J Optim Theory Appl (2007) 133: 1–18
DOI 10.1007/s10957-007-9178-0

Extensions of Stochastic Optimization Results to Problems with System Failure Probability Functions

J.O. Royset · E. Polak

Published online: 20 April 2007

Abstract We derive an implementable algorithm for solving nonlinear stochastic optimization problems with failure probability constraints using sample average approximations. The paper extends prior results dealing with a failure probability expressed by a single measure to the case of failure probability expressed in terms of multiple performance measures. We also present a new formula for the failure probability gradient. A numerical example addressing the optimal design of a reinforced concrete highway bridge illustrates the algorithm.

Keywords Stochastic optimization · Sample average approximations · Monte Carlo simulation · Reliability-based optimal design

1 Introduction

This paper focuses on a class of decision-making problems frequently arising in design and maintenance optimization of mechanical structures such as bridges, building frames and aircraft wings. Let $x \in \mathbb{R}^n$ be a vector of design variables, e.g. related to the size and form of the structure, let $c : \mathbb{R}^n \to \mathbb{R}$ be a deterministic, continuously

Communicated by P. Tseng.

This work was sponsored by the Research Associateship Program, National Research Council. The authors are grateful for the valuable insight from Professors Alexander Shapiro, Evarist Gine, and Jon A. Wellner. The authors also thank Professor Tito Homem-de-Mello for commenting on an early draft of this paper.

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

E. Polak
Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, CA, USA

differentiable cost function, and let $p : \mathbb{R}^n \to [0, 1]$ be a failure probability (to be defined below). Then, the optimal design problem is defined as a chance-constrained stochastic program in the form

$$(\mathbf{P}) \quad \min_{x \in \mathbb{R}^n} \{c(x) \mid p(x) \leq q,\ x \in X\}, \tag{1}$$

where q is a bound on the failure probability, $X = \{x \in \mathbb{R}^n \mid f_j(x) \leq 0,\ j \in \mathbf{J}\}$, and $f_j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{J} = \{1, 2, \ldots, J\}$, are deterministic, continuously differentiable functions.

Mechanical structures are assessed using one or more performance measures, e.g., displacement and stress levels at various locations in the structure. In [1], we focused on the case were failure probability is defined as the probability of one performance measure being unsatisfactory. In this paper, we consider the general case of "system failure probability", where failure probability is defined by a collection of performance measures. Here, "failure" occurs when specific combinations of the performance measures are unsatisfactory.

Let $g_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}, k \in \mathbf{K} = \{1, 2, \ldots, K\}$, be a collection of performance functions describing the relevant performance measures. The functions $g_k(\cdot, \cdot)$ depend on the design $x \in \mathbb{R}^n$ and a standard normal random m-vector u. This random vector incorporates the uncertainty in the structure and its environment. Note that random vectors governed by distributions such as the multivariate normal (possibly with correlation) and lognormal distributions can be transformed into a standard normal vector using a smooth bijective mapping. Hence, the limitation to a multivariate standard normal distribution is in many applications not restrictive (see e.g. Chap. 7 of [2] and [3–5].

By convention, $g_k(x, u) \leq 0$ represents unsatisfactory performance of the kth measure. Formally, let the probability space $(\mathbb{R}^m, \mathcal{R}^m, \mathbb{P})$ be defined in terms of the sample space $\mathbb{R}^m$, the Borel sets on $\mathbb{R}^m$, denoted $\mathcal{R}^m$, and the multivariate standard normal distribution $\mathbb{P}$ of u. Assuming that $g_k(x, \cdot)$ is measurable for all $x \in \mathbb{R}^n$ and that $k \in \mathbf{K}$, we define the failure probability of the structure by $p(x) = \mathbb{P}[\mathcal{F}(x)]$, where the failure domain

$$\mathcal{F}(x) = \bigcup_{i \in \mathbf{I}} \bigcap_{k \in \mathbf{C}_i} \{u \in \mathbb{R}^m \mid g_k(x, u) \leq 0\}, \tag{2}$$

with $\mathbf{C}_i \subset \mathbf{K}$ and $\mathbf{I} = \{1, 2, \ldots, I\}$ defining the combinations of performance measures that leads to structural failure. In [1], we focused on the case $K = 1$, $\mathbf{C}_1 = \{1\}$, and $I = 1$.

Solution approaches for stochastic programs tend to be based on either interior or exterior sampling. Interior sampling approaches aim to solve stochastic programs directly and resort to sampling techniques whenever the algorithm requires the evaluation of probability functions, or more generally, expectations. Usually, different samples are generated each time an evaluation is necessary. In this group of approaches we find stochastic quasigradient methods [6–8]. These methods are difficult to apply to problems involving failure probability constraints. In principle, such constraints can be removed by penalty or barrier terms in the objective function. However, the details of an implementable algorithm for nonlinear problems based on this principle do not appear to have been worked out.

Exterior sampling techniques construct and solve a sample average approximation without further sampling during the optimization. The results available for exterior sampling techniques include the fact that minimizers and minimum values of sample average approximations converge with probability one to the minimizers and the minimum value of the original problem, respectively, as the number of samples goes to infinity (see Chap. 6 of [9] and references therein). For techniques for checking whether a given point is close to stationarity see e.g. Sect. 6.4 of [9]. These results provide guidance for the selection of approximating problems to be solved using some deterministic optimization algorithm.

The authors of [10] present a framework based on sample average approximations, for proving convergence of nonlinear programs involving probabilities or expectations. Only a rather weak sense of convergence is presented in [10]: every accumulation point of the sequence of function values generated by the algorithm is bounded from above by the largest function value at any stationary point. In addition, the details of an implementable algorithm for problems with failure probability constraints are not provided.

Recently, elements of interior and exterior sampling techniques have been combined. In [11], sample average approximations are used to derive a gradient-type search direction for nonlinear stochastic programs. For each iteration, resampling is performed to generate a new search direction. Under the assumption of convex, twice-differentiable functions, it is shown that the expected distance to a Karush–Kuhn–Tucker point vanishes with increasing iterations when the search direction is combined with sufficiently small stepsizes. However, it is not clear how to implement a satisfactory stepsize rule. In addition, the literature contains a large number of approximate or heuristic approaches for solving (**P**) and similar, more specialized problems. A review of such results is found in [12].

The failure probability is continuously differentiable under broad conditions when $\mathcal{F}(x)$ is bounded and given by a union of events [13]. However, the derivative formula in [13] is difficult to use in estimation because it may involve surface integrals. In [14] (see also [8]), an integral transformation is presented, which, when it exists, leads to a simple formula for $\nabla p(x)$. However, it is not clear under what conditions the transformation exists. As in [13], Ref. [15] assumes that $\mathcal{F}(x)$ is bounded and given by a union of events. With this restriction, a formula for $\nabla p(x)$ involving integration over a simplex is derived. In principle, this integral can be evaluated by Monte Carlo methods. However, to the authors' knowledge, there is no computational experience with estimation of failure probabilities for highly reliable mechanical structures using this formula. In Sect. 9.2 of [2], a formula for $\nabla p(x)$ is suggested, without a complete proof, for the case $K = 1$. This formula is based on an expression for $p(x)$ that has been found computationally efficient in applications.

In the next section, we generalize the special-case formula for $\nabla p(x)$ found in [2] and provide a proof. We also present estimators for $p(x)$ and $\nabla p(x)$ and discuss their properties. For completeness, Sect. 3 presents Algorithm Model 3.3.27 from [16], which we use to develop our algorithm for solving (**P**). Section 4 derives an implementation of the algorithm model by generalizing the results of [1]. Section 5 presents a numerical example.

2 Failure Probability

As indicated in Sect. 1, a significant difficulty is associated with deriving a tractable formula for the gradient of $p(x)$.[1] To overcome this difficulty, we decompose the vector u into a direction w and a positive length r as described in further detail below (see [17] for the first application of such an decomposition).

We need the component failure domain $\mathcal{F}_k(x)$ defined by $\mathcal{F}_k(x) = \{u \in \mathbb{R}^m \mid g_k(x,u) \le 0\}$, $k \in \mathbf{K}$, and the surface of the unit hypersphere denoted by $\mathbb{B} = \{w \in \mathbb{R}^m \mid \|w\| = 1\}$. The following assumption is sufficient to ensure equivalence between $p(x)$ and its alternative expression in Proposition 2.1 below.

Assumption 2.1 We assume that:

(i) The distribution $\mathbb{P}$ of the random m-vector u is standard normal.
(ii) For a given $S \subset \mathbb{R}^n$, $\mathcal{F}_k(x)^c$ $(= \mathbb{R}^m - \mathcal{F}_k(x))$ is star-shaped with respect to the origin for all $k \in \mathbf{K}$ and $x \in S$, i.e., for all $x \in S$, $0 \in \mathcal{F}_k(x)^c$ and for every $w \in \mathbb{B}$ there either exists a unique $r > 0$ such that $g_k(x, rw) = 0$ or $g_k(x, rw) \ne 0$ for all $r > 0$.

In view of Assumption 2.1(i) and the fact that mechanical structures have small failure probabilities, we conclude that $\mathcal{F}_k(x)$ tends to be located far from the "high probability region" close to $0 \in \mathbb{R}^m$. Hence, the condition that $0 \in \mathcal{F}_k(x)^c$ is almost always satisfied for mechanical structures. The second part of Assumption 2.1(ii) is satisfied when $g_k(x,\cdot)$ is affine for all $x \in S$ and $k \in \mathbf{K}$, which is approximately true for many mechanical structures (Sects. 4.1 and 5.2 of [2]). However, in general, it can be difficult to verify the second part of Assumption 2.1(ii) analytically. This is especially the case when $g_k(\cdot,\cdot)$, $k \in \mathbf{K}$, are given by the solutions of some (differential) equations. Nevertheless, it is possible to obtain numerical indications of the validity of Assumption 2.1(ii) by estimating the failure probability using an alternative estimator as explained below. We also note that equivalent assumptions were adopted by [15, 18, 19] and Sect. 9.2 of [2].

Let P and E be the uniform distribution on $\mathbb{B}$ and the corresponding expectation, respectively. Furthermore, we define $r_k : \mathbb{R}^n \times \mathbb{B} \to [0, \infty]$, $k \in \mathbf{K}$, to be the smallest nonnegative solution of $g_k(x, rw) = 0$ for a given $x \in \mathbb{R}^n$ and $w \in \mathbb{B}$, i.e.,

$$r_k(x,w) = \begin{cases} \min\limits_r \{r \mid g_k(x,rw) \le 0, r \ge 0\}, & \text{if } \{r \mid g_k(x,rw) \le 0, r \ge 0\} \ne \emptyset, \\ \infty, & \text{otherwise.} \end{cases} \tag{3}$$

Note that, under Assumption 2.1(ii), $0 \in \mathcal{F}_k(x)^c$ and hence $r_k(x,w) > 0$.

Proposition 2.1 *If Assumption 2.1 holds at $x \in \mathbb{R}^n$, then*

$$p(x) = E\big[\phi(x,w)\big], \tag{4}$$

[1] We will not make use of this fact, but note that $p(x)$ can be estimated using standard Monte Carlo techniques independently of the results in this section.

where

$$\phi(x,w) = \max_{i\in\mathbf{I}} \min_{k\in\mathbf{C}_i} \left\{1 - \chi_m^2\left(r_k^2(x,w)\right)\right\} \tag{5}$$

and $\chi_m^2(\cdot)$ *is the chi-square cumulative distribution function with* m *degrees of freedom.*

Proof As in [17] (see alternatively [18, 19], and Sect. 9.2 of [2]), we observe that, if the standard normal random vector $u = rw$ and r^2 is chi-square distributed with m degrees of freedom, then w is a random vector, independent of r, uniformly distributed over the surface of the m-dimensional unit hypersphere. Hence, using an equivalent minimax expression of (2) and the total probability rule we obtain that

$$p(x) = E\left[P_r\left[\left\{\min_{i\in\mathbf{I}} \max_{k\in\mathbf{C}_i} g_k(x, rw) \le 0\right\} \mid w\right]\right], \tag{6}$$

where P_r is the chi-square distribution. In view of Assumption 2.1, the expression inside the expectation in (6) equals $1 - \chi_m^2(r^2(x,w))$, where $r(x,w) = \min_{i\in\mathbf{I}} \max_{k\in\mathbf{C}_i} r_k(x,w)$. Geometrically, $r(x,w)$ is the minimum distance in direction w from $0 \in \mathbb{R}^m$ to $\mathcal{F}(x)$. Since $\chi_m^2(\cdot)$ and the square function (positive domain) are strictly increasing, the result follows. □

Informally, we note that $\phi(x,w)$ is the probability of a failure event in the direction of w for a given x. We also observe that when Assumption 2.1(ii) is not satisfied, (4) may overestimate the failure probability. Hence, it is conservative to assume that Assumption 2.1(ii) is satisfied. For a given $x \in \mathbb{R}^n$, it is possible to get an indication of whether Assumption 2.1(ii) holds by computing an estimate $\sum_{j=1}^{N} I_{\mathcal{F}(x)}(u_j)/N$ of $p(x)$, where $u_1, u_2, \ldots, u_N$ are independent, identically distributed standard normal vectors and $I_{\mathcal{F}(x)}(u_j) = 1$ if $u_j \in \mathcal{F}(x)$, and zero otherwise. If this estimate is significantly smaller than the one of (4), then Assumption 2.1(ii) is not satisfied.

For a given $x \in \mathbb{R}^n$ and $w \in \mathbb{B}$, let the set of active performance functions

$$\hat{\mathbf{K}}(x,w) = \left\{k \in \mathbf{K} \mid k \in \hat{\mathbf{C}}_i(x,w), i \in \hat{\mathbf{I}}(x,w)\right\}, \tag{7}$$

where

$$\hat{\mathbf{I}}(x,w) = \left\{i \in \mathbf{I} \mid \min_{i'\in\mathbf{I}} \max_{k\in\mathbf{C}_{i'}} r_k(x,w) = \max_{k\in\mathbf{C}_i} r_k(x,w)\right\}, \tag{8}$$

$$\hat{\mathbf{C}}_i(x,w) = \left\{k \in \mathbf{C}_i \mid \max_{k'\in\mathbf{C}_i} r_{k'}(x,w) = r_k(x,w)\right\}. \tag{9}$$

Furthermore, let $\mathcal{A}_k(x)$ be the w-directions where k is active, i.e.,

$$\mathcal{A}_k(x) = \left\{w \in \mathbb{B} \mid k \in \hat{\mathbf{K}}(x,w)\right\}. \tag{10}$$

We now show that $p(\cdot)$ is continuously differentiable under the following assumptions.

Assumption 2.2 We assume that, for a given set $S \subset \mathbb{R}^n$, the following hold:

(i) There exists a constant $C_1 < \infty$ such that $\min_{i\in\mathbf{I}} \max_{k\in\mathbf{C}_i} r_k(x, w) \le C_1$ for all $x \in S$ and $w \in \mathbb{B}$.
(ii) The performance functions $g_k(x, u), k \in \mathbf{K}$, are continuously differentiable in both arguments for all $x \in S, u \in \mathbb{R}^m$.
(iii) There exists a constant $C_2 > 0$ such that $|\nabla_u g_k(x, r_k(x, w)w)^T w| \ge C_2$ for all $x \in S$, $w \in \mathcal{A}_k(x)$, and $k \in \mathbf{K}$.
(iv) $P[\mathcal{A}_k(x) \cap \mathcal{A}_l(x)] = 0$ for all $x \in S$ and $k, l \in \mathbf{K}, k \ne l$.

The particular set of performance functions $g_k(\cdot,\cdot), k \in \mathbf{K}$, arising in an application may not satisfy Assumption 2.2(i). However, it is always possible to define an artificial performance function $g_{K+1}(x, u) = \rho - \|u\|$, with a sufficiently large $\rho > 0$, replace I by $I + 1$, and set $\mathbf{C}_I = \{K + 1\}$. Then, $\mathcal{F}(x)$ satisfies Assumption 2.2(i). This is equivalent to enlarging the failure domain. The probability associated with the enlarged failure domain is slightly larger than the one associated with the original failure domain. The difference, however, is no greater than $1 - \chi^2_m(\rho^2)$ and therefore is negligible for sufficiently large ρ. Consequently, Assumption 2.2(i) is not restrictive in practice.

Assumption 2.2(iii) can be difficult to verify. However, it is our experience that the values of performance functions arising in practice tend to change as one moves from $\mathcal{F}(x)^c$ into the interior of the failure domain for a fixed x. If this was not the case, a perturbation of a scenario would have resulted in no change in the performance measure, which is unlikely in mechanical structures. Assumption 2.2(iv) states that only one performance function is active at each point on the boundary of the failure domain almost surely. Since the performance functions represent different performance measures in a structure, they tend to have quite different forms. If two performance functions are identical on significant subsets, then one of them is redundant and remodelling is appropriate.

Lemma 2.1 *Suppose that Assumptions 2.1 and 2.2 hold on a sufficiently large subset of $\mathbb{R}^n$ containing a compact set X_0. Then*:

(i) *There exists a constant $C < \infty$ such that $|\phi(x, w) - \phi(x', w)| \le C\|x - x'\|$ for all $x, x' \in X_0$ and $w \in \mathbb{B}$.*
(ii) *For each fixed $x \in X_0$, $\phi(\cdot, w)$ is continuously differentiable at x for P-almost all $w \in \mathbb{B}$.*
(iii) *The collections $\{\phi(x, w)\}_{x\in X_0}$ and $\{\nabla_x \phi(x, w)\}_{x\in X_0}$ are uniformly integrable, i.e.,*

$$\lim_{\gamma\to\infty} \sup_{x\in X_0} \int_{\{w\in\mathbb{B}||\phi(x,w)|\ge\gamma\}} \big|\phi(x, w)\big| P(dw) = 0 \tag{11}$$

and similarly with $|\phi(x, w)|$ replaced by $\|\nabla_x \phi(x, w)\|$.

Proof First, consider (i). By Assumption 2.2(i), $r_k(x, w) < \infty$ for all $x \in X_0$, $w \in \mathbb{B}$, and $k \in \hat{\mathbf{K}}(x, w)$. Hence, in view of Assumptions 2.2(ii, iii) and 2.1(ii), the implicit function theorem gives that $r_k(\cdot, w)$ is continuously differentiable for all $x \in X_0$, $w \in \mathbb{B}$, and $k \in \hat{\mathbf{K}}(x, w)$, and

$$\nabla_x r_k(x, w) = -\nabla_x g_k\big(x, r_k(x, w)w\big)/\nabla_u g_k\big(x, r_k(x, w)w\big)^T w, \tag{12}$$

when defined. Consequently, it can be deduced from Theorem 5.4.5 in [16] that for all $w \in \mathbb{B}$, $\phi(x, w)$ has directional derivatives on $x \in X_0$ in all directions with respect to its first argument and the directional derivative in the direction $h \in \mathbb{R}^n$ is defined by

$$d\phi(x, w; h) = \max_{i \in \hat{\mathbf{I}}(x,w)} \min_{k \in \hat{\mathbf{C}}_i(x,w)} -2 f_{\chi^2_m}\big(r^2_k(x, w)\big) r_k(x, w) \nabla_x r_k(x, w)^T h, \tag{13}$$

where $f_{\chi^2_m}(\cdot)$ is the probability density function of the chi-square distribution. Using Assumption 2.2(iii) and the fact that $r_k(x, w), k \in \hat{\mathbf{K}}(x, w)$ are bounded on X_0 for all $w \in \mathbb{B}$, (12) is bounded for all $x \in X_0$, $w \in \mathbb{B}$, and $k \in \hat{\mathbf{K}}(x, w)$. Since the max-min in (13) is only over $i \in \hat{\mathbf{I}}(x, w)$ and $k \in \hat{\mathbf{C}}_i(x, w)$, it follows from the definition of $\hat{\mathbf{K}}(x, w)$ and the boundedness of (12) that (13), given an $h \in \mathbb{R}^n$, is bounded, for all $x \in X_0$ and $w \in \mathbb{B}$. Hence, $|\phi(x, w) - \phi(x', w)| \le C\|x - x'\|$, for all $x, x' \in X_0$ and $w \in \mathbb{B}$.

Now, consider (ii). Let $x' \in X_0$ be arbitrary. For P-almost all $w \in \mathbb{B}$, it follows from Assumption 2.2(iv) that the ray from $0 \in \mathbb{R}^m$ in the direction of w does not intersect $\mathcal{A}_k(x') \cap \mathcal{A}_l(x')$ for any $k, l \in \mathbf{K}, k \ne l$. Let $w' \in \mathbb{B}$ be such that the corresponding ray has this property. Consequently, the set $\hat{\mathbf{K}}(x', w')$ has cardinality one. Suppose that $k' = \hat{\mathbf{K}}(x', w')$. Then, $g_{k'}(x', r(x', w')w') = 0$, where $r(x, w) = \min_{i \in \mathbf{I}} \max_{k \in \mathbf{C}_i} r_k(x, w)$. By Assumption 2.1(ii) and continuity, there exists a neighborhood $X'_0 \subset X_0$ of x' such that $k' = \hat{\mathbf{K}}(x, w')$ and $g_{k'}(x, r(x, w')w') = 0$ for all $x \in X'_0$. Hence, $\phi(x, w') = 1 - \chi^2_m(r^2_{k'}(x, w'))$ for all $x \in X'_0$ and, due to the smoothness of $r_{k'}(\cdot, w')$, $\phi(x, w')$ is continuously differentiable at x' with

$$\nabla_x \phi(x', w') = 2 f_{\chi^2_m}\big(r^2_{k'}(x', w')\big) r_{k'}(x', w') \frac{\nabla_x g_{k'}(x', r_{k'}(x', w')w')}{\nabla_u g_{k'}(x', r_{k'}(x', w')w')^T w'}. \tag{14}$$

Finally, consider (iii). Clearly, $|\phi(x, w)| \le 1$ for all $x \in X_0$ and $w \in \mathbb{B}$ and (11) holds. By Assumption 2.2(i–iii) and the fact that X_0 is compact, it follows that for all $w \in \mathbb{B}$, the right-hand side of (14) is uniformly bounded on X_0. Hence, $\{\nabla_x \phi(x, w)\}_{x \in X_0}$ is uniformly integrable. □

In view of Lemma 2.1, the next result follows directly from Proposition 2.1 in [20].

Proposition 2.2 *If Assumptions 2.1 and 2.2 hold on a sufficiently large subset of* $\mathbb{R}^n$ *containing a convex and compact set* X_0, *then* $p(\cdot)$ *is continuously differentiable on* X_0 *and its gradient is given by*

$$\nabla p(x) = E\big[\nabla_x \phi(x, w)\big], \tag{15}$$

where $\nabla_x \phi(x, w)$ *is given in* (14) *with* $k' \in \hat{\mathbf{K}}(x, w)$.

We estimate the expectations in (4) and (15) by Monte Carlo sampling. Consider an infinite sequence of sample points, each generated by independent sampling from P. Let $\overline{\mathbb{B}} = \mathbb{B} \times \mathbb{B} \times \cdots$ and let $\overline{P}$ be the probability distribution on $\overline{\mathbb{B}}$ generated by P.

Let subelements of $\overline{w} \in \overline{\mathbb{B}}$ be denoted $w_j \in \mathbb{B}$, $j = 1, 2, \ldots$, i.e., $\overline{w} = (w_1, w_2, \ldots)$. For every $\overline{w} \in \overline{\mathbb{B}}$, we define the estimator of (4),

$$p_N(x, \overline{w}) = \sum_{j=1}^{N} \phi(x, w_j)/N. \tag{16}$$

The asymptotic property of this estimator is given by the next well-known result (see e.g. [21] for a proof which holds under weaker assumptions than those stated here.).

Proposition 2.3 *If Assumptions 2.1 and 2.2 hold on a sufficiently large subset of* $\mathbb{R}^n$ *containing a compact set* X_0, *then, for* $\overline{P}$*-almost all* $\overline{w} \in \overline{\mathbb{B}}$,

$$\lim_{N\to\infty} \sup_{x\in X_0} \left| p_N(x, \overline{w}) - p(x) \right| = 0. \tag{17}$$

Since $\phi(\cdot, w)$ is only Lipschitz continuous (Lemma 2.1(i)), we observe that $p_N(\cdot, \overline{w})$ is generally nonsmooth. However, as the next proposition shows, $p_N(\cdot, \overline{w})$ has directional derivatives and a nonempty subgradient[2] for all $\overline{w} \in \overline{\mathbb{B}}$.

Proposition 2.4 *Suppose that Assumptions 2.1 and 2.2 hold on a sufficiently large subset of* $\mathbb{R}^n$ *containing a compact set* X_0. *Then, for all* $\overline{w} \in \overline{\mathbb{B}}$, $p_N(\cdot, \overline{w})$ *is Lipschitz continuous on* X_0 *and has a nonempty subgradient* $\partial p_N(x, \overline{w})$ *defined by*

$$\partial p_N(x, \overline{w}) = \sum_{j=1}^{N} \partial \phi(x, w_j)/N, \tag{18}$$

where

$$\partial \phi(x, w_j) = \operatorname*{conv}_{k\in\hat{\mathbf{K}}(x,w_j)} \left\{ 2 f_{\chi^2_m}\big(r_k^2(x, w_j)\big) r_k(x, w_j) \frac{\nabla_x g_k(x, r_k(x, w_j) w_j)}{\nabla_u g_k(x, r_k(x, w_j) w_j)^T w_j} \right\}, \tag{19}$$

$f_{\chi^2_m}$ *being the chi-square probability density function.*

Proof It follows from Lemma 2.1(i), that for all $\overline{w} \in \overline{\mathbb{B}}$, $p_N(\cdot, \overline{w})$ is Lipschitz continuous on X_0. It can be deduced from Theorem 5.4.5 in [16] that for all $w \in \mathbb{B}$, $\phi(\cdot, w)$ has directional derivatives with respect to its first argument in all directions. Hence, it follows from Lemma 1 in [23] that for all $\overline{w} \in \overline{\mathbb{B}}$, $p_N(x, \overline{w})$ has directional derivatives in all directions for all $x \in X_0$. Furthermore, a slight generalization of Corollary 5.4.6 in [16] yields that the directional derivative of $p_N(x, \overline{w})$ is identical to the (Clarke) generalized directional derivative of $p_N(x, \overline{w})$. Hence, $\partial p_N(x, \overline{w})$ is identical to the (Clarke) generalized gradient of $p_N(x, \overline{w})$, which is nonempty. A slight extension of Corollary 5.4.6 in [16], yields the formula above. □

By Lemma 2.1, the next result follows directly from Proposition 2.2 in [20].

[2] See e.g. Definition 5.1.31 in [16]. This type of subgradient is sometimes referred to as a regular subgradient (Definition 8.3 in [22]).

Proposition 2.5 *Suppose that Assumptions 2.1 and 2.2 hold on a sufficiently large subset of $\mathbb{R}^n$ containing a convex and compact set X_0. Then, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, $\partial p_N(x, \overline{w})$ converges uniformly to $\nabla p(x)$, i.e.,*

$$\lim_{N\to\infty} \sup_{x\in X_0} \sup_{d\in\partial p_N(x,\overline{w})} \left\| d - \nabla p(x) \right\| = 0. \tag{20}$$

Using (16), we define a sequence of approximating problems. For any $\overline{w} \in \overline{\mathbb{B}}$ and $N \in \mathbb{N} = \{1, 2, \ldots\}$, let the sample average approximating problem $(\mathbf{P}_N(\overline{w}))$ be defined by

$$(\mathbf{P}_N(\overline{w})) \quad \min_{x\in\mathbb{R}^n} \left\{ c(x) \mid p_N(x, \overline{w}) \le q,\ x \in X \right\}. \tag{21}$$

Intuitively, $(\mathbf{P}_N(\overline{w}))$ becomes a better "approximation" to $(\mathbf{P})$ as N increases. In fact, epi-convergence characterizes this effect more precisely, as we see in the next proposition (see e.g. Theorems 3.3.2–3.3.3 in [16] for a proof), which requires a constraint qualification:

Assumption 2.3 Given $\overline{w} \in \overline{\mathbb{B}}$, we assume that for every $x \in X$ satisfying $p(x) \le q$, there exists a sequence $\{x_N\}_{N=1}^{\infty} \subset X$, with $p_N(x_N, \overline{w}) \le q$, such that $x_N \to x$, as $N \to \infty$.

Proposition 2.6 *Consider the sequence of approximate problems $\{\mathbf{P}_N(\overline{w})\}_{N=1}^{\infty}$. Suppose that Assumptions 2.1 and 2.2 hold on a sufficiently large convex and compact subset of $\mathbb{R}^n$. Then, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, the following holds*:

(i) *If Assumption 2.3 is satisfied at $\overline{w} \in \overline{\mathbb{B}}$, then $\{\mathbf{P}_N(\overline{w})\}_{N=1}^{\infty}$ epi-converges to $\mathbf{P}$.*
(ii) *If Assumption 2.3 is satisfied at $\overline{w} \in \overline{\mathbb{B}}$ and $\{\hat{x}_N\}_{N=1}^{\infty}$ is a sequence of global minimizers of $\{\mathbf{P}_N(\overline{w})\}_{N=1}^{\infty}$, then every accumulation point of $\{\hat{x}_N\}_{N=1}^{\infty}$ is a global minimizer of $\mathbf{P}$.*

Before presenting an algorithm, we need to strengthen the result in Proposition 2.3.

Proposition 2.7 *Suppose that Assumptions 2.1 and 2.2 hold on a sufficiently large subset of $\mathbb{R}^n$ containing a compact set X_0. Then, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, there exists a constant $C < \infty$ such that, for all $x \in X_0$ and $N \in \mathbb{N}$,*

$$\left| p_N(x, \overline{w}) - p(x) \right| \le C\sqrt{(\log\log N)/N}. \tag{22}$$

Proof Let G be defined by $G(x, w) = \phi(x, w) - p(x)$ for all $x \in X_0, w \in \mathbb{B}$. By Lemma 2.1 and (4), G is centered and $G(\cdot, w) \in C(X_0)$, where $C(X_0)$ is the space of continuous functions on X_0. Furthermore, by Assumptions 2.1 and 2.2(ii, iii) and by the implicit function theorem, $r_k(\cdot, \cdot)$ is continuous for all $x \in X_0$, $w \in \mathbb{B}$, and $k \in \hat{\mathbf{K}}(x, w)$. Hence, it can be deduced from Corollary 5.4.4 in [16] that $r(\cdot, \cdot) = \min_{i\in\mathbf{I}} \max_{k\in\mathbf{C}_i} r_k(\cdot, \cdot)$ is continuous on $X_0 \times \mathbb{B}$. Since X_0 and $\mathbb{B}$ are compact, it follows that $r(\cdot, \cdot)$ is uniformly continuous on $X_0 \times \mathbb{B}$. Let $\tilde{r} : \mathbb{B} \to C(X_0)$ be defined by $\tilde{r}(w) = r(\cdot, w)$. Then, it follows that $\tilde{r}$ is continuous on $\mathbb{B}$ and hence measurable

with respect to the Borel sets in $C(X_0)$. Consequently, G is also measurable. Hence, G is a random variable with values in a separable Banach space. Define $\|G\|_\infty = \sup_{x\in X_0} |G(x,w)|$. A corollary of Theorem 8.11 in [24, p. 217], is that if G is a centered Banach space valued random variable such that $E(\|G\|_\infty^2/\log\log\|G\|_\infty) < \infty$ and such that the central limit theorem holds for G then the law of the iterated logarithm also holds for G. Since G is bounded, $E(\|G\|_\infty^2/\log\log\|G\|_\infty) < \infty$. Hence, it only remains to show that the central limit theorem holds for G.

Let $N(\epsilon, X_0, \|\cdot\|)$ be the covering number, i.e., the minimal number of open balls $B^o(x,\epsilon) = \{x' \in \mathbb{R}^n | \|x' - x\| < \epsilon\}$ needed to cover X_0. Since X_0 is compact, there exists a constant $\eta < \infty$ such that $N(\epsilon, X_0, \|\cdot\|) \le (\eta/\epsilon)^n$ for all $\epsilon > 0$. Hence, the entropy integral

$$\int_0^\infty \sqrt{\log N(\epsilon, X_0, \|\cdot\|)}d\epsilon \le \int_0^\eta \sqrt{n(\log\eta - \log\epsilon)}d\epsilon < \infty. \tag{23}$$

By Lemma 2.1(i), G has Lipschitz continuous sample paths with a square-integrable Lipschitz constant. Hence, by the Jain–Marcus theorem (see [25], Example 2.11.13), it follows that the central limit theorem holds. Consequently, the law of iterated logarithm holds for G. Hence, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and all $x \in X_0$, (22) holds for some C and sufficiently large N. By increasing C, the result can be made to hold for all N. □

3 Algorithm Model

Before presenting an algorithm model, we present the optimality conditions for (**P**). Under Assumptions 2.1 and 2.2, (**P**) is a nonlinear program involving continuously differentiable functions, with stationary points defined by the F. John conditions. We find it convenient to express the F. John conditions (see Theorems 2.2.4 and 2.2.8 in [16]) by means of a nonpositive, continuous, optimality function $\theta : \mathbb{R}^n \to \mathbb{R}$, which is defined by

$$\theta(x) = -\min_{z\in Z}\{z^T b(x) + z^T B(x)^T B(x) z/(2\delta)\}, \tag{24}$$

with $Z = \{z \in \mathbb{R}^{J+2} \mid \sum_{l=1}^{J+2} z^{(l)} = 1,\ z^{(l)} \ge 0,\ l = 1,\ldots,J+2\}$,

$$b(x) = \big(\gamma\psi(x)_+, \psi(x)_+ - p(x) + q, \psi(x)_+ - f_1(x), \ldots, \psi(x)_+ - f_J(x)\big)^T, \tag{25}$$

and $B(x) = (\nabla c(x), \nabla p(x), \nabla f_1(x), \ldots, \nabla f_J(x))$, where $\psi(x) = \max\{p(x) - q, \max_{j\in\mathbf{J}} f_j(x)\}$, $\psi(x)_+ = \max\{0, \psi(x)\}$, and $\gamma, \delta > 0$. For any $\hat{x} \in X$ such that $p(\hat{x}) \le q$, the F. John conditions hold if and only if $\theta(\hat{x}) = 0$ (Theorem 2.2.8 in [16]).

Since neither $p(x)$ nor $\nabla p(x)$ can be evaluated exactly in finite computing time, an algorithm for (**P**) involving the evaluations of $p(x)$ and $\nabla p(x)$ is conceptual. We construct an implementable algorithm by using Algorithm Model 3.3.27 in [16]. For completeness, the algorithm model is presented below. The algorithm model makes use of an approximate algorithm map $A_{N,\overline{w}} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ involving $p_N(x, \overline{w})$ and

$\partial p_N(x, \overline{w})$. Note that the algorithm map can be set-valued. The algorithm model also uses the function $F_N : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ in a precision-adjustment rule, where

$$F_N(x', x'') = \max\{c(x'') - c(x') - \gamma \psi_N(x')_+, \psi_N(x'') - \psi_N(x')_+\}, \quad (26)$$

$$\psi_N(x) = \max\left\{ p_N(x, \overline{w}) - q, \max_{j \in \mathbf{J}} f_j(x) \right\}, \quad (27)$$

$\psi_N(x)_+ = \max\{0, \psi_N(x)\}$, and $\gamma > 0$.

Algorithm Model for Solving (P) (Adapted from Algorithm Model 3.3.27, [16])
Parameters $\tau \in (0, 1)$, $\eta > 0$.
Data $x_0 \in \mathbb{R}^n$, an unbounded set $\mathcal{N}$ of positive integers, and a collection $\overline{w} = (w_1, w_2, \ldots) \in \overline{\mathbb{B}}$ of independent sample points from P.
Step 0 Set $i = 0$ and $N = \min \mathcal{N}$.
Step 1 Compute $y \in A_{N,\overline{w}}(x_i)$.
Step 2 If

$$F_N(x_i, y) \leq -\eta \left(\sqrt{(\log \log N)/N} \right)^{\tau}, \quad (28)$$

then set $x_{i+1} = y$, $N_i = N$, replace i by $i + 1$, and go to Step 1.
Else, replace N by $\min\{N' \in \mathcal{N} | N' > N\}$, and go to Step 1.

Note that the algorithm model uses a precision-adjustment rule (28) to ensure that the error in function evaluations is sufficiently small in comparison to the algorithmic progress. Given τ and η as well as a set of sample sizes $\mathcal{N}$, the rule determines how fast the sample size is increased. Empirical evidence from the areas of optimal control and semi-infinite optimization indicates that such a feedback rule is computationally more efficient than using a predetermined schedule for increasing the sample size. To ensure convergence of the algorithm model, we adopt the following assumption regarding the algorithm map in Step 1. For brevity, for any $x \in \mathbb{R}^n$ and $\rho > 0$, let $B(x, \rho) = \{x' \in \mathbb{R}^n | \|x - x'\| \leq \rho\}$.

Assumption 3.1 Given $S \subset \mathbb{R}^n$, we assume that the algorithm map $A_{N,\overline{w}} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ satisfies the following property for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$:

For every $x \in S$ with $\theta(x) < 0$, there exist $\delta_x > 0$, $N_x \in \mathbb{N}$, and $\rho_x > 0$ such that $F_N(x', y) \leq -\delta_x$ for all $N \geq N_x$, $x' \in B(x, \rho_x)$, and $y \in A_{N,\overline{w}}(x')$.

In the next section, we present one particular algorithm map that satisfies Assumption 3.1. The convergence of the algorithm model is given by the next theorem, which can be proven using the same arguments as in the proof of Theorem 3.3.29 in [16].

Theorem 3.1 *Suppose that Assumptions 2.1, 2.2, and 3.1 hold on a sufficiently large subset of $\mathbb{R}^n$ and that the algorithm model for solving* (**P**) *has constructed a bounded sequence $\{x_i\}_{i=0}^{\infty}$. If $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, then $\hat{x}$ is a stationary point for* (**P**), *i.e.*, $\theta(\hat{x}) = 0$, $\overline{P}$*-almost surely.*

4 Implementation

Consider the following set-valued algorithm map, which is a generalization of the Polak–He algorithm, see Sect. 2.6 in [16]. For any $x_i \in \mathbb{R}^n$, $N \in \mathbb{N}$, and $\overline{w} \in \overline{\mathbb{B}}$, we define

$$A_{N,\overline{w}}(x_i) = \{x_i + \lambda_N(x_i, d)h_N(x_i, d) \mid d \in \partial p_N(x_i, \overline{w})\}, \tag{29}$$

where the Armijo stepsize is given by

$$\lambda_N(x_i, d) = \max_{k\in\{0,1,2,\ldots,\}} \{\beta^k \mid F_N(x_i, x_i + \beta^k h_N(x_i, d)) \le \beta^k \alpha \theta_N(x_i, d)\}, \tag{30}$$

with $F_N(\cdot,\cdot)$ as in (26),

$$\theta_N(x, d) = -\min_{z\in Z}\{z^T b_N(x) + z^T B_N(x, d)^T B_N(x, d)z/(2\delta)\}, \tag{31}$$

where $\delta > 0$, Z is defined as in (24),

$$b_N(x) = \big(\gamma\psi_N(x)_+, \psi_N(x)_+ - p_N(x, \overline{w}) + q, \psi_N(x)_+ - f_1(x), \ldots, \psi_N(x)_+ - f_J(x)\big)^T, \tag{32}$$

$$B_N(x, d) = \big(\nabla c(x), d, \nabla f_1(x), \ldots, \nabla f_J(x)\big), \tag{33}$$

$\alpha \in (0, 1]$, and $\beta \in (0, 1)$. Finally, the search direction

$$h_N(x_i, d) = -B_N(x_i, d)\hat{z}/\delta, \tag{34}$$

where $\hat{z}$ is any solution of (31). The parameter γ in (26) and (31) should be set equal to γ in $F_N(\cdot,\cdot)$ in the algorithm model. The problem in (31) is quadratic and can be solved in a finite number of iterations by a standard QP-solver (e.g. Quadprog [26]).

Our next result shows that (29) satisfies Assumption 3.1. Hence, this algorithm map combined with the algorithm model result in a convergent implementable algorithm.

Proposition 4.1 *Suppose that Assumptions 2.1 and 2.2 hold on an open set $S \subset \mathbb{R}^n$. For any $N \in \mathbb{N}$ and $\overline{w} \in \overline{\mathbb{B}}$, let the algorithm map $A_{N,\overline{w}}(\cdot)$ be defined by (29), with the same values of the parameters α, β, δ, and γ for all $N \in \mathbb{N}$. Then, $A_{N,\overline{w}}(\cdot)$ satisfies Assumption 3.1 on any convex and compact subset of S for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$.*

Proof Let $X_0 \subset S$ be convex and compact. For $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, the search direction $h_N(x, d)$ is bounded for all $x \in X_0$ and $d \in \partial p_N(x, \overline{w})$ because it is defined as a linear combination of bounded vector-valued functions (see (34), (33), Proposition 2.4, and Assumption 2.2). For $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, the bound is independent of N due to Proposition 2.5. Since S is open, there exists a $\lambda_1 \in (0, 1]$ such that $x + \lambda h_N(x, d) \in S$ for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and for all $x \in X_0$, $\lambda \in (0, \lambda_1]$, $N \in \mathbb{N}$, and $d \in \partial p_N(x, \overline{w})$.

It is seen from Proposition 2.4 and its proof that for all $\overline{w} \in \overline{\mathbb{B}}$ and $x \in S$, $p_N(x, \overline{w})$ is Lipschitz continuous, with a directional derivative equal to the (Clarke) generalized directional derivative. Hence, the Lebourg mean-value theorem (see e.g. Theorem 5.4.13b in [16]) is applicable. By the Lebourg mean-value theorem and the fact

that $\lambda_1 \le 1$, we obtain that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and for all $x \in X_0$, $\lambda \in (0, \lambda_1]$, $d \in \partial p_N(x, \overline{w})$, and $N \in \mathbb{N}$, there exist some $s, s_j \in [0,1]$, $j = 0, 1, 2, \ldots, J$, and $d' \in \partial p_N(x + s\lambda h_N(x,d), \overline{w})$ such that

$$
\begin{aligned}
&F_N\big(x, x + \lambda h_N(x,d)\big) \\
&\quad \le \lambda \max\Big\{ -\gamma \psi_N(x)_+ + \nabla c(x)^T h_N(x,d) \\
&\qquad\qquad + \big(\nabla c\big(x + s_0 \lambda h_N(x,d)\big) - \nabla c(x)\big)^T h_N(x,d), \\
&\qquad\qquad p_N(x,\overline{w}) - q - \psi_N(x)_+ + d^T h_N(x,d) + (d' - d)^T h_N(x,d), \\
&\qquad\qquad \max_{j \in \mathbf{J}} \big\{ f_j(x) - \psi_N(x)_+ + \nabla f_j(x)^T h_N(x,d) \\
&\qquad\qquad + \big(\nabla f_j\big(x + s_j \lambda h_N(x,d)\big) - \nabla f_j(x)\big)^T h_N(x,d) \big\} \Big\}. \qquad (35)
\end{aligned}
$$

For any $\epsilon > 0$, it follows from Proposition 2.5 that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ there exists a $N_\epsilon \in \mathbb{N}$ such that, for all $N \ge N_\epsilon$

$$
\sup_{x \in S} \sup_{d \in \partial p_N(x,\overline{w})} \big\| d - \nabla p(x) \big\| \le \epsilon/3. \qquad (36)
$$

Furthermore, $\nabla c(\cdot)$, $\nabla f_j(\cdot)$, $j \in \mathbf{J}$, and $\nabla p(\cdot)$ are uniformly continuous on compact sets and $h_N(x,d)$ is bounded almost surely. Hence, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ there exists a $\lambda_\epsilon \le \lambda_1$ such that for all $\lambda \in (0, \lambda_\epsilon]$, $x \in X_0$, $d \in \partial p_N(x, \overline{w})$, and $N \in \mathbb{N}$ we have

$$
\big\| \nabla p\big(x + s\lambda h_N(x,d)\big) - \nabla p(x) \big\| \le \epsilon/3, \qquad (37)
$$

$$
\big\| \nabla c\big(x + s_0 \lambda h_N(x,d)\big) - \nabla c(x) \big\| \le \epsilon, \qquad (38)
$$

$$
\big\| \nabla f_j\big(x + s_j \lambda h_N(x,d)\big) - \nabla f_j(x) \big\| \le \epsilon, \ \ j \in \mathbf{J}. \qquad (39)
$$

Hence, using (36) and (37), we obtain that $\|d' - d\| \le \epsilon$. From (35), we then obtain that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and for all $x \in X_0$, $\lambda \in (0, \lambda_\epsilon]$, $d \in \partial p_N(x, \overline{w})$, and $N \ge N_\epsilon$,

$$
\begin{aligned}
&F_N\big(x, x + \lambda h_N(x,d)\big) \\
&\quad \le \lambda \max\Big\{ -\gamma \psi_N(x)_+ + \nabla c(x)^T h_N(x,d), \\
&\qquad\qquad p_N(x,\overline{w}) - q - \psi_N(x)_+ + d^T h_N(x,d), \\
&\qquad\qquad \max_{j \in \mathbf{J}} \big\{ f_j(x) - \psi_N(x)_+ + \nabla f_j(x)^T h_N(x,d) \big\} \Big\} + \lambda\epsilon \big\| h_N(x,d) \big\|. \qquad (40)
\end{aligned}
$$

We can deduce from Theorem 2.2.8 in [16] that

$$
\begin{aligned}
\theta_N(x,d) = \max\Big\{ &-\gamma \psi_N(x)_+ + \nabla c(x)^T h_N(x,d), \\
&p_N(x,\overline{w}) - q - \psi_N(x)_+ + d^T h_N(x,d),
\end{aligned}
$$

$$\max_{j\in\mathbf{J}}\left\{f_j(x)-\psi_N(x)_+ +\nabla f_j(x)^T h_N(x,d)\right\}\Big\}$$

$$+\delta\left\|h_N(x,d)\right\|^2/2. \tag{41}$$

Hence, by adding and subtracting $\lambda\delta\|h_N(x,d)\|^2/2$ from (40), we obtain

$$F_N\big(x,x+\lambda h_N(x,d)\big)\le\lambda\theta_N(x,d)+\lambda\big(\epsilon-\delta\left\|h_N(x,d)\right\|/2\big)\left\|h_N(x,d)\right\| \tag{42}$$

for $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$ and for all $x\in X_0$, $\lambda\in(0,\lambda_\epsilon]$, $d\in\partial p_N(x,\overline{w})$, and $N\ge N_\epsilon$.

Now, suppose that $x^*\in X_0$ is such that $\theta(x^*)<0$. Without loss of generality, we assume that x^* is in the interior of X_0. Let $h(x)$ be given by

$$h(x)=-B(x)^T\hat{z}/\delta, \tag{43}$$

where $B(x)$ is defined as in (24), $\delta>0$ is as in (24), and $\hat{z}$ is any solution of (24). Since $h(x^*)=0$ implies $\theta(x^*)=0$, $\|h(x^*)\|\ne 0$. Then, by continuity of $\theta(\cdot)$ and $h(\cdot)$ (see Theorem 2.2.8 in [16]), there exist $\delta_1>0$ and $\rho_{x^*}>0$ such that $\theta(x)\le-\delta_1$ and $\|h(x)\|\ge\delta_1$ for all $x\in B(x^*,\rho_{x^*})$. Set $\epsilon^*=\delta\delta_1/4$. By Proposition 7.1 (see Appendix), for $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$ there exists an $N_{x^*}\ge N_{\epsilon^*}$ such that $\theta_N(x,d)\le-\delta_1/2$ and $\|h_N(x,d)\|\ge\delta_1/2$ for all $x\in B(x^*,\rho_{x^*})$ and $d\in\partial p_N(x,\overline{w})$. Since $\epsilon^*-\delta\|h_N(x,d)\|/2)\le 0$ for $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$ and for all $x\in B(x^*,\rho_{x^*})$ and $d\in\partial p_N(x,\overline{w})$, it now follows from (42) that $F_N(x,x+\lambda h_N(x,d))\le\lambda\theta_N(x,d)\le 0$ for $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$ and for all $x\in B(x^*,\rho_{x^*})$, $\lambda\in(0,\lambda_{\epsilon^*}]$, $d\in\partial p_N(x,\overline{w})$, and $N\ge N_{x^*}$. Hence, for the algorithm parameter $\alpha\in(0,1]$,

$$F_N\big(x,x+\lambda h_N(x,d)\big)-\lambda\alpha\theta_N(x,d)\le\lambda(1-\alpha)\theta_N(x,d)\le 0 \tag{44}$$

for $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$ and for all $x\in B(x^*,\rho_{x^*})$, $\lambda\in(0,\lambda_{\epsilon^*}]$, $d\in\partial p_N(x,\overline{w})$, and $N\ge N_{x^*}$. Consequently, for any $x\in B(x^*,\rho_{x^*})$ and $N\ge N_{x^*}$, the algorithm map $A_{N,\overline{w}}(\cdot)$ has stepsize $\lambda_N(x,d)\ge\beta\lambda_{\epsilon^*}$ for any $d\in\partial p_N(x,\overline{w})$ and for $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$. Hence, for any $x\in B(x^*,\rho_{x^*})$, $d\in\partial p_N(x,\overline{w})$, and $N\ge N_{x^*}$,

$$F_N(x,y)\le\alpha\lambda_N(x,d)\theta_N(x,d)\le\alpha\beta\lambda_{\epsilon^*}\theta_N(x,d)\le-\alpha\beta\lambda_{\epsilon^*}\delta_1/2, \tag{45}$$

for all $y\in A_{N,\overline{w}}(x')$ and $\overline{P}$-almost all $\overline{w}\in\overline{\mathbb{B}}$. This completes the proof. □

Usually, the one-dimensional root finding problems in the evaluation of $r_k(x,w)$, $k\in\hat{\mathbf{K}}(x,w)$, cannot be solved exactly in finite computing time. One possibility is to introduce a precision parameter that ensures a gradually better accuracy in the root finding as the algorithm progresses. Alternatively, we can prescribe a rule saying that the root finding algorithm should terminate after CN_i iterations, with C being some constant. For simplicity, we have not discussed the issue of root finding. In fact, this issue is not problematic in practice. The root finding problems can be solved in a few iterations with high accuracy using standard algorithms. Hence, the root finding problems are solved with a fixed precision for all iterations in the algorithm giving a negligible error.

5 Numerical Example

We consider the same highway bridge as in [12]. The objective is to design a reinforced concrete girder for minimum cost using nine design variables ($n = 9$). Uncertainty is modeled using eight random variables ($m = 8$). We assume that the girder can fail in four different modes. Failure occurs if any of the four failure modes occur. This gives rise to four performance functions. To ensure that $\mathcal{F}(x)^c$ is bounded (see Assumption 2.2(i)), we define an artificial performance function $g_5(x, u) = 8 - \|u\|$. Note that this implies an enlargement of the failure domain, but the increase in the failure probability is less than 10^{-10}. This leads to five performance functions and $\mathbf{C}_i = \{i\}$, $i \in \mathbf{I} = \{1, 2, \ldots, 5\}$ (see (2)). We impose the constraint $p(x) \leq q = 0.001350$, as well as 23 other deterministic, nonlinear constraints. For details about the example, we refer to [12]. It should be noted that the performance functions are nonlinear and sufficiently differentiable. We are unable to verify analytically that the performance functions satisfy Assumption 2.1(ii). However, we estimate the failure probability for the first and last iterations using the estimator $\sum_{j=1}^{N} I_{\mathcal{F}(x)}(u_j)/N$ of $p(x)$, where $u_1, u_2, \ldots, u_N$ are independent, identically distributed standard normal vectors and $I_{\mathcal{F}(x)}(u_j) = 1$ if $u_j \in \mathcal{F}(x)$, and zero otherwise, and find it not significantly different from the estimates obtained using (16). Furthermore, we do not experience numerically difficulties, which could have been expected in an example not satisfying Assumption 2.1(ii). Hence, it is reasonable to believe that Assumption 2.1(ii) is satisfied over a sufficiently large subset.

The resulting instance of (**P**) is implemented in Matlab 6.5 [26] and solved using our algorithm model with the algorithm map defined in (29). The evaluation of $r_k(x, w)$ is performed using the Matlab root-finder Fzero, with tolerance 1×10^{-5}, and (31) is solved using Matlab's Quadprog. The parameters are $\tau = 0.9999$, $\eta = 0.002$, $\gamma = 2$, $\alpha = 0.5$, $\beta = 0.8$, and $\delta = 1$. Furthermore, $\mathcal{N} = \{200, 1600, 5400, 12800, 25000, \ldots\}$. The computations are terminated when the algorithm model reaches a sample size greater than 25000.

After an application of our new algorithm, we obtain an optimized structure with cost 13.288. In comparison, the design obtained in [12] has a somewhat larger cost of 13.664. In this example, a less reliable structure is also cheaper. As expected, when our algorithm was terminated the constraint $p(x) \leq 0.001350$ was (approximately) active: estimated failure probability is 0.001350 with coefficient of variation 0.02. An examination of the design from [12] shows that its failure probability is 0.001310 with coefficient of variation 0.01. Hence, a 95% confidence interval of the failure probability is (0.001284, 0.001336) which is outside the constraint limit 0.001350. From this analysis we conclude that the algorithm in [12] may give excessively safe designs. The algorithm in [12] is based on the heuristic updating of first-order approximations of the failure probability and is not expected to lead to the same accuracy level as our new algorithm. However, the algorithm in [12] appears to involve fewer evaluations of the performance functions and their gradients. Note also that the algorithm in [12] is limited to the special case of $\mathbf{C}_i$ having only one element for all $i \in \mathbf{I}$.

6 Conclusions

We construct an implementable algorithm for nonlinear stochastic programming problems with system failure probability constraints. First, we generalize an expression for the failure probability and show that it is continuously differentiable. Second, we prove a uniform strong law of large numbers for the estimators of the failure probability and its gradient. We also establish a uniform law of the iterated logarithm for the estimator of the failure probability. Third, we construct an algorithm map that satisfies the assumptions of an algorithm model. Preliminary numerical testing on a realistic design problem demonstrates the potential for sampling-based optimization algorithms in structural engineering. In particular, the high accuracy of such algorithms compared to frequently used heuristics is promising.

Appendix

Proposition 7.1 *Suppose that Assumptions 2.1 and 2.2 hold on a convex and compact set $X_0 \subset \mathbb{R}^n$. Then, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$,*

$$\lim_{N\to\infty} \sup_{x\in X_0} \sup_{d\in\partial p_N(x,\overline{w})} \left|\theta_N(x,d) - \theta(x)\right| = 0, \tag{46}$$

$$\lim_{N\to\infty} \sup_{x\in X_0} \sup_{d\in\partial p_N(x,\overline{w})} \left\|h_N(x,d) - h(x)\right\| = 0, \tag{47}$$

where $\theta(\cdot)$, $\theta_N(\cdot,\cdot)$, $h(\cdot)$, and $h_N(\cdot,\cdot)$ are defined in (24), (31), (43), and (34), respectively.

Proof Let $\epsilon > 0$ be arbitrary. We deduce from Propositions 2.3 and 2.5 that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ there exists a $N_\epsilon \in \mathbb{N}$ such that, for all $N \geq N_\epsilon$ and $h \in \mathbb{R}^n$

$$\sup_{x\in X_0} \left|\psi(x)_+ - \psi_N(x)_+\right| \leq \epsilon, \tag{48}$$

$$\sup_{x\in X_0} \sup_{d\in\partial p_N(x,\overline{w})} \left|d^T h - \nabla p(x)^T h\right| \leq \epsilon\|h\|. \tag{49}$$

Consequently, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and all $N \geq N_\epsilon$ and $h \in \mathbb{R}^n$

$$\begin{aligned}\sup_{x\in X_0} \sup_{d\in\partial p_N(x,\overline{w})} \big| & p_N(x,\overline{w}) - q - \psi_N(x)_+ + d^T h \\ & - (p(x) - q - \psi(x)_+ + \nabla p(x)^T h)\big| \leq 2\epsilon + \epsilon\|h\|.\end{aligned} \tag{50}$$

Let

$$\begin{aligned}\tilde{\psi}_N(x, x+h, d) = \max\Big\{ & -\gamma\psi_N(x)_+ + \nabla c(x)^T h,\ p_N(x,\overline{w}) - q - \psi_N(x)_+ + d^T h, \\ & \max_{j\in\mathbf{J}}\{f_j(x) - \psi_N(x)_+ + \nabla f_j(x)^T h\}\Big\} + \delta\|h\|^2/2,\end{aligned} \tag{51}$$

$$\begin{aligned}\tilde{\psi}(x, x+h) = \max\Big\{ & -\gamma\psi(x)_+ + \nabla c(x)^T h,\ p(x) - q - \psi(x)_+ + \nabla p(x)^T h, \\ & \max_{j\in\mathbf{J}}\{f_j(x) - \psi(x)_+ + \nabla f_j(x)^T h\}\Big\} + \delta\|h\|^2/2.\end{aligned} \tag{52}$$

Using (48), (50), and the fact that $c(\cdot)$ and $\nabla c(\cdot)$ are bounded functions on X_0, we find that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and all $N \geq N_\epsilon$ and $h \in \mathbb{R}^n$,

$$\sup_{x \in X_0} \sup_{d \in \partial p_N(x,\overline{w})} \left|\tilde{\psi}_N(x, x+h, d) - \tilde{\psi}(x, x+h)\right| \leq \max\left\{\gamma\epsilon, 2\epsilon + \epsilon\|h\|\right\}. \tag{53}$$

Next, $h(x)$ is bounded for all $x \in X_0$ because it is defined as a linear combination of bounded vector-valued functions. Using the same argument and Proposition 2.5 we have that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, $h_N(x,d)$ is bounded for all $x \in X_0$, $d \in \partial p_N(x,\overline{w})$, and $N \in \mathbb{N}$. Hence, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ there exists a $C^* < \infty$ such that $\|h(x)\| \leq C^*$ and $\|h_N(x,d)\| \leq C^*$ for all $x \in X_0$, $d \in \partial p_N(x,\overline{w})$, and $N \in \mathbb{N}$. From Theorem 2.2.8 of [16] we deduce that

$$\theta_N(x,d) = \tilde{\psi}_N\big(x, x + h_N(x,d), d\big) = \min_{h \in \mathbb{R}^n} \tilde{\psi}_N(x, x+h, d), \tag{54}$$

$$\theta(x) = \tilde{\psi}\big(x, x + h(x)\big) = \min_{h \in \mathbb{R}^n} \tilde{\psi}(x, x+h). \tag{55}$$

Let $\epsilon^* = \max\{\gamma\epsilon, 2\epsilon + \epsilon C^*\}$. We now have that for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$ and all $x \in X_0$, $d \in \partial p_N(x,\overline{w})$, and $N \geq N_{\epsilon^*}$,

$$\theta(x) \leq \tilde{\psi}\big(x, x + h_N(x,d)\big) \leq \tilde{\psi}_N\big(x, x + h_N(x,d)\big) + \epsilon^* \leq \theta_N(x) + \epsilon^*, \tag{56}$$

$$\theta(x) = \tilde{\psi}\big(x, x + h(x)\big) \geq \tilde{\psi}_N\big(x, x + h(x)\big) - \epsilon^* \geq \theta_N(x) - \epsilon^*. \tag{57}$$

Hence, (46) holds. We now address (47). For the sake of a contradiction, suppose that (47) is not valid. Then, there exists a subset $\overline{\mathbb{B}}_0 \subset \overline{\mathbb{B}}$ with $\overline{P}[\overline{\mathbb{B}}_0] > 0$ such that for every $\overline{w} \in \overline{\mathbb{B}}_0$ there exist $\epsilon > 0$, $\{N_i\}_{i=1}^\infty$, $N_i \to \infty$, as $i \to \infty$, $\{x_i\}_{i=1}^\infty \subset X_0$, $\{d_i\}_{i=1}^\infty$, $d_i \in \partial p_{N_i}(x_i, \overline{w})$, with the property that

$$\left\|h_{N_i}(x_i, d_i) - h(x_i)\right\| \geq \epsilon \tag{58}$$

for all $i \in \mathbb{N}$. As stated above, for $\overline{P}$-almost all $\overline{w} \in \overline{\mathbb{B}}$, $h_N(x,d)$ is bounded for all $x \in X_0$, $d \in \partial p_N(x,\overline{w})$, and $N \in \mathbb{N}$. Consider an $\overline{w} \in \overline{\mathbb{B}}_0$ with this boundedness property. Then, there exist an infinite subset $I_0 \subset \mathbb{N}$, an $x^* \in \mathbb{R}^n$ and an $h^* \in \mathbb{R}^n$ such that $x_i \to x^*$, $h_{N_i}(x_i, d_i) \to h^*$, as $i \to \infty$, $i \in I_0$. The continuity of $\tilde{\psi}(\cdot,\cdot)$ and (53) imply that

$$\lim_{i \to \infty, i \in I_0} \left|\tilde{\psi}_{N_i}\big(x_i, x_i + h_{N_i}(x_i, d_i), d_i\big) - \tilde{\psi}(x^*, x^* + h^*)\right| = 0. \tag{59}$$

Hence, it follows Theorem 3.3.2 in [16] that the problems

$$\min_{h \in \mathbb{R}^n} \tilde{\psi}_{N_i}(x_i, x_i + h, d_i) \tag{60}$$

epi-converge to the problem

$$\min_{h \in \mathbb{R}^n} \tilde{\psi}(x^*, x^* + h), \tag{61}$$

as $i \to \infty$, $i \in I_0$. Since $\{h_{N_i}(x_i, d_i)\}_{i=1}^\infty$ is a sequence of global minimizers of (60), it follows from Theorem 3.3.3 in [16] that h^* must be a global minimizer of (61). Since the problem in (61) is strictly convex, it has a unique global minimizer. Hence, $h^* = h(x^*)$, which contradicts (58). This completes the proof. □

References

1. Royset, J.O., Polak, E.: Implementable algorithm for stochastic programs using sample average approximations. J. Optim. Theory Appl. **122**, 157–184 (2004)
2. Ditlevsen, O., Madsen, H.O.: Structural Reliability Methods. Wiley, New York (1996)
3. Liu, P.-L., Kuo, C.-Y.: Safety evaluation of the upper structure of bridge based on concrete nondestructive tests. In: Der Kiureghian, A., Madanat, S., Pestana, J.M. (eds.) Applications of Statistics and Probability in Civil Engineering, pp. 1683–1688. Millpress, Rotterdam (2003)
4. Akgul, F., Frangopol, D.M.: Probabilistic analysis of bridge networks based on system reliability and Monte Carlo simulation. In: Der Kiureghian, A., Madanat, S., Pestana, J.M. (eds.) Applications of Statistics and Probability in Civil Engineering, pp. 1633–1637. Millpress, Rotterdam (2003)
5. Holicky, M., Markova, J.: Reliability analysis of impacts due to road vehicles. In: Der Kiureghian, A., Madanat, S., Pestana, J.M. (eds.) Applications of Statistics and Probability in Civil Engineering, pp. 1645–1650. Millpress, Rotterdam (2003)
6. Ermoliev, Y.: Stochastic quasigradient methods. In: Ermoliev, Y., Wets, R.J.-B. (eds.) Numerical Techniques for Stochastic Optimization. Springer, New York (1988)
7. Benveniste, A., Metivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations. Springer, New York (1990)
8. Marti, K.: Stochastic Optimization Methods. Springer, Berlin (2005)
9. Ruszczynski, A., Shapiro, A.: Stochastic Programming. Elsevier, New York (2003)
10. Shapiro, A., Wardi, Y.: Convergence analysis of stochastic algorithms. Math. Oper. Res. **21**, 615–628 (1996)
11. Sakalauskas, L.L.: Nonlinear stochastic programming by Monte-Carlo estimators. Eur. J. Oper. Res. **137**, 558–573 (2002)
12. Royset, J.O., Der Kiureghian, A., Polak, E.: Optimal design with probabilistic objective and constraints. J. Eng. Mech. **132**, 107–118 (2006)
13. Uryasev, S.: Derivatives of probability functions and some applications. Ann. Oper. Res. **56**, 287–311 (1995)
14. Marti, K.: Differentiation formulas for probability functions: the transformation method. Math. Program. **75**, 201–220 (1996)
15. Tretiakov, G.: Stochastic quasi-gradient algorithms for maximization of the probability function. A new formula for the gradient of the probability function. In: Marti, K. (ed.) Stochastic Optimization Techniques, pp. 117–142. Springer, New York (2002)
16. Polak, E.: Optimization Algorithms and Consistent Approximations. Springer, New York (1997)
17. Deak, I.: Three digit accurate multiple normal probabilities. Numer. Math. **35**, 369–380 (1980)
18. Ditlevsen, O., Oleson, R., Mohr, G.: Solution of a class of load combination problems by directional simulation. Struct. Saf. **4**, 95–109 (1987)
19. Bjerager, P.: Probability integration by directional simulation. J. Eng. Mech. **114**, 1288–1302 (1988)
20. Shapiro, A.: Asymptotic properties of statistical estimators in stochastic programming. Ann. Stat. **17**, 841–858 (1989)
21. Rubinstein, R.Y., Shapiro, A.: Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. Wiley, New York (1993)
22. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, New York (1998)
23. Polak, E., Qi, L., Sun, D.: First-order algorithms for generalized semi-infinite min-max problems. Comput. Optim. Appl. **13**, 137–161 (1999)
24. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, New York (1991)
25. van der Vaart, A., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, New York (1996)
26. Matlab, Version 6.5, Mathworks Inc., Natick, MA (2002)