Topics on Monte Carlo Simulation-Based Methods for Stochastic Optimization Problems: Stochastic Constraints and Variance Reduction Techniques

Tito Homem-de-Mello and Güzin Bayraksan

Abstract We provide an overview of two select topics in Monte Carlo simulation-based methods for stochastic optimization: problems with stochastic constraints and variance reduction techniques. While Monte Carlo simulation-based methods have been successfully used for stochastic optimization problems with deterministic constraints, there is a growing body of work on its use for problems with stochastic constraints. The presence of stochastic constraints brings new challenges in ensuring and testing optimality, allocating sample sizes, etc., especially due to difficulties in determining feasibility. We review results for general stochastic constraints and also discuss special cases such as probabilistic and stochastic dominance constraints. Next, we review the use of variance reduction techniques (VRT) in a stochastic optimization setting. While this is a well-studied topic in statistics and simulation, again, the use of VRT in stochastic optimization requires a more thorough analysis. We discuss asymptotic properties of the resulting approximations and their use within Monte Carlo simulation-based solution methods. This chapter is intended to provide an overview of these two topics to introduce them to students and researchers.

1 Introduction

In this chapter we consider stochastic optimization problem of the form

\[
\min_{x \in \mathcal{X}} \{ g_0(x) := \mathbb{E}[G_0(x, \xi)] \mid g_k(x) := \mathbb{E}[G_k(x, \xi)] \leq 0, \ k \in \mathcal{K} \}, \quad (SP)
\]
where $G_k$, $k \in \{0\} \cup \mathcal{K}$ are extended real-valued functions with inputs being the decision vector $x$ and a random vector $\xi$. The set of stochastic constraints is determined by the set $\mathcal{K}$. Typically, $\mathcal{K}$ contains a finite number of constraints; that is, $\mathcal{K} = \{1, 2, \ldots, K\}$; however, this does not have to be the case and it can contain uncountably many constraints (see, e.g., Section 2.3). In this chapter, we use $\mathcal{K} = \emptyset$ to denote (SP) without any stochastic constraints and we set $\mathcal{K} = 1$ for the case with a single stochastic constraint. The set of deterministic constraints $x$ must satisfy is denoted by $X \subseteq \mathbb{R}^{d_x}$ and $\Xi \subseteq \mathbb{R}^{d_\xi}$ denotes the support of $\xi$, where $d_x$ and $d_\xi$ are the dimensions of the vectors $x$ and $\xi$, respectively. We assume that $\xi$ has a known distribution, $P$, that is independent of $x$, and the expectations in (SP), taken with respect to the distribution of $\xi$, are well-defined and finite for all $x \in X$. We will refer to (SP) as the “true” optimization problem (as opposed to the approximating problems to be discussed in the sequel).

A wide variety of problems can be cast as (SP) depending on $\mathcal{K}$, $X$ and $G_k$, $k \in \{0\} \cup \mathcal{K}$. For example, in a two-stage stochastic linear program with recourse, $\mathcal{K} = \emptyset$, $X = \{Ax = b, x \geq 0\}$, and $G_0(x, \xi) = cx + h(x, \xi)$, where $h(x, \xi)$ is the optimal value of the linear program

$$h(x, \xi) = \min_y \tilde{q}y$$

$$\text{s.t. } \tilde{W}y = \tilde{r} - \tilde{T}x, \quad y \geq 0.$$  

Here, $\xi$ is a random vector that is comprised of random elements of $\tilde{q}, \tilde{W}, \tilde{R}$ and $\tilde{T}$.

In contrast, in a stochastic linear program with a single probabilistic constraint (i.e., $P(\tilde{A}'x \geq \tilde{b}') \leq \alpha$), we have $K = 1$, $G_0(x, \xi) = cx$ and

$$G_1(x, \xi) = \mathbb{I}\{\tilde{A}'x \geq \tilde{b}'\} - \alpha,$$

where $\mathbb{I}\{E\}$ denotes the indicator function that takes value 1 if the event $E$ happens and 0 otherwise and $\alpha \in (0, 1)$ is a desired probability level. In this case, $\xi$ is comprised of random elements of $\tilde{A}'$ and $\tilde{b}'$. Here, the decision maker requires that the relationship $\tilde{A}'x \geq \tilde{b}$ be satisfied with probability no more than $\alpha$.

Monte Carlo simulation-based methods have been successfully used in many different applications of stochastic optimization. The appeal of such methods results from the fact that they often approximate well, with a small number of samples, problems that have very large number of scenarios; see, for instance, Linderoth et al. [60] for numerical reports.

There are multiple ways to use Monte Carlo methods in problem (SP). A generic way of describing them is to construct an approximating problem as follows. Consider a family $\{g_{k,N_k}(\cdot)\}$ of random approximations of the function $g_k(\cdot)$, each $g_{k,N_k}(\cdot)$ being defined as

$$g_{k,N_k}(x) := \frac{1}{N_k} \sum_{j=1}^{N_k} G_k(x, \xi^j_k),$$  

(1)
where \( \{\xi_k^1, \ldots, \xi_k^{N_k}\} \) is a sample of size \( N_k \) from the distribution\(^1\) of \( \xi \), for \( k \in \{0\} \cup \mathcal{K} \). When \( \xi_k^1, \ldots, \xi_k^{N_k} \) are mutually independent, the quantity \( g_{k,N_k}(x) \) is called a (standard or crude) Monte Carlo estimator of \( g_k(x) \). Given the family of estimators \( \{g_{k,N_k}(\cdot)\} \) defined in (1), one can construct the corresponding approximating program

\[
\min_{x \in X} \left\{ g_{0,N_0}(x) \mid g_{k,N_k}(x) \leq 0, \ k \in \mathcal{K} \right\}.
\]

(2)

In this chapter we adopt the following notation: \( N \) denotes the vector of sample sizes \( [N_k]_{k \in \{0\} \cup \mathcal{K}} \); \( \nu^* \) and \( \nu_N \) denote the optimal value of (SP) and the optimal value of its approximation (2), respectively; similarly, \( x^* \) and \( x_N \) denote an optimal solution to (SP) and (2), respectively. For a particular realization \( \{\hat{\xi}_k^1, \ldots, \hat{\xi}_k^{N_k}\} \) of \( \{\xi_k^1, \ldots, \xi_k^{N_k}\}, k \in \{0\} \cup \mathcal{K} \), we use the notation \( \hat{x}_N \) and \( \hat{\nu}_N \). Note that for each \( x \in X \) the quantity \( g_{k,N_k}(x) \) is a random variable, since it depends on the sample \( \{\xi_k^1, \ldots, \xi_k^{N_k}\} \). So, the optimal solution(s) \( \hat{x}_N \) and the optimal value of (2) \( \hat{\nu}_N \) are random as well. Nevertheless, it is possible to study convergence and statistical properties of the approximating problem (2). Note that “convergence” here indicates the asymptotic behavior of (2) as the sample sizes \( N_k, k \in \{0\} \cup \mathcal{K} \), go to infinity. We will sometimes abuse the notation and write \( N \to \infty \) as an abbreviated form of the latter condition. The idea of replacing the expectations with sample averages and solving the resulting formulation has a long history but it has recently become popular under the name of Sample Average Approximation (SAA) approach.

Much of the existing work in the literature focuses on the case where \( \mathcal{K} = \emptyset \) in (SP), i.e., the constraints in the problem are deterministic; see, for instance, Shapiro [91], Shapiro et al. [94], Homem-de-Mello and Bayraksan [37] and also the chapter by Kim et al. [46] in this book for reviews of that literature. In this chapter we review some of the existing works that deal with Monte Carlo simulation-based approximations for problems with stochastic constraints. As we shall see, some new issues arise in that context, and the extension of the methods designed for deterministic constraints not only is involved but in some cases is not even known.

We also review some alternative ways to construct the approximating problem (2) which do not rely on standard Monte Carlo sampling. The goal of such methods is to obtain estimates that are more accurate than those obtained with standard Monte Carlo sampling, thus allowing the user to obtain a good solution from (2) with fewer samples. Given that the computational effort required to solve (2) typically grows fast with the sample sizes, the use of such alternative methods becomes crucial for the solution of certain problems.

We close this introduction by acknowledging that this chapter is largely based on Homem-de-Mello and Bayraksan [37], which is a more extensive survey of Monte Carlo methods for stochastic optimization. Our treatment here is more detailed with respect to assessment of solution quality and selection of sample sizes for problems with stochastic constraints. We also have an expanded discussion on the asymptotic properties of Monte Carlo sampling-based approximations of stochastic optimiza-

---

\(^1\) Throughout this paper, we will use the terminology “sample [of size \( N \)] from the distribution of \( \xi \)” to indicate a set of \( N \) random variables with the same distribution as \( \xi \).
tion problems using alternative sampling techniques. Finally, we remark that the purpose of this chapter is to give an overview of two topics in Monte Carlo methods for stochastic optimization—namely, problems with stochastic constraints and variance reduction techniques—with the goal of introducing them to students and researchers. Aiming to be accessible to a wider audience, we outline the main results and provide references for a more comprehensive treatment.

The remainder of this chapter is organized as follows. In Section 2 we study problems with stochastic constraints. We first discuss the general case of expected-value constraints in Section 2.1. Next, we examine in more detail two special classes of stochastic constraints, namely probabilistic and stochastic dominance constraints, in Sections 2.2 and 2.3, respectively. In our review of Monte Carlo methods for stochastic optimization problems with stochastic constraints, we start with formulations, then discuss some properties of the SAA approach, survey solution methods, review choice of sample sizes when appropriate, and finally discuss how to assess the quality of approximate solutions. Monte Carlo methods are often enhanced by the use of variance reduction techniques; the use of such methods in the context of sampling-based stochastic optimization is reviewed in Section 3. We review antithetic variates, Latin hypercube sampling, quasi-Monte Carlo, importance sampling and the use of control variates. Finally, we end the chapter with some concluding remarks and future research directions in Section 4.

2 Problems with stochastic constraints

We start by considering formulation (SP) and the corresponding approximation (2) in which $\mathcal{K}$ is a non-empty finite set. Such problems arise naturally in applications where the decision $x$ must satisfy inequalities but the functions defining the inequalities depend on a random parameter such as demand or future prices. For example, Atlason et al. [3] study call center staffing problems where the constraints ensure that the expected number of answered calls must be at least a certain percentage of the expected total number of received calls. The expected number of calls is approximated by sampling. Krokhmal et al. [51] study a portfolio optimization problem where the constraints are defined by the conditional value-at-risk (CVaR) of the random returns. CVaR constraints can also be used to provide convex approximations to chance constrained problems, as discussed in Nemirovski and Shapiro [71]. There are, however, some classes of problems where the structure of the stochastic constraints plays a fundamental modeling role, as we shall see below in Sections 2.2–2.3. Since specialized methods have been developed for these classes of problems, we will study them separately.
2.1 Problems with general expected-value constraints

In this section, we discuss some general results that do not exploit any particular structure of the constraints.

2.1.1 The SAA Approach

We first consider the properties of the SAA approach, which as discussed earlier consists of solving (2). One of the main concerns is the convergence of optimal values and limit points of optimal solutions to SAA problems in the general case as the sample sizes tend to infinity. This issue has been studied two decades ago, for instance, in papers by Shapiro [90] and King and Rockafellar [48]. While these asymptotic results are very useful, in practice one is naturally concerned with what happens with finite sample sizes; Wang and Ahmed [99] provide some results for this case.

The main issue that arises when using sampling approximations within the constraints is that of feasibility. For example, consider a single constraint of the form $E[G_1(x, \xi)] \leq 0$, and let $\bar{x}$ be a point on the boundary of the feasibility set so that $E[G_1(\bar{x}, \xi)] = 0$. Moreover, suppose that $G_1(\bar{x}, \xi)$ is normally distributed. Consider the sampling approximation $\frac{1}{N_1} \sum_{j=1}^{N_1} G_1(\bar{x}, \xi_j)$. Clearly, by the strong law of large numbers, this quantity converges to zero with probability one (w.p.1). However, no matter how large $N_1$ is, there is always a 50% chance that $\frac{1}{N_1} \sum_{j=1}^{N_1} G_1(\bar{x}, \xi_j) > 0$—in which case $\bar{x}$ is infeasible to the approximating problem. To circumvent the problem, we can replace the constraint $E[G_1(x, \xi)] \leq 0$ with

$$E[G_1(x, \xi)] \leq \varepsilon, \quad (3)$$

where $\varepsilon \in \mathbb{R}$. A positive value of $\varepsilon$ provides a relaxation of the problem, whereas a negative value tightens the problem. Let $U_\varepsilon$ denote the set defined by the set of $x \in X$ satisfying (3), and let $U^\varepsilon_{N_1}$ be the corresponding set defined by the sampling approximation. Then, it is possible show that, under proper assumptions—which include compactness of $X$ along with other conditions on $G_1(x, \xi)$ such as Lipschitz continuity—one has that, for any $\varepsilon > 0$,

$$P \left( U^{-\varepsilon} \subseteq U^0_{N_1} \subseteq U^\varepsilon \right) \geq 1 - Me^{-\beta \varepsilon^2 N_1}, \quad (4)$$

for some constants $M > 0$, $\beta > 0$. In other words, the probability that the feasibility set of the approximating problem is “sandwiched” between $U^{-\varepsilon}$ and $U^\varepsilon$ goes to one exponentially fast. The rate of convergence, of course, depends on the value of $\varepsilon$. We refer to Wang and Ahmed [99] for details.
2.1.2 Solution methods and choice of sample sizes

Under the SAA approach, given a realization \( \{\hat{\xi}^1, \hat{\xi}^2, \ldots \} \) of \( \{\xi^1, \xi^2, \ldots \} \), one then solves (2) for that realization. Any appropriate deterministic algorithm can be used to solve the corresponding problem, depending on the underlying structure (discrete, convex, etc.). In particular, one can use a penalization technique to bring the constraints into the objective function; see, for instance, Liu et al. [61] for a discussion. We are not aware of other approaches designed specifically for general problems with expected-value constraints in mathematical programming.

In contrast to mathematical programming, in the simulation literature often no assumptions regarding the structure of \( G_k \) are made. The approximations are obtained through a “black box” response to a simulated system for a given decision \( x \). When \( |X| \) is finite and typically of small size, a growing number of recent papers have been studying the stochastic optimization problems with general expected value constraints. Andradóttir and Kim [2] and Batur and Kim [6] extend classical ranking and selection procedures (see, e.g., Kim and Nelson [47]) to the cases where there are single and multiple stochastic constraints, respectively. Ranking and selection procedures aim to find the best solution among a discrete set of decisions with a pre-specified probability of selecting the correct solution. In the presence of stochastic constraints, this also includes determining the feasibility of the decisions.

In contrast, Lee et al. [57] analyze asymptotically optimal sample size allocations among different alternatives by extending the optimal computing budget allocation (see e.g., Chen and Lee [16]) to also consider feasibility of the alternatives. Hunter and Pasupathy [43] provide an alternative way of determining asymptotically optimal sample sizes by using large deviations theory. Other ongoing work in this area includes looking at relaxation of independent sampling for some procedures (Healey et al. [33, 32]), cases when \( |X| \) is large (Pujowidianto et al. [85]) or when stochastic constraints have some additional structure (e.g., Park and Kim [80]).

2.1.3 Assessing solution quality

As mentioned above, optimal solutions of (2) converge to their “true” corresponding values under some conditions. More specifically, the distance between \( x_N \) —an optimal solution of (2)—and the set \( S^* \) of optimal solutions of the “true” problem (SP) goes to zero as \( N \to \infty \) w.p.1. Thus, we expect \( x_N \) to be a good solution for the original problem. It is necessary, however, to assess the quality of such solution in rigorous terms. One of the classic approaches for assessing solution quality in optimization is to bound the candidate solution’s optimality gap. If the bound on the optimality gap is sufficiently small, then the candidate solution is of high quality.

In stochastic optimization, Monte Carlo sampling can be used to obtain (statistical) lower bounds. A whole research area has been developed to deal with that issue, starting with the work of Norkin et al. [73] and Mak, Morton, and Wood [64] (we refer the reader to Bayraksan and Morton [7] and Homem-de-Mello and Bayraksan [37] for reviews). The basic idea is as follows. Consider problem (SP)
with $\mathcal{X} = \emptyset$—i.e., there are no stochastic constraints—and let $\hat{x}$ be a “candidate” solution, whose quality we want to assess (for example, we can take $\hat{x}$ as an optimal solution to an independently generated SAA problem). The optimality gap of $\hat{x}$ is given by $g(\hat{x}) - v^*$. Of course, typically we cannot compute exactly either $g(\hat{x})$ or $v^*$. We can however estimate the gap by calculating

$$\mathcal{G}_N(\hat{x}) := g_N(\hat{x}) - v_N,$$  \hspace{1cm} (5)

using $v_N$, the optimal value of the approximating problem (2) (we write $N$ instead of $N_0$ since there is no ambiguity). Note that when viewed as an estimator of optimality gap, $\mathcal{G}_N(\hat{x})$ is biased, i.e., $\mathbb{E}[\mathcal{G}_N(\hat{x})] \geq g(\hat{x}) - v^*$. Such bias follows from the fact that $\mathbb{E}[v_N] \leq v^*$. While there are different ways to calculate the above optimality gap estimator, a basic version uses the same independent and identically distributed (i.i.d.) observations $\xi^1, \xi^2, \ldots, \xi^N$ from the distribution of $\xi$ for both terms in (5).

Then, multiple independent estimators $\mathcal{G}_N^k(\hat{x})$ are generated using $N_\mu$ “batches” of observations $\xi^{k1}, \xi^{k2}, \ldots, \xi^{kN}$, $k = 1, 2, \ldots, N_\mu$, and these $\mathcal{G}_N^k(\hat{x})$ are averaged to obtain a point estimator of the optimality gap

$$\bar{\mathcal{G}}(\hat{x}) := \frac{1}{N_\mu} \sum_{k=1}^{N_\mu} \mathcal{G}_N^k(\hat{x}).$$  \hspace{1cm} (6)

The sample variance is calculated in the usual way, $s^2 := \frac{1}{N_\mu - 1} \sum_{k=1}^{N_\mu} (\mathcal{G}_N^k(\hat{x}) - \bar{\mathcal{G}}(\hat{x}))^2$. An approximate $(1 - \alpha)$-level confidence interval estimator on the optimality gap of $\hat{x}$ is then obtained by

$$\left[ 0, \bar{\mathcal{G}}(\hat{x}) + \frac{z_{\alpha}s_{\mathcal{G}}}{\sqrt{N_\mu}} \right],$$  \hspace{1cm} (7)

where $z_{\alpha}$ denotes a $1 - \alpha$ quantile from a standard Normal distribution. The resulting point estimator (6) and interval estimator (7) are called the Multiple Replications Procedure (MRP) estimators. A major advantage of MRP is its wide applicability, as it does not require much structure from the original problem (SP). The main assumption is that $G(x, \xi)$ have finite second moments for all $x \in X$. The framework to implement MRP is relatively simple and step by step instructions can be found, for instance, in Bayraksan and Morton [7].

The approach discussed above for the case $\mathcal{X} = \emptyset$ can be extended to the setting of finite $\mathcal{X} \neq \emptyset$ with some adjustments. Suppose there is a single expected-value constraint; i.e., $K = 1$. Wang and Ahmed [99] suggest the following approach to derive a lower bound. Notice that the optimal value of $\min_{x \in X} \{ \mathbb{E}[G_0(x, \xi)] | \mathbb{E}[G_1(x, \xi)] \leq 0 \}$ is bounded from below by the optimal value of $\min_{x \in X} \{ \mathbb{E}[G_0(x, \xi)] + \lambda \mathbb{E}[G_1(x, \xi)] \}$ for any $\lambda \geq 0$. Since the latter problem does not have stochastic constraints, the methods derived for problems with deterministic constraints can be used to obtain a lower bound for its optimal value. For example, by solving $M$ independent approximations of the form
\[ v_N(\lambda) := \min_{x \in X} \left\{ \frac{1}{N} \sum_{j=1}^{N} G_0(x, \xi^j) + \lambda G_1(x, \xi^j) \right\} \]

one can construct confidence intervals for \( E[v_N(\lambda)] \), which in turn is a lower bound for \( \min \{ E[G_0(x, \xi)] + \lambda E[G_1(x, \xi)] \} \). Of course, the quality of the overall bound will depend on the value of \( \lambda \). One possible choice is to take \( \lambda \) as the optimal dual multiplier of the problem

\[ \min_{x \in X} \left\{ \frac{1}{N} \sum_{j=1}^{N} G_0(x, \xi^j) \left| \frac{1}{N} \sum_{j=1}^{N} G_1(x, \xi^j) \leq 0 \right\}, \]

where the sample is independent of the samples drawn in each of the \( M \) replications. Note that in this case we use the same sample to estimate both the objective function and the constraint. Note also that when the original problem is not convex (e.g., when the feasibility set is finite) such a lower bound can be loose even if \( N \) is large.

As in the case of problems with deterministic constraints, an upper bound for the optimal value of (SP) can be obtained simply by evaluating the objective function at any feasible solution. However, as we saw above, in case of stochastic constraints feasibility cannot be guaranteed since the functions defining the constraints cannot be evaluated exactly. One way (suggested by Shapiro et al. [94]) to ensure feasibility of a given \( x \in X \) with a given confidence is to fix \( 0 < \beta < 1 \) and construct a one-sided 100(1 - \( \beta \))% confidence interval for \( E[G_1(x, \xi)] \) as

\[ \frac{1}{N} \sum_{j=1}^{N} G_1(x, \xi^j) + z_{\beta} \sqrt{\frac{s_2^2(x)}{N}}, \]

where \( s_2^2(x) \) is the sample variance of the sample \( \{ G_1(x, \xi^1), \ldots, G_1(x, \xi^N) \} \). If the above quantity is less than or equal to zero, then we are 100(1 - \( \beta \))% confident that \( x \) is feasible and consequently the objective value at \( x \) yields upper bound for the optimal value of (SP).

Recall that Karush-Kuhn-Tucker (KKT) conditions provide necessary and sufficient conditions for optimality for a class of problems. In the class of problems we consider, the expectations in (SP) and their (sub)gradients can only be approximated by simulation-based methods. Several researchers have looked at determining if a solution is optimal or near optimal by testing KKT conditions. For instance, Bettonvil et al. [8] develop a series of hypotheses tests for verification of KKT conditions. Royset [86] uses a so-called optimality function that has two parts, one that is related to feasibility of the candidate solution with respect to the stochastic constraints and the other used for tesing the more general Fritz-John conditions (similar to KKT conditions) for optimality.
2.2 Problems with probabilistic constraints

This class of problems can be written as

\[
\min_{x \in X} \{ \mathbb{E}[G_0(x, \xi)] \mid P(H_k(x, \xi) \leq 0) \geq 1 - \alpha_k, \ k \in \mathcal{K} \}. \tag{8}
\]

Clearly, such a problem falls into the framework of (SP) since we can write

\[
P(H_k(x, \xi) \leq 0) \geq 1 - \alpha_k
\]

as

\[
\mathbb{E}[(1 - \alpha_k) - I\{H_k(x, \xi) \leq 0\}] \leq 0,
\]

where as before \(I\{E\}\) denotes the indicator function of the event \(E\). While a probabilistic constraint (also called chance constraint) is a special case of expected-value constraint, it is possible to exploit its special properties for its analysis and solution. The probabilistic constraints in (8) are used in situations where violation of the constraint inside the probability has a qualitative instead of a quantitative nature—that is, it matters whether the constraints are violated or not; the amount of violation is less important. Such constraints often used to model service level or reliability restrictions, such as “demand must be satisfied in at least 95% of the cases.”

It is important to distinguish between the case of separate chance constraints as in (8) and that of joint constraints of the form

\[
P(H_1(x, \xi) \leq 0, \ldots, H_K(x, \xi) \leq 0) \geq 1 - \alpha
\]

by defining \(H(x, \xi) := \max\{H_1(x, \xi), \ldots, H_K(x, \xi)\}\), although some properties such as differentiability may be lost in such representation. Still, using this representation, we can assume that \(K = 1\), which is a particular case of (8).

Problems with probabilistic constraints have been studied for decades, starting with the work of Charnes and Cooper [15]. As pointed out by Ahmed and Shapiro [1], the two major difficulties with such problems are that (i) evaluating \(P(H_k(x, \xi) \leq 0) \geq 1 - \alpha\) can be difficult (e.g., it may involve multidimensional integrals), and (ii) the sets \(\{x : P(H_k(x, \xi) \leq 0) \geq 1 - \alpha_k\}\) may be nonconvex. This subject is very rich; we refer to Prékopa [84] for a thorough discussion, and to Ahmed and Shapiro [1] for a more recent view. For the purposes of this survey, we will review work that uses Monte Carlo methods to approximate the chance constraints.

2.2.1 The SAA Approach

We start by discussing a direct SAA approach for this class for problems. To simplify the discussion, we shall assume that there is only one chance constraint, i.e., \(K = 1\), and we drop the index \(k\) from the notation. That is, we consider the problem

\[
\min_{x \in X} \left\{ \frac{1}{N} \sum_{j=1}^{N} G_0(x, \xi_j) \middle| \frac{1}{N} \sum_{j=1}^{N} I\{H(x, \xi_j) \leq 0\} \geq 1 - \gamma \right\} \tag{9}
\]

and as before consider the behavior of the optimal value and optimal solutions of (9) as function of \(N\). Note that we have replaced the term \(1 - \alpha\) on the right hand side with \(1 - \gamma\), where \(\gamma\) is a parameter—as before, by allowing \(\gamma\) to be different from \(\alpha\).
we obtain a problem that is either more relaxed or more tight than the original one, which is important when considering feasibility issues.

It is important to point out that the analysis discussed in Section 2.1 cannot be used here since the function \( \mathbb{1}\{H(\cdot, \xi_j) \leq 0\} \) is not continuous. Nevertheless, some convergence results can be derived in this setting as well. For example, Pagnoncelli et al. [78] show that, if (i) both \( G_0 \) and \( H \) are continuous in \( x \), (ii) the set \( X \) is compact and (iii) there exists an optimal solution \( x^* \) such that, given any neighborhood of \( x^* \), there exists some \( x \) in that neighborhood such that \( P(H(x, \xi) \leq 0) > 1 - \alpha \), then, we have that

- \( v_N \to v^* \) w.p.1 and
- \( \text{dist}(x_N, S^*) \to 0 \) w.p.1,

where \( \text{dist}(x, A) \) denotes the distance from a point \( x \) to a set \( A \), defined as \( \inf_{a \in A} \| x - a \| \). Luedtke and Ahmed [62] show that, under certain assumptions—such as compactness of \( X \) and Lipschitz continuity of \( H \) with respect to \( x \), or finiteness of \( X \)—an exponential convergence of the type (4) holds in this setting as well.

### 2.2.2 Solution methods and choice of sample sizes

A natural question that arises is, how to solve the approximating problem (9)? Let \( \{\hat{\xi}^1, \hat{\xi}^2, \ldots\} \) be a realization of \( \{\xi^1, \xi^2, \ldots\} \), and consider problem (9) defined for that specific realization. Here we need to distinguish between the two cases \( \gamma = 0 \) and \( \gamma > 0 \). When \( \gamma = 0 \), problem (9) can be equivalently written as

\[
\min_{x \in X} \left\{ \frac{1}{N} \sum_{j=1}^{N} G_0(x, \hat{\xi}^j) : H(x, \hat{\xi}^j) \leq 0, \quad j = 1, \ldots, N \right\}.
\]  

(10)

Depending on the structure of \( H \)—for example, when \( H \) is linear or convex in \( x \)—this can be a very tractable problem. The downside, of course, is that the replacement of \( \alpha > 0 \) with \( \gamma = 0 \) yields a more conservative model. Still, Campi et al. [14] improve upon an original result by Calafiore and Campi [12] and show that, given \( 0 < \delta \leq 1 \), by choosing

\[
N \geq \frac{2}{\alpha} \left( \log \frac{1}{\delta} + d_x \right)
\]

(recall that \( d_x \) is the dimension of \( x \) in (SP)), the optimal solution to (10) is feasible to the original problem (8) with probability at least \( 1 - \delta \), regardless of the distribution of \( \hat{\xi} \) when \( H(\cdot, \hat{\xi}) \) is convex. Nemirovski and Shapiro [70] show that a better value for \( N \), which grows as \( \log(1/\alpha) \) instead of \( 1/\alpha \), can be obtained under further assumptions on \( H \) and the distribution of \( \hat{\xi} \). Campi and Garatti [13] and Pagnoncelli et al. [79] discuss approaches to remove some of the sampled constraints in order to obtain a less conservative problem.

The situation is rather different when \( \gamma > 0 \) in (9). It is easy to see that convexity is lost because of the presence of the indicator functions even if \( H \) has nice properties. Note however that (9) is still a chance-constrained problem, where the underlying
distribution is the empirical distribution defined by \( \{ \hat{\xi}^1, \ldots, \hat{\xi}^N \} \). Thus, any method proposed for chance-constrained problems with finite number of scenarios can be used to solve (9). For example, Dentcheva et al. [22] aim to find the so-called \( p \)-efficient points corresponding to the constraints, whereas Luedtke et al. [63] provide a strengthened integer programming formulation for the problem when \( H(x, \xi) \) is of the form \( \max_i \{ \xi_i - h_i(x) \} \), i.e., one can separate the random component from the function.

A different sampling-based approach for chance constrained problems is proposed by Hong et al. [41] for the situation where, given \( x \in X \), the function \( H(x, \xi) \) is differentiable at \( x \) with probability one. Note that when \( H \) represents joint chance constraints such an assumption typically will not hold when \( \xi \) has discrete distribution, because of the “kinks” of the max function. When this assumption (and some others) do hold, Hong et al. [41] show that the constraint \( P(H(x, \xi) \leq 0) \geq 1 - \alpha \) can be written as a difference of convex (DC) functions. As a result, the original problem can be approximated (by successive linearization of one of the functions in the DC formulation) by a sequence of convex problems, and in the limit one obtains a KKT point for the original problem. Because the functions in the DC formulation cannot be evaluated exactly, a sampling-based approach is proposed to solve each of these convex problems.

Recently, problems with probabilistic constraints have also been studied from the perspective of ranking and selection procedures; see Hong and Nelson [40].

### 2.2.3 Assessing solution quality

It is possible to derive statistical lower and upper bounds for the optimal value of (8), \( \nu^* \). For a given \( x \in X \) consider the estimator

\[
\hat{p}(x) := \frac{1}{N} \sum_{j=1}^{N} I\{H(x, \xi^j) > 0\}
\]

of \( p(x) := P(H(x, \xi) > 0) \). As discussed before, any feasible \( x \) yields an upper bound for the optimal value of the problem. A given \( x \in X \) is feasible for (8) if \( p(x) \leq \alpha \). In our context, computing \( p(x) \) is impractical so we would like to use the estimator \( \hat{p}(x) \). One way of doing this is to construct a one-sided \( 100(1 - \beta) \% \) confidence interval for \( p(x) \) similarly to the idea described for general expected-value constrained problems, i.e., given a realization \( \{ \hat{\xi}^1, \hat{\xi}^2, \ldots, \} \) of \( \{ \xi^1, \xi^2, \ldots, \} \), compute the value of \( \hat{p}(x) \) corresponding that sample and check whether

\[
\hat{p}(x) + z_{\beta} \sqrt{\frac{\hat{p}(x)(1 - \hat{p}(x))}{N}} \leq \alpha,
\]

where we used the fact that \( \sum_{j=1}^{N} I\{H(x, \xi^j) > 0\} \) has binomial distribution with parameters \( N \) and \( p(x) \) and assumed that \( N \) is sufficiently large to ensure that the binomial distribution can be well approximated by a normal distribution with mean
$Np(x)$ and variance $Np(x)(1 - p(x))$. If (11) is satisfied, we are 100$(1 - \beta)\%$ confident that $x$ is feasible.

More efficient techniques have been developed by exploiting the structure of the probabilistic constraints. Following Nemirovski and Shapiro [71], let $B(k; p, n)$ denote the cumulative distribution function of the binomial distribution with parameters $n$ and $p$. Given $0 < \beta < 1$, define the quantity $U(x) := \sup \{p \in [0, 1] | B(N\hat{p}(x); \rho, N) \geq \beta\}$ (note that $U(x)$ is random). Then, it is possible to show that $P(p(x) < U(x)) \geq 1 - \beta$. This suggests the following procedure: given a realization $\{\hat{\xi}^1, \hat{\xi}^2, \ldots\}$ of $\{\xi^1, \xi^2, \ldots\}$, compute the values of $\hat{p}(x)$ and $U(x)$ corresponding to that sample; if $U(x) \leq \alpha$, then we are 100$(1 - \beta)\%$ confident that $x$ is feasible.

The calculation of lower bounds for the optimal value of (8) can in principle follow the method described in Section 2.1, but again the lack of convexity implies that Lagrangian-based bounds are not useful due to the existence of an optimality gap. The following alternative approach is suggested in Nemirovski and Shapiro [71] and Pagnoncelli et al. [78]. Consider problem (8), and suppose the objective function is probabilistic constraints. Following Nemirovski and Shapiro [71], let $\nu$ be the statistical lower bound for the optimal value of (8), as it can be shown that $\nu \leq \nu^\star$(note that $\nu$ is random). Then, it is possible to show that $P(p(x) < U(x)) \geq 1 - \beta$. This suggests the following procedure: given a realization $\{\hat{\xi}^1, \hat{\xi}^2, \ldots\}$ of $\{\xi^1, \xi^2, \ldots\}$, compute the values of $\hat{p}(x)$ and $U(x)$ corresponding to that sample; if $U(x) \leq \alpha$, then we are 100$(1 - \beta)\%$ confident that $x$ is feasible.

2.3 Problems with stochastic dominance constraints

We turn now to the class of optimization problems with stochastic dominance constraints. Stochastic dominance is used to compare the distributions of two random variables (e.g., see Müller and Stoyan [66]), thus providing a way to measure risk. Dentcheva and Ruszczyński [20, 21] first introduced optimization problems with stochastic dominance constraints as an attractive approach for managing risks in an optimization setting. While pursuing expected profits, one avoids high risks by choosing options that are preferable to a random benchmark. Recently, optimization
models using stochastic dominance have increasingly been the subject of theoretical considerations and practical applications in areas such as finance, energy, and transportation.

2.3.1 Review of stochastic dominance

For completeness, we briefly review the main concepts of stochastic dominance. Given a real-valued random variable \( Z \), we write the cumulative distribution function of \( Z \) as 
\[
F_1(Z; \eta) = P(Z \leq \eta).
\]
Furthermore, for \( n \geq 2 \), define recursively the functions 
\[
F_n(Z; \eta) := \int_{-\infty}^{\eta} F_{n-1}(Z; t) \, dt,
\]
assuming that the first \( n-1 \) moments of \( Z \) are finite. We then say that \( Z \) stochastically dominates another random variable \( Y \) in \( n \)th order (denoted \( Z \succeq (n) Y \)) if 
\[
F_n(Z; \eta) \leq F_n(Y; \eta) \quad \text{for all } \eta \in \mathbb{R}.
\]

Let \((a)_+ = \max\{a, 0\}\). It is useful to note the equivalence given in Ogryczak and Ruszczyński [74] for \( n \geq 2 \)
\[
F_j(Z; \eta) = \frac{1}{(j-1)!} \mathbb{E} \left[ ((\eta - Z)_+)^{j-1} \right], \quad j = 2, \ldots, n,
\]
which implies that the condition that the first \( n-1 \) moments of \( Z \) are finite suffices to ensure that \( F_n(Z; \eta) < \infty \) for all \( \eta \).

The concept of stochastic dominance is also related to utility theory (von Neumann and Morgenstern [98]), which hypothesizes that for each rational decision maker there exists a utility function \( u \) such that the (random) outcome \( Z \) is preferred to the (random) outcome \( Y \) if \( \mathbb{E}[u(Z)] \geq \mathbb{E}[u(Y)] \). Often the decision maker’s exact utility function is not known; in such cases one would say that \( Z \) is preferred to \( Y \) if \( \mathbb{E}[u(Z)] \geq \mathbb{E}[u(Y)] \) for all \( u \) belonging to a certain set of functions. This set of functions is determined by the risk attitude—for example, a risk-averse decision maker’s utility function is nondecreasing and concave. To see the connection with the notions of stochastic dominance defined above (for \( n = 1, 2 \)), let \( \mathcal{U}_1 \) be the set of all nondecreasing functions \( u : \mathbb{R} \rightarrow \mathbb{R} \) and let \( \mathcal{U}_2 \) be the set of all nondecreasing concave functions \( u : \mathbb{R} \rightarrow \mathbb{R} \). Then, it is well known that
\[
Z \succeq (n) Y \iff \mathbb{E}[u(Z)] \geq \mathbb{E}[u(Y)], \quad \forall u \in \mathcal{U}_n,
\]
whenever the expectations exist. Stochastic dominance is also closely related to concepts of stochastic ordering; for example, the condition \( \mathbb{E}[u(Z)] \geq \mathbb{E}[u(Y)] \) for all \( u \in \mathcal{U}_2 \) is called stochastic increasing concave order (see, e.g. Shaked and Shanthikumar [89]).
2.3.2 Basic properties and reformulations

Using the above concepts, an optimization model with stochastic dominance constraints can then be formulated as follows (Dentcheva and Ruszczyński [20, 21]):

\[
\begin{align*}
\min & \quad g_0(x) \\
\text{s.t.} & \quad H(x, \xi) \succeq (n) Y
\end{align*}
\]

\(x \in X.\)  

The cases that have received most attention in the literature are \(n = 1\) and \(n = 2.\) The difficulties with the \(n = 1\) case are similar to those arising with probabilistic constraints, notably nonconvexity. The case \(n = 2,\) on the other hand, is a convex problem so long as \(g_0(\cdot)\) is convex, \(H(\cdot, \xi)\) is concave and the set \(X\) is convex—indeed, the equivalence (12) allows us to write the problem with expected value constraints, yielding

\[
\begin{align*}
\min & \quad g_0(x) \\
\text{s.t.} & \quad \mathbb{E} \left[ \left( \eta - H(x, \xi) \right)_+ \right] \leq \mathbb{E} \left[ \left( \eta - Y \right)_+ \right] \quad \forall \eta \in \mathbb{R}
\end{align*}
\]

\(x \in X,\) which is a convex program.

In principle, problem (15) falls into the general framework of Section 2.1, as it contains expected-value constraints. A major difference, however, is the fact that (15) has one constraint for each \(\eta \in \mathbb{R}\)—that is, it has uncountably many constraints. This issue is circumvented when the random variable \(Y\) has finitely many outcomes \(y_1, \ldots, y_r;\) in that case, Dentcheva and Ruszczyński [20] show that it suffices to write the constraints in (15) only for \(\eta = y_j, \ j = 1, \ldots, r,\) thus yielding a problem with finitely many expected-value constraints. When the distribution of \(\xi\) also has finite support, the expectations in (15) can be written as sums, so the problem becomes deterministic. When \(Y\) has infinitely many outcomes, it is natural to resort to sampling methods; note however that the analysis is more delicate than that described in Section 2.1 since it involves not only approximating the expectations but also sampling over the set of (uncountably many) constraints. We will discuss this issue further shortly.

It is also useful to consider the case when the function \(H\) in (14) is vector-valued, which we write as \((H_1(x, \xi), \ldots, H_m(x, \xi)).\) This situation occurs in many practical settings—for example, when \(H(x, \xi)\) is a linear function of the form \(A(\xi)x,\) where \(A(\xi)\) indicates a random matrix. Of course, in this case, \(Y\) is also an \(m-\)dimensional random vector. We have then two alternatives: one is to write the problem with \(m \) one-dimensional stochastic dominance constraints; i.e., \(H_j(x, \xi) \succeq (n) Y_j, \ j = 1, \ldots, m.\) Even though such a formulation provides a direct extension of the unidimensional case seen above, it disregards the dependence among the components \(H_j\) of \(H.\) Alternatively, we can use concepts of multivariate stochastic dominance. One such concept is that of convex dominance introduced by Hu et al. [42]. Given
m-dimensional random vectors Z and Y and a convex set \( \mathcal{C} \subset \mathbb{R}^m \), we say that Z dominates Y in nth order linearly with respect to \( \mathcal{C} \) if

\[
v^T Z \succeq_{(n)} v^T Y \quad \text{for all } v \in \mathcal{C}.
\]  

(16)

Note that the notion of convex dominance includes as a particular case the concept of positive linear dominance (see, e.g., Müller and Stoyan [66]), which corresponds to \( \mathcal{C} = \mathbb{R}^m_+ \). Under convex dominance problem (15) is then written as

\[
\min_{x \in X} g_0(x)
\]

s.t. 1
\[
\mathbb{E} \left[ \left( \eta - v^T H(x, \xi) \right)_+ \right] \leq \mathbb{E} \left[ \left( \eta - v^T Y \right)_+ \right] \quad \forall \eta \in \mathbb{R}, \forall v \in \mathcal{C}
\]

(17)

Homem-de-Mello and Mehrotra [38] extend the aforementioned results of Dentcheva and Ruszcyński [20] and show that, when Y has finitely many outcomes \( y_1, \ldots, y_r \) and the set \( \mathcal{C} \) is polyhedral, the dominance relationship (16) for \( n = 2 \) can be written as

\[
\mathbb{E} \left[ (v^T_k y_j - v^T_k H(x, \xi))_+ \right] \leq \mathbb{E} \left[ (v^T_k y_j - v^T_k Y)_+ \right] \quad j = 1, \ldots, r, k = 1, \ldots, K,
\]

where \( v_1, \ldots, v_K \) are certain vectors in the set \( \mathcal{C} \). Thus, in that case the problem still has finitely many expected-value constraints.

2.3.3 The SAA approach and solution methods

Hu et al. [42] provide an analysis of sampling approximations to problem (17) (which includes (15) as a particular case). The corresponding sample average approximation is written as

\[
\min_{x \in X} g_0(x)
\]

s.t. 1
\[
\frac{1}{N} \sum_{j=1}^{N} \left( (v^T_k y^j - (v^T_k H(x, \xi^j)))_+ \right) \leq \frac{1}{N} \sum_{j=1}^{N} \left( (v^T_k y^j - (v^T_k Y^j))_+ \right) 
\]

(18)

In the above formulation, \{ (\xi^j, Y^j) \}, \( j = 1, \ldots, N \) is a sample of size \( N \) from \( (\xi, Y) \), and the \{v^j_k\}, \( i = 1, \ldots, N \) are certain vectors in \( \mathcal{C} \). Typically, the \( v^j_k \) vectors are unknown in advance; to remedy the problem, a cutting-surface algorithm based on the ideas in Homem-de-Mello and Mehrotra [38] is proposed. Hu et al. [42] show that the algorithm converges in finitely many iterations to an optimal solution of (18). Moreover, the feasibility set \( U_N \) of (18) satisfies
\( P \left( U^{-\varepsilon} \subseteq U_N \subseteq U^\varepsilon \right) \geq 1 - Me^{-Be^{2N}} \)

where \( U^\varepsilon \) is the feasibility set corresponding to the dominance constraint in (15) perturbed by \( \varepsilon \) on the right-hand side.

### 2.3.4 Assessing solution quality

Statistical lower and upper bounds (e.g., to assess solution quality) can also be derived for the approximation (18). Hu et al. [42] propose procedures which are based on similar ideas to those discussed in Section 2.1—more specifically, a Lagrangian-based relaxation for the lower bound, and the objective value of a feasible solution for the upper bound—but with the necessary adaptation to the setting of (17). Zhang and Homem-de-Mello [100] discuss an alternative procedure for the case where \( H \) is real-valued (rather than vector-valued) but the set \( X \) is nonconvex (for example, discrete). The basic idea is to formulate a hypothesis test to check feasibility of a given solution, and then use a multiple-replication procedure similar to that described in Section 2.1—but modified to discard certain replications—to calculate an optimality gap. We refer to that paper for details.

### 3 Variance reduction techniques

Monte Carlo sampling-based approximations and algorithms can be significantly improved by reducing the variability of the estimates they generate. Variance reduction techniques have a long history in the simulation and statistics literature. The main goal of such methods is to provide estimators of values associated with a random variable that have better properties than the standard Monte Carlo estimators. Consider for example the quantity \( g_0(x) \) defined in (SP) for a fixed \( x \in X \) and its sample average estimator defined in (1). When the sample \( \{\xi^1, \ldots, \xi^N\} \) is independent and identically distributed, its variance is given by

\[
\text{Var}[g_N(x)] = \frac{\text{Var}[G_0(x, \xi)]}{N}.
\]

Although \( \text{Var}[G_0(x, \xi)] \) is typically unknown, it can be estimated by a sample variance as follows:

\[
S_N^2(x) := \frac{\sum_{i=1}^{N}(G_0(x, \xi^i) - g_N(x))^2}{N - 1}.
\]

The above estimator is unbiased, i.e., \( \mathbb{E}[S_N^2(x)] = \text{Var}[G_0(x, \xi)] \).

Of course, it is desirable to have estimators with as small variance as possible. While this is the case in the context of pointwise estimation, it is even more so in the case of optimization, since poor estimates of the objective (or of its derivatives) may lead to slow convergence of an algorithm. Clearly, if \( \text{Var}[G_0(x, \xi)] \) is large then
Var[\(g_N(x)\)] will be large as well, unless the sample size can be chosen to counterbalance that effect. In many cases, however, choosing a large sample size is not practical, as the evaluation of \(G_0(x, \xi)\) for a given \(\xi\) can be costly.

The goal of variance reduction techniques is to derive estimators \(g_N(x)\), \(\nu_N\), etc. with smaller variance than those obtained with standard Monte Carlo. While in some cases this is accomplished by exploiting the structure of the problem, some general techniques do exist. We discuss next some variance reduction methods, mostly in the context of stochastic optimization problems though we do provide some background for the case of pointwise estimators (i.e., for a fixed \(x \in X\)).

In our presentation below, we revert to the case where there are no stochastic constraints (\(\mathcal{K} = \emptyset\) in (SP)) and we drop the subscript 0 from the objective function in (SP). Also, for some of the methods discussed below we assume that the vector \(\xi\) has independent components. When such an assumption does not hold one can often write the components of \(\xi\) as functions of some independent uniform random variables; see, for instance, Biller and Ghosh [9].

### 3.1 Antithetic Variates

Antithetic Variates (AV) aim to reduce variance by inducing correlations. Suppose \(N\) is even and components of \(\xi\) are independent. Instead of using \(N\) i.i.d. random variates, the AV estimator aims to use \(N/2\) negatively correlated pairs \((\xi_j, \overline{\xi}_j)\), \(j = 1, \ldots, N/2\). This is typically achieved by generating \(N/2\) i.i.d. random vectors \(U^1, \ldots, U^{N/2}\) of dimension \(d_\xi\) distributed uniformly over \([0, 1]^{d_\xi}\), and their corresponding antithetic variates from the opposite end of the distribution (taken component-wise), \(1 - U^1, \ldots, 1 - U^{N/2}\), which are also i.i.d. distributed uniformly over \([0, 1]^{d_\xi}\) but \((U^j, 1 - U^j), j = 1, \ldots, N/2\) are negatively correlated. Then, regular variate generation techniques are utilized to generate the pairs \((\xi_j, \overline{\xi}_j)\) using \((U^j, 1 - U^j)\). For the case with dependent components of \(\xi\), we refer to Rubinstein et al. [87].

In contrast to the standard Monte Carlo estimator of \(\mathbb{E}[G(x, \xi)]\) with variance \(N^{-1}\sigma^2(x) := N^{-1} \text{Var}[G(x, \xi)]\), the antithetic variates estimator

\[
g_{N, \text{AV}}(x) = \frac{1}{N/2} \sum_{j=1}^{N/2} \left( G(x, \xi_j) + G(x, \overline{\xi}_j) \right) / 2
\]

has variance \(N^{-1}\sigma^2(x) + N^{-1} \text{Cov}(G(x, \xi_j), G(x, \overline{\xi}_j))\). Therefore, as long as \(\text{Cov}(G(x, \xi_j), G(x, \overline{\xi}_j)) < 0\), the antithetic estimator has a smaller variance than its crude Monte Carlo counterpart and they both produce unbiased estimators of the expectation.

The degree of variance reduction depends on the extent to which the negative correlation between the pair \((\xi_j, \overline{\xi}_j)\) is preserved after \(G(x, \cdot)\) is applied to this pair. Higle [35] argues that the negative correlation can be preserved for a class of two-
stage stochastic linear programs with stochasticity only on the right-hand side and presents computational results. This is because $G(x, \cdot)$ is a monotone function of the right-hand side of the second stage problem for this class of problems. In general, if $G(x, \cdot)$ is a bounded and monotone function in each of its arguments that is not constant in the interior of its domain, variance reduction can be achieved using AV (Lemieux [58]). Koivu [49] and Freimer et al. [27] expand the work of Higle [35] by also considering $\nu_N$, the optimized sample means. Freimer et al. [27] analytically show the extent of variance reduction using AV on a newsvendor problem and present computational results on two-stage stochastic linear programs. The computations in Koivu [49] indicate that when the monotonicity in the objective function is lost, AV can increase (e.g., double) the variance. However, when AV is effective, combination of AV with other variance reduction techniques such as randomized quasi-Monte Carlo is found to be very effective. These papers indicate that antithetic variates can result in modest variance reduction with minimal computational effort for a class of stochastic optimization problems.

3.2 Latin Hypercube Sampling

A fairly general way of obtaining estimators with smaller variance is based on the concept of stratified sampling (see, for instance, Fishman [26] and references therein). Generally speaking, the idea is to partition the sample space and fix the number of samples on each component of the partition, which should be proportional to the probability of that component. This way we ensure that the number of sampled points on each region will be approximately equal to the expected number of points to fall in that region. It is intuitive that such a procedure yields smaller variance than crude Monte Carlo; for proofs see Fishman [26]. Notice, however, that, though theoretically appealing, implementing such a procedure is far from trivial, since the difficulty is to determine the partition as well as to compute the corresponding probabilities.

There are many variants of this basic method; a classical one is the so-called **Latin Hypercube Sampling** (LHS) approach, introduced in McKay et al. [65]. The LHS method operates as follows. Suppose we want to draw $N$ samples from a random vector $\xi$ with $d_\xi$ independent components, each of which has a Uniform(0,1) distribution. The algorithm consists repeating the two steps below for each dimension $j = 1, \ldots, d_\xi$:

1. Generate
   \[ Y^1 \sim U \left(0, \frac{1}{N}\right), \quad Y^2 \sim U \left(\frac{1}{N}, \frac{2}{N}\right), \ldots, \quad Y^N \sim U \left(\frac{N-1}{N}, 1\right) ; \]

2. Let $\xi_j := Y^{\pi(i)}$, where $\pi$ is a random permutation of $1, \ldots, N$. 
In McKay et al. [65], it is shown that each sample $\xi_i$ (viewed as a random variable) has the same distribution as $\xi_j$, which in turn implies the estimators generated by the LHS method are unbiased. In case of arbitrary distributions, the above procedure is easily modified by drawing the sample as before and applying an inversion method.

It is also shown in McKay et al. [65] that, under some conditions, the LHS method does indeed reduce the variance compared to standard Monte Carlo. Stein [96] and Owen [76] show that, asymptotically (i.e., as the sample size $N$ goes to infinity), LHS is never worse than standard Monte Carlo, even without the assumptions of McKay et al. [65]. More specifically, $V_{LHS} \leq N/(N-1)V_{MC}$, where $V_{LHS}$ and $V_{MC}$ are respectively the variances under LHS and standard Monte Carlo.

Thanks to such properties (and to the simplicity of the method), the LHS technique has been widely used in simulation, showing up even in off-the-shelf spreadsheet-based software.

One drawback of using the LHS method is that, by construction, the generated samples are not independent—indeed, variance is reduced precisely because of the correlation introduced by the method. Lack of independence implies that classical statistical results such as the Central Limit Theorem do not apply directly to the resulting estimator; consequently, confidence intervals cannot be built in the standard way. It is worthwhile mentioning that Owen [75] proves a version of the CLT for LHS estimators, which is useful from the perspective of rates of convergence but not necessarily for the construction of confidence intervals as the result involves some quantities that are difficult to estimate. In practice, in order to derive confidence intervals one typically performs multiple independent replications (for example, $m$ replications, each with a sample of size $N$) and applies the classical theory to the data set consisting of the LHS average estimator from each replication.

The use of LHS in stochastic optimization, while not as widespread as in simulation, has demonstrated the benefits of that approach, as reported in papers such as Bailey et al. [4], Shapiro et al. [93], Linderoth et al. [60], Homem-de-Mello et al. [39], Zhang and Homem-de-Mello [100]. Freimer et al. [27] study in detail the effect of using LHS in the context of the newsvendor model, and draw some general conclusions about the effectiveness of that approach for stochastic optimization.

It is possible to study convergence of estimators of optimal values and optimal solutions of (2) with sampling that is not necessarily i.i.d. under a general theoretical framework. Throughout the discussion below, assume that $X$ is compact. Suppose that for each $x \in X$ we have that

$$g_N(x) \rightarrow g(x) \quad \text{w.p.1.}$$

Suppose also that the function $G(\cdot, \xi)$ is Lipschitz. Then, we have that

(i) $g_N(x) \rightarrow g(x)$ uniformly on $X$ w.p.1;
(ii) $\nu_N \rightarrow \nu^*$ w.p.1;
(iii) $\text{dist}(x_N, S^*) \rightarrow 0$ w.p.1.

Such a result is well known; see, e.g., Rubinstein and Shapiro [88, p.67-70]. Under further assumptions on $G(\cdot, \xi)$ and $X$ (for example, $G(\cdot, \xi)$ is piecewise linear and
the support of the distribution of $\xi$ is finite, or the feasible set $X$ is finite) the conclusion (iii) can be replaced with $x_N \in S^*$ w.p.1 for $N$ large enough. For reference, we shall call that condition the “piecewise-linear/finite-set assumption” for short.

The study of rates of convergence of estimators of optimal values and optimal solutions in this context is introduced by Homem-de-Mello [36]. In what follows we summarize the results from that paper, to where we refer the reader for details. Suppose that for each $x \in X$, there exist a number $C_x > 0$ and a function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$, $\gamma_x(z) > 0$ if $z > 0$, and

$$P(|g_N(x) - g(x)| \geq \delta) \leq C_x e^{-N\gamma_x(\delta)}$$

for all $N \geq 1$ and all $\delta > 0$. (20)

That is, the probability that the deviation between $g_N(x)$ and $g(x)$ is bigger than $\delta$ goes to zero exponentially fast with $N$ (notice that (20) implies that $g_N(x)$ converges in probability to $g(x)$). Suppose also that the function $G(\cdot, \xi)$ is Lipschitz. Then, it can be shown that given $\epsilon > 0$, there exist constants $K > 0$ and $\alpha > 0$ such that

$$P(\text{dist}(x_N, S^*) \geq \epsilon) \leq Ke^{-\alpha N}$$

for all $N \geq 1$. (21)

In other words, if an exponential rate of convergence holds for the pointwise estimators $g_N(x)$, then it will hold for the estimators of optimal solutions. The result in (21) can be strengthened in case under the piecewise-linear/finite-set assumption, as it can be shown that

$$P(x_N \notin S^*) \leq Ke^{-\alpha N}$$

for all $N \geq 1$.

Results for convergence of estimators of optimal value can be derived under the piecewise-linear/finite-set assumption. A simplified description of the results from Homem-de-Mello [36] goes as follows. Suppose that the “true” problem (SP) has a unique optimal solution $x^*$, and that

$$\frac{g_N(x^*) - g(x^*)}{\sigma_N(x^*)} \overset{d}{\to} \text{Normal}(0, 1).$$

(22)

where $\sigma_N^2(x) := \text{Var}[g_N(x)]$ and $\overset{d}{\to}$ denotes convergence in distribution. Then,

$$\frac{v_N - v^*}{\sigma_N(x^*)} \overset{d}{\to} \text{Normal}(0, 1).$$

(23)

Similarly to the result for estimators of optimal solution, the above results states that if a Central Limit Theorem holds for the pointwise estimators $g_N(x)$, then it will hold for the estimators of optimal values.

The above framework allows us to study convergence of estimators of optimal values and optimal solutions of (2) under LHS. Indeed, it can be shown that condition (19) holds under finiteness of second moment of $G(x, \xi)$; condition (20) holds under certain monotonicity assumptions on $G(x, \cdot)$; and condition (22) holds if the random variable $G(x^*, \xi)$ is bounded, and the function $G(x^*, \cdot)$ is not additive.
3.3 Quasi-Monte Carlo

Quasi-Monte Carlo (QMC) methods have a long history as tools to approximate integrals, and as such have been widely used in many areas. Describing all the nuances and properties of such methods would fall out of the scope of this paper; thus, we only provide a brief discussion. We refer to Niederreiter [72], Lemieux [58] and Dick and Pillichshammer [23] for comprehensive treatments of QMC concepts. To set the stage, consider again the function $G(x, \xi)$ and assume that $\xi$ is a random vector with independent components, each with uniform distribution on $[0, 1]^d$. Consider the problem of estimating $g(x) := \mathbb{E}[G(x, \xi)]$ for a fixed $x$.

The basic idea of QMC is to calculate a sample average estimate as in the standard Monte Carlo but, instead of drawing a random sample from the uniform distribution on $[0, 1]^d$, a certain set of points $\xi^1, \ldots, \xi^N$ on space $[0, 1]^d$ is carefully chosen. The deterministic estimate

$$g_{N, \text{QMC}}(x) := \frac{1}{N} \sum_{i=1}^{N} G(x, \xi^i)$$

is constructed. A key result is the so-called Koksma-Hlawka inequality which, roughly speaking, states that the quality of the approximation given by $g_{N, \text{QMC}}(x)$ depends on the quality of the chosen points (measured by the difference between the corresponding empirical measure and the uniform distribution, which is quantified by the so-called star-discrepancy) as well as on the nature of the function $G(x, \cdot)$ (measured by its total variation). A great deal of the research on QMC methods aims at determining ways to construct low-discrepancy sequences, i.e., sequences of points $\xi^1, \xi^2, \ldots$ for which the star-discrepancy is small for all $N$. Particular types of sequences that have proven valuable are the so-called digital nets and also lattice rules. We briefly describe them next.

Let $b \geq 2$ be an arbitrary integer, called the base. An elementary interval in base $b$ (in dimension $s$) is a subinterval $E$ of $[0, 1]^s$ of the form

$$E = \prod_{j=1}^{s} \left[ \frac{a_j}{b^{d_j}}, \frac{a_j+1}{b^{d_j}} \right]$$

for nonnegative integers $\{a_j\}$ and $\{d_j\}$ such that $a_j < b^{d_j}$ for all $j$. The volume of $E$ is $b^{-\sum d_j}$. Next, let $t$ and $m$ be nonnegative integers such that $t \leq m$. A finite sequence of $b^m$ points is a $(t, m, s)$-net in base $b$ if every elementary interval in base $b$ of volume $b^{-m}$ contains exactly $b^t$ points of the sequence. A sequence of points $u^1, u^2, \ldots$ is a $(t, s)$-sequence in base $b$ if, for all integers $k \geq 0$ and $m > t$, the set of points consisting of the $u^n$ such that $kb^m \leq n < (k+1)b^m$ is a $(t, m, s)$-net in base $b$. Such $(t, m, s)$-nets can be shown to yield approximations such that the integration error is of order $(\log N)^t / N$.

Many of the lattice rules in $[0, 1]^3$ used in practice are generated as follows: $\xi_i = (i/N)\nu_0$ mod $1$ for $i = 0, \ldots, N-1$ where $\nu_0$ is the generating vector, whose components are integers between 0 and $N-1$. There are some general principles
to choose the generating vector. L’Ecuyer and Munger [56] discuss a new approach whereby the lattice rules are tailored to the function being integrated. They show that such method leads to convergence rates that sometimes approach the theoretical optimal rate of $1/N^2$, provided the function is sufficiently smooth.

Despite the theoretical attractiveness of QMC methods with respect to error rates, an issue that arises when using such techniques in practice is the fact that the bounds provided by the Koksma-Hlawka inequality involve difficult-to-compute quantities such as the total variation of $G(x, \cdot)$. In other words, they yield qualitative rather than quantitative results; hence, obtaining a good estimate of the error may be difficult. A common way to overcome this issue is to incorporate some randomness into the choice of QMC points. By doing so, errors can be estimated using standard methods, e.g., via multiple independent replications. Some choices for randomizing the points of the QMC sequence include the so-called Cranley-Patterson procedure—where every number in the sequence is perturbed (modulo 1) by a single number generated from a Uniform(0,1) distribution—and scrambling the digits of each number in the sequence in a particular way; we refer to L’Ecuyer and Lemieux [54] for details.

Even with the randomization, one should be cautious when using QMC methods since oftentimes these techniques “backfire” in problems with moderate or large dimensionality if not used properly. This can be explained by the fact that the error rates depend on the dimensionality—for example, the $(\log N)^{d(2)/N}$ rate seen above. In such cases, one may try to determine the effective dimension of the problem, i.e., the number of variables that account for most of the variability, and then apply a QMC strategy only for those variables. Such notion can be made precise, see for instance Owen [76, 77]. Moreover, the theoretical rates derived for QMC often rely on smoothness of the integrand, which may not always be present. Still, when used properly such techniques can be highly valuable, resulting in estimators that are orders of magnitude better than standard Monte Carlo.

Similarly to what happens with standard random numbers, generating a good QMC sequence may not be simple. Some sequences are easy to generate but the more powerful ones require sophisticated methods. Fortunately, public software is available—here we mention Friedel and Keller [28], Lemieux et al. [59], L’Ecuyer [53], L’Ecuyer and Munger [55], where libraries can be found.

A few papers study the application of QMC methods to stochastic optimization problems. In Kalagnanam and Diwekar [45], empirical results are provided for the use of Hammersley sequences (one form of QMC). Drew and Homem-de-Mello [25] use the aforementioned concept of effective dimension to develop an algorithm for two-stage stochastic programs that attempt to determine the “important variables” in the problem based on dual information. The remaining variables are “padded” with either Monte Carlo or Latin Hypercube sampling; a rigorous analysis of such strategy can be found in Drew [24]. As in the case of LHS discussed in Section 3.2, theoretical results on convergence are harder to obtain than in the Monte Carlo case due to the loss of the i.i.d. property. In Pennanen [81] and Pennanen and Koivu [82], the authors show that, under mild assumptions, the estimator function $g_N$ constructed with QMC points epiconverges to the true function $g$, which guarantees convergence of optimal values and optimal solutions under appropriate further
conditions. In Koivu [49] those results are applied to the case where the QMC sequence is randomized with the Cranley-Patterson procedure. The numerical results in those papers also suggest considerable gains in terms of rates of convergence when using QMC methods. Homem-de-Mello [36] studies the rates of convergence of estimators of optimal values under randomized QMC, using the general framework for non-i.i.d. sampling described in Section 3.2. Results are provided for a specific QMC method for which a Central Limit Theorem exists, so that (22) holds. As it turns out, in that case it is possible to show that the estimators of optimal values converge at the rate \($\log N/d_\xi - 1/N^3\)^{1/2}$. One particular difficulty that arises when using QMC methods for stochastic programming problems lies in the fact that such problems do not have smooth integrands. Recent work by Heitsch et al. [34] sheds new light on that issue by showing that, under certain assumptions, all terms of the so-called ANOVA decomposition of such functions except the residual term are actually infinitely differentiable.

3.4 Importance Sampling

Importance Sampling (IS) aims to reduce variance by “shifting” the samples to the most important regions. Suppose again the aim is to estimate $E[G(x, \xi)]$ for a given $x \in X$ and suppose $\xi$ has density $f$. Then, $g(x) = E[G(x, \xi)] = \int_\Xi G(x, \xi)f(\xi)d\xi$. Now consider another density $q$ over $\Xi$ with the property that $f(E) = 0$ for every set $E$ for which $q(E) = 0$, and rewrite $E[G(x, \xi)] = \int_\Xi G(x, \xi)L(\xi)q(\xi)d\xi$. Here, $L(\xi) = \frac{f(\xi)}{q(\xi)}$ is the likelihood ratio, which we assume is well-defined (for this, we may set $L$ to zero whenever both $f$ and $q$ are zero). Instead of the usual Monte Carlo estimator $g_N(x) = \frac{1}{N}\sum_{j=1}^{N} G(x, \xi_j)$ that uses an i.i.d. sample $\xi_1, \xi_2, \ldots, \xi_N$ from the density $f$, the importance sampling estimator

$$g_{N, IS}(x) = \frac{1}{N} \sum_{j=1}^{N} G(x, \tilde{\xi}_j)L(\tilde{\xi}_j)$$

uses an i.i.d. sample $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_N$ from the new density $q$. Note that both estimators are unbiased. However, not all choices of density $q$ will lead to a reduction in variance. In fact, if $q$ is not chosen appropriately, it might lead to an increase in variance. Therefore, finding an appropriate density $q$ is critical to IS and much of the research in this area is directed to this issue.

To understand how to find $q$, consider the following facts. First, $E[G(x, \xi)L(\xi)] = E[G(x, \xi)]$ and $E[G^2(x, \xi)L^2(\xi)] = E[G^2(x, \xi)L(\xi)]$. Therefore, the variance of the IS estimator is given by

$$\text{Var}[g_{N, IS}(x)] = \frac{1}{N} \left[ E[G^2(x, \xi)L(\xi)] - (E[G(x, \xi)])^2 \right],$$

(25)
and the variance is reduced if and only if $\mathbb{E}[G^2(x, \xi)\mathcal{L}'(\xi)] < \mathbb{E}[G^2(x, \xi)]$. If we want to minimize the variance—i.e., have zero variance—in (25), we need to have $q$ such that $\mathbb{E}[G^2(x, \xi)\mathcal{L}'(\xi)] = (\mathbb{E}[G(x, \xi)])^2$. When $G(x, \cdot)$ is nonnegative, this results in the optimal zero-variance density

$$q^*(\xi) = \frac{f(\xi)G(x, \xi)}{\mathbb{E}[G(x, \xi)]}.$$ 

This optimal density, however, requires the knowledge of an unknown quantity, $\mathbb{E}[G(x, \xi)]$, and is therefore unattainable. Nevertheless, this gives the intuition that $q$ should be approximately proportional to $f(\xi)G(x, \xi)$ in order to achieve variance reduction even though the proportionality constant may not be known. Here we see the difficulty of using IS in the optimization context: it is clear that the above expression for $q^*(\xi)$ cannot hold for all $x \in X$, so the same IS density may be good for some $x$ but bad for others.

In the context of pointwise estimation, there is significant work on how to determine the IS distribution in the literature as this is critical to the success of IS. We briefly mention some of that work without getting into much detail. In exponential titling, the IS distribution is restricted to belong to an exponential family of distributions (Glasserman [30], Siegmund [95]). In Rubinstein and Shapiro [88], a method is discussed which parameterizes the IS distribution and solves a stochastic optimization problem to determine the parameters of this distribution in order to minimize variance. This optimization can be done via a sampling-based method and perturbation analysis; see, e.g., Fu [29]. Other approaches in the literature to obtain the IS distribution include nonparametric methods (Neddermeyer [67], Zhang [101]) and minimization of the Kullback-Leibler distance to the optimal distribution (de Boer et al. [19]). Importance sampling techniques have been very successful in the estimation of rare event probabilities, especially due to the connection with large deviations theory; see, for instance, Glasserman et al. [31] and also Blanchet and Lam [11] for a recent survey of that area.

Importance sampling is one of the earliest variance reduction methods applied to two- and multi-stage stochastic linear programs (Dantzig and Glynn [18], Infanger [44], Dantzig and Infanger [17]). Dantzig and Glynn [18] suggest using an additive approximation of $G(x, \xi)$ given by $G(x, \xi) + \sum_{i=1}^{d\xi} \Delta_i G(x, \xi)$, where $\xi$ is a base case and each $\Delta_i G(x, \xi)$ gives the marginal effect of the $i$th element of $\xi$. They suggest finding the marginal effects by $\Delta_i G(x, \xi) = G(x, \xi_{i}) - G(x, \xi)$, where $\xi_{i}$ agrees with the base case $\xi$ in all elements except for the $i$th element, which agrees with $\xi$. This results in solving $d\xi + 1$ linear programs to determine $q$, one for each marginal and one for the base case. The authors argue that in the context of power generation, IS can capture rare events such as power supply down and high demands better than crude Monte Carlo, which are supported by the computations in Infanger [44]. Higle [35] applies this method to a wider range of problems, and the computations indicate that the method can be effective in reducing variance of $\mathbb{E}[G(x, \xi)]$ estimates for some problems, but it can also actually increase variance for some other problems.
As mentioned earlier, a major difficulty that arises when employing IS methods in the context of iterative algorithms for more general stochastic optimization problems is that changes in $x$ throughout an optimization routine can result in different IS distributions for different $x$. Shapiro and Homem-de-Mello [92] discuss the idea of using trust regions on which the same IS distribution can be used, but the results are inconclusive. Barrera et al. [5] employ IS techniques for a chance-constrained problem in which the violation probabilities $\alpha_k$ in (8) are very small, so the methods for choice of sample sizes discussed in Section 2.2 are not practical. They show that, for the application they study, there exist IS distributions that are good for all $x \in X$. Recent work of Kozmík and Morton [50] apply IS within the stochastic dual dynamic programming algorithm for multi-stage stochastic programs with nested mean-CVaR objectives and show promising results. As only relatively few scenarios under random sampling contribute to estimating CVaR, an IS scheme provides better estimators by concentrating the sampling to the important regions. A thorough study of sequential IS methods in the context of optimization remains an open research area (Birge [10]).

### 3.5 Control Variates

Like antithetic variates, Control Variates (CV) aim to reduce variance by inducing correlations. In the case of control variates, though, this is achieved by introducing a control variable that can either be negatively or positively correlated with $G(x, \xi)$. Let $C$ denote the control variable and $\lambda$ be a scalar. Suppose $\mathbb{E}[C] = 0$. Note that if the mean of the control variable is known, which is often the case, then it can be subtracted from it to obtain a variable with zero mean. The control variate estimator of $\mathbb{E}[G(x, \xi)]$ is given by

$$ g_{N, CV}(x) = \frac{1}{N} \sum_{j=1}^{N} \left( G(x, \xi^j) + \lambda C^j \right). $$

For any given $\lambda$, $g_{N, CV}(x)$ is an unbiased estimator with variance

$$ \frac{1}{N} \left( \sigma^2(x) + \lambda^2 \text{Var}[C] + 2\lambda \text{Cov}[G(x, \xi), C] \right). \quad (26) $$

We can minimize this variance by setting $\lambda$ to $\lambda^* = -\frac{\text{Cov}[G(x, \xi), C]}{\text{Var}[C]}$. Plugging $\lambda^*$ back in (26), we see that as long as $C$ and $G(x, \xi)$ are correlated, the variance of the CV estimator

$$ \text{Var}[g_{N, CV}(x), \lambda^*] = \frac{1}{N} \left( \sigma^2(x) - \frac{\text{Cov}^2[G(x, \xi), C]}{\text{Var}[C]} \right) $$

is less than the variance of the crude MC estimator, $N^{-1} \sigma^2(x)$. Notice that even though \text{Var}[C] may be known, Cov$[G(x, \xi), \xi]$ is unknown but can be estimated, resulting in an estimator of $\lambda^*$. Unfortunately, when an estimator of $\lambda^*$ is used,
$g_{N, CV}(x)$ is no longer unbiased. However, this can still yield significant variance reduction and the resulting CV estimator obeys a CLT of the form

$$\sqrt{N}(g_{N, CV}(x) - \mathbb{E}[G(x, \xi)]) \overset{d}{\rightarrow} \text{Normal}(0, \text{Var}[g_{N, CV}(x), \lambda^*])$$

due to a result of Nelson [69].

Higle [35] presents a number of control variates to estimate $\mathbb{E}[G(x, \xi)]$ that are cheap to compute and are quite effective across a number of test problems. Shapiro and Homem-de-Mello [92] use linear control variates to obtain more accurate estimators of the gradient, Hessian, and the value of $\mathbb{E}[G(x, \xi)]$ at a current solution point $x$ in each iteration of a Monte Carlo sampling-based method to solve two-stage stochastic linear programs. Similarly, Pierre-Louis et al. [83] use a subgradient-inequality-based linear control variate within stratified sampling to estimate $\mathbb{E}[G(x, \xi)]$ at each iteration’s solution $x$ of an algorithm for a class of two-stage stochastic convex programs. Both papers show that control variates significantly reduce variance (up to more than 1,000 times in some cases) and allow these Monte Carlo sampling-based solution procedures to be numerically more viable. Control variates have also been successfully used in ranking and selection procedures. These are analyzed by Nelson and Staum [68] and Tsai and Nelson [97] with promising computational results.

4 Conclusions

In this chapter, we have surveyed the current landscape of Monte Carlo simulation-based methods for stochastic optimization problems with stochastic constraints and the use of variance reduction techniques in Monte Carlo approximations of stochastic optimization problems. For stochastically constrained problems, we reviewed asymptotic and finite sample size properties, and discussed solution methods and assessment of solution quality. Variance reduction techniques can significantly improve the quality of the Monte Carlo simulation-based estimators. We reviewed their use in the stochastic optimization setting, discussing asymptotic properties and their use within Monte Carlo simulation-based solution methods. In particular, we discussed the use of antithetic variates, Latin hypercube sampling, quasi-Monte Carlo, importance sampling, and control variates. One topic we skipped in VRT is the use of common random numbers (CRN), which aims to reduce variance by inducing positive correlations via correlated observations and typically applies when we are comparing two or more alternative systems (see, e.g., Law [52]). In the context of stochastic optimization, CRN techniques can be efficiently used, for instance, in assessing solution quality (see, e.g., Mak et al. [64]) or when ranking and selecting a system among a discrete set of alternatives (Kim and Nelson [47]).

We note that both topics covered in this chapter are active areas of research and new work is frequently appearing to address unresolved issues. Being able to efficiently solve problems with stochastic constraints brings us closer to solving
many real-world problems with multiple performance measures that involve randomness. Variance reduction techniques help make Monte Carlo sampling-based methods considerably more reliable and efficient; therefore, they are very important for the practical success of Monte Carlo methods in stochastic optimization. These two important topics are being tackled by the mathematical programming and simulation communities. Broadly speaking, in mathematical programming one typically exploits certain structures found in the problem, whereas the simulation literature typically works with problems that do not have much structure and can only be evaluated with expensive simulations. Both approaches have their merits and can be preferable in certain situations. That said, the line between the two has been becoming blurry, as this book illustrates. One area of future research is to further explore and exploit the connections between the simulation and the more structured mathematical programming approaches. For the increased success and applicability of these methods, data-driven methods need to be further developed along with software that link real-world data to stochastic optimization models to Monte Carlo simulation-based methods.

Acknowledgements The authors express their gratitude to Michael Fu for the invitation to write this chapter. This work is supported in part by the National Science Foundation under Grant CMMI-1151226, and by Conicyt-Chile under grant Fondecyt 1120244.

References


