^j\}$, and

$$\hat{\mathbf{q}}_w^+(\overline{\zeta}) \triangleq \begin{cases} \hat{\mathbf{q}}_w(\overline{\zeta}) \cup \{0\} & \text{if } w(\overline{\zeta}) = 0 \\ \hat{\mathbf{q}}_w(\overline{\zeta}) & \text{if } w(\overline{\zeta}) > 0 \\ \{0\} & \text{otherwise.} \end{cases}$$

Moreover, let $\hat{\mathcal{M}}_\phi(\overline{\zeta}) \triangleq \{\mu \in \mathcal{M} \mid \phi(\overline{\zeta}) = -\sum_{j\in\mathbf{q}} \mu^j \zeta_{-1}^j + \frac{1}{2}\|\sum_{j\in\mathbf{q}_0} \mu^j \zeta_j\|^2\}$.

It follows from Danskin Theorem; see, for example, Theorem 7.21 in [41], that $w(\cdot)$ and $\phi(\cdot)$ are locally Lipschitz continuous and directional differentiable with directional derivatives at $\overline{\zeta} \in \mathbb{R}^{q+(q+1)n}$ in the direction $\overline{\xi} \in \mathbb{R}^{q+(q+1)n}$ given by $dw(\overline{\zeta}; \overline{\xi}) = \max_{j\in\hat{\mathbf{q}}_w^+(\overline{\zeta})} \xi_{-1}^j$, with $\xi_{-1}^0 \triangleq 0$, and

$$d\phi(\overline{\zeta}; \overline{\xi}) = \min_{\mu\in\hat{\mathcal{M}}_\phi(\overline{\zeta})} \left\{ -\sum_{j\in\mathbf{q}} \mu^j \xi_{-1}^j + \sum_{j\in\mathbf{q}_0} \mu^j \left\langle \sum_{k\in\mathbf{q}_0} \mu^k \zeta_k, \xi_j \right\rangle \right\}.$$

Consequently, $g(\cdot)$ is locally Lipschitz continuous and directional differentiable with directional derivatives at $\overline{\zeta} \in \mathbb{R}^{q+(q+1)n}$ in the direction $\overline{\xi} \in \mathbb{R}^{q+(q+1)n}$ given by

$$dg(\overline{\zeta}; \overline{\xi}) = -\max_{j\in\hat{\mathbf{q}}_w^+(\overline{\zeta})} \xi_{-1}^j - \min_{\mu\in\hat{\mathcal{M}}_\phi(\overline{\zeta})} \left\{ -\sum_{j\in\mathbf{q}} \mu^j \xi_{-1}^j + \sum_{j\in\mathbf{q}_0} \mu^j \left\langle \sum_{k\in\mathbf{q}_0} \mu^k \zeta_k, \xi_j \right\rangle \right\}.$$

Hence, it follows from Proposition 7.57 in [41] that $g(\cdot)$ is Hadamard directional differentiable.

Let $f(x) \triangleq (f^1(x), f^2(x), \ldots, f^q(x))'$, $f_N(x) \triangleq (f_N^1(x), f_N^2(x), \ldots, f_N^q(x))'$, $\nabla\overline{f}(x) \triangleq (\nabla f^0(x)', \nabla f^1(x)', \ldots, \nabla f^q(x)')'$, and $\nabla\overline{f}_N(x) \triangleq (\nabla f_N^0(x)', \nabla f_N^1(x)', \ldots, \nabla f_N^q(x)')'$. Then, by a vector-valued central limit theorem (e.g. Theorem 29.5 in [7]) and Delta Theorem 7.59 in [41], we obtain that

$$N^{1/2}(g((f_N(x), \nabla\overline{f}_N(x)')') - g((f(x), \nabla\overline{f}(x)')')) \Rightarrow dg((f(x), \nabla\overline{f}(x)')'; \overline{Y}(x)).$$

The result now follows from the facts that $g((f_N(x), \nabla\overline{f}_N(x)')') = \theta_N(x)$, $g((f(x), \nabla\overline{f}(x)')') = \theta(x)$, $\hat{\mathbf{q}}_w^+((f(x), \nabla\overline{f}(x)')') = \hat{\mathbf{q}}^+(x)$, and $\hat{\mathcal{M}}_\phi((f(x), \nabla\overline{f}(x)')') = \hat{\mathcal{M}}(x)$ and from rearranging terms. □

Proof of Proposition 6 Let $j \in \mathbf{q}_0$. First, we consider (i). Let $\delta > 0$ be arbitrary. By Theorem 7.48 in [41], $f_N^j(x)$ converges to $f^j(x)$ uniformly on X, as $N \to \infty$, almost surely. Hence, there exists $N_0 \in \mathbb{N}$ such that for all $x \in X$ and $N \geq N_0$, $|f_N^j(x) - f^j(x)| \leq \delta/2$, almost surely. In view of (26), there exists an $N_1 \geq N_0$ such that for all $x \in \mathbb{R}^n$ and $N \geq N_1$, $0 \leq f_{N\epsilon_N}^j(x) - f_N^j(x) \leq \delta/2$, for every $\{\omega_l\}_{l=1}^\infty$, with $\omega_l \in \Omega, l \in \mathbb{N}$. Consequently, for all $x \in X$ and $N \geq N_1$, $|f_{N\epsilon_N}^j(x) - f^j(x)| \leq$

$|f^j_{N\epsilon_N}(x) - f^j_N(x)| + |f^j_N(x) - f^j(x)| \leq \delta$, almost surely, which completes the proof of (i).

Second, we consider (ii) and adopt a similar argument as in Theorems 4.3 and 4.4 of [48] (see also Theorem 2 in [43]). We define the set-valued random function $\mathcal{G} : \mathbb{R}^n \times [0, 1] \times \Omega \to 2^{\mathbb{R}^n}$ by

$$\mathcal{G}^j(x, \epsilon, \omega) \stackrel{\triangle}{=} \begin{cases} \nabla_x F^j_\epsilon(x, \omega), & \text{if } \epsilon > 0 \\ \text{co}_{k \in \hat{\mathbf{r}}^j(x,\omega)} \{\nabla_x g^{jk}(x, \omega)\}, & \text{if } \epsilon = 0, \end{cases}$$

where co$\{\cdot\}$ denotes the convex hull. From (25), we find that for any $k \in \mathbf{r}^j$,

$$\mu^{jk}_\epsilon(x, \omega) = \frac{\exp[(g^{jk}(x, \omega) - F^j(x, \omega))/\epsilon]}{\sum_{k' \in \mathbf{r}^j} \exp[(g^{jk'}(x, \omega) - F^j(x, \omega))/\epsilon]}. \tag{37}$$

Let $\{x_i\}_{i=1}^\infty \subset S$, $\{\epsilon_i\}_{i=1}^\infty \subset (0, 1]$, and $\hat{x} \in S$ be such that $x_i \to \hat{x}$ and $\epsilon_i \to 0$, as $i \to \infty$. Also, let $\omega \in \Omega$ be such that $g^{jk}(\cdot, \omega)$, $k \in \mathbf{r}^j$, are continuously differentiable on S. From (37) we see that if $k \notin \hat{\mathbf{r}}^j(\hat{x}, \omega)$, then $\mu^{jk}_{\epsilon_i}(x_i, \omega) \to 0$, as $i \to \infty$. Moreover, since $\mu^{jk}_{\epsilon_i}(x_i, \omega) \subset (0, 1)$ and $\sum_{k \in \mathbf{r}^j} \mu^{jk}_{\epsilon_i}(x_i, \omega) = 1$ for all $i \in \mathbb{N}$, it follows from (24) that the outer limit of $\{\nabla_x F^j_{\epsilon_i}(x_i, \omega)\}_{i=1}^\infty$ in the sense of Painleve-Kuratowski is contained in $\text{co}_{k \in \hat{\mathbf{r}}^j(\hat{x},\omega)}\{\nabla_x g^{jk}(\hat{x}, \omega)\}$. Hence, it follows that $\mathcal{G}^j(\cdot, \cdot, \omega)$ is outer semi-continuous in the sense of Rockafellar-Wets for almost every $\omega \in \Omega$.

Next, let $\{x_N\}_{N=1}^\infty \subset S$, $\{\epsilon_N\}_{i=1}^\infty \subset (0, 1]$, and $\hat{x} \in S$ be such that $x_N \to \hat{x}$ and $\epsilon_N \to 0$, as $N \to \infty$. Then using the fact that $\mathcal{G}^j(\cdot, \cdot, \omega)$ is outer semi-continuous for almost every $\omega \in \Omega$ and the proofs of Theorems 4.3 and 4.4 in [48], we obtain that $\{\nabla_x f^j_{N\epsilon_N}(x_N)\}$ tends to $E[\text{co}_{k \in \hat{\mathbf{r}}^j(\hat{x},\omega)}\{\nabla_x g^{jk}(\hat{x}, \omega)\}]$, as $N \to \infty$, almost surely. In view of Assumption 3 and Proposition 5, we find that $E[\text{co}_{k \in \hat{\mathbf{r}}^j(\hat{x},\omega)}\{\nabla_x g^{jk}(\hat{x}, \omega)\}] = \{\nabla f^j(\hat{x})\}$ and the result follows. □

References

1. Basova, H.G.: Reliability-based design optimization using buffered failure probability. Master's thesis, Naval Postgraduate School, Monterey, California (2010)
2. Bastin, F., Cirillo, C., Toint, P.L.: An adaptive Monte Carlo algorithm for computing mixed logit estimators. Comput. Manage. Sci. **3**(1), 55–79 (2006)
3. Bastin, F., Cirillo, C., Toint, P.L.: Convergence theory for nonconvex stochastic programming with an application to mixed logit. Math. Program. **108**(2–3), 207–234 (2006)
4. Bayraksan, G., Morton, D.P.: Assessing solution quality in stochastic programs via sampling. In: Tutorials in Operations Research, pp. 102–122. INFORMS (2009)
5. Bayraksan, G., Morton, D.P.: A sequential sampling procedure for stochastic programming. Oper. Res. (2010, to appear)
6. Bettonvil, B., del Castillo, E., Kleijnen, J.P.C.: Statistical testing of optimality conditions in multiresponse simulation-based optimization. Eur. J. Oper. Res. **199**, 448–458 (2009)
7. Billingsley, P.: Probability and Measure. Wiley, New York (1995)
8. Gill, P., Murray, W., Saunders, M.: User's guide for SNOPT 5.3: A Fortran package for large-scale nonlinear programming. Technical Report SOL-98-1, System Optimization Laboratory, Stanford University, Stanford, California (1998)

9. Gürkan, G., Özge, A., Robinson, S.M.: Sample-path solution of stochastic variational inequalities. Math. Program. **84**(2), 313–333 (1999)
10. Higle, J.L., Sen, S.: Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming. Springer, Berlin (1996)
11. Higle, J.L., Sen, S.: Statistical verification of optimality conditions for stochastic programs with recourse. Ann. Oper. Res. **30**, 215–240 (1991)
12. Higle, J.L., Sen, S.: Duality and statistical tests of optimality for two stage stochastic programs. Math. Program. **75**, 257–275 (1996)
13. Holmstrom, K.: Tomlab optimization. http://tomopt.com (2009)
14. Homem-de-Mello, T.: Variable-sample methods for stochastic optimization. ACM Trans. Model. Comput. Simul. **13**(2), 108–133 (2003)
15. Homem-de-Mello, T., Shapiro, A., Spearman, M.L.: Finding optimal material release times using simulation-based optimization. Manage. Sci. **45**(1), 86–102 (1999)
16. Kall, P., Meyer, J.: Stochastic Linear Programming, Models, Theory, and Computation. Springer, Berlin (2005)
17. Kim, S.H., Nelson, B.L.: Selecting the best system. In: Simulation, pp. 501–534. Elsevier, Amsterdam (2006)
18. Kort, B.W., Bertsekas, D.P.: A New Penalty Function Algorithm for Constrained Minimization. In: Proceedings 1972 IEEE Conference on Decision and Control. New Orlean, Louisiana (1972)
19. Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, 2 edn. Springer, Berlin (2003)
20. Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM J. Optim. **19**, 674–699 (2008)
21. Mak, W.K., Morton, D.P., Wood, R.K.: Monte Carlo bounding techniques for determining solution quality in stochastic programs. Oper. Res. Lett. **24**, 47–56 (1999)
22. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
23. Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. SIAM Optim. **17**(4), 969–996 (2006)
24. Norkin, V.I., Pflug, G.C., Ruszczynski, A.: A branch and bound method for stochastic global optimization. Math. Program. **83**, 425–450 (1998)
25. Pasupathy, R.: On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. Oper. Res. **58**(4), 889–901 (2010)
26. Pee, E.Y., Royset, J.O.: On solving large-scale finite minimax problems using exponential smoothing. J. Optim. Theory Appl. **148**(2), 390–421 (2011)
27. Polak, E.: On the mathematical foundations of nondifferentiable optimization in engineering design. SIAM Rev. **29**, 21–89 (1987)
28. Polak, E.: On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems. Math. Program. Ser. B **62**(1–3), 385–414 (1993)
29. Polak, E.: Optimization. Algorithms and Consistent Approximations. Springer, New York (1997)
30. Polak, E., Royset, J.O.: Efficient sample sizes in stochastic nonlinear programming. J. Comput. Appl. Math. **217**, 301–310 (2008)
31. Rao, S.S.: Engineering Optimization Theory and Practice, 4th edn. Wiley, London (2009)
32. Rockafellar, R.T., Royset, J.O.: On buffered failure probability in design and optimization of structures. Reliabil. Eng. Syst. Safety **95**, 499–510 (2010)
33. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Finance **26**, 1443–1471 (2002)
34. Royset, J.O., Der Kiureghian, A., Polak, E.: Reliability-based optimal design of series structural systems. J. Eng. Mech. **127**(6), 607–614 (2001)
35. Royset, J.O., Polak, E.: Extensions of stochastic optimization results from problems with simple to problems with complex failure probability functions. J. Optim. Theory Appl. **133**(1), 1–18 (2007)
36. Sakalauskas, L.L.: Nonlinear stochastic programming by Monte-Carlo estimators. Eur. J. Oper. Res. **137**, 558–573 (2002)
37. Sakalauskas, L.L.: Towards implementable nonlinear stochastic programming. In: Coping with Uncertainty, pp. 257–279. Springer, Berlin (2006)
38. Samson, S., Thoomu, S., Fadel, G., Reneke, J.: Reliable design optimization under aleatory and epistemic uncertainty. In: Proceedings of ASME 2009 International Design Engineering Technical Conferences, pp. DETC2009-86473 (2009)

39. Shapiro, A.: Asymptotic behavior of optimal solutions in stochastic programming. Math. Oper. Res. **18**(4), 829–845 (1993)
40. Shapiro, A.: Testing KKT conditions. Private Communication, June 2 (2003)
41. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. Society of Industrial and Applied Mathematics, Philadelphia (2009)
42. Shapiro, A., Homem-de-Mello, T.: A simulation-based approach to two-stage stochastic programming with recourse. Math. Program. **81**, 301–325 (1998)
43. Shapiro, A., Xu, H.: Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. J. Math. Anal. Appl. **325**, 1390–1399 (2007)
44. Spall, J.C.: Introduction to Stochastic Search and Optimization. Wiley, New York (2003)
45. Tong, X., Qi, L., Wu, F., Zhou, H.: A smoothing method for solving portfolio optimization with CVaR and applications in allocation of generation asset. Appl. Math. Comput. **216**, 1723–1740 (2010)
46. Wallace, S.W., Ziemba, W.T.: Applications of Stochastic Programming. Society for Industrial and Applied Mathematics, Philadelphia (2005)
47. Wang, W., Ahmed, S.: Sample average approximation of expected value constrained stochastic programs. Oper. Res. Lett. **36**(5), 515–519 (2008)
48. Xu, H., Zhang, D.: Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. Math. Program. **119**, 371–401 (2009)