

# Stochastic Programming by Monte Carlo Simulation Methods \*

Alexander Shapiro

School of Industrial and Systems Engineering,  
Georgia Institute of Technology,  
Atlanta, Georgia 30332-0205, USA

## Abstract

We consider in this paper stochastic programming problems which can be formulated as an optimization problem of an expected value function subject to deterministic constraints. We discuss a Monte Carlo simulation approach based on sample average approximations to a numerical solution of such problems. In particular, we give a survey of a statistical inference of the sample average estimators of the optimal value and optimal solutions of the true problem. We also discuss stopping rules and a validation analysis for such sample average approximation optimization procedures and give some illustration examples.

---

\*This work was supported, in part, by grant Grant DMI-9713878 from the National Science Foundation.

# 1 Introduction

We consider in this paper optimization problems of the form

$$\text{Min}_{x \in S} \{g(x) := \mathbb{E}_P G(x, \omega)\}. \quad (1.1)$$

Here  $x \in \mathbb{R}^m$  is a (finite dimensional) vector of decision variables,  $S$  is a closed subset of  $\mathbb{R}^m$  representing feasible solutions of the above problem,  $(\Omega, \mathcal{F}, P)$  is a probability space and  $G : \mathbb{R}^m \times \Omega \rightarrow \mathbb{R}$  is a real valued function. We assume throughout the paper that for every  $x \in S$  the expected value function  $g(x)$  is well defined, i.e. the function  $G(x, \cdot)$  is  $\mathcal{F}$ -measurable and  $P$ -integrable.

The above optimization problem gives an abstract formulation of a situation where the considered system involves data which are subject to random variations, uncertainty, lack of information etc., and one wants to optimize that system on *average*. We assume that the probability measure (distribution)  $P$  is known, although may be not given explicitly, or at least can be estimated from available data. For the most part of this paper one can think about  $G(x, W)$  as a (known) function of  $x$  and a (finite dimensional) random vector  $W = W(\omega)$ , whose distribution can be continuous or discrete. We will use notation  $G(x, \omega)$  or  $G(x, W)$ , which one will be clear from the context.

The purpose of this paper is to give a survey of some recent developments in an approach to a numerical solution of (1.1) based on Monte Carlo simulation techniques. The basic idea of such methods is simple indeed, a random (or rather pseudorandom) sample  $\omega^1, \dots, \omega^N$  is generated and consequently problem (1.1) is approximated by

$$\text{Min}_{x \in S} \left\{ \hat{g}_N(x) := N^{-1} \sum_{i=1}^N G(x, \omega^i) \right\}. \quad (1.2)$$

We refer to (1.1) and (1.2) as the true (or expected value) and the sample average approximation (SAA) problems, respectively. Let  $\hat{v}_N$  be the optimal value and  $\hat{x}_N$  be an optimal solution of the SAA problem (1.2). We discuss, in particular, statistical inferences of  $\hat{v}_N$  and  $\hat{x}_N$  considered as estimators of their “true” counterparts  $v^*$  and  $x^*$ , respectively.

We assume that the feasible set  $S$  is given explicitly, typically by smooth (or even linear) constraints. It is possible to extend the presented theory to situations where constraint functions, defining the feasible set, are also given in a form of expected values and have to be estimated say by the corresponding sample average functions. Quite often such constraints can be incorporated into the objective function in a form of penalty or barrier terms. Therefore, for the sake of simplicity, we restrict our attention to stochastic problems with an explicitly given feasible set.

If for, every  $x \in S$ , the function  $G(x, W)$  is linear in  $W$ , then we have that  $g(x) = G(x, \mu)$ , where  $\mu := \mathbb{E}\{W\}$ . In that case, provided the mean vector  $\mu$  is known, the objective function of problem (1.1) is given explicitly, and hence it becomes a deterministic optimization problem. In general, however, an optimal solution of (1.1) can be different from an optimal solution obtained by replacing the involved random variables with their means.

There is an extensive literature, in the statistics and optimization areas, dealing with various aspects of the sample average approximation approach. The classical Maximum

Likelihood (ML) method of estimation can be considered in the framework of the SAA problem (1.2) (see example 2.1 below), with an essential difference that in statistical estimation procedures the random sample is provided by observed data rather than generated in the computer. It is difficult to point out who first suggested the SAA approach to solving stochastic problems of the form (1.1). The idea is quite simple and natural, and it seems that variants of the SAA method were discovered and rediscovered by various authors over the years. In a context of simulation models a variant of the SAA method, based on Likelihood Ratio transformations, was suggested in Rubinstein and Shapiro [33, 34]. Independently, and more or less at the same time, similar ideas were employed for calculating ML estimators by Monte Carlo techniques based on Gibbs sampling (see Geyer and Thompson [12], Geyer [13] and references therein). Ad hoc algorithms, based on Monte Carlo simulation, for two stage stochastic problems with recourse were developed by Hight and Sen [14] and Infanger [19]. Statistical inference of  $\hat{v}_N$  and  $\hat{x}_N$  was incorporated into numerical algorithms, for purposes of error estimates, stopping rules and validation analysis, in Shapiro and Homem-de-Mello [40].

Let us remark that the terminology “sample average approximation” is not uniform in the literature. For example, the term “sample-path optimization” was used in Plambeck, Fu, Robinson and Suri [29]. Let us also note that an alternative approach to solving (1.1) is based on the Stochastic Approximation (SA) method combined with Monte Carlo simulation (see Benveniste, Métivier and Priouret [4] and Kushner and Clark [24] for a basic discussion of the SA method and Chong and Ramadge [8] and L’Ecuyer, Giroux and Glynn [25], for example, for applications to queueing systems).

This paper is organized as follows. In the next section we give several illustrational examples which motivate our discussion. In section 3 we discuss differentiability properties of the expected value function. Section 4 is devoted to the Likelihood Ratio method. Statistical inference of the SAA estimators is discussed in section 5. Finally, in section 6, we give a brief discussion of stopping rules and validation analysis for the SAA method.

We use the following notation and terminology throughout the paper. For a set  $S \subset \mathbb{R}^m$ ,  $\text{dist}(x, S) := \inf_{z \in S} \|x - z\|$  denotes the distance from a point  $x$  to the set  $S$ . For a real valued function  $g : \mathbb{R}^m \rightarrow \mathbb{R}$ , by  $\nabla g(x) := (\partial g(x)/\partial x_1, \dots, \partial g(x)/\partial x_m)^T$  we denote its gradient and by  $\nabla^2 g(x) := [\partial^2 g(x)/\partial x_i \partial x_j]$  its Hessian matrix of second order partial derivatives. For a number  $a$ , we denote  $[a]^+ := \max\{a, 0\}$ . The sign of a number  $a \neq 0$  is defined as  $\text{sign}(a) := 1$  if  $a > 0$  and  $\text{sign}(a) := -1$  if  $a < 0$ .

## 2 Examples

In this section we introduce several examples which motivate and illustrate the subsequent discussion.

**Example 2.1** Our first example is motivated by the classical Maximum Likelihood (ML) method of estimation. That is, let  $f(y, \theta)$  be a family of probability density functions (pdf), parameterized by the parameter vector  $\theta \in \Theta \subset \mathbb{R}^m$ , and let  $Y^1, \dots, Y^N$  be an i.i.d. random sample with a probability distribution  $P$ . Define

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

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

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

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

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

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

Note that  $\theta^*$  is an *unconstrained* minimizer of  $g(\theta)$ , in the sense that it minimizes  $g(\theta)$  over its domain, even if  $\theta^*$  lies on the boundary of the feasible region  $\Theta$ .

There exists a vast literature on the ML method, and the above derivation of optimality of  $\theta^*$  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  $\hat{g}_N(\theta)$ , over  $\Theta$ , is viewed as the ML estimator of the “true” value  $\theta^*$  of the parameter vector. There are also various extensions of the ML method, in particular the method of  $M$ -estimators introduced by Huber [16, 18].

Somewhat different type of examples is motivated by a Monte Carlo simulation approach to numerical solutions of stochastic programming problems of the form (1.1). The probability distribution  $P$  is supposed to be known, although may be not given explicitly. However, the expected value function  $g(x)$  cannot be calculated in a closed form and has to be approximated. Monte Carlo simulation techniques provide such an approximation by averaging a generated random sample. In the next example we consider a problem which, on one hand, is sufficiently simple and can be solved analytically, on the other hand it demonstrates various properties which hold in considerably more complex situations.

**Example 2.2** Let  $W$  be a real valued random variable with cumulative distribution function (cdf)  $F(w) := P(W \leq w)$  and let  $G(x, w) := |x - w|$ , where  $x, w \in \mathbb{R}$ . For any  $w \in \mathbb{R}$ , the function  $G(\cdot, w)$  is convex, and hence the corresponding expected value function

$$g(x) := \mathbb{E}\{|x - W|\} = \int_{-\infty}^{+\infty} |x - w|dF(w) \tag{2.1}$$

is also convex. A minimizer  $x^*$  of  $g(x)$ , over  $\mathbb{R}$ , is given by the median  $x^* = F^{-1}(1/2)$  of the distribution of  $W$ , i.e.  $x^*$  is such that  $F(x^*_-) \leq 1/2$  and  $F(x^*) \geq 1/2$ , where  $F(x^*_-)$  denotes the left side limit of  $F(x)$  at  $x = x^*$ . If  $F(w)$  is continuous, then  $x^*$  is a median iff  $F(x^*) = 1/2$ , while if the event  $\{W = x^*\}$  can happen with positive probability, then  $F(x^*)$  can be bigger than half. Note also that it can happen that the minimizer  $x^*$  is not unique.

If we replace  $W$  by its mean  $\mu$ , then clearly the minimizer of the function  $G(x, \mu) = |x - \mu|$ , over  $x \in \mathbb{R}$ , is given by  $\mu$ . In general, the mean  $\mu$  can be different from the

median  $F^{-1}(1/2)$ . Note that in this example a minimizer  $\hat{x}_N$  of the corresponding sample average function  $\hat{g}_N(x)$ , over  $\mathbb{R}$ , is given by sample median  $\hat{F}_N^{-1}(1/2)$ , where  $\hat{F}_N$  denotes the empirical cdf based on the considered sample.

Following is an example of stochastic programming with recourse. The above median problem (2.1) can be considered as a particular case of such programs.

**Example 2.3** Consider the optimization problem

$$\text{Min}_{x \in S} c^T x + \mathbb{E}\{Q(x, h(\omega))\}, \quad (2.2)$$

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

$$\text{Min}_{y \geq 0} q^T y \text{ subject to } Wy = h - Ax, \quad (2.3)$$

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 (2.3) are deterministic.) This is the so-called two-stage stochastic programming problem with recourse, which originated in works of Beale [3] and Dantzig [9]. If the random vector  $h$  has a discrete distribution, then the expected value function  $\mathbb{E}\{Q(\cdot, h)\}$  is representable as a weighted sum of values of  $Q(\cdot, h)$ , and problem (2.2) 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. An interested reader is referred to recent books by Kall and Wallace [20] and Birge and Louveaux [5], 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  $h^1, \dots, h^N$  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 sample average function  $\hat{Q}_N(x) := N^{-1} \sum_{i=1}^N Q(x, h^i)$ . Consequently the “true” problem (2.2) is approximated by the SAA problem

$$\text{Min}_{x \in S} c^T x + \hat{Q}_N(x). \quad (2.4)$$

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

By the Law of Large Numbers we have that the SAA function  $\hat{Q}_N(x)$  converges, pointwise, to  $\mathbb{E}\{Q(x, h)\}$  with probability one, as  $N \rightarrow \infty$ . 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 (2.3). This allows to apply, reasonably efficient, deterministic algorithms in order to solve the SAA problem (2.4).

Let us make the following observations. The above example is different from the ML example 2.1 in several respects. In the above example the corresponding random sample

is generated in the computer and can be controlled to some extent. 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 an unconstrained minimizer of the objective function. There is no reason for such behavior of an optimal solution of the optimization problem (2.2). 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 sample 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 2.4** Consider stochastic process  $I_t$ ,  $t = 1, 2, \dots$ , governed by the recursive equation

$$I_t = [I_{t-1} + R(x_t, V_t) - D_t]^+, \quad (2.5)$$

with initial value  $I_0$ . Here  $V_t$  are random vectors,  $D_t$  are random numbers,  $R(\cdot, \cdot)$  is a real valued function of two vector variables, and vectors  $x_t$  represent decision variables. The above process  $I_t$  can describe the waiting time of  $t$ -th customer in a  $G/G/1$  queue, where  $D_t$  is the interarrival time between the  $(t-1)$ -th and  $t$ -th customers and  $R(x_t, V_t)$  is the service time of  $(t-1)$ -th customer. Alternatively, we may view  $I_t$  as an inventory of a certain product at time  $t$ , with  $D_t$  and  $R(x_t, V_t)$  representing the demand and production (or reordering) of the product at time  $t$ .

Suppose that the process is considered over a finite horizon at periods  $t = 1, \dots, T$ . Our goal then is to minimize (or maximize) the expected value of an objective function involving  $I_1, \dots, I_T$ . For instance, one may be interested in maximizing the expected value of a profit given by (cf. Albritton, Shapiro and Spearman [1])

$$G(x, W) := \sum_{t=1}^T \{\pi_t \min[I_{t-1} + R(x_t, V_t), D_t] - h_t I_t\}. \quad (2.6)$$

Here  $x := (x_1, \dots, x_T)$  is a vector of decision variables,  $W := (V_1, \dots, V_T, D_1, \dots, D_T)$  is a random vector of the involved random variables, and  $\pi_t$  and  $h_t$  are non negative parameters representing the marginal profit and the holding cost, respectively, of the product at period  $t$ . Note that the profit function  $G(x, W)$  can be also written in the form

$$\begin{aligned} G(x, W) &= \sum_{t=1}^T \{\pi_t (D_t + \min[I_{t-1} + R(x_t, V_t) - D_t, 0]) - h_t I_t\} = \\ &= \sum_{t=1}^T \{\pi_t [D_t + (I_{t-1} + R(x_t, V_t) - D_t) - I_t] - h_t I_t\} = \\ &= \sum_{t=1}^T \pi_t R(x_t, V_t) + \sum_{t=1}^{T-1} (\pi_{t+1} - \pi_t - h_t) I_t + \pi_1 I_0 - (\pi_T + h_T) I_T. \end{aligned} \quad (2.7)$$

It is also possible to consider a stationary distribution of the process  $I_t$  (if it exists) and to optimize an associated objective function. Typically, probability measure of such

stationary distribution cannot be written in a closed form. This introduces additional technical difficulties into the problem. In this paper we mainly deal with problems over a finite horizon where involved probability distributions are governed by finite dimensional random vectors.

If the initial value  $I_0$  is sufficiently large, then with probability close to one variables  $I_1, \dots, I_T$  stay above the zero. If, moreover,  $R(x_t, V_t)$  are linear in  $V_t$ , then  $I_1, \dots, I_T$  become linear functions of the random data vector  $W$ . In that case components of the random vector  $W$  can be replaced by their means. In many practical situations, however, the process  $I_t$  hits the zero with high probability over the considered horizon  $T$ . In such cases the corresponding expected value function  $g(x) := \mathbb{E} G(x, W)$  cannot be written in a closed form and one needs to use say a Monte Carlo simulation procedure in order to evaluate  $g(x)$ .

### 3 Perturbation Analysis

In order to design an efficient numerical optimization algorithm, one needs to study a differential structure of the expected value function. Let us observe that for a given realization  $w$  of the random vector  $W$ , the profit function  $G(\cdot, w)$ , defined in (2.6), is not everywhere differentiable in  $x$  even if the function  $R(\cdot, v)$  is differentiable for all  $v$ . This is because the operations of taking maximum or minimum do not preserve differentiability of the involved functions. For example, if  $R(x_t, v_t) := x_t$ , then  $I_t$  is a piecewise linear convex function of  $x_1, \dots, x_t$  which is not everywhere differentiable. Nevertheless  $G(\cdot, w)$  is directionally differentiable in all directions. Such behavior of the objective function also happens in examples 2.2 and 2.3, and is typical in many interesting applications. Let us make a quick detour into the theory of directional differentiability.

Consider a mapping (function)  $F : \mathbb{R}^m \rightarrow \mathbb{R}^n$ . It is said that  $F$  is directionally differentiable at a point  $x^* \in \mathbb{R}^m$  if the limit

$$F'(x^*, h) := \lim_{t \downarrow 0} \frac{F(x^* + th) - F(x^*)}{t} \quad (3.1)$$

exists for all  $h \in \mathbb{R}^m$ . In that case  $F'(x^*, h)$  is called the directional derivative of  $F(x)$  at  $x^*$  in the direction  $h$ . Note that  $F'(x^*, h)$  is positively homogeneous in  $h$ , i.e.  $F'(x^*, th) = tF'(x^*, h)$  for any  $t \geq 0$ . If  $F(x)$  is directionally differentiable at  $x^*$  and  $F'(x^*, h)$  is *linear* in  $h$ , then it is said that  $F(x)$  is Gâteaux differentiable at  $x^*$ . Equation (3.1) can be also written in the form

$$F(x^* + h) = F(x^*) + F'(x^*, h) + r(h), \quad (3.2)$$

where the remainder term  $r(h)$  is such that  $r(th)/t \rightarrow 0$ , as  $t \downarrow 0$ , for any fixed  $h \in \mathbb{R}^m$ . If, moreover,  $F'(x^*, h)$  is linear in  $h$  and the remainder term  $r(h)$  is “uniformly small” in the sense that  $r(h)/\|h\| \rightarrow 0$  as  $h \rightarrow 0$ , then it is said that  $F(x)$  is differentiable at  $x^*$  in the sense of Fréchet, or simply differentiable at  $x^*$ .

Clearly Fréchet differentiability implies Gâteaux differentiability. The converse of that is not necessarily true. However, for locally Lipschitz continuous mappings both concepts do coincide. Recall that  $F(x)$  is said to be locally Lipschitz continuous near  $x^*$  if there is a

positive constant  $K$  such that  $\|F(x_1) - F(x_2)\| \leq K\|x_1 - x_2\|$  for all  $x_1, x_2$  in a neighborhood of  $x^*$ .

Let now  $F : \mathbb{R}^m \rightarrow \mathbb{R}$  be a real valued *convex* function. In that case  $F(x)$  is always directionally differentiable and locally Lipschitz continuous, and moreover the directional derivatives can be written in the form

$$F'(x^*, h) = \inf_{t>0} \frac{F(x^* + th) - F(x^*)}{t}. \quad (3.3)$$

Therefore  $F(x)$  is differentiable at  $x^*$  iff  $F'(x^*, h)$  is linear in  $h$ . It is said that a vector  $a \in \mathbb{R}^m$  is a *subgradient* of  $F(x)$  at  $x^*$  if

$$F(x) - F(x^*) \geq a^T(x - x^*) \quad (3.4)$$

for all  $x \in \mathbb{R}^m$ . The set of all subgradients of  $F(x)$ , at  $x^*$ , is called the *subdifferential* and denoted  $\partial F(x^*)$ . By duality theory of convex analysis we have that

$$F'(x^*, h) = \sup_{a \in \partial F(x^*)} a^T h. \quad (3.5)$$

Hence  $F(x)$  is differentiable at  $x^*$  iff  $\partial F(x^*)$  is a singleton, i.e. contains only one element, which then coincides with the gradient  $\nabla F(x^*)$  (see Rockafellar [32] for a thorough discussion of convex analysis).

Let us come back now to the expected value function  $g(x)$ , defined in (1.1). Unless stated otherwise all probabilistic statements will be made with respect to the considered probability measure  $P$ . We say that a property holds for almost every (a.e.)  $\omega$  if it holds for all  $\omega \in \Omega$  except possibly on a set of  $P$ -measure zero. Another way of saying that a property holds for almost every  $\omega$ , is to say that the property holds with probability one (w.p.1). We sometimes write  $G_\omega(\cdot)$  for the function  $G(\cdot, \omega)$  and denote by  $G'_\omega(x^*, h)$  the directional derivative of  $G_\omega(\cdot)$  at the point  $x^*$  in the direction  $h$ .

**Proposition 3.1** *Suppose that the expected value function  $g(x)$  is well defined in a neighborhood of a point  $x^*$ , that for almost every  $\omega$  the function  $G_\omega(\cdot) := G(\cdot, \omega)$  is directionally differentiable at  $x^*$ , and that there exists a positive valued random variable  $K(\omega)$  such that  $\mathbb{E}\{K(\omega)\}$  is finite and for all  $x_1, x_2$  in a neighborhood of  $x^*$  and almost every  $\omega \in \Omega$  the following inequality holds*

$$|G(x_1, \omega) - G(x_2, \omega)| \leq K(\omega)\|x_1 - x_2\|. \quad (3.6)$$

*Then the expected value function  $g(x)$  is Lipschitz continuous in a neighborhood of  $x^*$ , directionally differentiable at  $x^*$ , and*

$$g'(x^*, h) = \mathbb{E}\{G'_\omega(x^*, h)\}. \quad (3.7)$$

*Moreover, if in addition the function  $G(\cdot, \omega)$  is differentiable at  $x^*$  w.p.1, then  $g(x)$  is differentiable at  $x^*$  and*

$$\nabla g(x^*) = \mathbb{E}\{\nabla_x G(x^*, \omega)\}. \quad (3.8)$$

Let us quickly outline a proof of the above proposition. First, Lipschitz continuity of  $g(x)$  follows directly from (3.6). Next, by the Lebesgue Dominated Convergence Theorem, the limit of the corresponding ratio, defining the directional derivative of  $g(x)$ , can be taken inside the expected value, and hence formula (3.7) follows. Finally, if  $G'_\omega(x^*, h)$  is linear in  $h$  for almost every  $\omega$ , i.e. the function  $G_\omega(\cdot)$  is differentiable at  $x^*$  w.p.1, then (3.7) implies that  $g'(x^*, h)$  is linear in  $h$ , and hence (3.8) follows (see, e.g., [34, p.70] for details). Note that since  $g(x)$  is locally Lipschitz continuous, we only need to verify linearity of  $g'(x^*, \cdot)$  in order to establish (Fréchet) differentiability of  $g(x)$  at  $x^*$ .

The above analysis shows that two basic conditions for interchangeability of the expectation and differentiation operators, i.e. for validity of formula (3.8), are: (i) locally Lipschitz continuity of the random function  $G_\omega(x)$ , and (ii) differentiability of  $G_\omega(x)$ , at the given point  $x^*$ , w.p.1. The required regularity conditions are simplified even further if the function  $G_\omega(\cdot)$  is convex for almost every  $\omega$ . In that case the ratio  $[G_\omega(x^* + th) - G_\omega(x^*)]/t$  is monotonically decreasing to the directional derivative  $G'_\omega(x^*, h)$  as  $t$  is monotonically decreasing to zero. Then by using the Monotone Convergence Theorem, instead of the Lebesgue Dominated Convergence Theorem, it is possible to prove the following result.

**Proposition 3.2** *Suppose that the expected value function  $g(x)$  is well defined in a convex neighborhood  $V$  of a point  $x^*$ , that for almost every  $\omega$  the function  $G_\omega(\cdot) := G(\cdot, \omega)$  is convex on  $V$ , and that  $\mathbb{E}\{G'_\omega(x^*, h)\}$  exists for every direction  $h$ . Then  $g(x)$  is convex on  $V$  and formula (3.7) holds. Moreover,  $g(x)$  is differentiable at  $x^*$  if and only if  $G_\omega(x)$  is differentiable at  $x^*$  w.p.1, in which case formula (3.8) holds.*

Let us observe that in the above convex case differentiability of  $G_\omega(x)$ , at  $x^*$ , w.p.1 is a necessary and sufficient condition for differentiability of the expected value function, and validity of the interchangeability formula (3.8). Necessity of that condition follows from formula (3.7) and since if  $G_\omega(\cdot)$  is convex, then the directional derivative  $G'_\omega(x^*, h)$  is convex in  $h$ . Therefore if  $G_\omega(x)$  is nondifferentiable at  $x^*$  on a set of positive measure, then the directional derivative  $g'(x^*, h)$  is not linear in  $h$ , and hence  $g(x)$  is not differentiable at  $x^*$ .

Suppose that the interchangeability formula (3.8) holds and consider the sample average function  $\hat{g}_N(x)$ , defined in (1.2). We have then

$$\mathbb{E}\{\nabla \hat{g}_N(x^*)\} = N^{-1} \sum_{i=1}^N \mathbb{E}\{\nabla_x G(x^*, \omega^i)\} = N^{-1} \sum_{i=1}^N \nabla_x \mathbb{E} G(x^*, \omega^i) = \nabla g(x^*),$$

i.e.  $\nabla \hat{g}_N(x^*)$  is an unbiased estimator of  $\nabla g(x^*)$ .

**Example 2.2 (continued)** The function  $G(x, w) := |x - w|$  is piecewise linear and differentiable at every  $x$  except at  $x = w$ , with  $\partial G(x, w)/\partial x = \text{sign}(x - w)$ . Therefore  $g(x)$  is differentiable at  $x^*$  iff the cdf  $F(\cdot)$  is continuous at  $x^*$ , in which case

$$\frac{dg(x^*)}{dx} = \int_{-\infty}^{+\infty} \text{sign}(x^* - w) dF(\omega) = 2F(x^*) - 1. \quad (3.9)$$

If  $F(\cdot)$  is discontinuous at  $x^*$ , i.e. the event  $\{W = x^*\}$  has a positive probability, then the right side derivative of  $g(x)$ , at  $x^*$ , is  $2F(x^*) - 1$ , while the corresponding left side derivative is  $2F(x^*_-) - 1$ , where  $F(x^*_-)$  denotes the left side limit of  $F(\cdot)$  at  $x^*$ . The gap between these

right and left side derivatives is given by twice the jump  $F(x^*) - F(x^*_-)$  of the cdf function at  $x = x^*$ , i.e. by the quantity  $2P(W = x^*)$ .

Note that if  $F(\cdot)$  is differentiable at  $x^*$ , then it follows from (3.9) that  $g(x)$  is twice differentiable at  $x^*$ . However, second order derivatives cannot be taken inside the integral in (2.1). This is because  $\partial G(x, w)/\partial x$  is discontinuous at  $x = w$ , and indeed  $\partial^2 G(x, w)/\partial x^2$  equals zero whenever it exists. The above differential behavior is typical for piecewise smooth (differentiable) functions.

**Example 2.3 (continued)** Consider the function

$$G(z) := \inf\{q^T y : Wy = z, y \geq 0\}. \quad (3.10)$$

Clearly the function  $Q(x, h)$ , given as the optimal value of the problem (2.3), can be written as  $Q(x, h) = G(h - Ax)$ . By duality arguments of linear programming we have that

$$G(z) = \sup\{z^T \xi : W^T \xi \leq q\}, \quad (3.11)$$

provided the set  $\{\xi : W^T \xi \leq q\}$  is non empty. So let us suppose, for the sake of simplicity, that this set is non empty and bounded. Then the function  $G(z)$  is a real valued piecewise linear convex function, and its subdifferential is given by the set of optimal solutions of the problem (3.11). It follows that the subdifferential of the function  $Q(\cdot, h)$ , at a point  $x$ , is given by the set of vectors  $-A^T \bar{\xi}$ , where

$$\bar{\xi} \in \arg \max_{W^T \xi \leq q} (h - Ax)^T \xi,$$

and that  $Q(\cdot, h)$  is differentiable at  $x$ , with  $\nabla_x Q(x, h) = -A^T \bar{\xi}$ , iff  $\bar{\xi}$  is unique. Moreover, the SAA function

$$\hat{g}_N(x) := c^T x + N^{-1} \sum_{i=1}^N Q(x, h^i)$$

is a piecewise linear convex function with

$$\partial \hat{g}_N(x) = c + N^{-1} \sum_{i=1}^N \partial_x Q(x, h^i).$$

Suppose now that the random vector  $h$  has a continuous distribution with a pdf  $f(\cdot)$ . Let us fix a point  $x \in \mathbb{R}^m$ . Since the function  $G(z)$  is convex, the set of points where it is not differentiable has Lebesgue measure zero. Since random vector  $h = h(\omega)$  has a density, it follows then that the function  $Q(\cdot, h)$  is differentiable at  $x$  w.p.1. It follows then by proposition 3.2 that the expected value function  $g(x)$  is differentiable at  $x$  and  $\nabla g(x) = c + \mathbb{E}\{\nabla_x Q(x, h)\}$ , and hence  $\nabla \hat{g}_N(x)$  is an unbiased estimator of  $\nabla g(x)$ .

On the other hand if the distribution of  $h$  is discrete, then  $g(x)$  is a convex piecewise linear function. In that case  $g(x)$  cannot be everywhere differentiable, except in a trivial case when it is linear. As we shall see later statistical properties of the SAA estimator  $\hat{x}_N$  are completely different in cases of continuous and discrete distributions of  $h$ .

**Example 2.4 (continued)** Suppose that for every  $v$  the function  $R(\cdot, v)$  is continuously differentiable. In that case  $I_t$  is a piecewise smooth function of  $x_1, \dots, x_t$ , for any given realization of the random data vector  $W$ . That is,  $I_t$  is differentiable except possibly at

such points where  $I_{\tau-1} + R(x_\tau, V_\tau) - D_\tau = 0$  for some  $\tau \leq t$ . It follows that  $I_t$  is differentiable w.p.1 if, for every  $\tau = 1, \dots, t$ , the event " $I_{\tau-1} + R(x_\tau, V_\tau) - D_\tau = 0$ " occurs with probability zero. In turn, probability of such event is zero if the conditional probability of " $I_{\tau-1} + R(x_\tau, V_\tau) - D_\tau = 0$ " given " $I_{\tau-1} = z$ " is zero for all  $z$  in the support of the distribution of  $I_{\tau-1}$ . This happens, for example, if the conditional distribution of  $R(x_\tau, V_\tau) - D_\tau$ , given  $I_{\tau-1}$  (or equivalently given  $V_1, \dots, V_{\tau-1}, D_1, \dots, D_{\tau-1}$ ), is continuous, and in particular if vectors  $V_1, \dots, V_\tau$  are independent of  $D_1, \dots, D_\tau$ , and  $D_1, \dots, D_\tau$  are mutually independent and have continuous distributions. Clearly we have, by formula (2.7), that the profit function  $G(\cdot, W)$  is differentiable w.p.1, at a certain point, if all  $I_t$ ,  $t = 1, \dots, T$ , are differentiable at that point w.p.1.

The condition (3.6) is also not difficult to verify. The corresponding Lipschitz constant, for the function  $I_t$ , is given by the maximum of  $\|\nabla R(x_\tau, V_\tau)\|$  over all  $x_\tau$ ,  $\tau = 1, \dots, t$ , in a neighborhood of the considered point. (Here and afterwards the gradient  $\nabla R(x_\tau, V_\tau)$  is calculated with respect to  $x_\tau$ .) Therefore we obtain that the corresponding expected value function is differentiable and the interchangeability formula (3.8) holds if  $\|\nabla R(\cdot, V_\tau)\|$  are locally bounded by positive variables having finite first order moments, and the conditional distribution of  $R(x_\tau, V_\tau) - D_\tau$ , given  $V_1, \dots, V_{\tau-1}, D_1, \dots, D_{\tau-1}$ , is continuous for all  $\tau = 1, \dots, T$ .

Note that  $I_t$  are convex functions of the decision variables, for any realization of the random vector  $W$ , if the function  $R(\cdot, v)$  is convex for any  $v$ . This follows from (2.5) since the operation of taking maximum preserves convexity of the involved functions. Moreover, if  $R(\cdot, v)$  is linear for all  $v$  and  $\pi_{t+1} - \pi_t - h_t \leq 0$ ,  $t = 1, \dots, T-1$ , then the profit function  $G(\cdot, w)$  is concave for all  $w$ . In that case the expected value of  $G(\cdot, W)$  is also a concave function, and hence an optimization problem of maximization of that expected value function, over a convex region, is a convex programming problem.

Let us calculate now derivatives of  $I_t$  in an explicit form. Denote by  $\tau_1$  the first time the process  $I_t$  hits zero, i.e.  $\tau_1 \geq 1$  is the first time  $I_{\tau_1-1} + R(x_{\tau_1}, V_{\tau_1}) - D_{\tau_1}$  becomes less than or equal to zero, and hence  $I_{\tau_1} = 0$ . Let  $\tau_2 > \tau_1$  be the second time  $I_t$  hits zero, etc. Note that if  $I_{\tau_1+1} = 0$ , then  $\tau_2 = \tau_1 + 1$ , etc. Let  $1 \leq \tau_1 < \dots < \tau_n \leq T$  be the sequence of hitting times. For a given time  $t \in \{1, \dots, T\}$ , let  $\tau_{i-1} \leq t < \tau_i$ . Suppose that the events  $I_{\tau-1} + R(x_\tau, V_\tau) - D_\tau = 0$ ,  $\tau = 1, \dots, T$ , occur with probability zero. We have then that, for almost every  $W$ , the gradient of  $I_s$  with respect to the components of vector  $x_t$  can be written as follows

$$\nabla_{x_t} I_s = \begin{cases} \nabla R(x_t, V_t), & \text{if } t \leq s < \tau_i \text{ and } t \neq \tau_{i-1}, \\ 0, & \text{otherwise.} \end{cases} \quad (3.12)$$

## 4 Likelihood Ratios Method

In many applications the objective function is piecewise smooth and its first order derivatives are discontinuous. Usually in such cases the corresponding second order derivatives cannot be taken inside the expected value. This happens, for instance, in the case of the objective functions considered in examples 2.2, 2.3 and 2.4. In such situations one cannot estimate second order derivatives of the expected value function by using the corresponding second order derivatives of the sample average function. In this section we briefly discuss the

Likelihood Ratios (LR) method which provides an alternative way of estimating derivatives of the expected value function. For a thorough discussion of that method we refer to [34].

Suppose that the expected value function can be represented in the form

$$g(x) = \mathbb{E}_{P_x}\{H(W)\}, \quad (4.1)$$

where the probability distribution  $P_x$ , of random vector  $W$ , is a function of vector  $x$  of decision variables, while  $H(\cdot)$  does not depend on  $x$ . In some cases such a representation either comes naturally or can be achieved by a suitable transformation. Suppose further that  $P_x$  has a probability density function (pdf)  $f(\cdot, x)$ , depending on vector  $x$ , and hence

$$g(x) = \int H(w)f(w, x)dw. \quad (4.2)$$

Let  $\psi(\cdot)$  be another probability density function such that the ratio

$$L(w, x) := \frac{f(w, x)}{\psi(w)} \quad (4.3)$$

is well defined. That is, if  $\psi(w) = 0$ , then  $f(w, x) = 0$  and by the definition  $\frac{0}{0} = 0$ , i.e. we do not divide a positive number by zero. We can write then

$$g(x) = \int H(w)L(w, x)\psi(w)dw = \mathbb{E}_\psi\{H(W)L(W, x)\}, \quad (4.4)$$

where  $\mathbb{E}_\psi$  denotes the expectation with respect to the pdf  $\psi$ . In the above representation (4.4) the probability distribution in the expectation operator is fixed (by the pdf  $\psi$ ) and only the objective function  $H(w)L(w, x)$  depends on  $x$ . The function  $L(w, x)$  is called the *likelihood ratio* (LR) function.

Usually the pdf  $f(w, x)$  is given in a closed form and is a smooth function of  $x$ . In that case  $L(w, x)$  is also a smooth function of  $x$ , and hence first and higher order derivatives can be taken inside the expected value under suitable regularity conditions. For instance, if  $L(w, \cdot)$  is  $s$ -times continuously differentiable and  $\|H(w)\nabla^k L(w, x)\| \leq K(w)$ ,  $k = 1, \dots, s$ , for all  $x$  in a neighborhood of  $x^*$  and such that  $\mathbb{E}_\psi\{K(W)\}$  is finite, then

$$\nabla^s g(x^*) = \mathbb{E}_\psi\{H(W)\nabla^s L(W, x^*)\}, \quad (4.5)$$

where all derivatives are taken with respect to  $x$  (this follows by the Lebesgue Dominated Convergence theorem).

Let  $W^1, \dots, W^N$  be an i.i.d. random sample with the common pdf  $\psi$ . Note that the pdf  $\psi$ , and hence this random sample, do not depend on  $x$ . It follows then by (4.4) that

$$\tilde{g}_N(x) := N^{-1} \sum_{i=1}^N H(W^i)L(W^i, x) \quad (4.6)$$

is an unbiased estimator of  $g(x)$ . We also have that if the interchangeability formula (4.5) holds, then

$$\nabla^k \tilde{g}_N(x) = N^{-1} \sum_{i=1}^N H(W^i)\nabla^k L(W^i, x) \quad (4.7)$$

is an unbiased estimator of  $\nabla^k g(x)$ . Similar derivations can be performed in a case of discrete distributions by replacing integrals with the corresponding sums.

Let us observe that although the mean (expected value) of  $\tilde{g}_N(x)$  is equal to  $g(x)$ , and hence is independent of  $\psi$ , its variance depends on a choice of the pdf  $\psi$  as well as on  $x$ . Therefore one may try to choose  $\psi$  in such a way as to minimize the variance of  $\tilde{g}_N(x)$ . This is related to a variance reduction technique known as *importance sampling*. It should be mentioned, however, that such an “optimal” choice of  $\psi$  is associated with a particular (considered) point  $x$ . In fact the LR (importance sampling) techniques typically are unstable, and an optimal (or rather reasonable) choice of  $\psi$  with respect to one value of  $x$  can produce a huge variance for a different value of  $x$ .

In order to improve accuracy of Monte Carlo simulation based estimators it is possible to employ various variance reduction methods. In particular, control variables techniques can be quite efficient when applied to the LR estimators  $\nabla^k \tilde{g}_N(x)$ . Note that  $\mathbb{E}_\psi\{L(W, x)\} = 1$  for any  $x$ , and hence  $\mathbb{E}_\psi\{\nabla L(W, x)\} = 0$ . It follows that for any  $\alpha \in \mathbb{R}$ ,

$$\nabla g(x) = \mathbb{E}_\psi\{[H(W) - \alpha]\nabla L(W, x)\}. \quad (4.8)$$

One can then choose  $\alpha$  such as to minimize, say, the sum of variances of the estimated components of  $\nabla g(x)$ . Similar derivations can be applied to higher order derivatives as well (see [34] for details). This method produces a considerable variance reduction if the main variability of the LR estimators come from the derivatives of the LR function, and hence  $H(W)\nabla L(W, x)$  and  $\nabla L(W, x)$  are highly correlated.

**Example 2.3 (continued)** Suppose that the random vector  $h$  has a pdf  $f(\cdot)$ . Since  $Q(x, h) = G(h - Ax)$ , where  $G(\cdot)$  is defined in (3.10), by using the transformation  $z = h - Ax$  we obtain

$$\mathbb{E}_f Q(x, h) = \int G(h - Ax)f(h)dh = \int G(z)f(z + Ax)dz = \mathbb{E}_\psi\{G(Z)L(Z, x)\}.$$

Here  $\psi$  is a chosen pdf,  $Z$  is a random vector having pdf  $\psi$ , and  $L(z, x) := f(z + Ax)/\psi(z)$  is the corresponding LR function. For a discussion of the above LR transformation and a numerical example see [40].

**Example 2.4 (continued)** Consider random variables  $Z_t := D_t - R(x_t, V_t)$ ,  $t = 1, \dots, T$ . We have that  $I_t = [I_{t-1} - Z_t]^+$ , and hence we can consider  $I_t$  as a function of the random vector  $\underline{Z}_t := (Z_1, \dots, Z_t)$ . Clearly the distribution of  $Z_t$  depends on  $x_t$  and hence the distribution of  $\underline{Z}_t$  depends on vector  $\underline{x}_t := (x_1, \dots, x_t)$ . Suppose that  $\underline{Z}_t$  has pdf  $\underline{f}_t(\underline{z}_t, \underline{x}_t)$ . We can write then

$$E_{\underline{f}_t}\{I_t\} = \int I_t(\underline{z}_t)\underline{f}_t(\underline{z}_t, \underline{x}_t)d\underline{z}_t = \int I_t(\underline{z}_t)L_t(\underline{z}_t, \underline{x}_t)\underline{\psi}_t(\underline{z}_t)d\underline{z}_t = E_{\underline{\psi}_t}\{I_t L_t(\underline{Z}_t, \underline{x}_t)\},$$

where  $\underline{\psi}_t$  is a chosen pdf and

$$L_t(\underline{z}_t, \underline{x}_t) := \frac{\underline{f}_t(\underline{z}_t, \underline{x}_t)}{\underline{\psi}_t(\underline{z}_t)}$$

is the corresponding LR function. If  $\underline{f}_t(\underline{z}_t, \cdot)$  is sufficiently smooth, it follows that

$$\nabla^k E_{\underline{f}_t}\{I_t\} = E_{\underline{\psi}_t}\{I_t \nabla^k L_t(\underline{Z}_t, \underline{x}_t)\}. \quad (4.9)$$

Of course in order to apply the above formulas we need to know the pdf  $f_t(\underline{z}_t, \underline{x}_t)$  in a closed form. Suppose, for instance, that  $R(x_t, v_t) := x_t$  and that each random variable  $D_t$  has pdf  $f_t(\cdot)$ ,  $t = 1, \dots, T$ , and that these variables are independent. Then  $Z_t = D_t - x_t$  has pdf  $f_t(z_t + x_t)$  and hence

$$\underline{f}_t(\underline{z}_t, \underline{x}_t) = \prod_{i=1}^t f_i(z_i + x_i).$$

Suppose further that  $D_t \sim N(\mu_t, \sigma_t^2)$  are normally distributed, i.e.

$$f_t(z_t) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp \left\{ \frac{-(z_t - \mu_t)^2}{2\sigma_t^2} \right\},$$

and take  $\underline{\psi}_t(\underline{z}_t) = \underline{f}_t(\underline{z}_t, \underline{x}_t^*)$  for some given (fixed)  $\underline{x}_t^*$ . Then, for  $r, s \leq t$ ,

$$\frac{\partial^2 L_t(\underline{Z}_t, \underline{x}_t^*)}{\partial x_r \partial x_s} = \begin{cases} (Z_r - \mu_r + x_r^*)(Z_s - \mu_s + x_s^*) / (\sigma_r^2 \sigma_s^2), & \text{if } r \neq s \\ -1/\sigma_r^2 + (Z_r - \mu_r + x_r^*)^2 / \sigma_r^4, & \text{if } r = s. \end{cases}$$

By using the corresponding sample averages, it is straightforward to construct unbiased estimates of second order derivatives of  $\mathbb{E}\{I_t\}$ , and hence of the expected value of the profit function.

## 5 Statistical Inference

In this section we discuss statistical properties of SAA estimators of the optimal value  $v^*$  and the set of optimal solutions, denoted  $\mathcal{S}^*$ , of the true problem (1.1). Unless stated otherwise,  $\hat{v}_N$  and  $\hat{x}_N$  denote the optimal value and an optimal solution, respectively, of the SAA problem (1.2).

### 5.1 Consistency

Suppose that, for any given  $x \in S$ , the sample averages  $\hat{g}_N(x)$  converge to their expected value  $g(x)$  w.p.1 as  $N \rightarrow \infty$ . Usually such convergence is ensured by the Law of Large Numbers (LLN). If the considered (generated) sample  $\omega^i$ ,  $i = 1, \dots$ , is i.i.d., then this is guaranteed by the classical LLN. There are also many extensions of the LLN for dependent sequences, e.g. for regenerative processes etc. It is natural to expect that such pointwise convergence of  $\hat{g}_N(x)$  to  $g(x)$  would imply convergence of  $\hat{x}_N$  to  $\mathcal{S}^*$ , and in particular to  $x^*$  if  $\mathcal{S}^* = \{x^*\}$  is a singleton. Unfortunately this is not true in general and some additional technical conditions are required. Nevertheless, usually the required additional conditions are mild and indeed in case the LLN can be applied pointwise (i.e. for any fixed  $x \in S$ ) one may expect convergence of the corresponding SAA estimators.

It is said that the *uniform* LLN holds, on the set  $S$ , if

$$\lim_{N \rightarrow \infty} \sup_{x \in S} |\hat{g}_N(x) - g(x)| = 0, \quad \text{w.p.1.} \quad (5.1)$$

Clearly we have that

$$|\hat{v}_N - v^*| \leq \sup_{x \in S} |\hat{g}_N(x) - g(x)|.$$

Therefore the uniform LLN implies that  $\hat{v}_N \rightarrow v^*$  w.p.1 as  $N \rightarrow \infty$ . It is also not difficult to show convergence of the corresponding SSA estimators  $\hat{x}_N$ .

Starting with a pioneering work of Wald [43], such consistency properties of the estimator  $\hat{x}_N$  were studied in numerous publications. In the context of stochastic programming, consistency of  $\hat{x}_N$  was investigated by tools of epi-convergence analysis in [11] and [31], for example. Following is a relatively simple consistency result which is already sufficient for many practical applications (e.g., [34]).

**Theorem 5.1** *Suppose that the uniform LLN (5.1) holds. Then  $\hat{v}_N \rightarrow v^*$  w.p.1 as  $N \rightarrow \infty$ . If, moreover, the set  $S$  is compact and the expected value function  $g(x)$  is continuous, then  $\text{dist}(\hat{x}_N, \mathcal{S}^*) \rightarrow 0$  w.p.1 as  $N \rightarrow \infty$ .*

Of course, if  $\mathcal{S}^* = \{x^*\}$  is a singleton, then the convergence  $\text{dist}(\hat{x}_N, \mathcal{S}^*) \rightarrow 0$  means that  $\hat{x}_N \rightarrow x^*$ . Note also that if  $g(x)$  is continuous and  $S$  is compact, then the set  $\mathcal{S}^*$ , of optimal solutions of true problem (1.1), is non empty and compact.

It is possible to give various conditions ensuring the uniform LLN. This is in particular simple in the convex case. Let us recall the following result from convex analysis. If a sequence of real valued convex functions converges pointwise on a dense subset of  $\mathbb{R}^m$ , then it converges *uniformly* on any compact subset of  $\mathbb{R}^m$  (e.g., [32, Theorem 10.8]). By taking a countable dense subset of  $\mathbb{R}^m$  we obtain that if the functions  $\hat{g}_N(\cdot)$  are convex and the LLN holds pointwise, then it holds uniformly on any compact subset of  $\mathbb{R}^m$ . It also follows that the function  $g(\cdot)$  is convex, and hence is continuous.

In the convex case it is also relatively easy to deal with situations where the feasible set  $S$  is not necessarily compact. That is, suppose that the SAA problem is convex, i.e. the set  $S$  and the function  $\hat{g}_N(\cdot)$  are convex, and that the set  $\mathcal{S}^*$  is non empty and bounded. For some  $\varepsilon > 0$  consider the neighborhood  $V := \{x : \text{dist}(x, \mathcal{S}^*) \leq \varepsilon\}$  of  $\mathcal{S}^*$ . Since  $\mathcal{S}^*$  is bounded it follows that  $V$  is compact. Consequently by the above discussion we have that restricted to the set  $S' := S \cap V$  a corresponding SAA estimator  $\tilde{x}_N$  converges to  $\mathcal{S}^*$  and hence  $\text{dist}(\tilde{x}_N, \mathcal{S}^*) < \varepsilon$  w.p.1 for  $N$  large enough. Since in the convex case a local minimizer is also a global minimizer, it follows that  $\tilde{x}_N = \hat{x}_N$  w.p.1 for  $N$  large enough. Therefore, in the convex case, we have that if the set  $\mathcal{S}^*$  is non empty and bounded and the LLN holds for every  $x$  in a neighborhood of  $\mathcal{S}^*$ , then  $\text{dist}(\hat{x}_N, \mathcal{S}^*) \rightarrow 0$  w.p.1 as  $N \rightarrow \infty$ .

For not necessarily convex problems consider the following conditions: (i) the set  $S$  is compact, (ii) for  $P$ -almost every  $\omega$  the function  $G(\cdot, \omega)$  is continuous on  $S$ , and (iii) the family  $G(x, \omega)$ ,  $x \in S$ , is dominated by a  $P$ -integrable function  $K(\omega)$ , i.e.  $\mathbb{E}_P K$  is finite and  $|G(x, \omega)| \leq K(\omega)$  for all  $x \in S$  and  $P$ -almost every  $\omega$ . It is possible to show that in the i.i.d. case, the above conditions ensure continuity of  $g(x)$  on  $S$  and the uniform LLN (5.1) (e.g., [34, pp. 67-68]).

In nonconvex situations there is an additional problem of local optima. That is, the SAA problem can be trapped in a locally optimal solution. Suppose, for example, that  $g(x) = x^2$ ,  $x \in \mathbb{R}$ , while  $\hat{g}_N(x) = x^2 + N^{-1} \sin(Nx^2)$ . We have then that  $|\hat{g}_N(x) - g(x)| \leq N^{-1}$ , for all  $x \in \mathbb{R}$ , and the global minimizer of  $\hat{g}_N(x)$ , over  $\mathbb{R}$ , converges to zero. On the other hand  $\hat{g}_N(x)$  has an infinite number of local minimizers which become dense on  $\mathbb{R}$  as  $N$  tends to infinity. The above example, of course, is pathological and such situations cannot happen if the (uniform) LLN holds for derivatives of the sample average functions and the feasible

set  $S$  is “sufficiently regular”. Let us remark that this problem can be non trivial and its thorough discussion will lead outside the scope of this paper.

## 5.2 Asymptotic Analysis of the Optimal Value

Consider the optimal values  $\hat{v}_N$  and  $v^*$  of the SAA and the true problems, respectively. We have that  $\min_{x \in S} \hat{g}_N(x) \leq \hat{g}_N(x)$  for any  $x \in S$ , and hence

$$\mathbb{E}\{\hat{v}_N\} = \mathbb{E}\left\{\min_{x \in S} \hat{g}_N(x)\right\} \leq \min_{x \in S} \mathbb{E}\{\hat{g}_N(x)\} = \min_{x \in S} \mathbb{E}\{g(x)\} = v^*. \quad (5.2)$$

Moreover, if the set  $\mathcal{S}^*$  of optimal solutions of the true problem is non empty, then the equality in (5.2) holds iff for almost every realization of the random sample there exists  $x^* \in \mathcal{S}^*$  such that  $\hat{g}_N(x) \geq \hat{g}_N(x^*)$  for all  $x \in S$ . This, of course, is unlikely to happen. Therefore typically the bias  $\mathbb{E}\{\hat{v}_N\} - v^*$  of the SAA estimator of the optimal value is negative.

In order to have a better understanding of the above bias problem let us consider the following construction. Suppose that for almost every  $\omega$  the function  $G(\cdot, \omega)$  is continuous and that the feasible set  $S$  is compact. Then the SAA function  $\hat{g}_N(x)$  is continuous on  $S$  and hence can be viewed as a point in the Banach space  $C(S)$ . (Recall that  $C(S)$  denotes the linear space of continuous functions  $\psi : S \rightarrow \mathbb{R}$  equipped with the sup-norm  $\|\psi\| := \sup_{x \in S} |\psi(x)|$ .) Moreover, under mild measurability conditions,  $\hat{g}_N(x)$  can be considered as a random element in the space  $C(S)$  equipped with its Borel sigma algebra. Suppose further that a functional Central Limit Theorem (CLT) holds for  $\hat{g}_N$ . That is, the random elements  $N^{1/2}(\hat{g}_N - g)$  converge in distribution to a random element  $Y$  of  $C(S)$ . In the i.i.d. case such functional CLT can be ensured by the following conditions (e.g., [2]):

- (A1) For every  $x \in S$ , the function  $G(x, \cdot)$  is measurable.
- (A2) For some point  $\bar{x} \in S$  the expectation  $\mathbb{E}_P\{G(\bar{x}, \omega)^2\}$  is finite.
- (A3) The Lipschitz continuity condition (3.6) holds for all  $x_1, x_2 \in S$  and almost every  $\omega$ , and the random variable  $K(\omega)$  has a finite second order moment.

Note that in such i.i.d. case, for any points  $x_1, \dots, x_k \in S$ , the random vector  $(Y(x_1), \dots, Y(x_k))$  has a multivariate normal distribution with covariance matrix given by the covariance matrix of the random vector  $(G(x_1, \omega), \dots, G(x_k, \omega))$ . In particular, for any  $x \in S$ , it follows that

$$N^{1/2}[\hat{g}_N(x) - g(x)] \Rightarrow N(0, \sigma^2(x)),$$

where  $\sigma^2(x) := \text{var}\{G(x, \omega)\}$  and “ $\Rightarrow$ ” denotes convergence in distribution. We have the following result (which can be proved by employing a Generalized Delta Theorem) [37].

**Theorem 5.2** *Let  $\{\hat{g}_N\}$  be a sequence of random elements in  $C(S)$ , and  $g \in C(S)$ . Suppose that  $N^{1/2}(\hat{g}_N - g)$  converges in distribution to a random element  $Y$  of  $C(S)$ . Then*

$$N^{1/2}(\hat{v}_N - v^*) \Rightarrow \min_{x \in \mathcal{S}^*} Y(x). \quad (5.3)$$

It follows that if the set  $\mathcal{S}^*$ , of optimal solutions of the true problem, is a singleton  $\mathcal{S}^* = \{x^*\}$ , and the above conditions (A1)-(A3) hold in the i.i.d. case, then

$$N^{1/2}(\hat{v}_N - v^*) \Rightarrow N(0, \sigma^2(x^*)). \quad (5.4)$$

In general convergence in distribution does not imply convergence of the corresponding expected values. Therefore we need an additional condition in order to conclude that (5.4) implies convergence to zero of the expected values of the random variables  $V_N := N^{1/2}(\hat{v}_N - v^*)$ . Such implication is ensured if we assume that the random variables  $V_N$  are uniformly integrable, which in turn follows, for example, from the condition that the second order moments of  $V_N$  are bounded (by a constant independent of  $N$ ) (e.g., [30]). Therefore, if the second order moments of  $V_N$  are bounded, then it follows from (5.4) that  $\mathbb{E}\{V_N\} \rightarrow 0$ , and hence  $\mathbb{E}\{\hat{v}_N\} - v^* = o(N^{-1/2})$ .

On the other hand if the set  $\mathcal{S}^*$  is not a singleton, then the random variable  $V := \min_{x \in \mathcal{S}^*} Y(x)$  has a distribution which is given by the distribution of a minimum of a number of (correlated) normally distributed random variables with zero means. In that case the expected value of  $V$  is negative. Assuming that the second order moments of  $V_N$  are bounded, we obtain from (5.3) that  $\mathbb{E}\{V_N\} \rightarrow \mathbb{E}\{V\}$ , and hence that the bias of  $\hat{v}_N$  is of order  $O(N^{-1/2})$ . This indicates that the (negative) bias  $\mathbb{E}\{\hat{v}_N\} - v^*$  tends to be bigger if the true problem has a large set of optimal or “almost optimal” solutions.

The asymptotic result (5.3) is based on a first order approximation of the optimal value function. Clearly it does not distinguish between feasible sets which produce the same set of optimal solutions. That is,

$$\hat{v}_N = \min_{x \in \mathcal{S}^*} \hat{g}_N(x) + o_p(N^{-1/2}), \quad (5.5)$$

i.e. the first order asymptotics of  $\hat{v}_N$  is the same if the feasible set  $S$  in (1.1) is replaced by the (smaller) set  $\mathcal{S}^*$ . In particular, if  $\mathcal{S}^* = \{x^*\}$  is a singleton, then  $\hat{v}_N = \hat{g}_N(x^*) + o_p(N^{-1/2})$ .

**Example 5.1** Consider the framework of the maximum likelihood example 2.1. Let  $\Theta_0$  and  $\Theta_1$  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

$$\ell_N := 2 \left[ \inf_{\theta \in \Theta_0} \sum_{i=1}^N \ln f(Y^i, \theta) - \inf_{\theta \in \Theta_1} \sum_{i=1}^N \ln f(Y^i, \theta) \right] \quad (5.6)$$

be the corresponding log-likelihood ratio test statistic. Suppose that

$$g(\theta) := -\mathbb{E}_P\{\ln f(Y, \theta)\} \quad (5.7)$$

has unique minimizers  $\theta_0$  and  $\theta_1$  over the sets  $\Theta_0$  and  $\Theta_1$ , respectively. Recall that if the distribution  $P$ , of the random sample, is given by a pdf  $f(y, \theta^*)$ , then  $\theta^*$  is an unconstrained minimizer of  $g(\theta)$ . Moreover, if the parameter vector  $\theta$  is identified at  $\theta^*$ , then  $\theta^*$  is such unique minimizer. We have by (5.5) that

$$N^{-1/2}\ell_N = 2N^{-1/2} \sum_{i=1}^N \left[ \ln f(Y^i, \theta_0) - \ln f(Y^i, \theta_1) \right] + o_p(1), \quad (5.8)$$

provided 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  $\ell_0$  and  $\sigma^2$  are the mean and the variance, respectively, of the random variable  $Z := 2 \ln \left[ \frac{f(Y, \theta_0)}{f(Y, \theta_1)} \right]$ .

Note that if  $\theta_0 = \theta_1$ , then this variable  $Z$  degenerates into  $Z \equiv 0$ . This happens if the distribution  $P$  is given by pdf  $f(y, \theta^*)$  with  $\theta^* \in \Theta_0 \cap \Theta_1$ . That is, if  $\Theta_0$  is a subset of  $\Theta_1$  and the model is correctly specified, then the asymptotic distribution of  $N^{-1/2}\ell_N$ , under  $H_0$ , degenerates into identical zero. Therefore in the likelihood ratio testing procedures a second order expansion of the optimal value function is needed in order to obtain useful asymptotics of  $\ell_N$ . However, in stochastic programming applications the asymptotic result (5.4) 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  $v^*$  (see [40]).

### 5.3 Second Order Expansion of the Optimal Value

As we have seen in the previous section, first order asymptotics of  $\hat{v}_N$  do not involve local structure of the feasible set  $S$ . In this section we discuss asymptotics of  $\hat{v}_N$  based on a second order expansion. In order to proceed with the second order analysis we need the following technical details.

Suppose that the function  $g(x)$  is twice continuously differentiable and that the true problem has unique optimal solution  $x^*$ . Then the following first order necessary conditions hold at the point  $x^*$ :

$$w^T \nabla g(x^*) \geq 0, \quad \text{for all } w \in T_S(x^*). \quad (5.9)$$

Here  $T_S(x)$  denotes the contingent (Bouligand) cone to the set  $S$  at  $x \in S$ , that is

$$T_S(x) := \{d \in \mathbb{R}^m : \exists t_n \downarrow 0 \text{ such that } \text{dist}(x + t_n d, S) = o(t_n)\}. \quad (5.10)$$

Note that if we replace in the above definition the condition: “there exists a sequence  $t_n \downarrow 0$ ”, by the condition: “for any sequence  $t_n \downarrow 0$ ”, we obtain the following cone

$$\overline{T}_S(x) := \{d \in \mathbb{R}^m : \text{dist}(x + td, S) = o(t), t \geq 0\}. \quad (5.11)$$

The cone  $\overline{T}_S(x)$  is known under various names. Clearly  $\overline{T}_S(x) \subset T_S(x)$ , and it can happen that  $\overline{T}_S(x)$  is strictly included in  $T_S(x)$ .

It is said that the second order growth condition holds, at  $x^*$ , if there exist a constant  $c > 0$  and a neighborhood  $U \subset \mathbb{R}^m$  of  $x^*$  such that

$$g(x) \geq g(x^*) + c\|x - x^*\|^2, \quad \text{for all } x \in S \cap U. \quad (5.12)$$

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

$$C(x^*) := \{w \in T_S(x^*) : w^T \nabla g(x^*) = 0\} \quad (5.13)$$

is called the *critical cone* of the problem (1.1). It represents those directions for which first order conditions (5.9) do not provide information about optimality of  $x^*$ . Note that

if  $\nabla g(x^*) = 0$ , then  $C(x^*) = T_S(x^*)$ . If the distribution  $P$ , in the maximum likelihood example 2.1, is given by a pdf  $f(y, \theta^*)$ ,  $\theta^* \in \Theta$ , then  $\theta^*$  is an unconstrained minimizer of  $g(\theta)$  and hence  $\nabla g(\theta^*) = 0$ . Therefore in that case the critical and contingent cones to the parameter set  $\Theta$  coincide at the point  $\theta^*$ .

It turns out that second order optimality conditions, as well as second order expansions of the optimal value, 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

$$T_S^2(x, d) := \left\{ w \in \mathbb{R}^m : \text{dist} \left( x + td + \frac{1}{2}t^2w, S \right) = o(t^2), t \geq 0 \right\} \quad (5.14)$$

is called the second order tangent set, to the set  $S$  at the point  $x$  in the direction  $d$ . Note that  $T_S^2(x, d)$  can be non empty only if  $x \in S$  and  $d \in \overline{T}_S(x)$ . Yet even if  $S$  is convex and  $x \in S$  and  $d \in \overline{T}_S(x)$ , it can happen that the corresponding second order tangent set is empty. Note also that for  $d = 0$  the second order tangent set  $T_S^2(x, 0)$  coincides with the cone  $\overline{T}_S(x)$ .

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

$$\lim_{n \rightarrow \infty} \text{dist} \left( w_n, T_S^2(x, d) \right) = 0. \quad (5.15)$$

If  $w_n \rightarrow w$ , then  $w \in T_S^2(x, d)$  by the definition of second order tangent sets, and hence (5.15) 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  $T_S^2(x, d)$  provides a ‘‘sufficiently tight’’ second order approximation of the set  $S$  at the point  $x$  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 [6]. Note that the second order regularity condition implies that the set  $T_S^2(x, d)$  is non empty, and that the contingent cone  $T_S(x)$  coincides with the cone  $\overline{T}_S(x)$ .

Under the second order regularity condition, at the point  $x^*$ , the following second order optimality conditions are *necessary* and *sufficient* for the second order growth condition (5.12) to hold ([6]):

$$d^T \nabla^2 g(x^*) d + \inf_{w \in T_S^2(x^*, d)} w^T \nabla g(x^*) > 0, \quad \text{for all } d \in C(x^*) \setminus \{0\}. \quad (5.16)$$

Apart from the quadratic term, corresponding to the second order Taylor expansion of the function  $g$ , an additional term, associated with the second order tangent set  $T_S^2(x^*, d)$ , appears in the left hand side of (5.16). This term vanishes if  $\nabla g(x^*) = 0$ . That is what happens in the maximum likelihood example 2.1.

**Example 5.2** Suppose that the set  $S$  is defined by equality and inequality constraints

$$S := \{x : h_i(x) = 0, i = 1, \dots, q; h_i(x) \leq 0, i = q + 1, \dots, p\}, \quad (5.17)$$

with the constraint functions  $h_i$ ,  $i = 1, \dots, p$ , being twice continuously differentiable. Let

$$L(x, \lambda) := g(x) + \sum_{i=1}^p \lambda_i h_i(x) \quad (5.18)$$

be the Lagrangian function of the true problem and

$$\mathcal{I}(x^*) := \{i : h_i(x^*) = 0, i = q + 1, \dots, p\} \quad (5.19)$$

be the set of active at  $x^*$  inequality constraints. Suppose that the following, Mangasarian-Fromovitz [27], constraint qualification holds, at the point  $x^*$ :

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

Then  $T_S(x^*) = \overline{T}_S(x^*)$  and

$$T_S(x^*) = \left\{ d \in \mathbb{R}^m : d^T \nabla h_i(x^*) = 0, i = 1, \dots, q; d^T \nabla h_i(x^*) \leq 0, i \in \mathcal{I}(x^*) \right\},$$

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

$$\nabla_x L(x^*, \lambda) = 0, \quad \lambda_i \geq 0, \quad \lambda_i h_i(x^*) = 0, \quad i = q + 1, \dots, p. \quad (5.20)$$

Under the Mangasarian-Fromovitz (MF) constraint qualification, the set  $\Lambda(x^*)$  of all Lagrange multipliers vectors  $\lambda$ , satisfying the above conditions (5.20), is non empty and bounded, and for any  $\lambda \in \Lambda(x^*)$  the critical cone can be written as

$$C(x^*) = \left\{ d : d^T \nabla h_i(x^*) = 0, i \in \{1, \dots, q\} \cup \mathcal{I}_+(\lambda), d^T \nabla h_i(x^*) \leq 0, i \in \mathcal{I}_0(\lambda) \right\}, \quad (5.21)$$

where

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

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

$$T_S^2(x^*, d) = \left\{ w \in \mathbb{R}^m : \begin{array}{l} w^T \nabla h_i(x^*) + d^T \nabla^2 h_i(x^*) d = 0, \quad i = 1, \dots, q, \\ w^T \nabla h_i(x^*) + d^T \nabla^2 h_i(x^*) d \leq 0, \quad i \in \mathcal{I}_1(x^*, d) \end{array} \right\}, \quad (5.22)$$

where

$$\mathcal{I}_1(x^*, d) := \left\{ i \in \mathcal{I}(x^*) : d^T \nabla h_i(x^*) = 0 \right\}. \quad (5.23)$$

It follows then by duality arguments that, under the MF-constraint qualification, the second order conditions (5.16) can be written in the following equivalent form:

$$\sup_{\lambda \in \Lambda(x^*)} d^T \nabla_{xx}^2 L(x^*, \lambda) d > 0, \quad \text{for all } d \in C(x^*) \setminus \{0\}. \quad (5.24)$$

We are prepared now to discuss second order expansions of the optimal value. We assume that the set  $S$  is compact and work in the Banach space  $C^1(S)$  of real valued continuously differentiable functions  $\psi(x)$ , defined on a neighborhood of the set  $S$ , and equipped with the norm

$$\|\psi\| := \sup_{x \in S} |\psi(x)| + \sup_{x \in S} \|\nabla \psi(x)\|.$$

The following result is obtained by employing a Generalized Delta Theorem together with a formula for a second order expansion of the optimal value function [42] (the corresponding second order expansions of the optimal value function are discussed in [6]).

**Theorem 5.3** *Let  $\{\hat{g}_N\}$  be a sequence of random elements in  $C^1(S)$ , and  $g \in C^1(S)$ . Suppose that: (i)  $N^{1/2}(\hat{g}_N - g)$  converges in distribution to a random element  $Y$  of  $C^1(S)$ , (ii) the true problem has a unique optimal solution  $x^*$ , (iii) the function  $g$  is twice continuously differentiable in a neighborhood of the point  $x^*$ , (iv) the second order growth condition (5.12) holds, (v) the set  $S$  is second order regular at  $x^*$ . Then*

$$\hat{v}_N = \hat{g}_N(x^*) + \frac{1}{2}N^{-1}\varphi(Z_N) + o_p(N^{-1}) \quad (5.25)$$

and

$$N[\hat{v}_N - \hat{g}_N(x^*)] \Rightarrow \frac{1}{2}\varphi(Z), \quad (5.26)$$

where  $Z := \nabla Y(x^*)$ ,  $Z_N := N^{1/2}[\nabla \hat{g}_N(x^*) - \nabla g(x^*)]$  and the function  $\varphi : \mathbb{R}^m \rightarrow \mathbb{R}$  is defined as follows

$$\varphi(\zeta) := \inf_{d \in C(x^*)} \left\{ 2d^T \zeta + d^T \nabla^2 g(x^*) d + \inf_{w \in T_S^2(x^*, d)} w^T \nabla g(x^*) \right\}. \quad (5.27)$$

Note that, because of the second order conditions (5.16), the function  $\varphi(\cdot)$ , defined in (5.27), is finite valued, continuous,  $\varphi(0) = 0$  and  $\varphi(\cdot)$  is positively homogeneous of degree 2, i.e.  $\varphi(t\zeta) = t^2\varphi(\zeta)$  for any  $t \geq 0$  and  $\zeta \in \mathbb{R}^m$ .

If the set  $S$  is defined by smooth constraints, as in (5.17), and the MF-constraint qualification holds, then the set  $S$  is second order regular at  $x^*$  and the second order growth condition (5.12) is equivalent to the second order optimality conditions (5.24). Recall also that, under the MF-constraint qualification, the set  $\Lambda(x^*)$  of Lagrange multipliers is non empty and bounded. Moreover, it is possible to show, by using formula (5.22) and duality arguments, that the second order expansion (5.27) can be written then in the following equivalent form

$$\varphi(\zeta) = \inf_{d \in C(x^*)} \left\{ 2d^T \zeta + \sup_{\lambda \in \Lambda(x^*)} d^T \nabla_{xx}^2 L(x^*, \lambda) d \right\}. \quad (5.28)$$

In that form formula (5.25) was derived in Shapiro [35] by a different method.

Clearly, if  $\nabla g(x^*) = 0$ , then the last term in the right hand side of (5.27) 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 such cases  $\nabla_{xx}^2 L(x^*, \lambda) = \nabla^2 g(x^*)$  for any  $\lambda \in \Lambda(x^*)$  and

$$\varphi(\zeta) = \inf_{d \in C(x^*)} \left\{ 2d^T \zeta + d^T \nabla_{xx}^2 g(x^*) d \right\}. \quad (5.29)$$

In general this term is related, through the second order tangent set  $T_S^2(x^*, d)$ , to the curvature of the set  $S$  at the point  $x^*$ .

In order to ensure that  $\hat{g}_N$  are random elements of the space  $C^1(S)$  we need to assume that the functions  $\hat{g}_N(\cdot)$  are continuously differentiable on  $S$  with probability one. This rules out many interesting applications where, in fact, the approximating functions are not everywhere differentiable. That is the case, for instance, in examples 2.2, 2.3 and 2.4. Nevertheless, formulas (5.25), (5.26) and (5.27) make sense if the expected value function  $g(\cdot)$  is twice differentiable at  $x^*$  and the SAA functions  $\hat{g}_N(\cdot)$  are differentiable at  $x^*$  with probability one. As it was discussed earlier this often happens if the underline probability

distribution is continuous. In such cases formulas (5.25) and (5.26) give correct asymptotics which can be proved by different methods (cf. [17],[38]).

Note that in the i.i.d. case it follows by the multivariate CLT that the random vectors  $N^{1/2} [\nabla \hat{g}_N(x^*) - \nabla g(x^*)]$  converge in distribution to multivariate normal  $N(0, \Sigma)$ , with the covariance matrix

$$\Sigma = \mathbb{E}_P \left\{ [\nabla_x G(x^*, \omega)] [\nabla_x G(x^*, \omega)]^T \right\} - \nabla g(x^*) \nabla g(x^*)^T, \quad (5.30)$$

provided the interchangeability formula (3.8) holds and this covariance matrix exists. In that case the random vector  $Z$  in formula (5.26) is distributed  $N(0, \Sigma)$ . Also under an additional condition, e.g. that random variables  $N [\hat{v}_N - \hat{g}_N(x^*)]$  have bounded second order moments, it follows from (5.26) that the expected values of these random variables converge to  $\frac{1}{2} \mathbb{E} \{ \varphi(Z) \}$ . In such case we obtain that

$$\mathbb{E} \{ \hat{v}_N \} - v^* = \frac{1}{2} N^{-1} \mathbb{E} \{ \varphi(Z) \} + o(N^{-1}). \quad (5.31)$$

**Example 5.1 (continued)** Consider the log-likelihood ratio statistic  $\ell_N$  defined in (5.6). Suppose that the distribution  $P$  is given by pdf  $f(y, \theta^*)$  with  $\theta^* \in \Theta_0 \cap \Theta_1$ . Then, under suitable regularity conditions, we have by the above discussion (see formulas (5.25) and (5.29) in particular) that

$$\ell_N = \inf_{z \in T_0} \left\{ 2z^T Z_N + z^T I(\theta^*) z \right\} - \inf_{z \in T_1} \left\{ 2z^T Z_N + z^T I(\theta^*) z \right\} + o_p(1), \quad (5.32)$$

where  $I(\theta^*) := \nabla^2 g(\theta^*)$  with  $g(\theta)$  being defined in (5.7),

$$Z_N := N^{-1/2} \sum_{i=1}^N \nabla \ln f(Y^i, \theta^*),$$

and  $T_0 := T_{\Theta_0}(\theta^*)$ ,  $T_1 := T_{\Theta_1}(\theta^*)$ .

Under standard regularity conditions, ensuring that second order derivatives can be taken inside the expected value in the definition of  $I(\theta^*)$ , we also have that

$$I(\theta^*) = -\mathbb{E} \{ \nabla^2 \ln f(Y, \theta^*) \} = \mathbb{E} \{ [\nabla \ln f(Y, \theta^*)] [\nabla \ln f(Y, \theta^*)]^T \},$$

i.e.  $I(\theta^*)$  is Fisher's information matrix, and that  $Z_N \Rightarrow N(0, I(\theta^*))$ . Assuming further that the matrix  $I(\theta^*)$  is nonsingular and substituting  $W_N := I(\theta^*)^{-1} Z_N$ , we can write (5.32) in the form

$$\ell_N = \inf_{z \in T_0} (W_N - z)^T I(\theta^*) (W_N - z) - \inf_{z \in T_1} (W_N - z)^T I(\theta^*) (W_N - z) + o_p(1). \quad (5.33)$$

Since  $W_N \Rightarrow N(0, I(\theta^*)^{-1})$ , it follows that

$$\ell_N \Rightarrow \inf_{z \in T_0} (W - z)^T I(\theta^*) (W - z) - \inf_{z \in T_1} (W - z)^T I(\theta^*) (W - z), \quad (5.34)$$

where  $W \sim N(0, I(\theta^*)^{-1})$ . This result is due to Chernoff [7].

Let us observe that from the point of view of general stochastic programming problems the above example is quite specific. This is because  $\theta^*$  is the unconstrained minimizer of  $g(\theta)$  and consequently  $\nabla g(\theta^*) = 0$ . Therefore the contingent cones coincide with the critical cones at the point  $\theta^* \in \Theta_0 \cap \Theta_1$ , and the term corresponding to the curvature of the sets  $\Theta_0, \Theta_1$ , at  $\theta^*$ , vanishes in the second order asymptotics (5.32) - (5.34). Also, because  $\nabla g(\theta^*) = 0$ , instead of second order regularity of the sets  $\Theta_0$  and  $\Theta_1$  we only need to assume that the contingent cones  $T_{\Theta_0}(\theta^*)$  and  $T_{\Theta_1}(\theta^*)$  coincide with the cones  $\bar{T}_{\Theta_0}(\theta^*)$  and  $\bar{T}_{\Theta_1}(\theta^*)$ , respectively.

## 5.4 Asymptotics of the Optimal Solutions

Let us discuss now asymptotics of optimal solutions  $\hat{x}_N$  of the SAA problem. It turns out that first order asymptotics of  $\hat{x}_N$  are closely related to second order expansions of the optimal value discussed in the previous section.

Let  $\bar{d}(\zeta)$  be a minimizer of the right hand side of (5.27). That is,

$$\bar{d}(\zeta) \in \arg \min_{d \in C(x^*)} \left\{ 2d^T \zeta + d^T \nabla^2 g(x^*) d + \inf_{w \in T_S^2(x^*, d)} w^T \nabla g(x^*) \right\}. \quad (5.35)$$

Suppose further that the minimizer  $\bar{d}(\zeta)$  is unique for any  $\zeta \in \mathbb{R}^m$ . We will discuss conditions for such uniqueness later.

**Theorem 5.4** *Suppose that the assumptions of theorem 5.3 hold and that for any  $\zeta \in \mathbb{R}^m$  the optimization problem in the right hand side of (5.35) has unique optimal solution  $\bar{d}(\zeta)$ . Then*

$$N^{1/2} (\hat{x}_N - x_0) = \bar{d}(Z_N) + o_p(1), \quad (5.36)$$

and

$$N^{1/2} (\hat{x}_N - x_0) \Rightarrow \bar{d}(Z), \quad (5.37)$$

where  $Z := \nabla Y(x^*)$  and  $Z_N := N^{1/2} [\nabla \hat{g}_N(x^*) - \nabla g(x^*)]$ .

Recall that  $Z_N \Rightarrow Z$  and that in the i.i.d. case  $Z \sim N(0, \Sigma)$ , where the covariance matrix  $\Sigma$  is given in (5.30). The optimal solution  $\bar{d}(\zeta)$  can be a nonlinear function of  $\zeta$  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. In the context of general stochastic programming problems asymptotics of SAA estimators were derived in King [21], King and Rockafellar [23], Shapiro [35, 38]. The above formulation is taken from [42].

For instance, consider the ML example 2.1. Suppose that the distribution  $P$  is given by pdf  $f(y, \theta^*)$  with  $\theta^* \in \Theta$ . Then, under suitable regularity conditions, we have that  $N^{1/2}(\hat{\theta}_N - \theta^*) \Rightarrow \bar{d}(W)$ , where  $W \sim N(0, I(\theta^*)^{-1})$  and  $\bar{d}(W)$  is the optimal solution of the problem

$$\min_{d \in T_{\Theta}(\theta^*)} (W - d)^T I(\theta^*) (W - d). \quad (5.38)$$

Note that this optimal solution  $\bar{d}(W)$  is unique for every  $W$  if Fisher's information matrix  $I(\theta^*)$  is nonsingular (and hence is positive definite) and the cone  $T_{\Theta}(\theta^*)$  is convex. If  $\theta^*$

lies on the boundary of  $\Theta$  and the contingent cone  $T_{\Theta}(\theta^*)$  is not a linear space, then the function  $\bar{d}(\cdot)$  is not linear and  $\hat{\theta}_N$  is not asymptotically normal.

Now let  $S$  be defined by constraints, as in (5.17), and suppose that the the gradient vectors  $\nabla h_i(x^*)$ ,  $i \in \{1, \dots, q\} \cup \mathcal{I}(x^*)$ , are linearly independent. Then  $\Lambda(x^*) = \{\lambda^*\}$  is a singleton and  $\varphi(\zeta)$  and  $\bar{d}(\zeta)$  are the optimal value and an optimal solution of the problem

$$\text{Min}_{d \in \mathbb{R}^m} 2d^T \zeta + d^T \nabla_{xx}^2 L(x^*, \lambda^*) d \text{ subject to } d \in C(x^*). \quad (5.39)$$

Recall that, in the present case, the critical cone  $C(x^*)$  can be written in the form (5.21), with  $\lambda = \lambda^*$ . Problem (5.39) is a quadratic programming problem. It has a unique optimal solution  $\bar{d}(\zeta)$  if the Hessian matrix  $\nabla_{xx}^2 L(x^*, \lambda^*)$  is positive definite over the linear space defined by the first  $q + |\mathcal{I}_+(\lambda^*)|$  (equality) linear constraints in (5.21). Also, because of the linear independence condition, this problem has a unique vector  $\bar{\alpha}(\zeta)$  of Lagrange multipliers associated with  $\bar{d}(\zeta)$ .

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

$$\begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} d \\ \alpha \end{bmatrix} = - \begin{bmatrix} \zeta \\ 0 \end{bmatrix}. \quad (5.40)$$

Here  $H := \nabla_{xx}^2 L(x^*, \lambda^*)$  and  $A$  is the  $m \times (q + |\mathcal{I}(x^*)|)$  matrix whose columns are formed by vectors  $\nabla h_i(x^*)$ ,  $i \in \{1, \dots, q\} \cup \mathcal{I}(x^*)$ . We obtain in that case, provided the block matrix in the left hand side of (5.40) is nonsingular, that  $N^{1/2}(\hat{x}_N - x^*, \hat{\lambda}_N - \lambda^*)$  converges in distribution to normal with zero mean and the covariance matrix

$$\begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} H & A \\ A^T & 0 \end{bmatrix}^{-1}. \quad (5.41)$$

It can happen that the critical cone  $C(x^*)$  consists of the single point 0, i.e.  $C(x^*) = \{0\}$ . In that case the functions  $\varphi(\cdot)$  and  $\bar{d}(\cdot)$  are identically zero, and hence the corresponding asymptotics (5.25), (5.26) and (5.36), (5.37) become trivial. For example, if the set  $S$  is defined by constraints, as in (5.17), and the MF-constraint qualification holds, then it follows from formula (5.21) that  $C(x^*) = \{0\}$  if, for some  $\lambda \in \Lambda(x^*)$ , the gradient vectors  $\nabla h_i(x^*)$ ,  $i \in \{1, \dots, q\} \cup \mathcal{I}_+(\lambda)$ , generate the space  $\mathbb{R}^m$ . In particular this happens if the number of active inequality constraints at  $x^*$  is  $m - q$  (i.e.,  $|\mathcal{I}(x^*)| = m - q$ ), the gradient vectors  $\nabla h_i(x^*)$ ,  $i \in \{1, \dots, q\} \cup \mathcal{I}(x^*)$ , are linearly independent and all Lagrange multipliers corresponding to the active inequality constraints are positive.

Suppose that  $C(x^*) = \{0\}$ . In that case there exists a neighborhood  $U$  of  $\nabla g(x^*)$  such that if  $\nabla \hat{g}_N(x^*) \in U$ , then the first order optimality conditions for the SAA problem hold at the point  $x^*$ , and  $x^*$  is a locally optimal solution of the SAA problem. By the strong Law of Large Numbers, we have that  $\nabla \hat{g}_N(x^*)$  converges to  $\nabla g(x^*)$  w.p.1. Consequently, w.p.1 for  $N$  large enough,  $\nabla \hat{g}_N(x^*) \in U$ , and hence  $x^*$  is a locally optimal solution of the SAA problem. It follows then that  $\hat{x}_N = x^*$  w.p.1 for  $N$  large enough. Moreover, by the Large Deviations theory (e.g., [10]) we have, under mild regularity conditions, that the probability of the event  $\nabla \hat{g}_N(x^*) \notin U$  tends to zero exponentially fast as  $N \rightarrow \infty$ . Therefore in such

cases we have that the probability of the event  $\{\hat{x}_N = x^*\}$  approaches one exponentially fast as  $N \rightarrow \infty$ , and the bias of  $\hat{v}_N$  approaches zero at an exponential rate.

Such asymptotic behavior of optimal solutions of SAA problems is typical in case of discrete distributions. That is, the following results hold (Shapiro and Homem-de-Mello [41]).

**Theorem 5.5** *Suppose that: (i) the set  $\Omega$  is finite (and hence the distribution  $P$  is discrete), (ii) for every  $\omega \in \Omega$  the function  $G(\cdot, \omega)$  is piecewise linear and convex, (iii) the feasible set  $S$  is closed, convex and polyhedral, (iv) the true problem (1.1) has a non empty bounded set  $\mathcal{S}^*$  of optimal solutions. Then the set  $\mathcal{S}^*$  is compact convex and polyhedral, and with probability one for  $N$  large enough the SAA problem (1.2) has a non empty set  $\hat{\mathcal{S}}_N$  of optimal solutions and  $\hat{\mathcal{S}}_N$  forms a face of the polyhedron  $\mathcal{S}^*$ . Moreover, there exists a constant  $\beta > 0$  such that*

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log [1 - P(\hat{\mathcal{S}}_N \subset \mathcal{S}^*)] \leq -\beta. \quad (5.42)$$

The upper bound (5.42) means that with probability approaching one exponentially fast, with increase of the sample size  $N$ , an optimal solution of the SAA problem provides an *exact* optimal solution of the true problem. That is what happens in examples 2.2-2.4 if the corresponding probability distribution is discrete. This shows that asymptotics of the optimal solutions can be completely different for continuous and discrete distributions.

## 6 Numerical Algorithms and Validation Analysis

In this section we discuss some numerical aspects of the SAA method. Recall that after a random sample is generated, the corresponding SAA problem becomes a deterministic problem and can be solved by an appropriate optimization algorithm. Asymptotic results of the previous section suggest that one can use somewhat different strategies in cases of continuous and discrete distributions. If the probability distribution is continuous, and consequently  $\hat{x}_N$  converges at a rate of  $O_p(N^{-1/2})$ , then there is no point in trying to solve the corresponding SSA problem very accurately. That is, it makes sense to stop the (iterative) deterministic procedure when the corresponding stochastic error starts to dominate a deterministic precision of the current iterate. On the other hand if the distribution is discrete, then one can try to solve the SAA problem exactly since in that case (under the assumptions of Theorem 5.5), with probability approaching one exponentially fast,  $\hat{x}_N$  is an exact optimal solution of the true problem.

Closely related to such stopping rules is validation analysis. That is, suppose that we are given a point  $\hat{x}$  which is suggested as a possible solution of the true problem. Can we evaluate (validate) accuracy of  $\hat{x}$  in that respect? Suppose that the assumptions of Proposition 3.1 are satisfied, so that the interchangeability formula (3.8) holds. Suppose also that  $x^*$  is an interior point of  $S$  and hence, at least locally, the true problem can be considered as unconstrained. Then by the first order optimality conditions we have that  $\nabla g(x^*) = 0$ . Consequently we can approach the problem of verifying validity of  $\hat{x}$  by testing the null hypothesis  $H_0 : \nabla g(\hat{x}) = 0$  against the alternative  $H_1 : \nabla g(\hat{x}) \neq 0$ . By generating

an i.i.d. random sample  $W^1, \dots, W^N$  we can estimate  $\nabla g(\hat{x})$  by

$$\gamma_N := \nabla \hat{g}_N(\hat{x}) = N^{-1} \sum_{i=1}^N \nabla_x G(\hat{x}, W^i). \quad (6.1)$$

By the CLT we have that  $N^{1/2} [\gamma_N - \nabla g(\hat{x})] \Rightarrow N(0, \Sigma)$ . The covariance matrix  $\Sigma$  can be consistently estimated, from the same sample, by the sample covariance matrix

$$\hat{\Sigma}_N := (N-1)^{-1} \sum_{i=1}^N [\nabla_x G(\hat{x}, W^i) - \gamma_N][\nabla_x G(\hat{x}, W^i) - \gamma_N]^T. \quad (6.2)$$

It follows then that under the null hypothesis  $H_0$ , the statistic

$$T := N \gamma_N^T \hat{\Sigma}_N^{-1} \gamma_N$$

has asymptotically a chi-square distribution with  $m$  degrees of freedom (e.g., [28]). Therefore we reject  $H_0$  if  $T$  is bigger than a critical value  $\chi_{\alpha, m}^2$ . Alternatively we can calculate the corresponding  $p$ -value of the test statistic  $T$ . In a similar way it is possible to construct asymptotically chi-square test statistics for testing the first order (KKT) optimality conditions in case the feasible set is defined by smooth constraints as in (5.23) (see [40] for details).

It should be understood that by accepting (i.e. by failing to reject)  $H_0$ , we do not claim that the gradient vector  $\nabla g(\hat{x})$  is exactly zero. By accepting  $H_0$  we rather assert that we cannot separate  $\nabla g(\hat{x})$  from zero given precision of the generated sample. That is, statistical error of the estimator  $\gamma_N$  is bigger than a possible difference between  $\nabla g(\hat{x})$  and the null vector. Also rejecting  $H_0$  does not necessarily mean that  $\hat{x}$  is a poor approximation of the true optimal solution  $x^*$ . A calculated value of the statistic  $T$  can be large simply because the estimated covariance matrix  $N^{-1} \hat{\Sigma}_N$  of  $\gamma_N$  is “small”, i.e.  $\gamma_N$  is an accurate estimator of  $\nabla g(\hat{x})$ . The above testing procedure could be combined with considering the corresponding  $100(1 - \alpha)\%$  confidence region

$$\left\{ z \in \mathbb{R}^m : (z - \gamma_N)^T \hat{\Sigma}_N^{-1} (z - \gamma_N) \leq \frac{\chi_{\alpha, m}^2}{N} \right\} \quad (6.3)$$

for  $\nabla g(\hat{x})$ . This confidence region can give an idea of how small or large  $\nabla g(\hat{x})$  could be.

The above statistical testing method can be also used as a stopping criterion in a numerical procedure. That is, given a current solution  $x^\nu$  at  $\nu$ -th iteration, of the SAA problem, a new (independent of the previous calculations) sample is generated and accuracy of  $x^\nu$  is evaluated by magnitude of the test statistic  $T$  and size of the corresponding confidence region (6.3). Then either the algorithm is stopped, if the user is satisfied with the achieved precision, or a larger sample is generated and a few iterations of the numerical procedure are performed for the obtained SAA problem. Confidence region (6.3) can give an idea of how large the sample size is needed in order to improve a current solution  $x^\nu$ . The sample size should be large enough such that, with a given confidence, vectors  $\gamma_N := \nabla \hat{g}_N(x^\nu)$  and  $\nabla g(x^\nu)$  form an acute angle and consequently  $-\gamma_N$  is a direction of descent, at the point  $x^\nu$ , for the “true” (expected value) function  $g(x)$ . This is guaranteed if the corresponding

confidence region does not include the null vector. This suggests the following formula for a required sample size  $N'$ :

$$N' \geq \chi_{\alpha, m}^2 / (\gamma_N^T \hat{\Sigma}_N^{-1} \gamma_N). \quad (6.4)$$

This analysis can be extended to situations where the feasible set is defined by smooth constraints as well (see [40] for details).

The above approach to stopping criteria has the following drawbacks. A new sample should be generated and the gradients  $\nabla_x G(x^\nu, W^i)$ ,  $i = 1, \dots, N$ , should be calculated every time the corresponding test is applied. In some cases formula (6.4) can give a sample size which is too large. And, finally, such testing of first order (KKT) optimality conditions cannot be applied in situations where the corresponding distributions are discrete and hence the expected value function  $g(x)$  can be non differentiable at the optimal solution.

An alternative approach to stopping a numerical algorithm is to compare successive values  $v_N^\nu := \hat{g}_N(x^\nu)$  calculated during an iterative procedure applied to the SAA problem. If  $v_N^{\nu+1}$  is not significantly smaller than  $v_N^\nu$ , then the algorithm may be stopped. The significance of the improvement can be tested by a paired  $t$ -test in case the same sample is used for calculation of both  $v_N^\nu$  and  $v_N^{\nu+1}$  (cf. [40]).

So far we did not discuss numerical algorithms which could be applied to the SAA problems. Choice of a particular numerical optimization technique is, of course, problem dependent. Nevertheless some general remarks are in order. As it was mentioned earlier, in many interesting applications the function  $G(\cdot, \omega)$  is piecewise smooth for almost every  $\omega \in \Omega$ . This is the case in examples 2.3 and 2.4. In such cases the objective function  $\hat{g}_N(\cdot)$  of the corresponding SAA problem is also piecewise smooth. In such situations first order methods, based on calculated gradients of  $\hat{g}_N(\cdot)$ , could be reasonably efficient. In particular, if  $G(\cdot, \omega)$  and hence  $\hat{g}_N(\cdot)$  are convex piecewise linear functions, then bundle or cutting type optimization algorithms work quite well (see, e.g., [15] for a discussion of bundle type methods).

Let us recall that in some cases, typically if the probability distributions are continuous, the expected value function  $g(\cdot)$  is twice differentiable and its second order derivatives can be estimated, say by the LR method (see section 4). In such cases one may think about employing second order (like Newton or quasi-Newton) type methods. In our experience Newton type methods did not work well, mainly because estimates of second order derivatives were not accurate enough. Anyway this requires further investigation.

Estimated gradients of the expected value function  $g(x)$  can be also employed in conjunction with Stochastic Approximation (SA) optimization techniques. Let us remark that SA algorithms are very sensitive to a choice of the involved step sizes and typically are unstable. In that respect the described above SAA techniques are more numerically robust. Moreover, the SAA method has an advantage of good stopping criteria. From a theoretical point of view both methods, SA with *optimal* stepsizes, and the SAA method converge *asymptotically* at the same rate provided the true optimization problem (1.1) is *smooth*, [39]. On the other hand in case the probability distributions are discrete, and consequently the function  $g(x)$  is not differentiable at the optimal solution, the SAA method can give an *exact* optimal solution for sample size large enough, while SA converges at an asymptotic rate determined by the corresponding choice of stepsizes. In such cases, whenever applicable, the SAA method would be preferable.

## References

- [1] M. Albritton, A. Shapiro and M.L. Spearman, “Optimal scheduling of a capacitated production facility subject to random demand,” manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1999.
- [2] A. Araujo and E. Giné, *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley, New York, 1980.
- [3] E. Beale, “On minimizing a convex function subject to linear inequalities”, *J. Roy. Statist. Soc., Ser. B*, 17 (1955), 173-184.
- [4] A. Benveniste, M. Métiver and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Sringer-Verlag, Berlin, 1990.
- [5] J.R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. Springer, New York, 1997.
- [6] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*. Springer-Verlag, to be published in 2000.
- [7] H. Chernoff, “On the distribution of the likelihood ratio”, *Ann. Math. Statist.*, 25 (1954), 573-578.
- [8] E.K.P. Chong and P.J. Ramadge, “Optimization of queues using an infinitesimal perturbation analysis-based stochastic algorithm with general update times”, *SIAM J. Control and Optimization*, 31 (1993), 698-732.
- [9] G. Dantzig, “Linear programming under uncertainty”, *Management Sci.*, 1 (1955), 197-206.
- [10] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd. ed., Springer-Verlag, New York, 1998.
- [11] 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.
- [12] C.J. Geyer and E. A. Thompson, “Constrained Monte Carlo maximum likelihood for dependent data, (with discussion),” *J. Roy. Statist. Soc. Ser. B*, 54 (1992), 657-699.
- [13] C.J. Geyer, “Practical Markov chain Monte Carlo (with discussion),” *Statist. Sci.*, 7 (1992), 473-511.
- [14] J.L. Higle and S. Sen, “Stochastic decomposition: an algorithm for two-stage linear programs with recourse”, *Mathematics of Operations Research*, 16 (1991), 650-669.
- [15] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

- [16] P.J. Huber, "Robust estimation of a location parameter", *Ann. Math. Statist.*, 35 (1964), 73-101.
- [17] 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.
- [18] P.J. Huber, *Robust Statistics*. Wiley, New York, 1981.
- [19] G. Infanger, *Planning under Uncertainty, Solving Large Scale Stochastic Linear Programs*, Boyd & Fraser Publishing Company, Massachusetts, 1994.
- [20] P. Kall and S.W. Wallace, *Stochastic Programming*. Wiley, Chichester, 1994.
- [21] 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.
- [22] A.J. King, "Generalized delta theorems for multivalued mappings and measurable selections", *Mathematics of Operations Research*, 14 (1989), 720-736.
- [23] 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.
- [24] H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, NY, 1978.
- [25] P. L'Ecuyer, N. Giroux and P.W. Glynn, "Stochastic optimization by simulation: numerical experiments with M/M/1 queue in steady-state", *Management Science*, 40 (1994), 1245-1261.
- [26] W.K. Mak, D.P. Morton and R.K. Wood, "Monte Carlo bounding techniques for determining solution quality in stochastic programs", *Operations Research Letters*, 24 (1999), 47-56.
- [27] 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.
- [28] R.J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.
- [29] 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.
- [30] C.R. Rao, *Linear Statistical Inference and Its Applications*. Wiley, New York, 1973.
- [31] S.M. Robinson, "Analysis of sample-path optimization, " *Math. Oper. Res.*, 21 (1996), 513-528.

- [32] R.T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.
- [33] R.Y. Rubinstein and A. Shapiro, "Optimization of static simulation models by the score function method", *Mathematics and Computers in Simulation*, 32 (1990), 373-392.
- [34] R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, New York, NY, 1993.
- [35] A. Shapiro, "Asymptotic Properties of Statistical Estimators in Stochastic Programming," *Annals of Statistics*, 17 (1989), 841-858.
- [36] A. Shapiro, "On concepts of directional differentiability," *Journal Optim. Theory and Appl.*, 66 (1990), 477-487.
- [37] A. Shapiro, "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, 30 (1991), 169-186.
- [38] A. Shapiro, "Asymptotic Behavior of Optimal Solutions in Stochastic Programming," *Mathematics of Operations Research*, 18 (1993), 829-845.
- [39] A. Shapiro, "Simulation based optimization - convergence analysis and statistical inference", *Stochastic Models*, 12 (1996), 425-454.
- [40] A. Shapiro and T. Homem-de-Mello, "A simulation-based approach to two-stage stochastic programming with recourse", *Mathematical Programming*, 81 (1998), 301-325.
- [41] A. Shapiro and T. Homem-de-Mello, "On rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs," preprint available at: *Stochastic Programming E-Print Series*. <http://dochost.rz.hu-berlin.de/speps/>
- [42] A. Shapiro, "Statistical inference of stochastic optimization problems", to appear in: *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev (Ed.), Kluwer, 2000.
- [43] A. Wald, "Note on the consistency of the maximum likelihood estimates", *Ann. Math. Statist.*, 20 (1949), 595-601.