Faculty and Researcher Publications

Faculty and Researcher Publications

2010-10-06

On Optimality Functions in Stochastic Programming and Applications

Royset, J.O.

http://hdl.handle.net/10945/41759

On Optimality Functions in Stochastic Programming and Applications

J.O. Royset[*]

Operations Research Department, Naval Postgraduate School, Monterey, CA 93943, USA

October 6, 2010

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 two algorithms for classes of stochastic programs that include CVaR-problems and utilize optimality functions to select sample sizes as well as "active" sample points in an active-set strategy. Numerical tests illustrate the procedures.

Keywords: Stochastic programming; nonlinear programming; optimality conditions; validation analysis; algorithms.

1 Introduction

Stochastic optimization problems arise in numerous contexts where decisions must be made in the presence of data uncertainty; see the books [18, 12, 31, 26, 62, 57] and references therein for algorithms, models, and applications. In this paper, we deal with a class of stochastic optimization problems defined in terms of expected values of random functions. Let $F^j : \mathbb{R}^n \times \Omega \to \mathbb{R}$, $j = 0, 1, 2, ..., q$, 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 $f^j : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, \infty\}$ be defined by

$$f^j(x) \stackrel{\triangle}{=} E[F^j(x, w)]$$

for all $j \in \mathbf{q}_0 \stackrel{\triangle}{=} \{0\} \cup \mathbf{q}$, with $\mathbf{q} \stackrel{\triangle}{=} \{1, 2, ..., q\}$, where E is the expectation with respect to $\mathcal{P}$. 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 $x \in \mathbb{R}^n$

[*] Tel.: + 1 831 656 2578, fax: +1 831 656 2595, email joroyset@nps.edu.

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 [57] 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 with recourse [26], investment portfolio optimization [49], inventory control [65], and engineering design optimization [51, 48]. Inventory control and engineering design optimization as well as estimation of mixed logit models [4] 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) [49, 48]. 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 [33] and [57], Chapter 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 $x \in \mathbb{R}^n$ for P, which we refer to as validation analysis. In that portion of the paper, we adopt assumptions that ensure a central limit theorem. 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. We are especially motivated by applications involving CVaR and in one algorithm take advantage of the special structure of P in such cases to develop an active-set strategy.

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, [58] 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 [19]. The results in [58] can be extended to constraints defined in terms of expectations [56]. 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, [10] 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 [52, 53], which also consider equality constraints.

Section 5.2 of [57] (see also [55, 17, 4]) 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 the 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 [4]. A similar result for the case with a nonsmooth objective function and deterministic constraints is found in [65]. We find also in Section 5.2 of [57] 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 in stochastic programming is based on estimating bounds on the optimal value of P; see [20, 37, 34, 5, 6]. Estimation of bounds in the case of constraints on expected value functions utilizes the Lagrangian function as described in [63] and [57], 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., [24, 18]), stochastic approximations (see, e.g., [13, 9, 31, 35]), other versions of stochastic search (see, e.g., [60]), and various algorithms based on sample average approximations (SAA) (see, e.g., [57]). 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 [57], Section 5.1 for an overview. The advantage of this approach is it 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 as for example discussed in [22, 4, 51, 39]. 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 [43, 39].

There are several algorithms for the special case of P arising in CVaR minimization and correspondingly constrained problems. In [49], a nonsmooth sample average problem is transcribed into a smooth problem using auxiliary variables. While the smooth problem may have special structure, it is typically large-scale and potentially difficulty to solve by standard algorithms [1]. Alternative approaches solve the sample average problem directly using nonsmooth algorithms

[8, 32, 25, 38], decomposition [14, 30], or smoothing [1, 65, 61].

In this paper, we propose optimization algorithms and validation analysis techniques for P based on *optimality functions*. Optimality functions are optimal values of certain 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 [40, 41, 42]. To the author's knowledge, optimality functions have not been applied previously for validation analysis and algorithm development in stochastic programming. As we see in this paper, 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 implementable algorithms for P based on approximately solving sequences of sample average problems.

The contributions of the paper are four-fold. (i) We introduce an optimality function to the area of stochastic programming and establish the properties of its estimator. (ii) 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. (iii) We construct validation analysis techniques for P based on the optimality function and the FJ conditions. (iv) We develop two implementable algorithms for classes of P and prove their convergence to FJ points. The first algorithm deals with the case when $F^j(\cdot,\cdot)$, $j \in \mathbf{q}_0$, are max-functions and the second considers a situation that arises for example in CVaR applications, which allows the development of an active-set strategy.

Section 2 defines optimality conditions for P in terms of an optimality function and show how that function relates to the distance to a local minimum of P. Section 3 constructs an estimator for the optimality function and derives its asymptotic distribution. Section 4 develops procedures for validation analysis. Section 5 derives two implementable algorithms for P. Section 6 gives illustrative 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. We observe that since $F^j(\cdot, \omega)$, $j \in \mathbf{q}_0$, are random functions, it follows by definition that $F^j(x, \cdot)$, $j \in \mathbf{q}_0$, are measurable for every $x \in \mathbb{R}^n$.

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 $|F^j(x,\omega)| \le C(\omega)$ 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*

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

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 [57]) and allows for certain classes of nonsmoothness in $F^j(\cdot,\omega)$ that may be satisfied in two-stage stochastic programs with recourse [26], CVaR problems [49], inventory control problems [65], and engineering design problems [51] 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 [57] 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 need the following notation. For any vector v, v^j denotes the vector's j-th component. Let

$$\Sigma_q^0 \overset{\triangle}{=} \left\{ \mu \in \mathbb{R}^{q+1} \;\middle|\; \sum_{j \in \mathbf{q}_0} \mu^j = 1, \mu^j \ge 0, j \in \mathbf{q}_0 \right\},$$

$\psi(x) \overset{\triangle}{=} \max_{j \in \mathbf{q}} f^j(x)$, and the constraint violation $\psi^+(x) \overset{\triangle}{=} \max\{0, \psi(x)\}$.

The FJ first-order necessary conditions for P take the following form.

Proposition 1. *If $\hat{x} \in \mathbb{R}^n$ is a local minimizer for P and Assumption 1 holds on an open set $S \subset \mathbb{R}^n$ containing $\hat{x}$, then there exists a multiplier vector $\hat{\mu} \in \Sigma_q^0$ such that*

$$\sum_{j \in \mathbf{q}_0} \hat{\mu}^j \nabla f^j(\hat{x}) = 0 \tag{2}$$

and

$$\sum_{j \in \mathbf{q}} \hat{\mu}^j f^j(\hat{x}) = 0. \tag{3}$$

□

We refer to a point $\hat{x} \in \mathbb{R}^n$ that satisfies (2) and (3) for some $\hat{\mu} \in \Sigma_q^0$ as a FJ point.

We follow [42], see p. 190, and express the FJ conditions by means of a continuous optimality function $\theta : \mathbb{R}^n \to (-\infty, 0]$ defined by

$$\theta(x) \overset{\triangle}{=} \min_{h \in \mathbb{R}^n} \left\{ \max\left\{ -\psi^+(x) + \langle \nabla f^0(x), h\rangle, \max_{j \in \mathbf{q}}\{f^j(x) - \psi^+(x) + \langle \nabla f^j(x), h\rangle\} \right\} + \tfrac{1}{2}\|h\|^2 \right\}. \tag{4}$$

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 (4) takes the following form after simplifications; see Theorem 2.2.8 in [42]:

$$\theta(x) = -\min_{\mu\in\Sigma_q^0} \left\{ \mu^0\psi^+(x) + \sum_{j\in\mathbf{q}} \mu^j[\psi^+(x) - f^j(x)] + \tfrac{1}{2}\Big\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j(x) \Big\|^2 \right\}. \tag{5}$$

It is clear that the optimality function equivalently expresses the FJ conditions in the sense stated next; see Theorem 2.2.8 in [42]. We let $X_\psi \overset{\triangle}{=} \{x \in \mathbb{R}^n \mid \psi(x) \le 0\}$ denote the feasible region of P.

Proposition 2. *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 there exists a multiplier vector $\hat{\mu} \in \Sigma_q^0$ such that (2) and (3) hold.* □

From Proposition 2 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 [42], p. 176, and for two-stage stochastic program with recourse in [19], but the present result is new. We need the notation $\mathbb{B}(x,\rho) \overset{\triangle}{=} \{x' \in \mathbb{R}^n \mid \|x' - x\| \le \rho\}$ for any $x \in \mathbb{R}^n$ and $\rho > 0$.

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*

$$\frac{\theta(x) - c\sqrt{-\theta(x)}}{m} \le f^0(\hat{x}) - f^0(x) \le \theta(x)/M \tag{6}$$

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

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

An examination of 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 (specifically (51) for all $x, x' \in \mathbb{R}^n$), then (6) holds for all $x \in X_\psi$ with $\hat{x}$ being a *global* minimizer.

In view of the above results, the optimality function offers a way of measuring the 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 (5)), 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 the sample average estimators for $f^j(x)$ and $\nabla f^j(x)$, $j \in \mathbf{q}_0$, that leads to validation analysis procedures.

3 Estimator of Optimality Function

Let $\omega_1, \omega_2, ...$ be an infinite sequence of independent random vectors each with value in Ω and distributed as $\mathcal{P}$. Let $\mathbb{N} \overset{\triangle}{=} \{1, 2, 3, ...\}$. We define for any $N \in \mathbb{N}$, $j \in \mathbf{q}_0$, and $x \in \mathbb{R}^n$, the estimators for $f^j(x)$, $\nabla f^j(x)$, $\psi(x)$, and $\psi^+(x)$ by $f_N^j(x) \overset{\triangle}{=} \frac{1}{N}\sum_{l=l}^N F(x, \omega_l)$, $\nabla f_N^j(x) \overset{\triangle}{=} \frac{1}{N}\sum_{l=l}^N \nabla_x F(x, \omega_l)$, $\psi_N(x) \overset{\triangle}{=} \max_{j\in\mathbf{q}} f_N^j(x)$, and $\psi_N^+(x) \overset{\triangle}{=} \max\{0, \psi_N(x)\}$, respectively. We refer to [15] for an overview of alternative approaches to estimating $\nabla f^j(x)$. In some situations it may be possible to use variance reduction techniques to define alternative estimators with smaller variance than those defined above; see for example Section 5.5 in [57]. However, such estimators are beyond the scope of the paper. Finally, we define the estimator of $\theta(x)$ by

$$\begin{aligned}\theta_N(x) \quad \overset{\triangle}{=} \quad \min_{h\in\mathbb{R}^n} \Bigg\{ \max\Big\{ -\psi_N^+(x) + \langle \nabla f_N^0(x), h\rangle, \\ \max_{j\in\mathbf{q}}\{f_N^j(x) - \psi_N^+(x) + \langle \nabla f_N^j(x), h\rangle\}\Big\} + \tfrac{1}{2}\|h\|^2\Bigg\}.\end{aligned}$$

As commonly done, we view $f_N^j(x)$, $j \in \mathbf{q}_0$, $\psi_N(x)$, $\psi_N^+(x)$, and $\theta_N(x)$ as random variables and $\nabla f_N^j(x)$, $j \in \mathbf{q}_0$, as random vectors defined on the product space generated by $(\Omega, \mathcal{F}, \mathcal{P})$ and denote the resulting probability measure by $\overline{\mathcal{P}}$; see Chapter 7 of [57] for further background. Similar to (5), we deduce from Theorem 2.2.8 of [42] the following equivalent and useful expression for $\theta_N(x)$:

$$\theta_N(x) = -\min_{\mu\in\Sigma_q^0} \Big\{\mu^0\psi_N^+(x) + \sum_{j\in\mathbf{q}} \mu^j[\psi_N^+(x) - f_N^j(x)] + \tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0} \mu^j \nabla f_N^j(x)\Big\|^2\Big\}. \tag{7}$$

We next derive properties of $\theta_N(x)$ using proof techniques found, for example, in Chapter 5 of [57]. We start by stating that $\theta_N(x)$ is a strongly consistent estimator of $\theta(x)$. This result is similar to classic results about almost sure convergence of optimal values of sample average problems to the optimal value of an original problem; see, e.g., [28, 47]. The proof follows standard arguments (see for example the proof of Proposition 5.2 in [57]) and we therefore omit it.

Proposition 3. *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 an appropriately shifted and scaled $\theta_N(x)$ for a given $x \in \mathbb{R}^n$ and need the following notation. Let for any $x \in \mathbb{R}^n$,

$$\hat{\Sigma}_q^0(x) \overset{\triangle}{=} \Big\{\mu \in \Sigma_q^0 \;\Big|\; \theta(x) = \mu^0\psi^+(x) + \sum_{j\in\mathbf{q}} \mu^j[\psi^+(x) - f^j(x)] + \tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j(x)\Big\|^2\Big\}, \tag{8}$$

$\hat{\mathbf{q}}(x) \overset{\triangle}{=} \{j \in \mathbf{q} \mid \psi(x) = f^j(x)\}$, and

$$\hat{\mathbf{q}}^+(x) \overset{\triangle}{=} \begin{cases} \hat{\mathbf{q}}(x) \cup \{0\} & \text{if } \psi(x) = 0 \\ \hat{\mathbf{q}}(x) & \text{if } \psi(x) > 0 \\ \{0\} & \text{otherwise.} \end{cases}$$

We use v' to denote the transpose of a vector v and define the following quantities:

$$f(x) \overset{\triangle}{=} (f^1(x), f^2(x), ..., f^q(x))',$$

$$f_N(x) \overset{\triangle}{=} (f_N^1(x), f_N^2(x), ..., f_N^q(x))',$$

$$\nabla \overline{f}(x) \overset{\triangle}{=} (\nabla f^0(x)', \nabla f^1(x)', ..., \nabla f^q(x)')',$$

and

$$\nabla \overline{f}_N(x) \overset{\triangle}{=} (\nabla f_N^0(x)', \nabla f_N^1(x)', ..., \nabla f_N^q(x)')'.$$

We need the following light-tail assumption to ensure a central limit theorem.

Assumption 2. *For a given $x \in \mathrm{IR}^n$, $E[F^j(x,\omega)^2] < \infty$ 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 $i = 1, 2, ..., n$.* □

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

For any $x \in \mathrm{IR}^n$, 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 $(F^1(x,\omega), F^2(x,\omega), ..., F^q(x,\omega), \nabla_x F^0(x,\omega)', \nabla_x F^1(x,\omega)', ..., \nabla_x F^q(x,\omega)')'$. 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 $\overline{Y}(x) = (Y_{-1}(x)', Y_0(x)', Y_1(x)', ..., Y_q(x)')'$.

We use $\Rightarrow$ to denote convergence in distribution. The following vector-valued central limit theorem is well known; see, for example, Theorem 29.5 in [11].

Proposition 4. *Suppose that Assumption 2 holds at a given $x \in \mathrm{IR}^n$ and that Assumption 1 holds on an open set containing $x \in \mathrm{IR}^n$. Then,*

$$N^{1/2}\left(\left(\begin{array}{c} f_N(x) \\ \nabla \overline{f}_N(x) \end{array}\right) - \left(\begin{array}{c} f(x) \\ \nabla \overline{f}(x) \end{array}\right)\right) \Rightarrow \overline{Y}(x),$$

as $N \to \infty$. □

We next examine the asymptotic distribution of a scaled and shifted $\theta_N(x)$. The proof follows by an application of the Delta Theorem 7.59 (see also Exercise 5.4, p. 249) in [57].

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

$$N^{1/2}(\theta_N(x) - \theta(x)) \Rightarrow -\min_{\mu \in \hat{\Sigma}_q^0(x)} \left\{ \mu^0 W(x) + \sum_{j \in \mathbf{q}} \mu^j [W(x) - Y_{-1}^j(x)] + \sum_{j \in \mathbf{q}_0} \mu^j \left\langle \sum_{k \in \mathbf{q}_0} \mu^k \nabla f^k(x), Y_j(x) \right\rangle \right\} \tag{9}$$

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

Proof: See Appendix. □

In general, the right-hand side in (9) is not a normal random variable. Hence, $\theta_N(x)$ cannot be expected to be approximately normal even for large N. In special cases, we find the following interesting corollaries.

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 $\hat{\Sigma}^0_q(x) = \{\hat{\mu}(x)\}$ is a singleton and*

$$\begin{aligned} & N^{1/2}(\theta_N(x) - \theta(x)) \\ \Rightarrow \quad & -\hat{\mu}^0(x)W(x) - \sum_{j\in\mathbf{q}} \hat{\mu}^j(x)[W(x) - Y^j_{-1}(x)] - \sum_{j\in\mathbf{q}_0} \hat{\mu}^j(x)\Big\langle \sum_{k\in\mathbf{q}_0} \hat{\mu}^k \nabla f^k(x), Y_j(x)\Big\rangle, \end{aligned} \tag{10}$$

as $N \to \infty$.

(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 $\hat{\Sigma}^0_q(x) = \{\hat{\mu}(x)\}$ is a singleton and*

$$N^{1/2}\theta_N(x) \Rightarrow -W(x) + \sum_{j\in\hat{\mathbf{q}}^+(x)} \hat{\mu}^j(x)Y^j_{-1}(x) \tag{11}$$

as $N \to \infty$. Moreover, if in addition $\hat{\mathbf{q}}(x) = \{j(x)\}$ is a singleton, then

$$N^{1/2}\theta_N(x) \Rightarrow \begin{cases} -\max\{0, Y^{j(x)}_{-1}\} + \hat{\mu}^{j(x)}(x)Y^{j(x)}_{-1}(x) & \text{if } f^{j(x)}(x) = 0 \\ 0 & \text{if } f^{j(x)}(x) < 0 \end{cases} \tag{12}$$

as $N \to \infty$.

Proof: If the vectors $\nabla f^j(x)$, $j \in \mathbf{q}_0$, are linearly independent, then the matrix $A(x) = (\nabla f^0(x), \nabla f^1(x), ..., \nabla f^q(x))$ has rank $q+1$. Hence, $A(x)'A(x)$ is positive definite and the objective function in (5) is strictly convex. Consequently, $\hat{\Sigma}^0_q(x)$ is a singleton and part (i) follows directly.

Next, consider part (ii). Since $x \in \mathbb{R}^n$ is a local minimizer of P, $\psi(x) \leq 0$ and, from Proposition 2, $\theta(x) = 0$. Hence, it follows from (5) that there exists a $\hat{\mu}(x) \in \hat{\Sigma}^0_q(x)$ 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, $\hat{\Sigma}^0_q(x)$ is a singleton. Since $Y^0_{-1}(x) = 0$ by definition, (9) reduces to (11). Finally, (12) follows from (11). □

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, i.e., $F^j(\cdot, \omega) = F^j(x)$, $j \in \mathbf{q}$, then*

$$N^{1/2}(\theta_N(x) - \theta(x)) \Rightarrow -\min_{\mu\in\hat{\Sigma}^0_q(x)} \mu^0\Big\langle \sum_{k\in\mathbf{q}_0} \mu^k \nabla f^k(x), Y_0(x)\Big\rangle, \tag{13}$$

as $N \to \infty$.

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

We see from (13) that $\theta_N(x)$ is approximately normal when $\hat{\Sigma}_q^0(x)$ is a singleton. Moreover, the limiting distribution degenerates to the constant zero when $\theta(x) = 0$.

The next corollary deals with the special case of no constraints.

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*

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

as $N \to \infty$, where $V_0(x)$ is the n-by-n variance-covariance matrix of $Y_0(x)$ (and $\nabla_x F^0(x, \omega)$) and $\mathcal{N}(0, \sigma^2)$ denotes a zero-mean normal random variable with variance σ^2.

Proof: This result follows from Theorem 2. It can also be shown using Delta Theorem 7.59 in [57] and the fact (see p. 6 in [42]) that in this case we obtain the simplifications

$$\theta(x) = -\tfrac{1}{2}\|\nabla f^0(x)\|^2 \tag{14}$$

and

$$\theta_N(x) = -\tfrac{1}{2}\|\nabla f_N^0(x)\|^2. \tag{15}$$

□

We next consider the bias $\overline{E}\theta_N(x) - \theta(x)$, where $\overline{E}$ denotes the expectation with respect to $\overline{\mathcal{P}}$. Convergence in distribution do not necessarily imply convergence of expectations. Under an uniform integrability property, however, the convergence of expectations is ensured; see for example p. 338 of [11]. The property holds under several assumptions, one of which is used in the next result.

Proposition 5. *Suppose that Assumption 2 holds at a given $x \in \mathbb{R}^n$ and that Assumption 1 holds on an open set containing $x \in \mathbb{R}^n$. Moreover, suppose that there exists an $\epsilon > 0$ such that*

$$\sup_{N \in \mathbb{N}} \overline{E}[|N^{1/2}(\theta_N(x) - \theta(x))|^{1+\epsilon}] < \infty.$$

Then,

$$\begin{aligned} & \overline{E}\theta_N(x) - \theta(x) \qquad\qquad (16)\\ = \;& N^{-1/2}\overline{E}\Big[-\min_{\mu \in \hat{\Sigma}_q^0(x)} \Big\{\mu^0 W(x) + \sum_{j \in \mathbf{q}} \mu^j [W(x) - Y_{-1}^j(x)] + \sum_{j \in \mathbf{q}_0} \mu^j \Big\langle \sum_{k \in \mathbf{q}_0} \mu^k \nabla f^k(x), Y_j(x) \Big\rangle \Big\}\Big] \\ + \;& o(N^{-1/2}). \end{aligned}$$

Moreover, if $\hat{\Sigma}_q^0(x)$ is a singleton, then

$$\overline{E}\theta_N(x) - \theta(x) = -N^{-1/2}\overline{E}[W(x)] + o(N^{-1/2}). \tag{17}$$

Proof: From Theorem 25.12 in [11] and Theorem 2, we directly obtain (16). Since Y_{-1}^j, $j \in \mathbf{q}$, and $Y_j(x)$, $j \in \mathbf{q}_0$, have zero mean and $\sum_{j\in\mathbf{q}_0} \mu^j = 1$ for all $\mu \in \Sigma_q^0$, (17) also holds. □

Conditions that ensure that $\hat{\Sigma}_q^0(x)$ is a singleton is given in Corollary 1. We observe that the bias identified above is similar to the well-known bias of the optimal value of $\min_{x\in X_\psi} f_N^0(x)$ relative to the optimal value of $\min_{x\in X_\psi} f^0(x)$; see, for example p. 167 in [57]. In that case, the bias is always nonpositive. In the present case, $\overline{E}\theta_N(x)$ may be larger than $\theta(x)$. However, in the absence of constraints in P, it follows directly from (14) and (15), and Jensen's inequality that for any $N \in \mathbb{N}$,

$$\overline{E}\theta_N(x) \leq \theta(x). \tag{18}$$

4 Validation Analysis

In this section, we develop procedures for assessing the quality of a candidate point $x \in \mathbb{R}^n$. Specifically, we develop confidence intervals and probabilistic bounds on $\theta(x)$ and $\psi(x)$. Using such bounds, we may claim with some confidence that x satisfies the conditions $\psi(x) \leq \delta$ and $\theta(x) \geq -\epsilon$ for given $\delta \geq 0$ and $\epsilon > 0$. We first consider the situation with no constraints in P, second deal with near feasibility, and third bound the optimality function of the full problem.

4.1 Unconstrained Optimization

Suppose that there are no constraints in P and let $x \in \mathbb{R}^n$ be a candidate solution. In view of Corollary 3, $\theta_N(x)$ is approximately normal with mean $\theta(x)$ and variance $\nabla f^0(x)'V_0(x)\nabla f^0(x)/N$ for large N. Hence, it is straightforward to construct a confidence interval for $\theta(x)$. Let

$$V_{0,N}(x) \overset{\triangle}{=} \frac{1}{N-1}\sum_{l=l}^{N}(\nabla_x F^0(x,\omega_l) - \nabla f_N^0(x))(\nabla_x F^0(x,\omega_l) - \nabla f_N^0(x))'.$$

be the standard unbiased estimator of V_0. Then for large N,

$$\left[\theta_N(x) - z_\alpha\sqrt{\nabla f_N^0(x)'V_{0,N}(x)\nabla f_N^0(x)/N},\ 0\right] \tag{19}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $\theta(x)$, where z_α is the standard normal α-quantile. In (19) and other confidence intervals below we use a quantile of the standard normal distribution instead of one of the t-distribution as the sample size is typically relatively large.

We observe that the approximate normality of $\theta_N(x)$ does not directly reflect the fact that $\theta_N(x) \leq 0$ almost surely. However, in practice, validation analysis is almost always carried out at an $x \in \mathbb{R}^n$ with $\theta(x) < 0$ in which case the truncation at zero is insignificant for large N. Our numerical experiments indicate that the normal model of $\theta_N(x)$ is quite accurate for both $\theta(x) < 0$ and $\theta(x) = 0$; see Section 6. The confidence interval (19) is one-sized, as are the confidence intervals derived below. While it is easy to convert (19) into a two-sided confidence interval, we believe that one-sided confidence intervals are more suitable in the present context as $\theta(x) \geq -\epsilon$ is a natural (though conceptual) criterion for stopping an algorithm applied to P. Hence, if (19) is contained in $[-\epsilon, 0]$, then we would be $100(1-\alpha)\%$ confident that $\theta(x) \geq -\epsilon$ is satisfied.

4.2 Near Feasibility in P

We next consider the full problem P and develop a procedure for determining whether $x \in \mathbb{R}^n$ is nearly feasible, i.e., $\psi(x) \leq \delta$ for some $\delta \geq 0$. We adopt a simple batching approach to estimate the value of $\psi(x)$. In the ranking and selection literature, we find more sophisticated and potentially more efficient ways of determining whether x is nearly feasible; see for example [27] and references therein. It is also possible to estimate $f^j(x)$ independently for each constraints $j \in \mathbf{q}$; see [53]. However, we do not explore those possibilities further.

By Jensen's inequality, we find that $\psi(x) \leq \overline{E}\psi_N(x)$. Hence, a confidence interval for $\overline{E}\psi_N(x)$ provides a conservative confidence interval for $\psi(x)$, which we construct next.

For given N and M, let $\psi_{N,k}(x)$, $k = 1, 2, ..., M$, be independent random variables distributed as $\psi_N(x)$. Then,

$$\overline{\psi}_{N,M}(x) \overset{\triangle}{=} \frac{1}{M} \sum_{k=1}^{M} \psi_{N,k}(x)$$

is an unbiased estimator of $\overline{E}\psi_N(x)$. If $E[F^j(x,\omega)^2] < \infty$ for all $j \in \mathbf{q}$, then a central limit theorem holds for $\overline{\psi}_{N,M}(x)$, i.e., $\overline{\psi}_{N,M}(x)$ is approximately normal with mean $\overline{E}\psi_N(x)$ and variance $Var[\psi_N(x)]/M$ for large M. Let $s^2_{\psi,N,M}(x)$ be the unbiased estimator of $Var[\psi_N(x)]$ given by

$$s^2_{\psi,N,M}(x) = \frac{1}{M-1} \sum_{k=1}^{M} (\psi_{N,k}(x) - \overline{\psi}_{N,M}(x))^2.$$

Then, it follows that

$$(-\infty, \overline{\psi}_{N,M}(x) + z_\alpha s_{\psi,N,M}(x)/\sqrt{M}] \tag{20}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $\overline{E}\psi_N(x)$ for large M and also a conservative $100(1-\alpha)\%$-confidence interval for $\psi(x)$.

4.3 Constrained Optimization

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 (20)), which assesses feasibility exclusively.

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

Proposition 6. *Suppose that Assumption 1 holds on an open set containing a given $x \in \mathbb{R}^n$. Then, for any $\mu \in \Sigma^0_q$,*

$$\theta(x) \geq \overline{E}\Big[-\mu^0 \psi^+_N(x) - \sum_{j \in \mathbf{q}} \mu^j (\psi^+_N(x) - f^j_N(x)) - \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f^j_N(x) \Big\|^2 \Big]. \tag{21}$$

Proof: For any $\mu \in \Sigma_q^0$, let $\tilde{\eta} : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ be defined by

$$\tilde{\eta}(\overline{\zeta}) \overset{\triangle}{=} \max\{0, \max_{j \in \mathbf{q}} \zeta_{-1}^j\} - \sum_{j \in \mathbf{q}} \mu^j \zeta_{-1}^j + \tfrac{1}{2}\Big\| \sum_{j \in \mathbf{q}_0} \mu^j \zeta_j \Big\|^2$$

for any $\overline{\zeta} = (\zeta_{-1}', \zeta_0', \zeta_1', ...\zeta_q') \in \mathbb{R}^{q+(q+1)n}$, with $\zeta_{-1} \in \mathbb{R}^q$ and $\zeta_j \in \mathbb{R}^n$, $j \in \mathbf{q}_0$. Since $\tilde{\eta}(\cdot)$ is convex, it follows from Jensen's inequality that

$$\overline{E}\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')') \geq \tilde{\eta}((f(x)', \nabla \overline{f}(x)')'). \tag{22}$$

From (5) and (22), we see that

$$\tilde{\eta}((f(x)', \nabla \overline{f}(x)')') = \mu^0 \psi^+(x) + \sum_{j \in \mathbf{q}} \mu^j (\psi^+(x) - f^j(x)) + \tfrac{1}{2}\Big\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f^j(x) \Big\|^2 \geq -\theta(x).$$

The result then follows from the fact that $\overline{E}\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')$ equals the negative of the right-hand side in (21). □

In view of Proposition 6, we construct a conservative confidence interval for $\theta(x)$ by computing a confidence interval for the right-hand side in (21). We adopt a batching approach and, for given N and M, let $\eta_{N,k}$, $k = 1, 2, ..., M$, be independent random variables distributed as $\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')$. Then,

$$\overline{\eta}_{N,M} \overset{\triangle}{=} \frac{1}{M} \sum_{k=1}^{M} \eta_{N,k}$$

is an unbiased estimator of $\overline{E}[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$. Under sufficient integrability assumptions for $(f_N(x)', \nabla \overline{f}_N(x)')$, a central limit theorem holds for $\overline{\eta}_{N,M}$ and, consequently, $\overline{\eta}_{N,M}$ is approximately normal with mean $\overline{E}[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ and variance $Var[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]/M$ for large M. Let $s_{\eta,N,M}^2$ be the standard unbiased estimator of $Var[\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ given by

$$s_{\eta,N,M}^2 = \frac{1}{M-1} \sum_{k=1}^{M} (\eta_{N,k} - \overline{\eta}_{N,M})^2.$$

Then, it follows that

$$[-\overline{\eta}_{N,M} - z_\alpha s_{\eta,N,M}(x)/\sqrt{M}, 0] \tag{23}$$

is an approximate $100(1-\alpha)\%$-confidence interval for $\overline{E}[-\tilde{\eta}((f_N(x)', \nabla \overline{f}_N(x)')')]$ for large M and also a conservative $100(1-\alpha)\%$-confidence interval for $\theta(x)$. To compute the above confidence interval, it is necessary to select a $\mu \in \Sigma_q^0$. In view of the proof of Proposition 6, we see that a tighter confidence interval can be expected when $\mu \in \hat{\Sigma}_q^0(x)$. Hence, we recommend to select μ as the optimal solution of (7) for some large N. We note, however, that even when using $\mu \in \hat{\Sigma}_q^0(x)$, the inequality in (21) 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

[36]; see also Section 5.7.2 in [57]. The approach requires a slightly different sampling scheme. While we above use common random numbers, i.e., $f_N^j(x)$, $\nabla f_N^j(x)$, $j \in \mathbf{q}_0$, $\psi_N(x)$, $\psi_N^+(x)$, and $\theta_N(x)$ are computed using the same sample, we now generate a sample of size N for each vector $(f_N^j(x), \nabla f_N^j(x)')$, $j \in \mathbf{q}_0$, independently, and also independently generate a sample of size N to compute $\psi_N(x)$. (Such independent sampling is for example discussed in [57], Chapter 5, Remark 9.) In contrast to the common random number scheme, 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 Section 6, we slightly abuse notation by using the same notation for both sampling schemes. We specifically state when the function-independent sampling scheme is applied.

It is beneficial to "decompose" the optimality function into feasibility and optimality parts. From (4) we see that $\theta(x) = -\psi^+(x) + u(x)$, where

$$u(x) \overset{\triangle}{=} \min_{(h,z)\in\mathbb{R}^{n+1}} \{z + \tfrac{1}{2}\|h\|^2 \mid \langle \nabla f^0(x), h\rangle \le z, f^j(x) + \langle \nabla f^j(x), h\rangle \le z, j \in \mathbf{q}\}. \tag{24}$$

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

$$u_N(x) \overset{\triangle}{=} \min_{(h,z)\in\mathbb{R}^{n+1}} \{z + \tfrac{1}{2}\|h\|^2 \mid \langle \nabla f_N^0(x), h\rangle \le z, f_N^j(x) + \langle \nabla f_N^j(x), h\rangle \le z, j \in \mathbf{q}\}. \tag{25}$$

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. Let $\overline{\mathcal{P}}^*$ denote the probability measure generated by this sampling scheme. Then,*

$$\liminf_{N\to\infty} \overline{\mathcal{P}}^*[u_N(x) \le u(x)] \ge \frac{1}{2^{q+1}}. \tag{26}$$

Proof: Suppose that $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is a feasible point in (24). We want to determine the probability, denoted $\hat{p}_N$, that $(\hat{h}, \hat{z})$ is feasible in (25). Since $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is feasible for (24), we obtain that

$$\begin{aligned} \hat{p}_N &\overset{\triangle}{=} \overline{\mathcal{P}}^*\Big[\Big\{\langle \nabla f_N^0(x), \hat{h}\rangle \le \hat{z}\Big\} \bigcap \Big(\bigcap_{j\in\mathbf{q}} \Big\{f_N^j(x) + \langle \nabla f_N^j(x), \hat{h}\rangle \le \hat{z}\Big\}\Big)\Big] \\ &\ge \overline{\mathcal{P}}^*\Big[\Big\{\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \le 0\Big\} \bigcap \\ &\qquad \Big(\bigcap_{j\in\mathbf{q}} \Big\{f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \le 0\Big\}\Big)\Big]. \end{aligned} \tag{27}$$

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

$$\hat{p}_N \ge \overline{\mathcal{P}}^*\Big[\langle \nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \le 0\Big] \prod_{j\in\mathbf{q}} \overline{\mathcal{P}}^*\Big[f_N^j(x) - f^j(x) + \langle \nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \le 0\Big].$$

By Proposition 4, $N^{1/2}\langle\nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle$ converges in distribution to a zero-mean normal random variable. Hence,

$$\lim_{N\to\infty} \overline{\mathcal{P}}^* \Big[\langle\nabla f_N^0(x) - \nabla f^0(x), \hat{h}\rangle \le 0\Big] \ge 1/2. \tag{28}$$

We observe that the limit in (28) is not equal to 1/2 as the zero-mean normal random variable may have zero variance. 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 $j \in \mathbf{q}$,

$$\lim_{N\to\infty} \overline{\mathcal{P}}^* \Big[f_N^j(x) - f^j(x) + \langle\nabla f_N^j(x) - \nabla f^j(x), \hat{h}\rangle \le 0\Big] \ge 1/2.$$

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 (24), it also holds for the optimal solution in (24). If $(\hat{h}, \hat{z}) \in \mathbb{R}^{n+1}$ is the optimal solution in (24) and it is also feasible in (25), then

$$u_N(x) \le \hat{z} + \tfrac{1}{2}\|\hat{h}\|^2 = u(x).$$

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 [36] in the context of chance constraints.

Let $u_{N,k}(x)$, $k = 1, 2, ..., K$, 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 $\tilde{u}_{N,1}, \tilde{u}_{N,2}, ..., \tilde{u}_{N,K}$, with $\tilde{u}_{N,k} \le \tilde{u}_{N,k+1}$, be this ordered sequence. That is, $\tilde{u}_{N,1}$ is the smallest value of $u_{N,k}(x)$, $k = 1, 2, ..., K$, $\tilde{u}_{N,2}$ is the second smallest, etc. Suppose that $\hat{\gamma}_N$ is a lower bound on $\overline{\mathcal{P}}^*[u_N(x) \le u(x)]$ and suppose that for a given $\beta \in (0, 1)$, K and L satisfy

$$\sum_{k=0}^{L-1} \binom{K}{k} \hat{\gamma}_N^k (1 - \hat{\gamma}_N)^{K-k} \le \beta. \tag{29}$$

Then, using the same arguments as in Section 5.7.2 of [57], we obtain that $\overline{\mathcal{P}}[\tilde{u}_{N,L} > u(x)] \le \beta$. Hence, $[\tilde{u}_{N,L}, 0]$ is a $100(1-\beta)\%$-confidence interval for $u(x)$. In view of Lemma 1 and its proof, we recommend a number slightly smaller than $1/2^{q+1}$ as an estimate of the lower bound $\hat{\gamma}_N$ when N is moderately large.

If the confidence interval for $\psi(x)$ in (20) is computed independently of the confidence interval for $u(x)$, then

$$\Big[-\max\{0, \overline{\psi}_{N,M}(x) + z_\alpha s_{\psi,N,M}(x)/\sqrt{M}\} + \tilde{u}_{N,L},\ 0\Big] \tag{30}$$

is an approximate $100(1-\alpha)(1-\beta)\%$-confidence interval for $\theta(x)$ for large M and N. We observe that the first approach to computing a confidence interval for $\theta(x)$ requires the solution of only one convex quadratic optimization problem to obtain $\mu \in \Sigma_q^0$. The second approach requires K such solutions. If $L = 1$, then $K \ge \log\beta/\log(1-\hat{\gamma}_N)$. Hence, K is typically moderate. For example, if $\beta = 0.01$ and $\hat{\gamma}_N = 0.49$, then $K = 7$ suffices.

5 Algorithms and Consistent Approximations

In this section, we use the optimality function $\theta(\cdot)$ and optimality functions of approximating problems to construct two implementable algorithms for P under additional assumptions on $F^j(\cdot,\omega)$, $j \in \mathbf{q}_0$. The first algorithm deals with the situation where $F^j(\cdot,\omega)$, $j \in \mathbf{q}_0$, are given by the maximum of continuously differentiable random functions. The second algorithm also considers max-functions, but focuses on a more specific class that arises, for instance, in problems involving CVaR. This specialization allows the development of an active-set strategy. 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*

$$F^j(x,\omega) = \max_{k\in\mathbf{r}^j} g^{jk}(x,\omega), j \in \mathbf{q}_0, \tag{31}$$

where $\mathbf{r}^j = \{1,2,...,r^j\}$, $r^j \in \mathbb{N}$, 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 $E[C^j(\omega)] < \infty$, $|g^{jk}(x,\omega)| \le C^j(\omega)$, and $\|\nabla_x g^{jk}(x,\omega)\| \le C^j(\omega)$ 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 $\hat{\mathbf{r}}^j(x,\omega) \overset{\triangle}{=} \{k \in \mathbf{r}^j \mid F^j(x,\omega) = g^{jk}(x,\omega)\}$ 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 (31) 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 7. *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 [42]. 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 [57]. $\square$

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 [57]. 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, following the smoothing approach in [29], 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 [29]; see also [1, 65, 61] 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

$$F_\epsilon^j(x,\omega) \overset{\triangle}{=} \epsilon \log \sum_{k \in \mathbf{r}^j} \exp[g^{jk}(x,\omega)/\epsilon]. \tag{32}$$

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

$$\nabla F_\epsilon^j(x,\omega) = \sum_{k \in \mathbf{r}^j} \mu_\epsilon^{jk}(x,\omega) \nabla_x g^{jk}(x,\omega), \tag{33}$$

where

$$\mu_\epsilon^{jk}(x,\omega) \overset{\triangle}{=} \frac{\exp[g^{jk}(x,\omega)/\epsilon]}{\sum_{k' \in \mathbf{r}^j} \exp[g^{jk'}(x,\omega)/\epsilon]}, k \in \mathbf{r}^j. \tag{34}$$

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

$$0 \le F_\epsilon^j(x,\omega) - F^j(x,\omega) \le \epsilon \log r^j. \tag{35}$$

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

$$f_{N\epsilon}^j(x) \overset{\triangle}{=} \frac{1}{N} \sum_{l=1}^N F_\epsilon^j(x,\omega_l). \tag{36}$$

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

$$P_{N\epsilon}: \quad \min_{x \in \mathbb{R}^n} \{f_{N\epsilon}^0(x) \mid f_{N\epsilon}^j(x) \le 0, j \in \mathbf{q}\}. \tag{37}$$

For given $\epsilon > 0$, $N \in \mathbb{N}$, and realization of $\{\omega_l\}_{l=1}^N$, $P_{N\epsilon}$ is a smooth problem and, hence, can be solved by standard nonlinear programming algorithms. Moreover, if $g^{jk}(\cdot,\omega)$, $j \in \mathbf{q}_0$, $k \in \mathbf{r}^j$, are convex for almost every $\omega \in \Omega$, then $f_{N\epsilon}^j(\cdot)$, $j \in \mathbf{q}_0$, are convex almost surely for any choice of $N \in \mathbb{N}$ and $\epsilon > 0$. 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.

One 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 nonlinear programming algorithm to $P_{N\epsilon}$. In the case of deterministic constraints in P, the results of [65] 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 [65] in one direction by considering a sequence of near-stationary points of $P_{N\epsilon}$ (with expectation constraints) as expressed by optimality functions. In the subsequent subsection, we utilize this result to obtain convergent algorithms that approximately solve 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; see [58, 23, 22, 3, 51, 43, 7, 39] for applications of this idea in SAA and [66, 44, 45] in the area of smoothing of max-functions. 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 as demonstrated in [44]. Hence, we proceed by considering a sequence of smoothed sample average problems with gradually higher precision.

5.2 Consistent Approximations

We analyze $P_{N\epsilon}$ within the framework of consistent approximations (see [42], Section 3.3), which allow 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 $P_{N\epsilon}$.

For any $N \in \mathbb{N}$ and $\epsilon > 0$, let $\theta_{N\epsilon} : \mathbb{R}^n \to (-\infty, 0]$ denote an optimality function for $P_{N\epsilon}$ defined by

$$\theta_{N\epsilon}(x) \overset{\triangle}{=} -\min_{\mu \in \Sigma_q^0} \Big\{ \mu^0 \psi_{N\epsilon}^+(x) + \sum_{j \in \mathbf{q}} \mu^j [\psi_{N\epsilon}^+(x) - f_{N\epsilon}^j(x)] + \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \nabla f_{N\epsilon}^j(x) \Big\|^2 \Big\}, \tag{38}$$

where $\psi_{N\epsilon}^+(x) \overset{\triangle}{=} \max\{\psi_{N\epsilon}(x), 0\}$, with $\psi_{N\epsilon}(x) = \max_{j \in \mathbf{q}} f_{N\epsilon}^j(x)$. Similar results as in Propositions 1 and 2 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 Section 3.3 in [42].

Definition 1. *The elements of the sequence $\{(P_{N\epsilon_N}, \theta_{N\epsilon_N}(\cdot)\}_{N=1}^\infty$ 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 sequence $\{x_N\}_{N=1}^\infty \subset \mathbb{R}^n$ with $x_N \to x$, as $N \to \infty$, $\limsup_{N\to\infty} \theta_{N\epsilon_N}(x_N) \leq \theta(x)$, *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 intermediate result.

Proposition 8. *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: Let $j \in \mathbf{q}_0$. First, we consider (i). Let $\delta > 0$ be arbitrary. By Theorem 7.48 in [57], $f^j_N(x)$ 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 $N \geq N_0$, $|f^j_N(x) - f^j(x)| \leq \delta/2$, almost surely. In view of (35), there exists an $N_1 \geq N_0$ such that for all $x \in \mathbb{R}^n$ and $N \geq N_1$, $0 \leq f^j_{N\epsilon_N}(x) - f^j_N(x) \leq \delta/2$, for every $\{\omega_l\}_{l=1}^\infty$, with $\omega_l \in \Omega$, $l \in \mathbb{N}$. Consequently, for all $x \in X$ and $N \geq N_1$,

$$|f^j_{N\epsilon_N}(x) - f^j(x)| \leq |f^j_{N\epsilon_N}(x) - f^j_N(x)| + |f^j_N(x) - f^j(x)| \leq \delta,$$

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 [65] (see also Theorem 2 in [59]). We define the set-valued random function $\mathcal{G} : \mathbb{R}^n \times [0,1] \times \Omega \to 2^{\mathbb{R}^n}$ by

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

where co$\{\cdot\}$ denotes the convex hull. From (34), we find that for any $k \in \mathbf{r}^j$,

$$\mu^{jk}_\epsilon(x,\omega) = \frac{\exp[(g^{jk}(x,\omega) - F^j(x,\omega))/\epsilon]}{\sum_{k'\in\mathbf{r}^j} \exp[(g^{jk'}(x,\omega) - F^j(x,\omega))/\epsilon]}. \tag{39}$$

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 (39) 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 (33) 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 $\mathrm{co}_{k\in\hat{\mathbf{r}}^j(\hat{x},\omega)}\{\nabla_x g^{jk}(\hat{x},\omega)\}$. 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 [65], we obtain that $\{\nabla_x f^j_{N\epsilon_N}(x_N)\}$ tends to $E[\mathrm{co}_{k\in\hat{\mathbf{r}}^j(\hat{x},\omega)}\{\nabla_x g^{jk}(\hat{x},\omega)\}]$, as $N \to \infty$, almost surely. In view of Assumption 3 and Proposition 7, we find that $E[\mathrm{co}_{k\in\hat{\mathbf{r}}^j(\hat{x},\omega)}\{\nabla_x g^{jk}(\hat{x},\omega)\}] = \{\nabla f^j(\hat{x})\}$ and the result follows. □

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 $\psi_N(x_N) \leq 0$, such that $x_N \to x$, as $N \to \infty$.* □

Theorem 3. *Suppose that Assumptions 3 and 4 hold on an open set $S \subset \mathbb{R}^n$, that $X \subset S$ is compact, and that $X_\psi \subset X$. 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 [42], it follows directly from Proposition 8(i) and Assumption 4 that $\{P_{N\epsilon_N}\}_{N=1}^\infty$ epi-converges to P, as $N \to \infty$, almost surely.

Next, we consider the optimality functions. Let $\eta : \Sigma_q^0 \times X \to \mathbb{R}$ and $\eta_{N\epsilon_N} : \Sigma_q^0 \times X \to \mathbb{R}$ be defined by

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

and

$$\eta_{N\epsilon_N}(\mu, x) \stackrel{\triangle}{=} \mu^0\psi^+_{N\epsilon_N}(x) + \sum_{j\in\mathbf{q}} \mu^j[\psi^+_{N\epsilon_N}(x) - f^j_{N\epsilon_N}(x)] + \tfrac{1}{2}\Big\| \sum_{j\in\mathbf{q}_0} \mu^j \nabla f^j_{N\epsilon_N}(x)\Big\|^2.$$

In view of Proposition 8, $\eta_{N\epsilon_N}(\mu, x)$ converges to $\eta(\mu, x)$ uniformly on $\Sigma_q^0 \times X$, as $N \to \infty$, almost surely. Since $\theta(x) = -\min_{\mu\in\Sigma_q^0} \eta(\mu, x)$ and $\theta_{N\epsilon_N}(x) = -\min_{\mu\in\Sigma_q^0} \eta_{N\epsilon_N}(\mu, x)$, we conclude that $\theta_{N\epsilon_N}(x)$ converges to $\theta(x)$ uniformly on X, as $N \to \infty$, almost surely, which completes the proof.□

As we see in the next section, this result directly leads to an implementable algorithm for P under Assumptions 3 and 4.

5.3 Algorithms

We next construct two algorithms for classes of instances of P that approximately solve 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 solution of $P_{N\epsilon_N}$ by means of the optimality function $\theta_{N\epsilon_N}(\cdot)$. When a point of sufficient precision is obtained for $P_{N\epsilon_N}$, then the algorithm starts solving $P_{N'\epsilon_{N'}}$, where N' is the next integer in $\mathcal{K}$ after N. We allow great flexibility in the choice of optimization algorithm for approximately solving $\{P_{N\epsilon_N}\}_{N\in\mathcal{K}}$. Essentially, all convergent nonlinear programming solvers can be used. For any realization $\{\omega_l\}_{l=1}^\infty$, $N \in \mathbb{N}$, and $\epsilon > 0$, let $A_{N\epsilon} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be an algorithm map that represents a specific number of iterations of a nonlinear programming solver as applied to $P_{N\epsilon}$. We assume that the algorithm map satisfies the following assumption.

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 a sequence $\{x_i\}_{i=0}^\infty$ generated by the algorithm map $A_{N\epsilon}(\cdot)$ using the recursion $x_{i+1} \in A_{N\epsilon}(x_i)$, $i = 0, 1, 2, \ldots$, satisfies $\theta_{N\epsilon}(\hat{x}) = 0$ and $\psi_{N\epsilon}(\hat{x}) \leq 0$.* □

The first algorithm, stated next, is a straightforward adaptation of Algorithm Model 3.3.14 in [42]. 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 $\Delta : \mathbb{N} \to (0, \infty)$ such that $\Delta(N) \to 0$, as $N \to \infty$; an ordered set $\mathcal{K}$ of strictly increasing positive integers with infinite cardinality; a sequence $\{\epsilon_N\}_{N \in \mathcal{K}}$, with $\epsilon_N > 0$ for all $N \in \mathcal{K}$ and $\epsilon_N \to^{\mathcal{K}} 0$, as $N \to \infty$; parameters $\delta_1, \delta_2 > 0$; $N_0 \in \mathcal{K}$; $x_0 \in \mathbb{R}^n$; 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 $x_{i+1} \in A_{N\epsilon_N}(x_i)$.

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 follows from Theorem 3.3.15 in [42]:

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.* □

Since $P_{N\epsilon_N}$ may be computationally expensive to solve for large N, we also construct a second algorithm that utilizes an active-set strategy. The second algorithm deals with a class of instances of P that arises in portfolio optimization with CVaR expressions, engineering design with the buffered failure probability, and other applications. In portfolio optimization, P may take the following form. Let $x^i \in \mathbb{R}$ be allocation of funds to asset i, $i = 1, 2, ..., n-1$, x^n be an auxiliary decision variable, and $\tilde{g} : \mathbb{R}^{n-1} \times \Omega \to \mathbb{R}$ be a continuously differentiable loss function that measures the performance of a portfolio $\tilde{x} \triangleq (x^1, x^2, ..., x^{n-1})$ under market condition $\omega \in \Omega$. Then, in view of [49], the CVaR minimization problem at confidence level $\alpha \in (0, 1)$ takes the form

$$\min_{x \in \mathbb{R}^n} \left\{ x^n + \frac{1}{1-\alpha} E[\max\{\tilde{g}(\tilde{x}, \omega) - x^n, 0\}] \;\middle|\; x \in X \right\}, \tag{40}$$

where $X \subset \mathbb{R}^n$ is a simple feasible region given by deterministic quantities only. Hence, this problem is a special case of P with a function of the form (31). Similarly in engineering design, $\tilde{x} = (x^1, x^2, ..., x^{n-1})$ may be a vector of design variables, x^n be an auxiliary design variable, $\tilde{g}^k : \mathbb{R}^{n-1} \times \Omega$, $k = 1, 2, ..., \tilde{r}$, be limit-state functions describing performance criteria for the system, and $c(\tilde{x})$ be the cost of design $\tilde{x}$. According to [48], the minimum cost problem subject to a buffered failure probability constraint then takes the form

$$\min_{x \in \mathbb{R}^n} \left\{ c(\tilde{x}) \;\middle|\; x^n + \frac{1}{1-\alpha} E[\max\{\max_{k=1,2,...,\tilde{r}} \{\tilde{g}^k(\tilde{x}, \omega) - x^n\}, 0\}] \leq 0, x \in X \right\}, \tag{41}$$

where $X \subset \mathbb{R}^n$ is a set given by deterministic quantities only and $\alpha \in (0,1)$ is a confidence level. While Algorithm 1 applies to these cases, we derive a second algorithm that takes advantage of the special structures that arise in these applications. We start by adopting the following assumption.

Assumption 6. *For all $j \in \mathbf{q}_0$ and $k \in \mathbf{r}^j$, $g^{jk}(\cdot, \cdot)$ takes the form*

$$g^{jk}(x, \omega) = \begin{cases} \phi^j(x) + \tilde{g}^{jk}(x, \omega), & \text{if } k \in \tilde{\mathbf{r}}^j \overset{\triangle}{=} \{1, 2, ..., r^j - 1\} \\ \phi^j(x), & \text{if } k = r^j, \end{cases}$$

where $\phi^j : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable functions and $\tilde{g}^{jk} : \mathbb{R}^n \times \Omega \to \mathbb{R}$, $k \in \tilde{\mathbf{r}}^j$, satisfy Assumption 3, with $g^{jk}(\cdot, \cdot)$ replaced by $\tilde{g}^{jk}(\cdot, \cdot)$. □

We note that under Assumptions 3 and 6, $f^j(x) = \phi^j(x) + E[\max\{\max_{k \in \tilde{\mathbf{r}}^j} \tilde{g}^{jk}(x, \omega), 0\}]$, $j \in \mathbf{q}_0$, and, consequently, Assumption 6 encapsulates (40) and (41).

The second algorithm utilizes the situation that under Assumption 6, we may have that $\tilde{g}^{jk}(x, \omega) < 0$ for all $k \in \tilde{\mathbf{r}}^j$, and many $x \in \mathbb{R}^n$ and $\omega \in \Omega$. This happens, for example, in design of highly reliable engineering systems. Hence, in that case, $f^j(x)$ essentially equals $\phi^j(x)$, which is a smooth deterministic function. Given realization $\{\omega_l\}_{l=1}^{\infty}$, we exploit this situation by partially ignoring any ω_l with $\tilde{g}^{jk}(x, \omega_l) < 0$ for all $k \in \tilde{\mathbf{r}}^j$ at a current point $x \in \mathbb{R}^n$. This may result in a significant reduction in the computing effort required per application of the algorithm map. To carry out this plan, we need the following notation.

For any $N \in \mathbb{N}$, $j \in \mathbf{q}_0$, $\epsilon > 0$, $\mathbf{N}^j \subset \{1, 2, ..., N\}$, and $\rho^j = \{\rho^{jl}\}_{l \in \mathbf{N}^j}$, with $\rho^{jl} \subset \tilde{\mathbf{r}}^j$, for all $l \in \mathbf{N}^j$, we define $\tilde{f}^j_{N\epsilon}(\cdot; \mathbf{N}^j, \rho^j) : \mathbb{R}^n \to \mathbb{R}$ as

$$\tilde{f}^j_{N\epsilon}(x; \mathbf{N}^j, \rho^j) \overset{\triangle}{=} \phi^j(x) + \frac{1}{N} \sum_{l \in \mathbf{N}^j} \tilde{F}^j_\epsilon(x, \omega_l; \rho^{jl}), \tag{42}$$

where for all $l \in \mathbf{N}^j$, $\tilde{F}^j_\epsilon(\cdot, \omega_l; \rho^{jl}) : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\tilde{F}^j_\epsilon(x, \omega_l; \rho^{jl}) \overset{\triangle}{=} \epsilon \log \left(1 + \sum_{k \in \rho^{jl}} \exp[\tilde{g}^{jk}(x, \omega_l)/\epsilon] \right).$$

We observe that $\tilde{F}^j_\epsilon(x, \omega_l; \rho^{jl})$ is the smooth approximation of $\max\{\max_{k \in \rho^{jl}} \tilde{g}^{jk}(x, \omega_l), 0\}$ using the exponential smoothing technique (32). If $\tilde{g}^{jk}(x, \omega_l) < 0$ for $k \notin \rho^{jl}$, then $\tilde{F}^j_\epsilon(x, \omega_l; \rho^{jl})$ is also a smooth approximation of $\max\{\max_{k \in \tilde{\mathbf{r}}^j} \tilde{g}^{jk}(x, \omega_l), 0\}$. If $\max_{k \in \tilde{\mathbf{r}}^j} \tilde{g}^{jk}(x, \omega_l) < 0$ for $l \notin \mathbf{N}^j$ and $\tilde{g}^{jk}(x, \omega_l) < 0$ for $k \notin \rho^{jl}$ and $l \in \mathbf{N}^j$, then the samples excluded in $\mathbf{N}^j$ do not contribute significantly. Hence, under a strong law of large numbers and with the proper construction of the "active" sets $\mathbf{N}^j$ and ρ^j, $\sum_{l\in\mathbf{N}^j}\tilde{F}^j_\epsilon(x,\omega_l;\rho^{jl})/N$ is a smooth approximation of $E[\max\{\max_{k\in\tilde{\mathbf{r}}^j}\tilde{g}^{jk}(x,\omega_l),0\}]$ and $\tilde{f}^j_{N\epsilon}(x;\mathbf{N}^j,\rho^j)$ is a smooth approximation of $f^j(x)$. The second algorithm approximately solves a sequence of problems involving such approximations.

For any realization $\{\omega_l\}_{l=1}^\infty$, $N\in\mathbb{N}$, $\epsilon>0$, $\mathbf{N}\overset{\triangle}{=}\{\mathbf{N}^j\}_{j\in\mathbf{q}_0}$, with $\mathbf{N}^j\subset\{1,2,...,N\}$ for all $j\in\mathbf{q}_0$, and $\rho\overset{\triangle}{=}\{\rho^{jl}\}_{l\in\mathbf{N}^j,j\in\mathbf{q}_0}$, with $\rho^{jl}\subset\tilde{\mathbf{r}}^j$ for all $l\in\mathbf{N}^j$ and $j\in\mathbf{q}_0$, we define the approximate problem

$$\tilde{P}_{N\epsilon}(\mathbf{N},\rho):\quad \min_{x\in\mathbb{R}^n}\{\tilde{f}^0_{N\epsilon}(x;\mathbf{N}^0,\rho^0)\mid \tilde{f}^j_{N\epsilon}(x;\mathbf{N}^j,\rho^j)\leq 0, j\in\mathbf{q}\}. \tag{43}$$

If $\mathbf{N}^j=\{1,2,...,N\}$ and $\rho^{jl}=\tilde{\mathbf{r}}^j$ for all $j\in\mathbf{q}_0$ and $l\in\mathbf{N}^j$, then $\tilde{P}_{N\epsilon}(\mathbf{N},\rho)$ is identical to $P_{N\epsilon}$. By considering strict subsets of $\{1,2,...,N\}$ and $\mathbf{r}^j$, $j\in\mathbf{q}_0$, we hope to reduce computational effort.

We define the optimality function $\tilde{\theta}_{N\epsilon}(\cdot;\mathbf{N},\rho):\mathbb{R}^n\to(-\infty,0]$ for $\tilde{P}_{N\epsilon}(\mathbf{N},\rho)$ to be given by

$$\begin{aligned}\tilde{\theta}_{N\epsilon}(x;\mathbf{N},\rho) \quad &\overset{\triangle}{=}\quad -\min_{\mu\in\Sigma^0_q}\Big\{\mu^0\tilde{\psi}^+_{N\epsilon}(x;\mathbf{N},\rho) \\ &+\quad \sum_{j\in\mathbf{q}}\mu^j[\tilde{\psi}^+_{N\epsilon}(x;\mathbf{N},\rho)-\tilde{f}^j_{N\epsilon}(x;\mathbf{N}^j,\rho^j)]+\tfrac{1}{2}\Big\|\sum_{j\in\mathbf{q}_0}\mu^j\nabla\tilde{f}^j_{N\epsilon}(x;\mathbf{N}^j,\rho^j)\Big\|^2\Big\},\end{aligned} \tag{44}$$

where $\tilde{\psi}^+_{N\epsilon}(x;\mathbf{N},\rho)\overset{\triangle}{=}\max\{0,\tilde{\psi}_{N\epsilon}(x;\mathbf{N},\rho)\}$, with $\tilde{\psi}_{N\epsilon}(x;\mathbf{N},\rho)\overset{\triangle}{=}\max_{j\in\mathbf{q}}\tilde{f}^j_{N\epsilon}(x;\mathbf{N}^j,\rho^j)$. In view of Proposition 2, we deduce that if $x\in\mathbb{R}^n$ is feasible in $\tilde{P}_{N\epsilon}(\mathbf{N},\rho)$, then $\tilde{\theta}_{N\epsilon}(\cdot;\mathbf{N},\rho)=0$ if and only if x is a FJ point of $\tilde{P}_{N\epsilon}(\mathbf{N},\rho)$.

We denote by $\tilde{A}_{N\epsilon}(\cdot;\mathbf{N},\rho):\mathbb{R}^n\to 2^{\mathbb{R}^n}$ an algorithm map that represents a specific number of iterations of a nonlinear programming solver as applied to $\tilde{P}_{N\epsilon}(\mathbf{N},\rho)$ and that satisfies the following assumption.

Assumption 7. *The following holds almost surely. For any $N\in\mathbb{N}$, $\epsilon>0$, $\mathbf{N}=\{\mathbf{N}^j\}_{j\in\mathbf{q}_0}$, with $\mathbf{N}^j\subset\{1,2,...,N\}$, $j\in\mathbf{q}_0$, and $\rho=\{\rho^{jl}\}_{l\in\mathbf{N}^j,j\in\mathbf{q}_0}$, with $\rho^{jl}\subset\tilde{\mathbf{r}}^j$, $l\in\mathbf{N}^j$ and $j\in\mathbf{q}_0$, every accumulation point $\hat{x}\in\mathbb{R}^n$ of a sequence $\{x_i\}_{i=0}^\infty$ generated by the algorithm map $\tilde{A}_{N\epsilon}(\cdot;\mathbf{N},\rho)$ using the recursion $x_{i+1}\in\tilde{A}_{N\epsilon}(x_i;\mathbf{N},\rho)$, $i=0,1,2,...$, satisfies $\tilde{\theta}_{N\epsilon}(\hat{x};\mathbf{N},\rho)=0$ and $\tilde{\psi}_{N\epsilon}(\hat{x};\mathbf{N},\rho)\leq 0$.*□

We are now ready to state the second algorithm, which generalizes Algorithm 1 by considering $\tilde{P}_{N\epsilon_N}(\mathbf{N},\rho)$ instead of $P_{N\epsilon_N}$ and by constructing and updating the "active sets" $\mathbf{N}$ and ρ.

Algorithm 2 (Solves P under Assumptions 3, 4, 6, and 7)

Input. Function $\Delta:\mathbb{N}\to(0,\infty)$ such that $\Delta(N)\to 0$, as $N\to\infty$; an ordered set $\mathcal{K}$ of strictly increasing positive integers with infinite cardinality; a sequence $\{\epsilon_N\}_{N\in\mathcal{K}}$, with $\epsilon_N>0$ for all $N\in\mathcal{K}$ and $\epsilon_N\to^{\mathcal{K}}0$, as $N\to\infty$; parameters $\delta_1,\delta_2>0$ and $\gamma_2\geq\gamma_1>0$; $N_{-1}\in\mathcal{K}$; $x_0\in\mathbb{R}^n$; and realizations $\{\omega_l\}_{l=1}^\infty$ obtained by independent sampling from $\mathcal{P}$.

Step 0. Set $i=-1$, $x^*_0=x_0$, $N=N_{-1}$, $\mathbf{N}^j_{-1}=\emptyset$, $\rho^{jl}_{-1}=\emptyset$, $j\in\mathbf{q}_0$, $l\in\{1,2,...,N\}$, and go to Step 2.

Step 1. Compute $x_{i+1} \in \tilde{A}_{N\epsilon_N}(x_i; \mathbf{N}_i, \rho_i)$.

Step 2. Compute active sets $\mathbf{N}_{i+1} = \{\mathbf{N}_{i+1}^j\}_{j \in \mathbf{q}_0}$, where $\mathbf{N}_{i+1}^j = \mathbf{N}_i^j \cup \hat{\mathbf{N}}_{i+1}^j$, with

$$\hat{\mathbf{N}}_{i+1}^j = \{l \in \mathbb{N} \mid \max_{k \in \tilde{\mathbf{r}}^j} \tilde{g}^{jk}(x_{i+1}, \omega_l) \geq -\gamma_1, l \leq N\}, j \in \mathbf{q}_0, \tag{45}$$

and $\rho_{i+1} = \{\rho_{i+1}^{jl}\}_{l \in \mathbf{N}_{i+1}^j, j \in \mathbf{q}_0}$, where $\rho_{i+1}^{jl} = \rho_i^{jl} \cup \hat{\rho}_{i+1}^{jl}$, with

$$\hat{\rho}_{i+1}^{jl} = \{k \in \tilde{\mathbf{r}}^j \mid \tilde{g}^{jk}(x_{i+1}, \omega_l) \geq -\gamma_2\}, j \in \mathbf{q}_0, l \in \mathbf{N}_{i+1}^j. \tag{46}$$

Step 3. If

$$\tilde{\theta}_{N\epsilon_N}(x_{i+1}; \mathbf{N}_{i+1}, \rho_{i+1}) \geq -\delta_1 \Delta(N) \tag{47}$$

and

$$\tilde{\psi}_{N\epsilon_N}(x_{i+1}; \mathbf{N}_{i+1}, \rho_{i+1}) \leq \delta_2 \Delta(N), \tag{48}$$

then set $x_N^* = x_{i+1}$, $\mathbf{N}_N^* \overset{\triangle}{=} \{\mathbf{N}_N^{*j}\}_{j \in \mathbf{q}_0}$, with $\mathbf{N}_N^{*j} = \mathbf{N}_{i+1}^j$, $j \in \mathbf{q}_0$, $\rho_N^* \overset{\triangle}{=} \{\rho_N^{*j}\}_{j \in \mathbf{q}_0}$, with $\rho_N^{*j} \overset{\triangle}{=} \{\rho_N^{*jl}\}_{l \in \mathbf{N}_N^{*j}}$ and $\rho_N^{*jl} = \rho_{i+1}^{jl}$, $l \in \mathbf{N}_N^{*j}$, $j \in \mathbf{q}_0$, replace N by $\mathcal{K}(N)$, and go to Step 4. Else, go to Step 5.

Step 4. Reset active sets by computing $\mathbf{N}_{i+1} = \{\mathbf{N}_{i+1}^j\}_{j \in \mathbf{q}_0}$, where $\mathbf{N}_{i+1}^j$ equals the right-hand side of (45) and $\rho_{i+1} = \{\rho_{i+1}^{jl}\}_{l \in \mathbf{N}_{i+1}^j, j \in \mathbf{q}_0}$, where ρ_{i+1}^{jl} equals the right-hand side of (46).

Step 5. Replace i by $i + 1$, and go to Step 1. □

In Algorithm 2, the active sets are monotonically increasing as long as the sample size N remains fixed; see Step 2. However, when the sample size is increased, the active sets are reset to only include those elements that are near-active at the current iterate; see Step 4. While the resetting is not required by the convergence proof below, we find it beneficial computationally as it reduces the sizes of the active sets. We need the following intermediate results before we state the convergence result for Algorithm 2.

Lemma 2. *Suppose that Assumptions 3 and 4 hold on a sufficiently large open subset of* $\mathbb{R}^n$ *and that Assumptions 6 and 7 are also satisfied. Moreover, suppose that Algorithm 2 has generated an infinite bounded sequence* $\{x_N^*\}_{N \in \mathcal{K}}$*. Then, for every accumulation point* $\hat{x}$ *of* $\{x_N^*\}_{N \in \mathcal{K}}$ *there exists a* $K \subset \mathcal{K}$ *such that* $x_N^* \to^K \hat{x}$*, as* $N \to \infty$*, and*

$$\tilde{f}_{N\epsilon_N}^j(x_N^*; \mathbf{N}_N^{*j}, \rho_N^{*j}) \to^K f^j(\hat{x}), j \in \mathbf{q}_0, \tag{49}$$

and

$$\nabla \tilde{f}_{N\epsilon_N}^j(x_N^*; \mathbf{N}_N^{*j}, \rho_N^{*j}) \to^K \nabla f^j(\hat{x}), j \in \mathbf{q}_0, \tag{50}$$

as $N \to \infty$*, almost surely.*

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

Lemma 3. *Suppose that Assumptions 3 and 4 hold on a sufficiently large open subset of* $\mathbb{R}^n$ *and that Assumptions 6 and 7 are also satisfied. Moreover, suppose that Algorithm 2 has generated an infinite bounded sequence* $\{x_N^*\}_{N\in\mathcal{K}}$*. Then, for every accumulation point* $\hat{x}$ *of* $\{x_N^*\}_{N\in\mathcal{K}}$ *there exists a* $K \subset \mathcal{K}$ *such that* $x_N^* \to^K \hat{x}$*, as* $N \to \infty$*, and*

$$\tilde{\theta}_{N\epsilon_N}(x_N^*; \mathbf{N}_N^*, \rho_N^*) \to^K \theta(\hat{x}),$$

as $N \to \infty$*, almost surely.*

Proof: In view of Lemma 2, the conclusion follows using similar arguments to those in the proof of Theorem 3. □

Convergence of Algorithm 2 is ensured by the next result.

Theorem 5. *Suppose that Assumptions 3 and 4 hold on a sufficiently large open subset of* $\mathbb{R}^n$ *and that Assumptions 6 and 7 are also satisfied. Moreover, suppose that Algorithm 2 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* $\psi(\hat{x}) \le 0$ *almost surely.*

Proof: Suppose that $\{x_N^*\}$ is a finite sequence. Then there exists $i_1 \in \mathbb{N}$ such that either (47) or (48) in Step 3 fail for all $i \ge i_1$. Hence, the sample size N is constant for all $i \ge i_1$. Also, for all $j \in \mathbf{q}_0$ and $i > i_1$, $\mathbf{N}_i^j$ and ρ_i^{jl}, $l \in \mathbf{N}_i^j$, are monotonically increasing. Since $\mathbf{N}_i^j \subset \{1, 2, ..., N\}$ and $\rho_i^{jl} \subset \tilde{\mathbf{r}}^j$, $l \in \mathbf{N}_i^j$, $j \in \mathbf{q}_0$, it follows that there exists an $i_2 > i_1$ such that $\mathbf{N}_i = \mathbf{N}_{i_2}$ and $\rho_i = \rho_{i_2}$ for all $i \ge i_2$. Since $\{x_i\}_{i=0}^{\infty}$ is bounded, it has an accumulation point $\hat{x} \in \mathbb{R}^n$ that, in view of Assumption 7, satisfies $\tilde{\theta}_{N\epsilon_N}(\hat{x}; \mathbf{N}_{i_2}, \rho_{i_2}) = 0$ and $\tilde{\psi}_{N\epsilon_N}(\hat{x}; \mathbf{N}_{i_2}, \rho_{i_2}) \le 0$. Hence, there exists $i_3 \ge i_2$ such that $\tilde{\theta}_{N\epsilon_N}(x_{i_3}; \mathbf{N}_{i_2}, \rho_{i_2}) \ge -\delta_1 \Delta(N)$ and $\tilde{\psi}_{N\epsilon_N}(x_{i_3}; \mathbf{N}_{i_2}, \rho_{i_2}) \le \delta_2 \Delta(N)$, which contradicts the fact that (47) and/or (48) fail for all $i \ge i_1$. Consequently, $\{x_N^*\}$ is an infinite sequence.

Suppose that $\hat{x} \in \mathbb{R}^n$ is an accumulation point of $\{x_N^*\}_{N\in\mathcal{K}}$. Let $K \subset \mathcal{K}$ be such that $x_N^* \to^K \hat{x}$, as $N \to \infty$. Then, by construction

$$\tilde{\theta}_{N\epsilon_N}(x_N^*; \mathbf{N}_N^*, \rho_N^*) \to^K 0,$$

as $N \to \infty$. Using this fact and Lemma 3, we obtain the conclusions that

$$|\theta(\hat{x})| \le |\theta(\hat{x}) - \tilde{\theta}_{N\epsilon_N}(x_N^*; \mathbf{N}_N^*, \rho_N^*)| + |\tilde{\theta}_{N\epsilon_N}(x_N^*; \mathbf{N}_N^*, \rho_N^*)| \to^K 0,$$

as $N \to \infty$, almost surely. A similar argument shows that $\psi(\hat{x}) \le 0$ almost surely. □

Algorithms 1 and 2 do not include a stopping criterion. One might run Algorithms 1 and 2 until a predetermined computing budget is exhausted and then carry out validation analysis on the candidate points $\{x_i\}$ or a subset thereof using a sample that is independent of the one used in Algorithms 1 and 2.

	Confidence intervals			
N	$\theta(\hat{x})$	$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
10^2	$[-259.0865, 0]$	$[-459.8441, 0]$	$[-696.60, 0]$	$[-6459, 0]$
10^3	$[-76.3627, 0]$	$[-31.1599, 0]$	$[-83.49, 0]$	$[-6070, 0]$
10^4	$[-2.8845, 0]$	$[-6.0065, 0]$	$[-67.19, 0]$	$[-5793, 0]$
10^5	$[-0.2897, 0]$	$[-0.6515, 0]$	$[-62.09, 0]$	$[-5774, 0]$
10^6	$[-0.0427, 0]$	$[-0.5771, 0]$	$[-57.98, 0]$	$[-5747, 0]$
10^7	$[-0.0043, 0]$	$[-0.4617, 0]$	$[-57.55, 0]$	$[-5743, 0]$
∞	0	-0.4420	-57.40	-5740

Table 1: 95%-confidence intervals in Example 1 for $\theta(\hat{x})$, $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (19) with varying sample size N. The last row gives the corresponding true values.

6 Numerical Examples

In this section, we present numerical tests of Algorithms 1 and 2 as well as the validation analysis procedures in Section 4 as applied to six examples. 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 Unconstrained Problem

We consider an instance of P where there are no constraints, $n = 20$, and $F^0(x, \omega) = \sum_{i=1}^{20} a^i (x^i - b^i \omega^i)^2$, where $a^i = i$, $b^i = 21 - i$, $i = 1, 2, ..., 20$, and $\omega = (\omega^1, \omega^2, ..., \omega^{20})'$ is a vector of independent and uniformly distributed random variables between 0 and 1. In this instance, both $\nabla f^0(x)$ and the unique global minimizer $\hat{x} = (10, 9.5, 9, 8.5, ..., 0.5)'$ are easily computed analytically. However, we still use the validation analysis of Section 4.1 and compare the resulting confidence interval of $\theta(x)$ with the true value of $\theta(x)$. Assumption 3 holds for this problem instance.

We consider four candidate points: the optimal solution $\hat{x}$, a near-optimal point x_1 = (10.0029, 9.4866, 9.0071, 8.5162, 7.9931, 7.5086, 7.0125, 6.4841, 5.9856, 5.5057, 4.9960, 4.5069, 4.0082, 3.5071, 3.0129, 2.5067, 2.0119, 1.4880, 0.9998, 0.4984)$'$ obtained by randomly perturbing $\hat{x}$, a further-away point $x_2 = (9.9, 9.4, 8.9, ..., 0.4)'$, and a relatively far-away point $x_3 = (9, 8.5, 8, ..., -0.5)'$.

Table 1 gives 95%-confidence intervals for $\theta(\hat{x})$, $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ in columns 2-5, respectively, using (19) with varying sample size N. The last row gives the corresponding true values. We observe that the confidence intervals cover the true 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 optimality function is close to zero, tightness can only be obtained by using a large sample size.

We also apply a hypothesis test based on a Chi-square statistic proposed in [58]. The test involves the null hypothesis that the current point satisfies the KKT conditions and the alternative hypothesis that they are not. For $\hat{x}$, we compute a p-value of 0.20 using a sample size of $N = 10^5$. Hence, with a test size of (for example) 0.05, we are unable to reject the null hypothesis. For

x_1, x_2, and x_3, we compute p-values of essentially zero. Hence, in those cases we reject the null hypothesis with high confidence. 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. While [58] provides expressions for a confidence region for $\nabla f^0(x)$ that can be computed and compared with a user-defined region containing $0 \in \mathbb{R}^n$, it is more natural and convenient to condense $\nabla f^0(x)$ into a single number as achieved with the optimality function. In view of Section 4, the approach based on the optimality function generalizes to constrained problems as illustrated next.

6.2 Example 2: Validation Analysis for Deterministically Constrained Problem

The next problem instance generalizes a classical problem arising in search and detection applications. Consider an area of interest divided into n cells. A stationary target is located in one of the cells. A priori information gives that the probability that the target is in cell i is p_i, $i = 1, 2, ..., n$, with $\sum_{i=1}^n p_i = 1$. The goal is to optimally allocate T time units of search effort such that the probability of not detecting the target is minimized (see, e.g., p. 5-1 in [64]). We generalize this problem and consider a random search effectiveness in cell i per time unit and minimize the expected probability of not detecting the target. We let $x \in \mathbb{R}^n$, with x^i representing the number of time units allocated to cell i, and let $\omega = (\omega^1, \omega^2, ..., \omega^n)'$ be independent lognormally distributed random variables (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 cell i. Then, the expected probability of not detecting the target is $f^0(x) = E[F^0(x, \omega)]$, where $F^0(x, \omega) = \sum_{i=1}^n p_i \exp(-\omega^i x^i)$. The decision variables are constrained by $\sum_{i=1}^n x^i \leq T$ and $x \geq 0$, where we use $T = 1$. We consider $n = 100$ cells. Assumption 3 holds for this problem instance. We consider three candidate solutions: $x_1 \in \mathbb{R}^{100}$, which is nearly optimal, $x_2 = (1/100,\ 1/100,\ ...,\ 1/100)' \in \mathbb{R}^{100}$, and $x_3 = (1/50, 1/50, ..., 1/50)' \in \mathbb{R}^{100}$, which is infeasible. Hence, $\psi(x_1) = \psi(x_2) = 0$ and $\psi(x_3) = 1$. We verify using long simulations ($N = 10^8$) 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 2.

We consider both confidence intervals (23) and (30). To compute (23), we first estimate μ by solving (7) using sample size N. Second, we compute $\overline{\eta}_{N,M}$ using the estimated μ with M replications. In (30), we use $L = 1$ which leads to $K = 5$ when $\beta = 0.05$; see (29).

Table 2 provides 95%-confidence intervals for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (23) (rows 3-6) and (30) (rows 7-10) with varying sample size N and replications M and K. We note that since $\psi(x_3) = 1$, $\theta(x_3)$ is near -1; see (4).

While the confidence intervals reported are from a single generation, we also verify the coverage

[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.

				Confidence Intervals		
Method	N	M	K	$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
(23)	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	-	$[-0.000006, 0]$	$[-0.007483, 0]$	$[-1.009602, 0]$
(30)	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"				$\approx 8 \cdot 10^{-7}$	≈ -0.00736	≈ -0.99318

Table 2: 95%-confidence intervals in Example 2 for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (23) (rows 3-6) and (30) (rows 7-10) with varying sample size N and replications M and K. The last row gives approximate values of $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$.

and variability of the confidence intervals across independent replications. Specifically, we confirm the confidence level in (30) by estimating coverage probabilities, i.e., the probability that the random confidence interval (30) includes $\theta(x)$. We find that 100%, 99%, 98% and 99% of 1000 (200 in the case of $N = 10^5$) independent replications of (30) 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 (23) and (30), 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 2. Hence, the variability of the confidence intervals is modest across independent replications.

We also apply the hypothesis test of [58] 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. As discussed in Section 6.1, we believe that results of the kind presented in Table 2 are more informative than such hypothesis tests.

6.3 Example 3: 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 [50]. The design variables are the width x^1 and depth x^2 of the column. In [51], 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 $0 \in \mathbb{R}^4$

		Confidence Intervals		
N	M	$\psi(x_1)$	$\psi(x_2)$	$\psi(x_3)$
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	$(-\infty, -0.0067]$	$(-\infty, 0.7898]$	$(-\infty, 9.9154]$

Table 3: 95%-confidence intervals in Example 3 for $\psi(x_1)$, $\psi(x_2)$, and $\psi(x_3)$ using (20) with varying sample size N and replications M.

				Confidence Intervals		
Method	N	M	K	$\theta(x_1)$	$\theta(x_2)$	$\theta(x_3)$
(23)	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	-	$[-0.0014, 0]$	$[-0.7816, 0]$	$[-9.8631, 0]$
(30)	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	$[-0.0003, 0]$	$[-0.8137, 0]$	$[-10.3143, 0]$

Table 4: 90%-confidence intervals in Example 3 for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$ using (23) (rows 3-6) and (30) (rows 7-10) with varying sample size N and replications M and K.

to a limit-state surface describing the performance of the column given design x and realization ω; see [50, 51]. 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 [50], 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 [51], Assumption 3 holds for this problem instance.

We consider three designs: $x_1 = (0.334, 0.586)'$ is the best point reported in [50]; $x_2 = (0.346, 0.553)'$ is an infeasible solution reported in [50], and $x_3 = (0.586, 0.334)'$ is the "mirror image" of x_1. Table 3 gives 95%-confidence intervals for $\psi(x_1)$, $\psi(x_2)$, and $\psi(x_3)$ for various sample sizes and replications. Table 4 presents confidence intervals for $\theta(x_1)$, $\theta(x_2)$, and $\theta(x_3)$, with $\alpha = 0.1$ in (23) and $\alpha = \beta = 0.05$ in (30). We see that (23) and (30) give comparable results. As observed earlier, a near optimal solution may require a relatively large sample size to ensure a tight confidence interval.

6.4 Example 4: Optimization and Validation Analysis for Full Problem

We illustrate Algorithm 1 by considering an extension of Example 1. Let $F^0(\cdot,\cdot)$ be as defined in that example and also define $F^1(\cdot,\cdot)$ and $F^2(\cdot,\cdot)$ 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 expression to construct constraints of the form $E[\sum_{i=1}^{20} a^i(x^i-$

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}^*	1600	66	$(-\infty, -0.0001]$	$[-0.7915, 0]$	[3453, 3482]
x_{3200}^*	3200	60	$(-\infty, -0.0107]$	$[-0.4043, 0]$	[3462, 3482]
x_{6400}^*	6400	75	$(-\infty, 0.0785]$	$[-0.2027, 0]$	[3466, 3481]
x_{12800}^*	12800	129	$(-\infty, 0.0125]$	$[-0.1082, 0]$	[3470, 3480]
x_{25600}^*	25600	79	$(-\infty, 0.0607]$	$[-0.1085, 0]$	[3467, 3474]
x_{51200}^*	51200	99	$(-\infty, 0.0499]$	$[-0.0609, 0]$	[3467, 3472]

Table 5: 95%-confidence intervals in Example 4 for $\psi(x_N^*)$ and $f^0(x_N^*)$, and 90%-confidence intervals for $\theta(x_N^*)$ for candidate points generated by Algorithm 1.

$b^i\omega^i)^2 - 100] \leq 0$. Hence, the resulting instance of P involves 20 decision variables, 60 independent random variables with uniform distribution each supported on $[0,1]$, an expected value objective function, and two expected value constraint functions.

We apply Algorithm 1 to this problem instance using $x_0 = 0 \in \mathrm{IR}^{20}$, $N_0 = 100$, $\Delta(N) = 1/\sqrt{N}$, and $\delta_1 = \delta_2 = 1$. Moreover, we let $\mathcal{K}(N) = 2N$. The algorithm map $A_{N\epsilon}(\cdot)$ is one iteration of the Polak-He Phase 1-Phase 2 algorithm; see Section 2.6 in [42]. 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}^*,..., x_{51200}^*. For each candidate point x_N^*, we compute the confidence intervals (20) and (30) using sample size $10N$ (1000 for x_0^*), replications $M = 30$ and $K = 23$, and $L = 1$; see Table 5. This selection of M, K, and L results in 95% confidence intervals for $\psi(x_N^*)$ and 90%-confidence intervals for $\theta(x_N^*)$. 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 sample average estimator; see column 6. The ten stages required 6900 seconds of run time. The verification analysis needed 3300 seconds.

6.5 Examples 5 and 6: Engineering Design Optimization

We also consider two more complex engineering design problems where the goal is to minimize the design cost subject to a buffered failure probability constraint and other constraints and hence is of the form (41) [48]. The first design example, referred to as Example 5, is taken from [54] and involves seven design variables ($\tilde{x} \in \mathrm{IR}^7$ in (41)), seven random variables ($\Omega \subset \mathrm{IR}^7$), and 10 limit-state functions ($\tilde{r} = 10$ in (41)). The second design example, referred to as Example 6, is

Algorithm	γ_1	Final sample size	Confidence Intervals $\psi(x_N^*)$	 $\theta(x_N^*)$
1	-	65624	$(-\infty, 0.0066]$	$[-0.0154, 0]$
2	0.1	2491871	$(-\infty, 0.0000]$	$[-0.0017, 0]$
2	1	2691871	$(-\infty, 0.0004]$	$[-0.0018, 0]$
2	10	2591871	$(-\infty, 0.0003]$	$[-0.0007, 0]$

Table 6: Final sample sizes, 95%-confidence intervals for $\psi(x_N^*)$, and 90%-confidence intervals for $\theta(x_N^*)$ in Example 5 by Algorithms 1 and 2 after one hour of computations for variable active-set strategy parameter γ_1.

taken from [46], pp. 472-473, and involves seven design variables, seven random variables, and nine limit-state functions. We refer to [2] for further details.

We apply Algorithms 1 and 2 to these problem instances setting x_0 equal to the variables upper bounds (see [2]), $N_0 = 1000$, $N_{-1} = 1000$, $\epsilon_N = 1000/N$, $\mathcal{K}(N) = N + \min\{10^4, \lfloor 0.5N \rfloor\}$, $\gamma_2 = \infty$, and γ_1 is varied in the interval $[0.1, 10]$ as indicated below. Since we set $\gamma_2 = \infty$, all functions $\tilde{g}^{jk}(\cdot, \cdot)$ are included; see (46). It appears that r^j must be large to incur substantial computational savings from setting $\gamma_2 < \infty$. Hence, we focus on reducing the number of "active" sample points by setting $\gamma_1 < \infty$; see (45).

Instead of defining $\Delta(\cdot)$ for the tests in Step 2 of Algorithm 1 and Step 3 of Algorithm 2, 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 both tests are satisfied. We use $\zeta = 0.1$ and 0.8 in Examples 5 and 6, respectively. Since Examples 5 and 6 are much more complex than Example 4, we utilize SNOPT [16] as implemented in TOMLAB [21] and set the algorithm map $\tilde{A}_{N\epsilon}(\cdot)$ equal to 20 iterations of that solver. These calculations are carried out using a desktop computer at 3.16 GHz with 3GB of RAM.

Tables 6 and 7 present final sample sizes (column 3), 95%-confidence intervals for $\psi(x_N^*)$ (column 4), and 90%-confidence intervals for $\theta(x_N^*)$ (column 5) at the last point obtained by Algorithms 1 and 2 after one hour of computations for variable active-set strategy parameter γ_1. The confidence intervals are based on (20) and (30) using sample size 10^6, replications $M = 30$ and $K = 5$, and $L = 1$. Table 6 shows results for Example 5 and illustrates that the active-set strategy of Algorithm 2 enables the use of much larger sample sizes than in Algorithm 1, where no active-set strategy is employed. Algorithm 2 obtains at least an order of magnitude smaller constraint violation and optimality function intervals across several values of the parameter γ_1. Table 7 shows similar results for Example 6. In this case, Algorithm 2 still utilizes larger sample sizes and obtain better results, but the performance is sensitive to the choice of γ_1. If the active-set strategy is aggressive (γ_1 close to zero), then the resulting small set of "active" sample points may not accurately approximate the full sample and result in additional iterations. In this case, Algorithm 2 with $\gamma_1 = 0.1$ is not significantly better than Algorithm 1. However, other choices of γ_1 yield an advantage over Algorithm 1.

Algorithm	γ_1	Final sample size	Confidence Intervals $\psi(x_N^*)$	 $\theta(x_N^*)$
1	-	65624	$(-\infty, 0.2132]$	$[-0.2408, 0]$
2	0.1	129721	$(-\infty, 0.2168]$	$[-0.2314, 0]$
2	1	129721	$(-\infty, 0.0402]$	$[-0.0588, 0]$
2	10	129721	$(-\infty, 0.0834]$	$[-0.1026, 0]$

Table 7: Similar results as in Table 6 but for Example 6.

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 see that optimality functions help determine sample sizes in each stage of the calculations of algorithms and therefore ensure convergence. The numerical results also indicate that when possible, it may be computationally beneficial to only consider a subset of sample points by means of an active-set strategy.

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.

Acknowledgement

The author acknowledges financial supported from Air Force Office of Scientific Research Young Investigator grant F1ATA08337G003 as well assistance with the numerical results from Mr. G.H. Basova, Turkish Army.

Appendix

Proof of Theorem 1. 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

$$m\|x' - x\|^2 \leq \langle x' - x, \nabla^2 f^j(x)(x' - x)\rangle \leq M\|x' - x\|^2, \tag{51}$$

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 (51) that for any $x \in \mathbb{B}(\hat{x}, \hat{\rho}) \cap X_\psi$, $x' \in \mathbb{B}(\hat{x}, \hat{\rho})$, and some $s^j \in [0, 1]$, $j \in \mathbf{q}_0$,

$$\begin{aligned} \tilde{\psi}(x, x') &= \max\Big\{\langle \nabla f^0(x), x' - x\rangle + \tfrac{1}{2}\langle x' - x, \nabla^2 f^0(x + s^0(x' - x))(x' - x)\rangle, \\ &\quad \max_{j \in \mathbf{q}}\{f^j(x) + \langle \nabla f^j(x), x' - x\rangle + \tfrac{1}{2}\langle x' - x, \nabla^2 f^j(x + s^j(x' - x))(x' - x)\rangle\}\Big\} \\ &\leq \frac{1}{M}\max\Big\{\langle \nabla f^0(x), M(x' - x)\rangle + \tfrac{1}{2}\|M(x' - x)\|^2, \\ &\quad \max_{j \in \mathbf{q}}\{f^j(x) + \langle \nabla f^j(x), M(x' - x)\rangle + \tfrac{1}{2}\|M(x' - x)\|^2\}\Big\}, \end{aligned} \tag{52}$$

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 (4), which according to Theorem 2.2.8 in [42] 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 (4) and the property $\psi^+(x) = 0$, the minimization of the right-hand side in (52) 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 (52) with respect to x' and the minimization in (4), 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 (52) with respect to x' over $\mathbb{B}(\hat{x}, \hat{\rho})$ that

$$\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x') \leq \theta(x)/M \tag{53}$$

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

$$\min_{x' \in \mathbb{B}(\hat{x}, \hat{\rho})} \tilde{\psi}(x, x') \geq \theta(x)/m \tag{54}$$

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

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 (53), 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 (6).

Second, we prove the left-most inequality and consider three cases. Let $x \in \mathbb{B}(\hat{x}, \rho) \cap X_\psi$ 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(54), $\theta(x)/m \leq f^0(\hat{x}) - f^0(x)$.

(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 (54), $\theta(x)/m \leq f^0(\hat{x}) - f^0(x)$. 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 $\mathbb{B}(\hat{x}, \hat{\rho})$,

$$\langle \nabla f^0(\hat{x}'), \hat{h} \rangle < (f^0(\hat{x}) - f^0(x)) - (f^0(\hat{x}') - f^0(x)) < 0. \tag{55}$$

Consequently,

$$d\psi(\hat{x}', \hat{h}) \geq 0. \tag{56}$$

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 (51) ,

$$f^{j'}(\hat{x}) \geq f^{j'}(\hat{x}') + \langle \nabla f^{j'}(\hat{x}'), \hat{h} \rangle + \tfrac{1}{2} m \|\hat{h}\|^2.$$

Hence, using (56) and (54), we obtain

$$\psi(\hat{x}) \geq f^{j'}(\hat{x}) \geq \psi(\hat{x}') + d\psi(\hat{x}'; \hat{h}) + \tfrac{1}{2} m \|\hat{h}\|^2 \geq \theta(x)/m + \tfrac{1}{2} m \|\hat{h}\|^2. \tag{57}$$

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 (55) and (54) 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 (57), we obtain that for any $x' \in \mathbb{B}(\hat{x}, \hat{\rho})$,

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

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 $f^0(\cdot)$ and (54) 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 (6) 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 [57]. Let $g : \mathbb{R}^{q+(q+1)n} \to \mathbb{R}$ be defined for any $\overline{\zeta} = (\zeta_{-1}, \zeta_0', \zeta_1', ..., \zeta_q') \in \mathbb{R}^{q+(q+1)n}$, with $\zeta_{-1} \in \mathbb{R}^q$, $\zeta_j \in \mathbb{R}^n$, $j \in \mathbf{q}_0$, by

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

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

$$\phi(\overline{\zeta}) \overset{\triangle}{=} \min_{\mu \in \Sigma_q^0} \Big\{ -\sum_{j \in \mathbf{q}} \mu^j \zeta_{-1}^j + \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \zeta_j \Big\|^2 \Big\}.$$

Let

$$\hat{\mathbf{q}}_w(\overline{\zeta}) \overset{\triangle}{=} \{ j \in \mathbf{q} \mid \max_{k \in \mathbf{q}} \zeta_{-1}^k = \zeta_{-1}^j \},$$

and

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

Moreover, let

$$\hat{\Sigma}_\phi(\overline{\zeta}) \overset{\triangle}{=} \Big\{ \mu \in \Sigma_q^0 \;\Big|\; \phi(\overline{\zeta}) = -\sum_{j \in \mathbf{q}} \mu^j \zeta_{-1}^j + \tfrac{1}{2} \Big\| \sum_{j \in \mathbf{q}_0} \mu^j \zeta_j \Big\|^2 \Big\}.$$

It follows from Danskin Theorem; see, for example, Theorem 7.21 in [57], 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 $\xi_{-1}^0 \overset{\triangle}{=} 0$, and

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

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

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

Hence, it follows from Proposition 7.57 in [57] that $g(\cdot)$ is Hadamard directional differentiable.

In view of Proposition 4, Delta Theorem 7.59 in [57] gives 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 $\hat{\Sigma}_\phi((f(x), \nabla \overline{f}(x)')') = \hat{\Sigma}_q^0(x)$ and from rearranging terms. □

Proof of Lemma 2: We first consider (49). For any $j \in \mathbf{q}_0$ and $N \in \mathcal{K}$,

$$|\tilde{f}^j_{N\epsilon_N}(x^*_N; \mathbf{N}^{*j}_N, \rho^{*j}_N) - f^j(\hat{x})| \leq |\tilde{f}^j_{N\epsilon_N}(x^*_N; \mathbf{N}^{*j}_N, \rho^{*j}_N) - f^j_{N\epsilon_N}(x^*_N)| + |f^j_{N\epsilon_N}(x^*_N) - f^j(\hat{x})|.$$

By Proposition 8 and continuity of $f^j(\cdot)$, $|f^j_{N\epsilon_N}(x^*_N) - f^j(\hat{x})| \to^K 0$, as $N \to \infty$, almost surely. Hence, we focus on the first term on the right-hand side. From (42), (36), and Assumption 6, we see that for any $j \in \mathbf{q}_0$ and $N \in \mathcal{K}$,

$$\begin{aligned}
&\tilde{f}^j_{N\epsilon_N}(x^{*j}_N; \mathbf{N}^{*j}_N, \rho^*_N) - f^j_{N\epsilon_N}(x^*_N) \\
= \;& \frac{1}{N}\left(\sum_{l\in\mathbf{N}^{*j}_N} \epsilon_N \log\left(1 + \sum_{k\in\rho^{*jl}_N} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right) - \sum_{l=1}^{N} \epsilon_N \log\left(1 + \sum_{k\in\tilde{\mathbf{r}}^j} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right)\right) \\
= \;& \frac{1}{N}\left(\sum_{l\in\mathbf{N}^{*j}_N} \left\{\epsilon_N \log\left(1 + \sum_{k\in\rho^{*jl}_N} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right) - \epsilon_N \log\left(1 + \sum_{k\in\tilde{\mathbf{r}}^j} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right)\right\}\right. \\
-\;& \left.\sum_{l\in\{1,\ldots,N\}-\mathbf{N}^{*j}_N} \epsilon_N \log\left(1 + \sum_{k\in\tilde{\mathbf{r}}^j} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right)\right). \qquad (58)
\end{aligned}$$

For any $l \in \mathbf{N}^{*j}_N$, we deduce from (35) that

$$0 \leq \epsilon_N \log\left(1 + \sum_{k\in\rho^{*jl}_N} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right) - \max\{\max_{k\in\rho^{*jl}_N} \tilde{g}^{jk}(x^*_N, \omega_l), 0\} \leq \epsilon_N \log r^j$$

and similarly with ρ^{*jl}_N replaced by $\tilde{\mathbf{r}}^j$. By construction of ρ^{*jl}_N,

$$\max\{\max_{k\in\rho^{*jl}_N} \tilde{g}^{jk}(x^*_N, \omega_l), 0\} = \max\{\max_{k\in\tilde{\mathbf{r}}^j} \tilde{g}^{jk}(x^*_N, \omega_l), 0\}.$$

Hence, for any $l \in \mathbf{N}^{*j}_N$,

$$0 \leq \epsilon_N \log\left(1 + \sum_{k\in\rho^{*jl}_N} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right) - \epsilon_N \log\left(1 + \sum_{k\in\tilde{\mathbf{r}}^j} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right) \leq \epsilon_N \log r^j.$$

Next, we consider $l \in \{1, 2, ..., N\} - \mathbf{N}^{*j}_N$. By construction of $\mathbf{N}^{*j}_N$, $\tilde{g}^{jk}(x^*_N, \omega_l) < -\gamma_1$ for all $k \in \tilde{\mathbf{r}}^j$ and $l \in \{1, 2, ..., N\} - \mathbf{N}^{*j}_N$. Hence, for $l \in \{1, 2, ..., N\} - \mathbf{N}^{*j}_N$,

$$0 \leq \epsilon_N \log\left(1 + \sum_{k\in\tilde{\mathbf{r}}^j} \exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]\right) \leq \epsilon_N \log(1 + \exp[-\gamma_1/\epsilon_N]).$$

It now follows from (58) that for any $j \in \mathbf{q}_0$ and $N \in \mathcal{K}$,

$$|\tilde{f}^j_{N\epsilon_N}(x^*_N; \mathbf{N}^*_N, \rho^*_N) - f^j_{N\epsilon_N}(x^*_N)| \leq \epsilon_N \log(r^j + \exp[-\gamma_1/\epsilon_N]).$$

Since $\epsilon_N \to 0$, as $N \to \infty$, we conclude that (49) holds.

We second consider (50). For any $j \in \mathbf{q}_0$ and $N \in \mathcal{K}$,

$$\begin{aligned} & \|\nabla \tilde{f}^j_{N\epsilon_N}(x^*_N; \mathbf{N}^*_N, \rho^*_N) - \nabla f^j(\hat{x})\| \\ \leq\ & \|\nabla \tilde{f}^j_{N\epsilon_N}(x^*_N; \mathbf{N}^*_N, \rho^*_N) - \nabla f^j_{N\epsilon_N}(x^*_N)\| + \|\nabla f^j_{N\epsilon_N}(x^*_N) - \nabla f^j(\hat{x})\|. \end{aligned}$$

By Proposition 8 and continuity of $\nabla f^j(\cdot)$, $\|\nabla f^j_{N\epsilon_N}(x^*_N) - \nabla f^j(\hat{x})\| \to^K 0$, as $N \to \infty$. Hence, we focus on the first term on the right-hand side. Since

$$\nabla \tilde{f}^j_{N\epsilon}(x; \mathbf{N}^j, \rho^j) = \nabla \phi^j(x) + \frac{1}{N} \sum_{l \in \mathbf{N}^j} \nabla_x \tilde{F}^j_\epsilon(x, \omega_l; \rho^{jl}),$$

where

$$\nabla \tilde{F}^j_\epsilon(x, \omega_l; \rho^{jl}) = \sum_{k \in \rho^{jl}} \tilde{\mu}^k_\epsilon(x, \omega_l; \rho^{jl}) \nabla_x \tilde{g}^{jk}(x, \omega_l), \tag{59}$$

with for any $k \in \rho^{jl}$,

$$\tilde{\mu}^k_\epsilon(x, \omega_l; \rho^{jl}) \triangleq \frac{\exp[\tilde{g}^{jk}(x, \omega_l)/\epsilon]}{1 + \sum_{k' \in \rho^{jl}} \exp[\tilde{g}^{jk'}(x, \omega_l)/\epsilon]}, \tag{60}$$

we find that, for any $j \in \mathbf{q}_0$ and $N \in \mathcal{K}$,

$$\begin{aligned} & \nabla \tilde{f}^j_{N\epsilon_N}(x^*_N; \mathbf{N}^*_N, \rho^*_N) - \nabla f^j_{N\epsilon_N}(x^*_N) \\ =\ & \frac{1}{N} \sum_{l \in \mathbf{N}^{*j}_N} \Big(\nabla_x \tilde{F}_{\epsilon_N}(x^*_N, \omega_l; \rho^{*jl}_N) - \nabla_x F_{\epsilon_N}(x^*_N, \omega_l) \Big) - \frac{1}{N} \sum_{l \in \{1,\ldots,N\} - \mathbf{N}^{*j}_N} \nabla_x F_{\epsilon_N}(x^*_N, \omega_l). \end{aligned} \tag{61}$$

We deal with the two terms on the right-hand side of (61) in turn. Using (59) and (33), we obtain that

$$\begin{aligned} & \frac{1}{N} \sum_{l \in \mathbf{N}^{*j}_N} \Big(\nabla_x \tilde{F}_{\epsilon_N}(x^*_N, \omega_l; \rho^{*jl}_N) - \nabla_x F_{\epsilon_N}(x^*_N, \omega_l) \Big) \\ =\ & \frac{1}{N} \sum_{l \in \mathbf{N}^{*j}_N} \left(\sum_{k \in \rho^{*jl}_N} \tilde{\mu}^k_{\epsilon_N}(x^*_N, \omega_l; \rho^{*jl}_N) \nabla_x \tilde{g}^{jk}(x^*_N, \omega_l) - \sum_{k \in \tilde{\mathbf{r}}^j} \mu^k_{\epsilon_N}(x^*_N, \omega_l) \nabla_x \tilde{g}^{jk}(x^*_N, \omega_l) \right), \end{aligned}$$

where $\mu^k_{\epsilon_N}(x^*_N, \omega_l)$ specializes under Assumption 6 (see (34)) to

$$\mu^k_{\epsilon_N}(x^*_N, \omega_l) = \frac{\exp[\tilde{g}^{jk}(x^*_N, \omega_l)/\epsilon_N]}{1 + \sum_{k' \in \tilde{\mathbf{r}}^j} \exp[\tilde{g}^{jk'}(x^*_N, \omega_l)/\epsilon_N]}. \tag{62}$$

Collecting terms, we obtain that

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{l\in\mathbf{N}_N^{*j}}\left(\nabla_x\tilde{F}_{\epsilon_N}(x_N^*,\omega_l;\rho_N^{*jl})-\nabla_xF_{\epsilon_N}(x_N^*,\omega_l)\right)\right\| \\
\leq\ & \frac{1}{N}\sum_{l\in\mathbf{N}_N^{*j}}\sum_{k\in\rho_N^{*jl}}|\tilde{\mu}_{\epsilon_N}^k(x_N^*,\omega_l;\rho_N^{*jl})-\mu_{\epsilon_N}^k(x_N^*,\omega_l)|\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\| \qquad (63)\\
+\ & \frac{1}{N}\sum_{l\in\mathbf{N}_N^{*j}}\sum_{k\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}}\mu_{\epsilon_N}^k(x_N^*,\omega_l)\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\|.
\end{aligned}$$

For all $l\in\mathbf{N}_N^{*j}$ and $k\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}$, we have by construction that $0\leq\exp[\tilde{g}^{jk}(x_N^*,\omega_l)/\epsilon_N]\leq\exp[-\gamma_2/\epsilon_N]$. Hence, since $\rho_N^{*jl}\subset\tilde{\mathbf{r}}^j$,

$$\begin{aligned}
0\ \leq\ & \sum_{k'\in\tilde{\mathbf{r}}^j}\exp[\tilde{g}^{jk'}(x_N^*,\omega_l)/\epsilon_N]-\sum_{k'\in\rho_N^{*jl}}\exp[\tilde{g}^{jk'}(x_N^*,\omega_l)/\epsilon_N]\\
=\ & \sum_{k'\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}}\exp[\tilde{g}^{jk'}(x_N^*,\omega_l)/\epsilon_N]\leq\sum_{k'\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}}\exp[-\gamma_2/\epsilon_N]\leq r^j\exp[-\gamma_2/\epsilon_N].
\end{aligned}$$

Consequently, in view of (62) and (60) there exists a sequence $\{\zeta_N\}_{N=1}^\infty$, such that $\zeta_N\to 0$, as $N\to\infty$, and

$$|\tilde{\mu}_{\epsilon_N}^k(x_N^*,\omega_l;\rho_N^{*jl})-\mu_{\epsilon_N}^k(x_N^*,\omega_l)|\leq\zeta_N$$

for all $N\in\mathbb{N}$, $l\in\mathbf{N}_N^{*j}$, and $k\in\rho_N^{*jl}$. Since $\rho_N^{*jl}\subset\tilde{\mathbf{r}}^j$ and $\mathbf{N}_N^{*j}\subset\{1,2,...,N\}$,

$$\begin{aligned}
& \frac{1}{N}\sum_{l\in\mathbf{N}_N^{*j}}\sum_{k\in\rho_N^{*jl}}|\tilde{\mu}_{\epsilon_N}^k(x_N^*,\omega_l;\rho_N^{*jl})-\mu_{\epsilon_N}^k(x_N^*,\omega_l)|\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\|\\
\leq\ & \zeta_N\sum_{k\in\tilde{\mathbf{r}}^j}\frac{1}{N}\sum_{l=1}^N\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\|
\end{aligned}$$

for all $N\in\mathcal{K}$. In view of Assumption 3(ii) and the uniform strong law of large number (see for example Theorem 7.52 in [57]), there exists a $C<\infty$ such that for all $k\in\tilde{\mathbf{r}}^j$,

$$\frac{1}{N}\sum_{l=1}^N\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\|\to^K E[\|\nabla_x\tilde{g}^{jk}(\hat{x},\omega)\|]\leq C, \qquad (64)$$

as $N\to\infty$, almost surely. Since $\zeta_N\to 0$, as $N\to\infty$, it follows that the first term on the right-hand side of (63) vanishes as $N\to\infty$ almost surely.

We next consider the second term on the right-hand side of (63). Since $0\leq\exp[\tilde{g}^{jk}(x_N^*,\omega_l)/\epsilon_N]\leq\exp[-\gamma_2/\epsilon_N]$ for all $k\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}$, $l\in\mathbf{N}_N^{*j}$, and $N\in\mathbb{N}$, it follows that $0\leq\mu_{\epsilon_N}^k(x_N^*,\omega_l)\leq\exp[-\gamma_2/\epsilon_N]$ for all $k\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}$, $l\in\mathbf{N}_N^{*j}$, and $N\in\mathbb{N}$. Hence,

$$\frac{1}{N}\sum_{l\in\mathbf{N}_N^{*j}}\sum_{k\in\tilde{\mathbf{r}}^j-\rho_N^{*jl}}\mu_{\epsilon_N}^k(x_N^*,\omega_l)\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\|\leq\exp[-\gamma_2/\epsilon_N]\sum_{k\in\tilde{\mathbf{r}}^j}\frac{1}{N}\sum_{l=1}^N\|\nabla_x\tilde{g}^{jk}(x_N^*,\omega_l)\|.$$

Using the same arguments as those leading to (64), we find that the second term on the right-hand side of (63) vanishes, as $N \to \infty$ almost surely. Hence, the left-hand side in (63) vanishes, as $N \to \infty$ almost surely. Consequently, the first term on the right-hand side in (61) vanishes, as $N \to \infty$ almost surely.

Finally, we consider the second term on the right-hand side in (61). Since $\max_{k \in \tilde{\mathbf{r}}^j} \tilde{g}^{jl}(x_N^*, \omega_l) < -\gamma_1$ for all $l \in \{1, 2, ..., N\} - \mathbf{N}_N^{*j}$ and $N \in \mathcal{K}$, we find that $0 \leq \mu_{\epsilon_N}^k(x_N^*, \omega_l) \leq \exp[-\gamma_1/\epsilon_N]$ for all $l \in \{1, 2, ..., N\} - \mathbf{N}_N^{*j}$, $k \in \tilde{\mathbf{r}}^j$, and $N \in \mathcal{K}$. Consequently,

$$\begin{aligned}
\left\| \frac{1}{N} \sum_{l \in \{1,...,N\} - \mathbf{N}_N^{*j}} \nabla_x F_{\epsilon_N}(x_N^*, \omega_l) \right\| &\leq \frac{1}{N} \sum_{l \in \{1,...,N\} - \mathbf{N}_N^{*j}} \sum_{k \in \tilde{\mathbf{r}}^j} \mu_{\epsilon_N}^k(x_N^*, \omega_l) \|\nabla_x \tilde{g}^{jk}(x_N^*, \omega_l)\| \\
&\leq \exp[-\gamma_1/\epsilon_N] \sum_{k \in \tilde{\mathbf{r}}^j} \frac{1}{N} \sum_{l=1}^{N} \|\nabla_x \tilde{g}^{jk}(x_N^*, \omega_l)\|.
\end{aligned}$$

Again using the same arguments as those leading to (64), we find that the second term on the right-hand side in (61) vanishes, as $N \to \infty$ almost surely. The conclusion then follows. □

References

[1] S. Alexander, T.F. Coleman, and Y. Li. Minimizing CVaR and VaR for a portfolio of derivatives. *J. Banking & Finance*, 30:583–605, 2006.

[2] H. G. Basova. *Reliability-based design optimization using buffered failure probability.* Master's thesis, Naval Postgraduate School, Monterey, California, 2010.

[3] F. Bastin, C. Cirillo, and P.L Toint. An adaptive Monte Carlo algorithm for computing mixed logit estimators. *Computational Management Science*, 3(1):55–79, 2006.

[4] F. Bastin, C. Cirillo, and P.L. Toint. Convergence theory for nonconvex stochastic programming with an application to mixed logit. *Mathematical Programming*, 108(2-3):207–234, 2006.

[5] G. Bayraksan and D.P. Morton. Assessing solution quality in stochastic programs. *Mathematical Programming*, 108:495–514, 2006.

[6] G. Bayraksan and D.P. Morton. Assessing solution quality in stochastic programs via sampling. In *Tutorials in Operations Research*, pages 102–122. INFORMS, 2009.

[7] G. Bayraksan and D.P. Morton. A sequential sampling procedure for stochastic programming. *Operations Research*, to appear, 2010.

[8] C. Beliakov and A. Bagirov. Non-smooth optimization methods for computation of the conditional value-at-risk and portfolio optimization. *Optimization*, 55(5-6):459–479, 2006.

[9] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer, New York, New York, 1990.

[10] B. Bettonvil, E. del Castillo, and J.P.C. Kleijnen. Statistical testing of optimality conditions in multiresponse simulation-based optimization. *European Journal of Operational Research*, 199:448–458, 2009.

[11] P. Billingsley. *Probability and Measure.* Wiley, New York, New York, 1995.

[12] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer, 2000.

[13] Y. Ermoliev. Stochastic quasigradient methods. In *Numerical Techniques for Stochastic Optimization*, Yu. Ermoliev and R.J-B. Wets (Eds.), New York, New York, 1988. Springer.

[14] C. I. Fabian. Handling CVaR objectives and constraints in two-stage stochastic models. *European J. Operational Research*, 191:888–911, 2008.

[15] M. C. Fu. Gradient estimation. In *Simulation*, pages 575–616, Amsterdam, Netherlands, 2006. Elsevier.

[16] P. Gill, W. Murray, and M. Saunders. 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.

[17] G. Gürkan, A. Özge, and S. M. Robinson. Sample-path solution of stochastic variational inequalities. *Mathematical Programming*, 84(2):313–333, 1999.

[18] J. L. Higle and S. Sen. *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming.* Springer, 1996.

[19] J.L. Higle and S. Sen. Statistical verification of optimality conditions for stochastic programs with recourse. *Annals of Operations Research*, 30:215–240, 1991.

[20] J.L. Higle and S. Sen. Duality and statistical tests of optimality for two stage stochastic programs. *Mathematical Programming*, 75:257–275, 1996.

[21] K. Holmstrom. Tomlab optimization. http://tomopt.com, 2009.

[22] T. Homem-de-Mello. Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation*, 13(2):108–133, 2003.

[23] T. Homem-de-Mello, A. Shapiro, and M.L. Spearman. Finding optimal material release times using simulation-based optimization. *Management Science*, 45(1):86–102, 1999.

[24] G. Infanger. *Planning under uncertainty: solving large-scale stochastic linear programs.* Thomson Learning, 1994.

[25] G. Iyengar and A. K. C. Ma. Fast gradient descent method for mean-CVaR optimization. *Operations Research Letters*, 2010.

[26] P. Kall and J. Meyer. *Stochastic Linear Programming, Models, Theory, and Computation.* Springer, 2005.

[27] S.H. Kim and B.L. Nelson. Selecting the best system. In *Simulation*, pages 501–534, Amsterdam, Netherlands, 2006. Elsevier.

[28] A. J. King and R. J. B. Wets. Epi-convergence of convex stochastic programs. *Stochastics and Stochastics Reports*, 34:83–92, 1991.

[29] B. W. Kort and D. P. Bertsekas. A New Penalty Function Algorithm for Constrained Minimization. In *Proceedings 1972 IEEE Conf. Decision and Control*, New Orlean, Louisiana, 1972.

[30] A. Kunzi-Bay and J. Mayer. Computational aspects of minimizing conditional value-at-risk. *Computational Management Science*, 3:3–27, 2006.

[31] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer, 2. edition, 2003.

[32] C. Lim, H. D. Sherali, and S. Uryasev. Portfolio optimization by minimizing conditional value-at-risk via nondifferentiable optimization. *Computational Optimization and Applications*, 2008.

[33] J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optimization*, 19:674–699, 2008.

[34] W. K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.

[35] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimization*, 19(4):1574–1609, 2009.

[36] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Optimization*, 17(4):969–996, 2006.

[37] V.I. Norkin, G.C. Pflug, and A. Ruszczynski. A branch and bound method for stochastic global optimization. *Mathematical Programming*, 83:425–450, 1998.

[38] W. Ogryczak and T. Sliwinski. On solving the dual for portfolio selection by optimizing conditional value at risk. *Computational Optimization and Applications*, 2010.

[39] R. Pasupathy. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58(4):889–901, 2010.

[40] E. Polak. On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Review*, 29:21–89, 1987.

[41] E. Polak. On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems. *Mathematical Programming, Series B*, 62(1-3):385–414, 1993.

[42] E. Polak. *Optimization. Algorithms and consistent approximations.* Springer, New York, New York, 1997.

[43] E. Polak and J. O. Royset. Efficient sample sizes in stochastic nonlinear programming. *J. Computational and Applied Mathematics*, 217:301–310, 2008.

[44] E. Polak, J. O. Royset, and R. S. Womersley. Algorithms with adaptive smoothing for finite minimax problems. *J. Optimization. Theory and Applications*, 119(3):459–484, 2003.

[45] E. Polak, R. S. Womersley, and H. X. Yin. An Algorithm Based on Active Sets and Smoothing for Discretized Semi-Infinite Minimax Problems. *J. Optimization Theory and Applications*, 138:311–328, 2008.

[46] S. S. Rao. *Engineering optimization theory and practice.* John Wiley & Sons, 4th edition, 2009.

[47] S. M. Robinson. Sample-path optimization of convex stochastic performance functions. *Mathematics of Operations Research*, 21(3):513–528, 1996.

[48] R. T. Rockafellar and J. O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, 95:499–510, 2010.

[49] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26:1443–1471, 2002.

[50] J. O. Royset, A. Der Kiureghian, and E. Polak. Reliability-based optimal design of series structural systems. *J. Engineering Mechanics*, 127(6):607–614, 2001.

[51] J. O. Royset and E. Polak. Extensions of stochastic optimization results from problems with simple to problems with complex failure probability functions. *J. Optimization. Theory and Application*, 133(1):1–18, 2007.

[52] L. L. Sakalauskas. Nonlinear stochastic programming by Monte-Carlo estimators. *European J. Operational Research*, 137:558–573, 2002.

[53] L. L. Sakalauskas. Towards implementable nonlinear stochastic programming. In *Coping with Uncertainty*, pages 257–279. Springer, 2006.

[54] S. Samson, S. Thoomu, G. Fadel, and J. Reneke. Reliable design optimization under aleatory and epistemic uncertainty. In *Proceedings of ASME 2009 International Design Engineering Technical Conferences*, pages DETC2009–86473, 2009.

[55] A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

[56] A. Shapiro. Testing KKT conditions. Private Communication, June 2, 2003.

[57] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory.* Society of Industrial and Applied Mathematics, 2009.

[58] A. Shapiro and T. Homem-de-Mello. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming*, 81:301–325, 1998.

[59] A. Shapiro and H. Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *J. Mathematical Analysis and Applications*, 325:1390–1399, 2007.

[60] J. C. Spall. *Introduction to stochastic search and optimization.* John Wiley and Sons, New York, New York, 2003.

[61] X. Tong, L. Qi, F. Wu, and H. Zhou. A smoothing method for solving portfolio optimization with CVaR and applications in allocation of generation asset. *Applied Mathematics and Computation*, 216:1723–1740, 2010.

[62] S. W. Wallace and W. T. Ziemba. *Applications of Stochastic Programming.* Society for Industrial and Applied Mathematics, 2005.

[63] W. Wang and S. Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5):515–519, 2008.

[64] A. R. Washburn. *Search and Detection.* INFORMS, Linthicum, Maryland, 4. edition, 2002.

[65] H. Xu and D. Zhang. Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Mathematical Programming*, 119:371–401, 2009.

[66] S. Xu. Smoothing method for minimax problems. *Computational Optimization and Applications*, 20:267–279, 2001.