Probablistic Constrained Optimization: Theory and Applications (S. P. Uryasev, Editor), pp. 91-116

Statistical inference of stochastic optimization problems

Alexander Shapiro (ashapiro@isye.gatech.edu)
Georgia Institute of Technology,
Atlanta, GA 30332-0205, USA

Abstract

We discuss in this paper statistical inference of Monte Carlo simulation based approximations of stochastic optimization problems, where the "true" objective function, and probably some of the constraints, are estimated, typically by averaging a random sample. The classical maximum likelihood estimation can be considered in that framework. Recently statistical analysis of such methods has been motivated by a development of simulation based optimization techniques. We investigate asymptotic properties of the optimal value and an optimal solution of the corresponding Monte Carlo simulation approximations by employing the so-called Delta method, and discuss some examples.

1 Introduction

Consider the optimization problem

$$\operatorname*{Min}_{x \in S} f(x), \tag{1.1}$$

where S is a subset of $\mathbb{R}^m$ and $f : S \to \mathbb{R}$. Suppose that the above optimization problem is approximated by a sequence of problems

$$\operatorname*{Min}_{x \in S} \hat{f}_N(x), \tag{1.2}$$

where $\hat{f}_N(x)$ are random functions converging, as $N \to \infty$, in some probabilistic sense to $f(x)$. We refer to (1.1) and (1.2) as the *true* and *approximating* problems, respectively. Typically the objective function $f(x)$ is given as the expected value function

$$f(x) := \mathbb{E}_P\{g(x,\omega)\} = \int_\Omega g(x,\omega)P(d\omega), \tag{1.3}$$

where $(\Omega, \mathcal{F}, P)$ is a probability space, and the approximating functions $\hat{f}_N(x)$ are constructed by averaging a random sample.

Let v_0, $\hat{v}_N$ and x_0, $\hat{x}_N$ be the optimal values and optimal solutions of the problems (1.1) and (1.2), respectively. In this paper we discuss asymptotic statistical inference of $\hat{v}_N$ and $\hat{x}_N$, as N tends to infinity. We also consider the cases where the feasible set S is subject to perturbations and is given by random constraints. Let us discuss some examples.

Example 1.1 Our first example is motivated by the classical maximum likelihood method of estimation. That is, let $g(y,\theta)$ be a family of probability density functions (pdf), parameterized by the parameter vector $\theta \in \Theta \subset \mathbb{R}^m$, and let $Y_1, ..., Y_N$ be an i.i.d. random sample with a probability distribution P. Define

$$\hat{f}_N(\theta) := -N^{-1}\sum_{i=1}^N \ln g(Y_i,\theta).$$

By the Law of Large Numbers we have that, for any fixed value of θ, $\hat{f}_N(\theta)$ converges to

$$f(\theta) := -\mathbb{E}_P\{\ln g(Y,\theta)\} = -\int \ln g(y,\theta)P(dy),$$

with probability one, as $N \to \infty$, provided of course that the above expectation exists. This leads to the "true" and "approximating" optimization problems of minimizing $f(\theta)$ and $\hat{f}_N(\theta)$, respectively, over the parameter set Θ.

In particular, suppose that the distribution P is given by a pdf $g(y,\theta_0)$, $\theta_0 \in \Theta$, from the above parametric family, i.e., the parametric model is correctly specified. Then θ_0 is an unconstrained minimizer of $f(\theta)$, and hence is an optimal solution of the "true" problem. Indeed, by using concavity of the logarithm function, we obtain

$$f(\theta_0) - f(\theta) = \int \ln\left[\frac{g(y,\theta)}{g(y,\theta_0)}\right] g(y,\theta_0)dy \le \int \left[\frac{g(y,\theta)}{g(y,\theta_0)} - 1\right] g(y,\theta_0)dy = 0.$$

There is a large literature on the maximum likelihood method, and the above derivation of optimality of θ_0 is known of course. We will come back to this example later. Let us note at this point that the corresponding random sample usually represents available data and the associated minimizer $\hat{\theta}_N$ of $\hat{f}_N(\theta)$, over Θ, is viewed as the maximum likelihood estimator of the "true" value θ_0 of the parameter vector. There are also various extensions of the maximum likelihood method, in particular the method of M-estimators introduced by Huber [13, 15].

Somewhat different type of examples is motivated by a Monte Carlo simulation approach to numerical solutions of stochastic programming problems. A goal of a such stochastic programming problem is to solve an optimization problem of the form (1.1) with the objective function $f(x)$ given as the expected value in the form (1.3). The probability distribution P is supposed to be known, although may be not given explicitly. However, the corresponding integral (expected value) cannot be calculated in a closed form and has to be approximated. Monte Carlo simulation techniques provide such an approximation by averaging a generated a random sample with an appropriate probability distribution. Let us discuss the following two examples of stochastic programming with recourse and a $GI/G/1$ queue.

Example 1.2 Consider the optimization problem

$$\operatorname*{Min}_{x \in S} c^T x + \mathbb{E}\{Q(x, h(\omega))\}, \tag{1.4}$$

where $c \in \mathbb{R}^m$ is a given vector, $Q(x, h)$ is the optimal value of the optimization problem

$$\operatorname*{Min}_{y \geq 0} q^T y \text{ subject to } Wy = h - Ax, \tag{1.5}$$

and $h = h(\omega)$ is a random vector with a known probability distribution. (For the sake of simplicity we assume that only vector h is random while other parameters in the linear programming problem (1.5) are deterministic.) This is the so-called two-stage stochastic programming problem with recourse, which originated in works of Beale [2] and Dantzig [8]. If the random vector h has a discrete distribution, then the expected value function $\mathbb{E}\{Q(x, h)\}$ is given in a form of summation and problem (1.4) can be written as a large linear programming problem. Over the years this approach was developed and various techniques were suggested in order to make it numerically efficient. The interested reader is referred to recent books by Kall and Wallace [16] and Birge and Louveaux [4], and references therein, for an extensive discussion of these methods.

However, the number of realizations of h (the number of discretization points in case the distribution of h is continuous) typically grows exponentially with the dimensionality of h. Consequently, this number can quickly become so large that even modern computers cannot cope with the required calculations. Monte Carlo simulation techniques suggest an approach to deal with this problem. That is, a random sample $h_1, ..., h_N$ of N independent realizations of the random vector h are generated, and the expected value function $\mathbb{E}\{Q(x, h)\}$ is estimated by the average function $\hat{Q}_N(x) := N^{-1} \sum_{i=1}^{N} Q(x, h_i)$. Consequently the "true" problem (1.4) is approximated by the problem

$$\operatorname*{Min}_{x \in S} c^T x + \hat{Q}_N(x). \tag{1.6}$$

By calculating an optimal solution $\hat{x}_N$ of the above approximating problem, one obtains an estimator of an optimal solution of the true problem.

By the Law of Large Numbers we have that the average function $\hat{Q}_N(x)$ converges, pointwise, to $\mathbb{E}\{Q(x, h)\}$ with probability one, as $N \to \infty$. The function $Q(\cdot, h)$, and hence the function $\hat{Q}_N(\cdot)$, are piecewise linear and convex. The function $Q(\cdot, h)$ is not given explicitly and in itself is an output of an optimization procedure. Nevertheless, its value and a corresponding subgradient can be calculated, at any given point x, by solving the linear program (1.5). This allows to apply, reasonably efficient, deterministic algorithms in order to solve the approximating problem (1.6). For a discussion of such algorithms and a numerical experience in solving two-stage stochastic programming problems by such methods we refer to Shapiro and Homem-de-Mello [33].

Let us make the following observations. The above example is different from the maximum likelihood example in several respects. In the above example the corresponding random sample is generated in the computer and can be controlled to some extend. The only limitation on the number N of generated points is the computational time and computer's memory capacity. It is also possible to implement various variance reduction techniques which in some cases considerably enhance the numerical performance of the algorithm. Usually the feasible set S is defined by constraints. In this respect inequality type constraints appear naturally in optimization problems. In the maximum likelihood example the optimal solution of the "true" problem is actually an unconstrained minimizer of the objective function. There is no reason for such behavior of an optimal solution of the optimization problem (1.4). As we shall see later this introduces an additional term in the asymptotic expansion of $\hat{x}_N$, associated with a curvature of the set S. Let us finally note that the average function $\hat{Q}_N(x)$ is not everywhere differentiable. If the distribution of h is discrete, this is carried over to the expected value function. On the other hand, if the distribution of h is continuous, then the expected value function is smooth (differentiable). This makes the asymptotics of $\hat{x}_N$ quite different in cases of discrete and continuous distributions of h. We shall discuss that later.

Example 1.3 As our last example we consider a $GI/G/1$ queue whose service times depend on a parameter vector x. Let Y_i be the time between arrivals of the $(i-1)^{th}$ and i^{th} customers, and for a given value of x, let $Z_i(x)$ be the service time of the i^{th} customer, $i = 1, 2, ...$ Let $G_i(x)$ denote the i^{th} sojourn time, i.e., the total time spent by the i^{th} customer in the queue. It is assumed that the interarrival and service times are random i.i.d., that the first customer arrives at an empty queue and that for every $x \in S$ the queue is regenerative with the expected number of customers served in one busy period (regenerative cycle) being finite. A recursive relation between the sojourn times is given by Lindley equation

$$G_i(x) = Z_i(x) + [G_{i-1}(x) - Y_i]_+. \tag{1.7}$$

Under standard regularity conditions (e.g., [36]), the long-run average functions

$$\hat{f}_N(x) := N^{-1} \sum_{i=1}^{N} G_i(x)$$

converge pointwise, with probability one, to the expected value (mean) steady state sojourn time $f(x)$. Consider the optimization problem

$$\operatorname*{Min}_{x \in S} f(x) + \psi(x), \tag{1.8}$$

where $\psi(x)$ is a (deterministic) cost function. The above "true" problem can be approximated by generating the i.i.d. sequences of the interarrival and service times and then calculating the sojourn times, by using Lindley equation (1.7), and replacing $f(x)$ with its average estimate $\hat{f}_N(x)$. Let us observe that the sojourn times, used in the averaging procedure, are not independent. The approximating functions $\hat{f}_N(x)$ are piecewise smooth. It is possible to extend the above example to more complex queueing systems. It is somewhat surprising that there are examples of simple queues with deterministic service times, depending on a parameter x belonging to an interval of the real line, such that the corresponding expected value steady state sojourn time is not differentiable at a dense set of points on that interval (Shapiro and Wardi [32]).

2 The Delta method

In order to investigate asymptotic properties of the estimators $\hat{v}_N$ and $\hat{x}_N$ it will be convenient to use the Delta method, which we discuss in this section. Let Y_N be a sequence of random vectors, converging in probability to a vector μ. Suppose that there exists a sequence τ_N of positive numbers, tending to infinity, such that $\tau_N(Y_N - \mu)$ converges in distribution to a random vector Y, denoted $\tau_N(Y_N - \mu) \Rightarrow Y$. Let $G(y)$ be a vector valued function, differentiable at μ. That is

$$G(y) - G(\mu) = M(y - \mu) + r(y), \tag{2.1}$$

where $M = \nabla G(\mu)$ is the Jacobian matrix (of first order partial derivatives) of G at μ, and the remainder $r(y)$ is of order $o(\|y - \mu\|)$, i.e., $r(y)/\|y - \mu\| \to 0$ as $y \to \mu$. It follows from (2.1) that

$$\tau_N[G(Y_N) - G(\mu)] = M[\tau_N(Y_N - \mu)] + \tau_N r(Y_N). \tag{2.2}$$

Since $\tau_N(Y_N - \mu)$ converges in distribution, it is bounded in probability, and hence $\|Y_N - \mu\|$ is of stochastic order $O_p(\tau_N^{-1})$. It follows that

$$r(Y_N) = o(\|Y_N - \mu\|) = o_p(\tau_N^{-1}),$$

and hence $\tau_N r(Y_N)$ converges in probability to zero. Consequently we obtain by (2.2) that

$$\tau_N[G(Y_N) - G(\mu)] \Rightarrow MY. \tag{2.3}$$

This formula is routinely employed in multivariate analysis and is known as the (finite dimensional) Delta Theorem (e.g., [24]).

We need to extend this method in several directions. The random functions $\hat{f}_N$ can be viewed as random elements in an appropriate functional space and the corresponding estimators $\hat{v}_N$ and $\hat{x}_N$ as functions of these random elements. This motivates us to extend formula (2.3) to a Banach space setting. Let B_1 and B_2 be two Banach (i.e., linear normed, complete) spaces, and $G : B_1 \to B_2$ be a mapping. Suppose that G is directionally differentiable at a considered point $\mu \in B_1$, i.e., the limit

$$G'_\mu(d) := \lim_{t\downarrow 0} \frac{G(\mu + td) - G(\mu)}{t} \tag{2.4}$$

exists for all $d \in B_1$. If, in addition, the directional derivative $G'_\mu : B_1 \to B_2$ is linear and continuous, then it is said that G is Gâteaux differentiable at μ. Note that, in any case, the directional derivative $G'_\mu(\cdot)$ is positively homogeneous, that is $G'_\mu(\alpha d) = \alpha G'_\mu(d)$ for any $\alpha \geq 0$ and $d \in B_1$.

It follows from (2.4) that

$$G(\mu + d) - G(\mu) = G'_\mu(d) + r(d)$$

with the remainder $r(d)$ being "small" along any fixed direction d, i.e., $r(td)/t \to 0$ as $t \downarrow 0$. This property is not sufficient, however, to neglect the remainder term in the corresponding asymptotic expansion and we need a stronger notion of directional differentiability. It is said that G is directionally differentiable at μ in the sense of Hadamard if the directional derivative $G'_\mu(d)$ exists for all $d \in B_1$ and, moreover,

$$G'_\mu(d) = \lim_{\substack{t\downarrow 0 \\ d'\to d}} \frac{G(\mu + td') - G(\mu)}{t}. \tag{2.5}$$

It is possible to show that if G is Hadamard directionally differentiable at μ, then the directional derivative $G'_\mu(\cdot)$ is continuous, although possibly is not linear. For a discussion of various concepts of directional differentiability see, e.g., [29].

Now let B_1 and B_2 be equipped with their Borel σ-algebras $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively. (Recall that Borel σ-algebra of a normed space is the σ-algebra generated by the family of its open sets.) An $\mathcal{F}$-measurable mapping from a probability space $(\Omega, \mathcal{F}, P)$ into B_1 is called a random element of B_1. Consider a sequence X_N of random elements of B_1. It is said that X_N converges in distribution (weakly) to a random element Y of B_1, and denoted $X_N \Rightarrow Y$, if the expected values $\mathbb{E}\{f(X_N)\}$ converge to $\mathbb{E}\{f(Y)\}$, as $N \to \infty$, for any bounded and continuous function $f : B_1 \to \mathbb{R}$ (see, e.g., Billingsley [3] for a discussion of weak convergence). Let us formulate now the first version of the Delta Theorem.

Theorem 2.1 *Let B_1 and B_2 be Banach spaces, equipped with their Borel σ-algebras, Y_N be a sequence of random elements of B_1, $G : B_1 \to B_2$ be a mapping, and τ_N be a sequence of positive numbers tending to infinity as $N \to \infty$. Suppose that the space B_1 is separable, the mapping G is Hadamard directionally differentiable at a point $\mu \in B_1$, and the sequence $X_N := \tau_N(Y_N - \mu)$ converges in distribution to a random element Y of B_1. Then*

$$\tau_N[G(Y_N) - G(\mu)] \Rightarrow G'_\mu(Y), \tag{2.6}$$

and

$$\tau_N[G(Y_N) - G(\mu)] = G'_\mu(X_N) + o_p(1). \tag{2.7}$$

Note that, because of the Hadamard directional differentiability of G, the mapping $G'_\mu : B_1 \to B_2$ is continuous, and hence is measurable with respect to the Borel σ-algebras of B_1 and B_2. The above infinite dimensional version of the Delta Theorem appeared in works of Gill [11], Grübel [12] and King [17, 18]. It can be proved easily by using the following Skorohod-Dudley almost sure representation theorem, e.g., [23, p.71]).

Representation Theorem. Suppose that a sequence of random elements X_N, of a separable Banach space B, converges in distribution to a random element Y. Then there exists a sequence X'_N, Y', defined on a single probability space, such that $X'_N \overset{\mathcal{D}}{=} X_N$, for all N, $Y' \overset{\mathcal{D}}{=} Y$ and $X'_N \to Y'$ with probability one.

Here $Y' \overset{\mathcal{D}}{=} Y$ means that the probability measures induced by Y' and Y coincide. We give now a proof of theorem 2.1 for the sake of completeness.

Proof of theorem 2.1. Consider the sequence $X_N := \tau_N(Y_N - \mu)$ of random elements of B_1. By the Representation Theorem, there exists a sequence X'_N, Y', defined on a single probability space, such that $X'_N \overset{\mathcal{D}}{=} X_N$, $Y' \overset{\mathcal{D}}{=} Y$ and $X'_N \to Y'$ w.p.1. Consequently for $Y'_N := \mu + \tau_N^{-1} X'_N$, we have $Y'_N \overset{\mathcal{D}}{=} Y_N$. It follows then from Hadamard directional differentiability of G that

$$\tau_N[G(Y'_N) - G(\mu)] \to G'_\mu(Y') \quad \text{w.p.1.} \tag{2.8}$$

Since convergence with probability one implies convergence in distribution and the terms in (2.8) have the same distributions as the corresponding terms in (2.6), the asymptotic result (2.6) follows.

Now since $G'_\mu(\cdot)$ is continuous and $X'_N \to Y'$ w.p.1, we have that

$$G'_\mu(X'_N) \to G'_\mu(Y') \quad \text{w.p.1.} \tag{2.9}$$

Together with (2.8) this implies that the difference between $G'_\mu(X'_N)$ and the left hand side of (2.8) tends w.p.1, and hence in probability, to zero. We obtain that

$$\tau_N[G(Y'_N) - G(\mu)] = G'_\mu[\tau_N(Y'_N - \mu)] + o_p(1),$$

which implies (2.7). ∎

Let us now formulate the second version of the Delta Theorem where the mapping G is restricted to a subset K of the space B_1. We say that G is Hadamard directionally differentiable at a point μ *tangentially* to the set K if for any sequence d_N of the form $d_N := (y_N - \mu)/t_N$, where $y_N \in K$ and $t_N \downarrow 0$, and such that $d_N \to d$, the following limit exists

$$G'_\mu(d) = \lim_{N\to\infty} \frac{G(\mu + t_N d_N) - G(\mu)}{t_N}. \tag{2.10}$$

Equivalently the above condition (2.10) can be written in the form

$$G'_\mu(d) = \lim_{\substack{t\downarrow 0 \\ d' \to_K d}} \frac{G(\mu + td') - G(\mu)}{t}, \tag{2.11}$$

where the notation $d' \to_K d$ means that $d' \to d$ and $\mu + td' \in K$.

Since $y_N \in K$, and hence $\mu + t_N d_N \in K$, the mapping G needs only to be defined on the set K. Recall that the contingent (Bouligand) cone to K at μ, denoted $T_K(\mu)$, is formed by vectors $d \in B$ such that there exist sequences $d_N \to d$ and $t_N \downarrow 0$ such that $\mu + t_N d_N \in K$. Note that $T_K(\mu)$ is nonempty only if μ belongs to the topological closure of K. By the above definitions we have that $G'_\mu(\cdot)$ is defined on the set $T_K(\mu)$. The following "tangential" version of the Delta Theorem can be easily proved in a way similar to the proof of theorem 2.1 (Shapiro [30]).

Theorem 2.2 *Let B_1 and B_2 be Banach spaces, K be a subset of B_1, $G : K \to B_2$ be a mapping, and Y_N be a sequence of random elements of B_1. Suppose that:* (i) *the space B_1 is separable,* (ii) *the mapping G is Hadamard directionally differentiable at a point μ tangentially to the set K,* (iii) *for some sequence τ_N of positive numbers tending to infinity, the sequence $X_N := \tau_N(Y_N - \mu)$ converges in distribution to a random element Y,* (iv) *$Y_N \in K$, with probability one, for all N large enough. Then*

$$\tau_N[G(Y_N) - G(\mu)] \Rightarrow G'_\mu(Y). \tag{2.12}$$

Moreover, if the set K is convex, then equation (2.7) *holds.*

Note that it follows from the assumptions (iii) and (iv) that the distribution of Y is concentrated on the contingent cone $T_K(\mu)$, and hence the distribution of $G'_\mu(Y)$ is well defined.

Our third variant of the Delta Theorem deals with a second order expansion of the mapping G. That is, suppose that G is directionally differentiable at μ and define

$$G''_\mu(d) = \lim_{\substack{t\downarrow 0 \\ d' \to d}} \frac{G(\mu + td') - G(\mu) - tG'_\mu(d')}{\frac{1}{2}t^2}. \tag{2.13}$$

If the mapping G is twice continuously differentiable, then this second order directional derivative $G''_\mu(d)$ coincides with the second order term in the Taylor expansion of

$G(\mu + d)$. The above definition of $G''_\mu(d)$ makes sense for directionally differentiable mappings. However, in interesting applications, where it is possible to calculate $G''_\mu(d)$, the mapping G is actually (Gâteaux) differentiable. We say that G is second order Hadamard directionally differentiable at μ if the second order directional derivative $G''_\mu(d)$, defined in (2.13), exists for all $d \in B_1$. We say that G is second order Hadamard directionally differentiable at μ tangentially to a set $K \subset B_1$ if for all $d \in T_K(\mu)$ the limit

$$G''_\mu(d) = \lim_{\substack{t \downarrow 0 \\ d' \to_K d}} \frac{G(\mu + td') - G(\mu) - tG'_\mu(d')}{\frac{1}{2}t^2} \tag{2.14}$$

exists.

Note that if G is first and second order Hadamard directionally differentiable at μ tangentially to K, then $G'_\mu(\cdot)$ and $G''_\mu(\cdot)$ are continuous on $T_K(\mu)$, and that $G''_\mu(\alpha d) = \alpha^2 G''_\mu(d)$ for any $\alpha \geq 0$ and $d \in T_K(\mu)$.

Theorem 2.3 *Let B_1 and B_2 be Banach spaces, K be a convex subset of B_1, Y_N be a sequence of random elements of B_1, $G : K \to B_2$ be a mapping, and τ_N be a sequence of positive numbers tending to infinity as $N \to \infty$. Suppose that:* (i) *the space B_1 is separable,* (ii) *G is first and second order Hadamard directionally differentiable at μ tangentially to the set K,* (iii) *the sequence $X_N := \tau_N(Y_N - \mu)$ converges in distribution to a random element Y of B_1,* (iv) *$Y_N \in K$ w.p.1 for N large enough. Then*

$$\tau_N^2 \left[G(Y_N) - G(\mu) - G'_\mu(Y_N - \mu) \right] \Rightarrow \tfrac{1}{2} G''_\mu(Y), \tag{2.15}$$

and

$$G(Y_N) = G(\mu) + G'_\mu(Y_N - \mu) + \tfrac{1}{2} G''_\mu(Y_N - \mu) + o_p(\tau_N^{-2}). \tag{2.16}$$

Proof. let X'_N, Y' and Y'_N be elements as in the proof of theorem 2.1. Recall that their existence is guaranteed by the Representation Theorem. Then by the definition of G''_μ we have

$$\tau_N^2 \left[G(Y'_N) - G(\mu) - \tau_N^{-1} G'_\mu(X'_N) \right] \to \tfrac{1}{2} G''_\mu(Y') \quad \text{w.p.1.}$$

Note that $G'_\mu(\cdot)$ is defined on $T_K(\mu)$ and, since K is convex, $X'_N = \tau_N(Y'_N - \mu) \in T_K(\mu)$. Therefore the expression in the left hand side of the above limit is well defined. Since convergence w.p.1 implies convergence in distribution, formula (2.15) follows. Since $G''_\mu(\cdot)$ is continuous on $T_K(\mu)$, and, by convexity of K, $Y'_N - \mu \in T_K(\mu)$ w.p.1, we have that $\tau_N^2 G''_\mu(Y'_N - \mu) \to G''_\mu(Y')$ w.p.1. Since convergence w.p.1 implies convergence in probability, formula (2.16) then follows. ∎

3 First order asymptotics of the optimal value

In this section we discuss asymptotics of the optimal value $\hat{v}_N$, of the approximating problem, based on first order expansions of the optimal value function. We assume that the feasible set S, of the true and approximating problems, is a *compact* subset of $\mathbb{R}^m$. In many interesting applications such assumption cannot be guaranteed and in fact S can be unbounded. Nevertheless, it can be often showed that an optimal solution of the approximating problem stays with probability one in a bounded subset of $\mathbb{R}^m$, and hence we can restrict the optimization procedure to a compact subset of $\mathbb{R}^m$.

Let us consider the Banach space $C(S)$ of continuous functions $y : S \to \mathbb{R}$ equipped with the sup-norm $\|y\| := \sup_{x \in S} |y(x)|$. We assume that the objective function $f(x)$, of the true problem (1.1), is continuous, and hence $f \in C(S)$, and that the approximating functions $\hat{f}_N$ are *random elements* of $C(S)$. Define the optimal value function $\vartheta : C(S) \to \mathbb{R}$ as $\vartheta(y) := \inf_{x \in S} y(x)$. We have then that $v_0 = \vartheta(f)$ and $\hat{v}_N = \vartheta(\hat{f}_N)$.

It is not difficult to see that the optimal value function ϑ is concave and Lipschitz continuous modulus one, i.e., $|\vartheta(y_1) - \vartheta(y_2)| \leq \|y_1 - y_2\|$ for any $y_1, y_2 \in C(S)$. Moreover, it is possible to show (e.g., [30]) that ϑ is Hadamard directionally differentiable at any point $\mu \in C(S)$ and for any $\delta \in C(S)$,

$$\vartheta'_\mu(\delta) = \inf_{x \in S^*(\mu)} \delta(x), \tag{3.1}$$

where $S^*(\mu) := \arg\min_{x \in S} \mu(x)$. Note that the set $S^*(\mu)$ is nonempty since $\mu(x)$ is continuous and S is compact. Together with theorem 2.1 this leads to the following asymptotic result (Shapiro [30]).

Theorem 3.1 *Suppose that, for a sequence τ_N of positive numbers converging to infinity, the sequence $\tau_N(\hat{f}_N - f)$, of random elements of $C(S)$, converges in distribution to a random element Y of $C(S)$. Then*

$$\tau_N(\hat{v}_N - v_0) \Rightarrow \inf_{x \in S^*(f)} Y(x), \tag{3.2}$$

where $S^*(f)$ is the set of optimal solutions of the true problem (1.1). *In particular, if the true problem has unique optimal solution x_0, then*

$$\tau_N(\hat{v}_N - v_0) \Rightarrow Y(x_0). \tag{3.3}$$

Let us specify the above, somewhat abstract, asymptotic result to the case where f is the expected value function, defined in (1.3), and the approximating functions $\hat{f}_N$ are constructed by averaging a random sample. That is,

$$\hat{f}_N(x) := N^{-1} \sum_{i=1}^{N} g(x, \omega_i), \tag{3.4}$$

where $\omega_1, ..., \omega_N$ is an i.i.d. random sample, in $(\Omega, \mathcal{F})$, with the probability distribution P. Let us make the following assumptions.

(A1) For every $x \in S$, the function $g(x, \cdot)$ is $\mathcal{F}$-measurable.

(A2) For some point $x \in S$, the expectation $\mathbb{E}_P\{g(x,\omega)^2\}$ is finite.

(A3) There exists an $\mathcal{F}$-measurable function $\kappa : \Omega \to \mathbb{R}$ such that $\mathbb{E}_P\{\kappa(\omega)^2\}$ is finite and

$$|g(x_1, \omega) - g(x_2, \omega)| \leq \kappa(\omega)\|x_1 - x_2\| \tag{3.5}$$

for all $x_1, x_2 \in S$ and P-almost all $\omega \in \Omega$.

The above assumptions (A1) - (A3) are sufficient for the Central Limit Theorem to hold in $C(S)$. That is, the sequence $N^{1/2}(\hat{f}_N - f)$, of random elements of $C(S)$, converges in distribution to a random element Y (see Araujo and Giné [1], for details). Note that for any fixed point $x_0 \in S$, $Y(x_0)$ is a real valued random variable having normal distribution with zero mean and variance $\sigma^2(x_0)$ equal to the variance of $g(x_0, \omega)$, i.e.,

$$\sigma^2(x_0) = \mathbb{E}_P\{g(x_0, \omega)^2\} - f(x_0)^2. \tag{3.6}$$

We obtain the following results [30].

Theorem 3.2 *Suppose that f and $\hat{f}_N$ are given in the form* (1.3) *and* (3.4)*, respectively, with ω_i being an i.i.d. random sample, that the above assumptions* (A1) - (A3) *hold, and that the true problem* (1.1) *has unique optimal solution x_0. Then it follows that $N^{1/2}(\hat{v}_N - v_0)$ converges in distribution to normal $N(0, \sigma^2)$, with variance $\sigma^2 = \sigma^2(x_0)$ given in* (3.6)*, and*

$$\hat{v}_N = \hat{f}_N(x_0) + o_p(N^{-1/2}). \tag{3.7}$$

Formula (3.7) shows that, under the assumptions of the above theorem, the optimal value $\hat{v}_N$ of the approximating problem (1.2) is equivalent, up to order $o_p(N^{-1/2})$, to the value of the problem with the same objective function $\hat{f}_N$ and the feasible set S reduced to the single point x_0. This indicates that the above (first order) asymptotics do not depend on the local structure of the set S near the point x_0. Note that since $\hat{f}_N(x_0)$ is an unbiased estimator of $v_0 = f(x_0)$ and that $\hat{v}_N \leq \hat{f}_N(x_0)$, the estimator $\hat{v}_N$ of v_0 typically has a negative bias. We will derive later an approximation of the asymptotic bias of $\hat{v}_N$, of order $O(N^{-1})$, by using a second order expansion of the optimal value function.

Consider the framework of the maximum likelihood example 1.1. Let Θ_0 and Θ_1 be subsets of $\mathbb{R}^m$ and suppose that we wish to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$. Let

$$\ell_N := 2\left[\inf_{\theta \in \Theta_0} \sum_{i=1}^{N} \ln g(Y_i, \theta) - \inf_{\theta \in \Theta_1} \sum_{i=1}^{N} \ln g(Y_i, \theta)\right] \tag{3.8}$$

be the corresponding log-likelihood ratio test statistic. Suppose that $f(\theta) := -\mathbb{E}_P\{\ln g(Y,\theta)\}$ has unique minimizers θ_0 and θ_1 over the sets Θ_0 and Θ_1, respectively. Recall that if the distribution P, of the random sample, is given by a pdf $g(x,\theta_0)$, then θ_0 is an unconstrained minimizer of $f(\theta)$. Moreover, if the parameter vector θ is identified at θ_0, then θ_0 is such unique minimizer. We have by (3.7) that

$$N^{-1/2}\ell_N = 2N^{-1/2}\sum_{i=1}^{N}\left[\ln g(Y_i,\theta_0) - \ln g(Y_i,\theta_1)\right] + o_p(1), \tag{3.9}$$

provided that the corresponding regularity assumptions (A1) - (A3) hold. It follows that
$N^{-1/2}(\ell_N - \ell_0)$ converges in distribution to normal $N(0,\sigma^2)$, where ℓ_0 and σ^2 are the mean and the variance, respectively, of the random variable $Z := 2\ln[g(Y,\theta_0)/g(Y,\theta_1)]$.

Note that if $\theta_0 = \theta_1$, then this variable Z degenerates into $Z \equiv 0$. Therefore in cases where vectors θ_0 and θ_1 are close to each other (and usually these are the cases we are interested in), the above normal approximation of the distribution of ℓ_N is not accurate. In fact it is possible to obtain a much better approximation of the distribution of ℓ_N by using a second order expansion of the optimal value function. However, in stochastic programming applications the asymptotic result (3.7) is very useful due to its simplicity and generality. The asymptotic variance $\sigma^2(x)$ can be consistently estimated at each iteration point $x = x^\nu$ of a simulation based optimization algorithm. This allows to incorporate t-test type procedures into such algorithms and to construct confidence intervals for the true optimal value v_0 (see [33]).

Let us consider now a situation where the feasible set is defined by constraints which are not given explicitly and should be estimated. That is,

$$S := \{x \in Q : h_i(x) \le 0,\ i = 1,...,k\}, \tag{3.10}$$

where Q is a closed subset of $\mathbb{R}^m$ and the constraint functions h_i are given as expected values, $h_i(x) := \mathbb{E}\{g_i(x,\omega)\}$, $i = 1,...,k$. Suppose that each constraint function $h_i(x)$ is real valued (i.e., the corresponding expectation exists), and that $h_i(x)$ can be estimated, say be a sample average, function $\hat{h}_{iN}(x)$. Then the true problem (1.1) can be approximated by the problem

$$\operatorname*{Min}_{x\in S_N} \hat{f}_N(x), \tag{3.11}$$

where

$$S_N := \left\{x \in Q : \hat{h}_{iN}(x) \le 0,\ i = 1,...,k\right\}. \tag{3.12}$$

It is possible to show that, under mild regularity conditions, the optimal value $\hat{v}_N$ and an optimal solution $\hat{x}_N$ of the above approximating problem (3.12) are consistent estimators of their "true" counterparts. Let us mention recent work of Dupačová and

Wets [10], King and Wets [19] and Robinson [25], where this consistency problem is studied from the point of view of epi-convergence analysis.

Recall that the true problem (1.1) is said to be *convex* if the set Q is convex and the objective function f and the constraint functions h_i, $i = 1, ..., k$, are convex. The Lagrangian function, associated with problem (1.1), is

$$L(x, \lambda) := f(x) + \sum_{i=1}^{k} \lambda_i h_i(x). \tag{3.13}$$

Suppose that the true problem (1.1) is convex and that the Slater condition holds, i.e., there exists a point $\bar{x} \in Q$ such that $h_i(\bar{x}) < 0$, $i = 1, ..., k$. Then with every optimal solution x_0 of (1.1) is associated a nonempty and bounded set $\Lambda(x_0)$ of Lagrange multipliers vectors $\lambda = (\lambda_1, ..., \lambda_k)$ satisfying the optimality conditions:

$$x_0 \in \arg\min_{x \in Q} L(x, \lambda), \quad \lambda_i \geq 0 \text{ and } \lambda_i h_i(x_0) = 0, \; i = 1, ..., k. \tag{3.14}$$

The set $\Lambda(x_0)$ coincides with the set of optimal solutions of the dual of (1.1) problem, and therefore is the same for any optimal solution of (1.1) (see Rockafellar [26]).

Let the set Q be a compact convex subset of $\mathbb{R}^m$ and consider the Banach space $B := C(Q) \times ... \times C(Q)$, given by the Cartesian product of $k + 1$ replications of the space $C(Q)$. Note that real valued convex functions are continuous and hence $(f, h_1, ..., h_k) \in B$. Denote by K the subset of B formed by $\xi = (\xi_0, ..., \xi_k) \in B$ such that each function $\xi_i(\cdot)$, $i = 0, ..., k$, is convex on Q. Since problem (1.1) is convex, we have that $(f, h_1, ..., h_k) \in K$. Note that the set K is closed and convex in B. Define the optimal value function

$$\vartheta(\xi) := \inf\{\xi_0(x) : x \in Q, \; \xi_i(x) \leq 0, \; i = 1, ..., k\}. \tag{3.15}$$

Clearly, for $\mu := (f, h_1, ..., h_k)$ and $Y_N := (\hat{f}_N, \hat{h}_{1N}, ..., \hat{h}_{kN})$, we have that $\vartheta(\mu) = v_0$ and $\vartheta(Y_N) = \hat{v}_N$.

It is possible to show that the optimal value function $\vartheta(\cdot)$ is Hadamard directionally differentiable at the point $\mu := (f, h_1, ..., h_k)$, tangentially to the set K, provided the Slater condition is satisfied, which together with the Delta Theorem 2.2 imply the following results (Shapiro [30]).

Theorem 3.3 *Suppose that the true problem is convex and that the Slater condition, for the true problem, is satisfied. Then the optimal value function ϑ is Hadamard directionally differentiable at the point $\mu := (f, h_1, ..., h_k)$ tangentially to the set K and for any $\delta \in T_K(\mu)$,*

$$\vartheta'_\mu(\delta) = \inf_{x \in S^*(\mu)} \sup_{\lambda \in \Lambda(\mu)} \left[\delta_0(x) + \sum_{i=1}^{k} \lambda_i \delta_i(x) \right], \tag{3.16}$$

where $S^*(\mu)$ and $\Lambda(\mu)$ are the sets of optimal solutions and Lagrange multipliers, respectively, of the true problem. If, moreover, $Y_N := (\hat{f}_N, \hat{h}_{1N}, ..., \hat{h}_{kN})$ are random *elements of the Banach space* B *such that with probability one* $Y_N \in K$*, i.e., the approximating problem* (3.11) *is convex, and* $N^{1/2}(Y_N - \mu)$ *converges in distribution to a random element* $Y = (Y_0, ..., Y_k)$ *of* B*, then*

$$N^{1/2}(\hat{v}_N - v_0) \Rightarrow \inf_{x \in S^*(\mu)} \sup_{\lambda \in \Lambda(\mu)} \left[Y_0(x) + \sum_{i=1}^{k} \lambda_i Y_i(x) \right]. \tag{3.17}$$

The above formula (3.17) indicates that in order to ensure asymptotic normality of $\hat{v}_N$, one needs to assume that the true problem has unique optimal solution x_0 to which corresponds unique Lagrange multipliers vector $\bar{\lambda} = (\bar{\lambda}_1, ..., \bar{\lambda}_k)$. In that case we obtain, assuming that conditions (A1) - (A3) hold for every function $g_i(x, \omega)$, that

$$N^{1/2}(\hat{v}_N - v_0) \Rightarrow N(0, \sigma^2), \tag{3.18}$$

with $\sigma^2 = var\left\{ g(x_0, \omega) + \sum_{i=1}^{k} \bar{\lambda}_i g_i(x_0, \omega) \right\}$.

Without the convexity assumption an asymptotic analysis of stochastic problems like (3.11) is more involved. It is still possible to derive asymptotic normality of the optimal value $\hat{v}_N$, as in (3.18), but under stronger regularity conditions. In particular, one needs to assume Lipschitz continuity of the involved functions and that assumptions like (A1) -(A3) hold for the corresponding Lipschitz constants as well (Shapiro [30]).

4 Second order expansions of the optimal value and asymptotics of optimal solutions

In this section we discuss second order expansions of the optimal value function, which (as we shall see) are closely related to asymptotics of optimal solutions of the approximating problems. We consider the case where the feasible set S is closed (not necessarily convex) and fixed (deterministic) and only the objective function f is subject to perturbations. Unless stated otherwise, we assume throughout this section that the function f is twice continuously differentiable and that the true problem (1.1) has unique optimal solution x_0. By $\nabla f(x)$ and $\nabla^2 f(x)$ we denote the gradient and the Hessian matrix (of second order partial derivatives), respectively, of f at x.

The following first order necessary conditions hold at the point x_0:

$$w^T \nabla f(x_0) \geq 0, \text{ for all } w \in T_S(x_0). \tag{4.1}$$

We say that the second order growth condition holds at x_0 if there exist a constant $c > 0$ and a neighborhood $U \subset I\!R^m$ of x_0 such that

$$f(x) \geq f(x_0) + c\|x - x_0\|^2, \text{ for all } x \in S \cap U. \tag{4.2}$$

This condition is closely related to second order optimality conditions. The set

$$C(x_0) := \Big\{ w \in T_S(x_0) : w^T \nabla f(x_0) = 0 \Big\} \tag{4.3}$$

is called the *critical cone* of the problem (1.1). It represents those directions for which first order conditions (4.1) do not provide information about optimality of x_0. Note that if $\nabla f(x_0) = 0$, then $C(x_0) = T_S(x_0)$. If the distribution P, in the maximum likelihood example 1.1, is given by a pdf $g(y, \theta_0)$, $\theta_0 \in \Theta$, then θ_0 is an unconstrained minimizer of $f(\theta)$ and hence $\nabla f(\theta_0) = 0$. Therefore in that case the critical and tangent cones to the parameter set Θ coincide at the point θ_0.

It turns out that second order optimality conditions, as well as second order expansions of the optimal value function, involve a term related to the curvature of the set S. There are several ways how the curvature of S can be measured. We approach that problem from the following point of view. The set

$$T_S^2(x, d) := \Big\{ w \in I\!\!R^m : \operatorname{dist}\Big(x + td + \tfrac{1}{2}t^2 w, S\Big) = o(t^2) \Big\} \tag{4.4}$$

is called the second order tangent set, to the set S at the point x in the direction d. Here $\operatorname{dist}(x, S) := \inf_{z \in S} \|x - z\|$ denotes the distance from a point x to the set S. Note that $T_S^2(x, d)$ can be nonempty only if $x \in S$ and $d \in T_S(x)$. Yet even if S is convex and $x \in S$ and $d \in T_S(x)$, it can happen that the corresponding second order tangent set is empty.

We also will need the following technical condition. We say that the set S is *second order regular* at the point x_0 if for any vector $d \in T_S(x_0)$ and any sequence $x_N \in S$ of the form $x_N := x_0 + t_N d + \frac{1}{2}t_N^2 w_N$, where $t_N \downarrow 0$ and $t_N w_N \to 0$, the following condition holds

$$\lim_{N \to \infty} \operatorname{dist}\Big(w_N, T_S^2(x_0, d)\Big) = 0. \tag{4.5}$$

If $w_N \to w$, then $w \in T_S^2(x_0, d)$ by the definition of second order tangent sets, and hence (4.5) holds. The sequence w_N, however, can be unbounded and it is only required that the term $t_N^2 w_N$, in the expansion of x_N, is of order $o(t_N)$. The above second order regularity condition ensures that $T_S^2(x_0, d)$ provides a "sufficiently tight" second order approximation of the set S in the direction d. This condition and a related second order analysis of optimization problems is extensively discussed in the forthcoming book by Bonnans and Shapiro [5]. Note that the second order regularity condition implies that the set $T_S^2(x_0, d)$ is nonempty, and that $\operatorname{dist}(x_0 + td, S) = o(t)$, $t > 0$, for any $d \in T_S(x_0)$.

Under the second order regularity condition, the following second order optimality conditions are *necessary* and *sufficient* for the second order growth condition (4.2) to hold at the point x_0 ([5]):

$$d^T \nabla^2 f(x_0) d + \inf_{w \in T_S^2(x_0, d)} w^T \nabla f(x_0) > 0, \quad \text{for all } d \in C(x_0) \setminus \{0\}. \tag{4.6}$$

Apart from the quadratic term, corresponding to the second order Taylor expansion of the function f, an additional term, associated with the second order tangent set

$T^2_S(x_0,d)$, appears in the left hand side of (4.6). This terms vanishes if $\nabla f(x_0) = 0$. That is what happens in the maximum likelihood example 1.1.

Example 4.1 Suppose that the set S is defined by equality and inequality constraints

$$S := \{x : h_i(x) = 0,\ i = 1, ..., q;\ h_i(x) \le 0,\ i = q+1, ..., p\}, \tag{4.7}$$

with the constraint functions h_i, $i = 1, ..., p$, being twice continuously differentiable. Let $L(x,\lambda) := f(x) + \sum_{i=1}^p \lambda_i h_i(x)$ be the Lagrangian function of the true problem. Suppose that the following, Mangasarian-Fromovitz [21], constraint qualification holds, at the point x_0:

- the gradient vectors $\nabla h_i(x_0)$, $i = 1, ..., q$, are linearly independent,
- there exists a vector $w \in I\!R^m$ such that $w^T \nabla h_i(x_0) = 0$, $i = 1, ..., q$, and $w^T \nabla h_i(x_0) < 0$, $i \in \mathcal{I}(x_0)$, where

  $$\mathcal{I}(x_0) := \{i : h_i(x_0) = 0,\ i = q+1, ..., p\} \tag{4.8}$$

  denotes the set of active at x_0 inequality constraints.

Then

$$T_S(x_0) = \left\{ d \in I\!R^m : d^T \nabla h_i(x_0) = 0,\ i = 1, ..., q;\ d^T \nabla h_i(x_0) \le 0,\ i \in \mathcal{I}(x_0) \right\}, \tag{4.9}$$

and first order (Kuhn-Tucker) necessary optimality conditions take the form: there exists a vector $\lambda = (\lambda_1, ..., \lambda_p)$ such that

$$\nabla_x L(x_0, \lambda) = 0,\ \ \lambda_i \ge 0,\ \ \lambda_i h_i(x_0) = 0,\ i = q+1, ..., p. \tag{4.10}$$

Under the Mangasarian-Fromovitz constraint qualification, the set $\Lambda(x_0)$ of all Lagrange multipliers vectors λ, satisfying the above conditions (4.10), is nonempty and bounded, and for any $\lambda \in \Lambda(x_0)$ the critical cone can be written as

$$C(x_0) = \left\{ d : d^T \nabla h_i(x_0) = 0,\ i \in \{1, ..., q\} \cup \mathcal{I}_+(\lambda),\ d^T \nabla h_i(x_0) \le 0,\ i \in \mathcal{I}_0(\lambda) \right\}, \tag{4.11}$$

where

$$\mathcal{I}_+(\lambda) := \{i \in \mathcal{I}(x_0) : \lambda_i > 0\} \text{ and } \mathcal{I}_0(\lambda) := \{i \in \mathcal{I}(x_0) : \lambda_i = 0\}.$$

Moreover, the set S is second order regular at x_0, and for $d \in T_S(x_0)$,

$$T^2_S(x_0,d) = \left\{ w \in I\!R^m : \begin{array}{l} w^T \nabla h_i(x_0) + d^T \nabla^2 h_i(x_0) d = 0,\ i = 1, \cdots, q, \\ w^T \nabla h_i(x_0) + d^T \nabla^2 h_i(x_0) d \le 0,\ i \in \mathcal{I}_1(x_0,d) \end{array} \right\}, \tag{4.12}$$

where

$$\mathcal{I}_1(x_0,d) := \left\{ i \in \mathcal{I}(x_0) : d^T \nabla h_i(x_0) = 0 \right\}. \tag{4.13}$$

It follows then by duality arguments that the second order conditions (4.6) can be written in the following equivalent form

$$\sup_{\lambda \in \Lambda(x_0)} d^T \nabla^2_{xx} L(x_0, \lambda) d > 0, \text{ for all } d \in C(x_0) \setminus \{0\}. \tag{4.14}$$

We are prepared now to discuss second order expansions of the optimal value function. We assume that the set S is compact and work in the Banach space $W^{1,\infty}(S)$ of Lipschitz continuous functions $y : S \to I\!R$ equipped with the norm

$$\|y\| := \sup_{x \in S} |y(x)| + \sup \left\{ \frac{|y(x') - y(x)|}{\|x' - x\|} : x, x' \in S,\ x' \neq x \right\}.$$

Since any function $y \in W^{1,\infty}(S)$ is Lipschitz continuous on S, the above norm of y is finite. Consider the optimal value function $\vartheta(y) := \inf_{x \in S} y(x)$, and let $\chi(y)$ be a corresponding optimal solution, i.e., $\chi(y) \in \arg\min_{x \in S} y(x)$. Note that, since it is assumed that the set S is compact, such optimal solution always exists, although possibly is not unique. Let K be the subset of $W^{1,\infty}(S)$ formed by (Fréchet) differentiable at x_0 functions, i.e., $y \in K$ if there exists $\nabla y(x_0) \in I\!R^m$ such that $y(x) = y(x_0) + (x - x_0)^T \nabla y(x_0) + o(\|x - x_0\|)$ for $x \in S$. Clearly K is a linear subspace of $W^{1,\infty}(S)$. We have then the following second order expansion of $\vartheta(\cdot)$ and first order expansion of $\chi(\cdot)$ in the space $W^{1,\infty}(S)$ tangentially to K (Bonnans and Shapiro [5]).

Theorem 4.2 *Suppose that:* (i) *the true problem has a unique optimal solution* x_0*,* (ii) *the function* f *is twice continuously differentiable in a neighborhood of the point* x_0*,* (iii) *the second order growth condition* (4.2) *holds,* (iv) *the set* S *is second order regular at* x_0*. Then the optimal value function* $\vartheta : W^{1,\infty}(S) \to I\!R$ *is first and second order Hadamard directionally differentiable at* f *tangentially to the space* K*, and for* $\delta \in K$ *it follows that* $\vartheta'_f(\delta) = \delta(x_0)$ *and*

$$\vartheta''_f(\delta) = \inf_{d \in C(x_0)} \left\{ 2d^T \nabla \delta(x_0) + d^T \nabla^2 f(x_0) d + \inf_{w \in T_S^2(x_0, d)} w^T \nabla f(x_0) \right\}. \tag{4.15}$$

Suppose, further, that: (v) *for any* $\delta \in K$ *the optimization problem in the right hand side of* (4.15) *has unique optimal solution* $\bar{d}(\delta)$*. Then the optimal solution function* $\chi(\cdot)$ *is Hadamard directionally differentiable at* f *tangentially to* K *and* $\chi'_f(\delta) = \bar{d}(\delta)$*.*

Clearly, if $\nabla f(x_0) = 0$, then the last term in the right hand side of (4.15) vanishes. Another situation where this term vanishes is if the set S is polyhedral, i.e., is defined by a finite number of linear constraints. In general this term is related, through the second order tangent set $T_S^2(x_0, d)$, to the curvature of the set S, at the point x_0. For two-stage stochastic programming problems with recourse expansion (4.15) was derived, and extended further to a case with multiple optimal solutions, in Dentcheva and Römisch [9].

In case the set S is defined by smooth constraints, as in (4.7), and the Mangasarian-Fromovitz constraint qualification holds, the set S is second order regular at x_0 and the second order growth condition (4.2) is equivalent to the second order optimality conditions (4.14). Moreover, it is possible to show, by using formula (4.12) and duality arguments, that the second order expansion (4.15) can be written then in the following equivalent form

$$\vartheta''_f(\delta) = \inf_{d \in C(x_0)} \left\{ 2d^T \nabla \delta(x_0) + \sup_{\lambda \in \Lambda(x_0)} d^T \nabla^2_{xx} L(x_0, \lambda) d \right\}. \tag{4.16}$$

Recall that, under the Mangasarian-Fromovitz constraint qualification, the set $\Lambda(x_0)$ of Lagrange multipliers is nonempty and bounded. Note also that the second order sufficient conditions (4.6) (second order sufficient conditions (4.14)) ensure that the infimum in the right hand side of (4.15) (in the right hand side of (4.16)) is attained, although it can be not unique. The optimization problem in the right hand side of (4.16) has a *unique* optimal solution if the function

$$\psi(d) := \sup_{\lambda \in \Lambda(x_0)} d^T \nabla^2_{xx} L(x_0, \lambda) d$$

is strictly convex on the linear space generated by the critical cone $C(x_0)$. In particular, this holds if the Hessian matrix $\nabla^2_{xx} L(x_0, \lambda)$ is positive definite for every $\lambda \in \Lambda(x_0)$.

The above second order expansion of the optimal value function $\vartheta(\cdot)$ and the corresponding first order approximation of the optimal solution mapping $\chi(\cdot)$, together with the Delta method, imply the following asymptotics of the optimal value $\hat{v}_N$ and an optimal solution $\hat{x}_N$ of the approximating problem.

Theorem 4.3 *Suppose that the assumptions* (i)-(iv) *of theorem* 4.2*, for the true problem, are satisfied. Let* τ_N *be a sequence of positive numbers tending to infinity, and* $\hat{f}_N$ *be a sequence of random elements of* $W^{1,\infty}(S)$ *such that the sequence* $\tau_N(\hat{f}_N - f)$ *converges in distribution to a random element* Y *of* $W^{1,\infty}(S)$ *and that* $\hat{f}_N(\cdot)$ *is (Fréchet) differentiable at* x_0 *w.p.1. Then*

$$\tau_N^2 \left[\hat{v}_N - \hat{f}_N(x_0) \right] \Rightarrow \tfrac{1}{2} \vartheta''_f(Y), \tag{4.17}$$

and

$$\hat{v}_N = \hat{f}_N(x_0) + \tfrac{1}{2} \vartheta''_f(\hat{f}_N - f) + o_p(\tau_N^{-2}), \tag{4.18}$$

where $\vartheta''_f(\cdot)$ *is given in* (4.15)*. Suppose, further, that the assumption* (v) *of theorem* 4.2 *holds and let* $\bar{d}(\cdot)$ *be the corresponding (unique) optimal solution function associated with the problem* (4.15)*. Then*

$$\tau_N \left(\hat{x}_N - x_0 \right) \Rightarrow \bar{d}(Y). \tag{4.19}$$

Regularity conditions which are required to ensure convergence in distribution of the sequence $\tau_N(\hat{f}_N - f)$ of random elements of the space $W^{1,\infty}(S)$ may be not satisfied in interesting nondifferentiable examples. Nevertheless, even in such cases formulas (4.17) - (4.19) often give correct asymptotics which can be proved by different methods.

Suppose now that the approximating functions $\hat{f}_N$ are constructed by averaging an i.i.d. random sample, as in (3.4). Suppose further that the function $g(\cdot,\omega)$ is Lipschitz continuous on S and (Fréchet) differentiable at x_0 for P-almost every ω. Moreover, suppose that first order derivatives of $f(x)$ can be taken inside the expected value, i.e., formula

$$\nabla f(x) = I\!E_P\{\nabla_x g(x,\omega)\} \tag{4.20}$$

holds. Then $N^{1/2}\left(\nabla \hat{f}_N(x_0) - \nabla f(x_0)\right)$ converges in distribution to multivariate normal $N(0,\Sigma)$, with the covariance matrix

$$\Sigma = I\!E_P\left\{[\nabla_x g(x_0,\omega)][\nabla_x g(x_0,\omega)]^T\right\} - \nabla f(x_0)\nabla f(x_0)^T, \tag{4.21}$$

provided that the second order moments of $\nabla_x g(x_0,\omega)$ do exist. We obtain therefore the following results.

Theorem 4.4 *Suppose that the assumptions* (i)-(iv) *of theorem* 4.2*, for the true problem, are satisfied. Let the approximating function $\hat{f}_N$ be constructed by averaging an i.i.d. random sample, and suppose that the function $g(\cdot,\omega)$ is Lipschitz continuous on S and (Fréchet) differentiable at x_0 w.p.1, that the interchangeability formula* (4.20) *holds, and that $N^{1/2}(\hat{f}_N - f)$ are random elements of $W^{1,\infty}(S)$ converging in distribution. Then*

$$N\left[\hat{v}_N - \hat{f}_N(x_0)\right] \Rightarrow \tfrac{1}{2}\varphi(Z) \tag{4.22}$$

and

$$\hat{v}_N = \hat{f}_N(x_0) + \tfrac{1}{2}\varphi(\zeta_N) + o_p(N^{-1}), \tag{4.23}$$

where $Z \sim N(0,\Sigma)$ is a random vector having multivariate normal distribution with the covariance matrix Σ given in (4.21)*, $\zeta_N := \nabla \hat{f}_N(x_0) - \nabla f(x_0)$, and*

$$\varphi(\zeta) := \inf_{d\in C(x_0)}\left\{2d^T\zeta + d^T\nabla^2 f(x_0)d + \inf_{w\in T_S^2(x_0,d)} w^T\nabla f(x_0)\right\}. \tag{4.24}$$

Suppose, further, that for any vector $\zeta \in I\!R^m$ the optimization problem in the right hand side of (4.24) *has a unique optimal solution, denoted $\bar{d}(\zeta)$. Then*

$$N^{1/2}\left(\hat{x}_N - x_0\right) \Rightarrow \bar{d}(Z). \tag{4.25}$$

In case the set S is defined by smooth constraints, as in (4.7), and the Mangasarian-Fromovitz constraint qualification holds, the function $\varphi(\cdot)$, defined in (4.24), can be written in the following equivalent from

$$\varphi(\zeta) = \inf_{d\in C(x_0)}\left\{2d^T\zeta + \sup_{\lambda\in\Lambda(x_0)} d^T\nabla^2_{xx}L(x_0,\lambda)d\right\}. \tag{4.26}$$

In that form formulas (4.23) and (4.25) were derived in Shapiro [28] by a different method. Asymptotics of the optimal solution $\hat{x}_N$ were also derived in King and Rockafellar [20] by the Delta method in a framework of variational inequalities (generalized equations).

Note that the "curvature term" (involving the second order tangent set $T_S^2(x_0, d)$) in the expansion (4.24) vanishes in two cases, namely if $\nabla f(x_0) = 0$ or if the set S is polyhedral. In such cases $\nabla^2_{xx} L(x_0, \lambda) = \nabla^2 f(x_0)$ for any $\lambda \in \Lambda(x_0)$.

Since $\hat{f}_N(x_0)$ is an unbiased estimator of v_0, we can view the term $\frac{1}{2}\mathbb{E}\{\varphi(Z)\}$, where $Z \sim N(0, \Sigma)$, as the asymptotic bias of $\hat{v}_N$, of order $O(N^{-1})$. Note that $\varphi(\zeta) \leq 0$ for any $\zeta \in \mathbb{R}^m$, and hence this asymptotic bias is negative.

The optimal solution $\bar{d}(\zeta)$ can be a nonlinear function of ζ even if this optimal solution is unique. In that case the distribution of $\bar{d}(Z)$ is not normal and hence $\hat{x}_N$ is not asymptotically normal (this was pointed out by King [17]). For example, let S be defined by constraints, as in (4.7), and suppose that the the gradient vectors $\nabla h_i(x_0)$, $i \in \{1, ..., q\} \cup \mathcal{I}(x_0)$, are linearly independent. Then $\Lambda(x_0) = \{\bar{\lambda}\}$ is a singleton and $\varphi(\zeta)$ and $\bar{d}(\zeta)$ are the optimal value and an optimal solution of the problem

$$\begin{array}{ll} \operatorname{Min}_{d \in \mathbb{R}^m} & 2d^T\zeta + d^T \nabla^2_{xx} L(x_0, \bar{\lambda}) d \\ \text{subject to} & d^T \nabla h_i(x_0) = 0,\ i \in \{1, ..., q\} \cup \mathcal{I}_+(\bar{\lambda}),\ d^T \nabla h_i(x_0) \leq 0,\ i \in \mathcal{I}_0(\bar{\lambda}). \end{array} \tag{4.27}$$

This is a quadratic programming problem. The above linear independence condition implies that it has a unique vector $\bar{\alpha}(\zeta)$ of Lagrange multipliers, and that it has a unique optimal solution $\bar{d}(\zeta)$ if the Hessian matrix $\nabla^2_{xx} L(x_0, \bar{\lambda})$ is positive definite over the linear space defined by the first $q + |\mathcal{I}_+(\bar{\lambda})|$ (equality) linear constraints in (4.27).

If, furthermore, the strict complementarity condition holds, i.e., $\bar{\lambda}_i > 0$ for all $i \in \mathcal{I}(x_0)$, or in other words $\mathcal{I}_+(\bar{\lambda}) = \mathcal{I}(x_0)$ and $\mathcal{I}_0(\bar{\lambda}) = \emptyset$, then $\bar{d}(\zeta)$ and $\bar{\alpha}(\zeta)$ can be obtained as solutions of the following system of linear equations

$$\begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} d \\ \alpha \end{bmatrix} = - \begin{bmatrix} \zeta \\ 0 \end{bmatrix}. \tag{4.28}$$

Here $H := \nabla^2_{xx} L(x_0, \bar{\lambda})$ and A is the $m \times (q + |\mathcal{I}(x_0)|)$ matrix whose columns are formed by vectors $\nabla h_i(x_0)$, $i \in \{1, ..., q\} \cup \mathcal{I}(x_0)$. We obtain in that case, provided the block matrix in the left hand side of (4.28) is nonsingular, that $N^{1/2}(\hat{x}_N - x_0, \hat{\lambda}_N - \bar{\lambda})$ converges in distribution to normal with zero mean and the covariance matrix

$$\begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix}^{-1}. \tag{4.29}$$

It can happen that the critical cone $C(x_0)$ consists of the single point 0, i.e., $C(x_0) = \{0\}$. In that case the functions $\varphi(\cdot)$ and $\bar{d}(\cdot)$ are identically zero and the corresponding asymptotics are different. For example, if the set S is defined by constraints, as in (4.7), and the Mangasarian-Fromovitz constraint qualification holds, then it follows from formula (4.11) that $C(x_0) = \{0\}$ if the gradient vectors $\nabla h_i(x_0)$, $i \in \{1, ..., q\} \cup \mathcal{I}_+(\lambda)$, generate the space $\mathbb{R}^m$. In particular this happens if the number of active inequality constraints at x_0 is $m - q$ (i.e., $|\mathcal{I}(x_0)| = m - q$), the gradient vectors $\nabla h_i(x_0)$, $i \in \{1, ..., q\} \cup \mathcal{I}(x_0)$, are linearly independent and all Lagrange multipliers corresponding to the active inequality constraints are positive.

Suppose that $C(x_0) = \{0\}$. In that case there exists a neighborhood U of $\nabla f(x_0)$ such that if $\nabla \hat{f}_N(x_0) \in U$, then the first order optimality conditions for the approximating problem hold at the point x_0, and x_0 is a locally optimal solution of the approximating problem. By the strong Law of Large Numbers, we have that $\nabla \hat{f}_N(x_0)$ converges to $\nabla f(x_0)$ w.p.1. Consequently, w.p.1 for N large enough, $\nabla \hat{f}_N(x_0) \in U$, and hence x_0 is a locally optimal solution of the approximating problem. It follows then that $\hat{x}_N = x_0$ w.p.1 for N large enough. Moreover, by the Large Deviations theory (e.g., [6]) we have, under mild regularity conditions, that the probability of the event $\nabla \hat{f}_N(x_0) \notin U$ tends to zero exponentially fast as $N \to \infty$, and hence the asymptotic bias of $\hat{v}_N$ approaches zero at an exponential rate.

Let us finally remark that it is also possible to derive similar asymptotics, of the optimal value and optimal solutions, in cases where the feasible set is defined by constraints and the constraint functions are estimated by corresponding sample averages (Rubinstein and Shapiro [27, section 6.6]).

5 Examples and a discussion

Consider the framework of the maximum likelihood example 1.1. Suppose that the parameter set Θ is compact and that the distribution P, of the corresponding random sample, is given by a pdf $g(y, \theta_0)$, $\theta_0 \in \Theta$, from the considered parametric family. Suppose also that for P-almost every y, the function $\ln g(y, \cdot)$ is continuously differentiable in a neighborhood of Θ, and that the corresponding assumptions (A1) - (A3) hold for the function $\ln g(y, \theta)$ and its first order partial derivatives $\partial \ln g(y, \theta)/\partial \theta_i$. Then, since θ_0 is an unconstrained minimizer of $f(\theta)$, and $\nabla f(\theta) = -\mathbb{E}_P\{\nabla_\theta \ln g(Y, \theta)\}$, we obtain that $\nabla f(\theta_0) = 0$. Suppose, further, that the expected value function $f(\theta)$ is twice continuously differentiable at θ_0 (note that this property does not follow from the above assumptions), that the parameter vector θ is identified at θ_0 (and hence the minimizer θ_0 is unique), that the second order growth condition holds at θ_0 and that the set Θ is "sufficiently regular" near θ_0. Then the corresponding asymptotic expansions given in theorem 4.3 hold.

Since $\nabla f(\theta_0) = 0$, we have here that $C(\theta_0) = T_\Theta(\theta_0)$ and the third term in the right hand side of (4.24) vanishes. The covariance matrix Σ is equal here to $I(\theta_0)$, where

$$I(\theta_0) := \mathbb{E}\left\{[\nabla_\theta \ln g(Y, \theta_0)][\nabla_\theta \ln g(Y, \theta_0)]^T\right\}$$

is Fisher's information matrix. As it is well known, under second order smoothness assumptions about the function $\ln g(y,\theta)$, we also have that $\nabla^2 f(\theta_0) = I(\theta_0)$. Consequently, the second order growth condition is ensured here by the condition: $d^T I(\theta_0) d > 0$ for all nonzero $d \in T_\Theta(\theta_0)$. In particular, this holds if $I(\theta_0)$ is nonsingular, and hence is positive definite. By (4.23) we obtain that

$$\sup_{\theta\in\Theta} \sum_{i=1}^N \ln g(Y_i,\theta) = \sum_{i=1}^N \ln g(Y_i,\theta_0) + \sup_{d\in T_\Theta(\theta_0)} \left\{ d^T Z_N - \tfrac{1}{2} d^T I(\theta_0) d \right\} + o_p(1),$$

where $Z_N := N^{-1/2} \sum_{i=1}^N \nabla_\theta \ln g(Y_i,\theta)$. Note that $Z_N \Rightarrow N(0, I(\theta_0))$.

Consider the log-likelihood ratio statistic ℓ_N, defined in (3.8), for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. Suppose that the true value θ_0 of the parameter vector belongs to both sets Θ_0 and Θ_1, that the information matrix $I(\theta_0)$ is nonsingular, and define $W_N := I(\theta_0)^{-1} Z_N$. Note that $W_N \Rightarrow N(0, I(\theta_0)^{-1})$. We obtain then the following expansion of ℓ_N,

$$\ell_N = \inf_{d\in T_{\Theta_0}(\theta_0)} (W_N - d)^T I(\theta_0)(W_N - d) - \inf_{d\in T_{\Theta_1}(\theta_0)} (W_N - d)^T I(\theta_0)(W_N - d) + o_p(1).$$

It also follows that if $\hat\theta_N$ is the maximum likelihood estimator of θ_0 under H_0 (under H_1), then $N^{1/2}(\hat\theta_N - \theta_0)$ converges in distribution to $\bar d(W)$, where $W \sim N(0, I(\theta_0)^{-1})$ and $\bar d(w)$ is the minimizer of $(w-d)^T I(\theta_0)(w-d)$ over $T_{\Theta_0}(\theta_0)$ (over $T_{\Theta_1}(\theta_0)$). This result goes back to Chernoff [7].

The above discussion shows that the example of maximum likelihood is quite specific from the point of view of general stochastic optimization problems. In that example the gradient of the objective function of the true problem is zero at the optimal solution, and consequently the "curvature term" vanishes from the corresponding second order expansions of the optimal value function.

Before proceeding further let us state the following useful proposition. It can be easily proved by using the Lebesgue dominated convergence theorem (e.g., [27, pp. 70,71]).

Proposition 5.1 *Let $f(x)$ be the expected value function defined in* (1.3). *Suppose that the expectation $\mathbb{E}_P\{g(x,\omega)\}$ exists for all x in a neighborhood of x_0, that for P-almost every ω the function $g(\cdot,\omega)$ is directionally differentiable at x_0, and that there exists a random variable $\kappa(\omega) \geq 0$ such that $\mathbb{E}_P\{\kappa(\omega)\}$ is finite and*

$$|g(x_1,\omega) - g(x_2,\omega)| \leq \kappa(\omega)\|x_1 - x_2\| \tag{5.1}$$

for all x_1, x_2 in a neighborhood of x_0 and P-almost all ω . Then the function $f(x)$ is Lipschitz continuous near x_0, directionally differentiable at x_0 and

$$f'(x_0,d) = \mathbb{E}_P\{g'_\omega(x_0,d)\}, \tag{5.2}$$

where $g'_\omega(x_0,d)$ denotes the directional derivative of $g(\cdot,\omega)$ at x_0 in the direction d. Moreover, if $g(\cdot,\omega)$ is differentiable at x_0 w.p.1, then the interchangeability formula (4.20) *holds at $x = x_0$.*

Let us discuss now the two-stage stochastic programming example 1.2. Consider the function

$$G(z) := \inf\{q^T y : Wy = z,\ y \geq 0\}.$$

Clearly the function $Q(x, h)$, given as the optimal value of the problem (1.5), can be written as $Q(x, h) = G(h - Ax)$. By duality arguments of linear programming we have that

$$G(z) = \sup\{\xi^T z : W^T \xi \leq q\},$$

provided the set $\{\xi : W^T\xi \leq q\}$ is nonempty. So let us suppose, for the sake of simplicity, that this set is nonempty and bounded. Then the function $G(z)$ is a real valued piecewise linear convex function. Suppose also that the expectation $\mathbb{E}\{Q(x, h)\}$ exists for all x.

It follows that the approximating function $\hat{f}_N(x) := c^T x + N^{-1} \sum_{i=1}^{N} Q(x, h_i)$ is a piecewise linear convex function, and hence is not everywhere differentiable. Therefore, the involved asymptotics are quite different depending on whether the distribution of the random vector h is continuous or discrete. Suppose first that the random vector h has a continuous distribution with a density function $g(\cdot)$. Let us fix a point $x_0 \in \mathbb{R}^m$. Since the function $G(z)$ is convex, the set of points where it is not differentiable has Lebesgue measure zero. Since h has a density, it follows then that the function $Q(\cdot, h)$ is differentiable at x_0 w.p.1. Together with (5.2) this implies that $f(x)$ is differentiable at x_0 and $\nabla f(x_0) = c + \mathbb{E}\{\nabla_x Q(x_0, h)\}$. If, moreover, the density function $g(\cdot)$ is continuous, then $f(x)$ is twice continuously differentiable (Wang [35]). In that case the asymptotic formulas (4.22), (4.23) and (4.25), of theorem 4.4, with the covariance matrix Σ of $Z \sim N(0, \Sigma)$ defined in (4.21), make sense. Under some mild assumptions about the density function $g(\cdot)$, these formulas can be proved by a different method, which is based on a stochastic mean value theorem due to Huber [14], (Shapiro [31]).

Let us finally mention that in case the random vector h has a discrete distribution the situation is quite different. It is possible to show that in such case and if the true problem has unique optimal solution x_0, then the probability of the event that $\hat{x}_N$ is exactly equal to x_0 approaches one exponentially fast (Shapiro and Homem-de-Mello [34]).

References

[1] A. Araujo and E. Giné, *The Central Limit Theorem for Real and Banach Valued Random Variables.* Wiley, New York, 1980.

[2] E. Beale, "On minimizing a convex function subject to linear inequalities", *J. Roy. Statist. Soc., Ser. B*, 17 (1955), 173-184.

[3] P. Billingsley, *Convergence of Probability Measures.* Wiley, New York, 1968.

[4] J.R. Birge and F. Louveaux, *Introduction to Stochastic Programming.* Springer, New York, 1997.

[5] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems.* Springer-Verlag, New York, 2000.

[6] J.A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1990.

[7] H. Chernoff, "On the distribution of the likelihood ratio", *Ann. Math. Statist.*, 25 (1954), 573-578.

[8] G. Dantzig, "Linear programming under uncertainty", *Management Sci.,* 1 (1955), 197-206.

[9] D. Dentcheva and W. Römisch, "Differential stability of two-stage stochastic programs", *SIAM J. Optimization*, to appear.

[10] J. Dupačová and R.J.B. Wets, "Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems," *The Annals of Statistics* 16 (1988) 1517-1549.

[11] R.D. Gill, "Non-and-semiparametric maximum likelihood estimators and the von Mises method (Part I)", *Scandinavian Journal of Statistics*, 16 (1989), 97-124.

[12] R. Grübel, "The length of the short", *Ann. Statist.*, 16 (1988), 619-628.

[13] P.J. Huber, "Robust estimation of a location parameter", *Ann. Math. Statist.*, 35 (1964), 73-101.

[14] P.J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions", *Proc. Fifth Berkeley Symp.Math. Statist. Probab.*, 1 (1967), 221-233, Univ. California Press.

[15] P.J. Huber, *Robust Statistics.* Wiley, New York, 1981.

[16] P. Kall and S.W. Wallace, *Stochastic Programming.* Wiley, Chichester, 1994.

[17] A.J. King, "Asymptotic behavior of solutions in stochastic optimization: Nonsmooth analysis and the derivation of non-normal limit distributions", Ph.D. dissertation, Dept. Applied Mathematics, Univ. Washington, 1986.

[18] A.J. King, "Generalized delta theorems for multivalued mappings and measurable selections", *Mathematics of Operations Research*, 14 (1989), 720-736.

[19] A.J. King and R.J.-B. Wets, "Epi-consistency of convex stochastic programs," *Stochastics*, 34 (1991), 83-92.

[20] A.J. King and R.T. Rockafellar, "Asymptotic theory for solutions in statistical estimation and stochastic programming", *Mathematics of Operations Research*, 18 (1993), 148-162.

[21] O.L. Mangasarian and S. Fromovitz, "The Fritz John necessary optimality conditions in the presence of equality and inequality constraints", *Journal of Mathematical Analysis and Applications*, 7 (1967), pp. 37-47.

[22] E.L. Plambeck, B.R. Fu, S.M. Robinson and R. Suri, "Sample-Path Optimization of Convex Stochastic Performance Functions", *Mathematical Programming*, vol. 75 (1996), no. 2, 137-176.

[23] D. Pollard, *Convergence of Stochastic Processes.* Springer-Verlag, New York, 1984.

[24] C.R. Rao, *Linear Statistical Inference and Its Applications.* Wiley, New York, 1973.

[25] S.M. Robinson, "Analysis of sample-path optimization, " *Math. Oper. Res.*, 21 (1996), 513-528.

[26] R.T. Rockafellar, *Convex Analysis.* Princeton University Press, 1970.

[27] R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* Wiley, New York, NY, 1993.

[28] A. Shapiro, "Asymptotic Properties of Statistical Estimators in Stochastic Programming," *Annals of Statistics*, 17 (1989), 841-858.

[29] A. Shapiro, "On concepts of directional differentiability," *Journal Optim. Theory and Appl.*, 66 (1990), 477-487.

[30] A. Shapiro, "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, 30 (1991), 169-186.

[31] A. Shapiro, "Asymptotic Behavior of Optimal Solutions in Stochastic Programming," *Mathematics of Operations Research*, 18 (1993), 829-845.

[32] A. Shapiro and Y. Wardi, "Nondifferentiability of the steady-state function in Discrete Event Dynamic Systems", *IEEE transactions on Automatic Control*, 39 (1994), 1707-1711.

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

[34] A. Shapiro and T. Homem-de-Mello, "On rate of convergence of Monte Carlo approximations of stochastic programs", *SIAM J. Optimization*, to appear.

[35] J. Wang, "Distribution sensitivity analysis for stochastic programs with complete recourse", *Mathematical Programming*, 31 (1985), 286-297.

[36] R.W. Wolff, *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ, 1989.