Chapter 8
A Guide to Sample Average Approximation

Sujin Kim, Raghu Pasupathy, and Shane G. Henderson

Abstract This chapter reviews the principles of sample average approximation (SAA) for solving simulation optimization problems. We provide an accessible overview of the area and survey interesting recent developments. We explain when one might want to use SAA and when one might expect it to provide good-quality solutions. We also review some of the key theoretical properties of the solutions obtained through SAA. We contrast SAA with stochastic approximation (SA) methods in terms of the computational effort required to obtain solutions of a given quality, explaining why SA "wins" asymptotically. However, an extension of SAA known as retrospective optimization can match the asymptotic convergence rate of SA, at least up to a multiplicative constant.

8.1 Introduction

How does one solve an optimization problem of the form

$$\min_{x \in \Theta} f(x), \tag{8.1}$$

where $\Theta \subseteq \mathbb{R}^d$ $(d < \infty)$ and the real-valued function $f(\cdot)$ cannot be computed exactly, but can be estimated through a (stochastic) simulation? The principle of Sample Average Approximation (SAA) allows one to tackle such problems through the use of sampling and optimization methods for deterministic problems. We introduce SAA, describe its properties through both examples and theory, and relate

S. Kim
National University of Singapore, Singapore
e-mail: iseks@nus.edu.sg

R. Pasupathy
Purdue University, West Lafayette, IN, USA
e-mail: pasupath@purdue.edu

S.G. Henderson (✉)
Cornell University, Ithaca, NY, USA
e-mail: sgh9@cornell.edu

M.C. Fu (ed.), *Handbook of Simulation Optimization*, International Series in Operations Research & Management Science 216, DOI 10.1007/978-1-4939-1384-8_8

SAA to established concepts in stochastic simulation. Our goal is to communicate the essence of the idea and the key results in the area, rather than to provide an exhaustive discussion of what is known about SAA. As such, this chapter is best viewed as a *guide* rather than a *survey*. Similar guides for the strongly related area of stochastic programming at a more introductory level can be found in [58].

Throughout, we assume that the function f cannot be observed or computed directly, but we know that $f(x) = \mathsf{E}[Y(x,\xi)]$, where ξ is a random element with a distribution that does not depend on x, and $Y(\cdot,\cdot)$ is a (deterministic) real-valued function. Implicit in this statement is that for each fixed $x \in \Theta$, $\mathsf{E}|Y(x,\xi)| < \infty$. We suppress measure-theoretic considerations unless they come into play at a practical level. Nevertheless, we attempt to state results precisely.

In SAA, we select and fix $\xi_1, \xi_2, \ldots, \xi_n$, all having the same distribution as ξ, and set

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y(x,\xi_i).$$

Given the (fixed) sample $\xi_1, \xi_2, \ldots, \xi_n$, the function $f_n(\cdot)$ is deterministic, and so we can apply deterministic optimization algorithms to solve the problem

$$\min_{x\in\Theta} f_n(x). \tag{8.2}$$

We then take an optimizer, X_n^* say, of (8.2) as an estimator of an optimal solution of (8.1). Unless otherwise stated, we assume that $(\xi_1, \xi_2, \ldots, \xi_n)$ form an independent and identically distributed (i.i.d.) sample. The independence assumption is sometimes relaxed, mostly in variance reduction schemes or randomized quasi-Monte Carlo schemes where dependence is deliberately introduced.

A popular example to illustrate SAA is the continuous newsvendor problem where we buy x units of some commodity at a cost $c > 0$ per unit, observe demand ξ, and sell as many units as we can at price $s > c$; see, e.g., [57, p. 330]. The goal is to choose x so as to maximize profit. (Of course, one can convert this problem to a minimization problem as in (8.1) simply by multiplying by -1.) The profit for a given realization ξ is $Y(x,\xi) = s\min\{x,\xi\} - cx$. This function is concave in x, has slope $s - c > 0$ for sufficiently small x and slope $-c < 0$ for sufficiently large x. It follows that the same is true for the approximating function $f_n(\cdot)$, which therefore achieves its maximum. In fact, it is straightforward to show that an optimizer X_n^* of $f_n(\cdot)$ occurs at the $1 - c/s$ quantile of the empirical distribution associated with the observed demands $\xi_1, \xi_2, \ldots, \xi_n$, i.e., the $\lceil n(1-c/s)\rceil$th smallest of the observed demands. If we assume that the distribution of ξ is continuous at its $1 - c/s$ quantile, which is optimal for the true problem (e.g., [41, p. 353]), then X_n^* converges to this value as $n \to \infty$ almost surely (a.s.). So in this case, SAA is successful, in that the sequence of optimizers $\{X_n^*\}$ converges to a true optimizer.

In general, is SAA a reasonable approach? What kinds of problems are such that SAA works, in the sense that X_n^* can be expected to converge to the set of optimizers

of (8.1) as $n \to \infty$ in some sense? What kinds of problems are such that (8.2) is amenable to deterministic optimization algorithms? Is this procedure competitive with alternative algorithms, in the sense that the solutions returned after a given computational effort are comparable in quality?

Most of these questions have been addressed in previous surveys of SAA, so what is different here? We emphasize the intuition behind SAA, developing concepts through a range of examples as well as through theory. Mostly we do not prove results here, but instead give references to complete proofs, and provide proof sketches where that helps build understanding. Many of those proofs can be found in the excellent surveys [55–57].

It is hard to pin down the origin of the SAA concept. Certainly there are strong ties to maximum likelihood estimation and M-estimation in statistics, but perhaps the strongest roots of the idea from an Operations Research perspective lie in variants called the stochastic counterpart method [47, 48] and sample-path optimization [20, 37, 43].

We focus on the unconstrained optimization problem (8.1), but SAA can also encompass constrained optimization problems, even when the constraints must also be evaluated using simulation; see [60, 61] and the next chapter. The SAA principle is very general, having been applied to settings including chance constraints [1], stochastic-dominance constraints [24] and complementarity constraints [18].

The rest of this chapter is organized as follows. Section 8.2 provides a set of examples that showcase when SAA is appropriate in the sense that the optimization problem (8.2) has reasonable structure that allows for numerical solution. Section 8.3 provides verifiable conditions under which one can expect the problems (8.1) and (8.2) to share important structural properties such as continuity and differentiability. This section also showcases the close connection between problems that are "SAA appropriate" and those that are amenable to infinitesimal perturbation analysis (IPA) [13, 17] for gradient estimation. This section can also be viewed as a review of IPA with particular emphasis on multidimensional problems. In Sect. 8.4, we review some key properties of SAA, particularly with regard to large-sample performance. Sects 8.5 and 8.6 delve into the selection of the sample size n in some detail. These sections relate the computational effort required to achieve a given solution quality in SAA to that of a competing method known as stochastic approximation. It turns out that SAA is not as efficient as stochastic approximation, at least in the asymptotic sense. (In the non-asymptotic world of small sample sizes, it is harder to make clear conclusions, although some results are known for idealized versions of both approaches; see, e.g., [54].) This leads one to the class of methods collectively known as retrospective optimization, which is an extension of SAA. We review some recent results on retrospective optimization that show that this class of methods can match stochastic approximation in terms of asymptotic efficiency.

8.2 When Is SAA Appropriate?

One hopes the X_n^* obtained by solving the SAA problem converges to a solution of the true problem x^*. The critical condition for convergence is a uniform version of the strong law of large numbers (ULLN), which takes the form

$$\sup_{x\in\Theta}|f_n(x)-f(x)| = \sup_{x\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}Y(x,\xi_i)-\mathsf{E}[Y(x,\xi)]\right| \to 0$$

as $n\to\infty$ a.s. The ULLN ensures that the optimal objective value of the SAA problem converges to that of the true problem. With additional conditions, the optimal SAA solution converges to the true optimal solution. When the sample function is convex (concave for a maximization problem), the pointwise law of large numbers ensures that the ULLN holds on a compact set. We will further discuss conditions under which the ULLN holds in Sect. 8.4.

In many problems, sample functions are not smooth and may have discontinuities. However, the true problem may still exhibit nice structure, being smooth and even convex. In such a case, if the ULLN holds, we may still be able to use a deterministic optimization technique to effectively solve the nonsmooth sample average problem and thereby obtain a good approximate solution.

In this section, we provide examples that illustrate how SAA works, what issues may arise when SAA is applied, and how we may deal with them in various settings. Henceforth, all vectors are assumed to be column vectors, and x^T denotes the transpose of x.

Example 8.1 (Multi-Dimensional Newsvendor Problem). Consider a firm that manufactures p products from q resources. For given resource type $i=1,\ldots,q$ and product type $j=1,\ldots,p$, let a_{ij} be the amount of resource i required to produce one unit of product j, and v_j be the unit margin for product j, i.e., revenue minus processing cost. Suppose that a manager must decide on a resource vector $x=(x_1,\ldots,x_q)$ before the product demand vector $\xi=(\xi_1,\ldots,\xi_p)$ is observed. After the demand becomes known, the manager chooses a production vector $y=(y_1,\ldots,y_p)$ so as to maximize the operating profit in the linear program

$$\begin{aligned}\mathscr{P}_{(x,\xi)}:\ \max_{y\in\mathbb{R}^p_+}\ & v^\mathsf{T}y\\ \text{s.t.}\quad & Ay\le x \quad \text{(capacity constraints)}\\ & y\le\xi \quad \text{(demand constraints)}.\end{aligned}$$

Here, A is a $(q\times p)$ matrix and a_{ij} is the (i,j) element of A. Let $\Pi(x,\xi)$ denote the maximal operating profit function for a given resource level vector x and a given demand vector ξ. This is precisely the optimal objective value of the problem $\mathscr{P}_{(x,\xi)}$. Then $\Pi(x,\xi)=v^\mathsf{T}y^*(x,\xi)$, where $y^*(x,\xi)$ is an associated optimal production vector.

Suppose that the demand ξ can be viewed as a random vector and the probability distribution of ξ is known. Let $\pi(x)$ denote the expected maximal operating profit, where

$$\pi(x) = \mathsf{E}\left[\Pi(x,\xi)\right],$$

for all $x \in \mathbb{R}^q_+$. Let $c_i, i = 1,\ldots,q$, be the unit investment cost for resource i. By incorporating the investment cost into the operating profit, the value of the firm is defined as $\Pi(x,\xi) - c^{\mathsf{T}}x$, for a fixed (x,ξ). The manager's objective is now to choose the resource level x so as to maximize the expected firm value. This leads to the following stochastic optimization problem:

$$\max_{x\in\mathbb{R}^q_+} f(x) = \pi(x) - c^{\mathsf{T}}x. \tag{8.3}$$

This problem is known as the *multi-dimensional newsvendor problem* [59]. For simplicity, we focus our attention on the single-period newsvendor model, but the structure of the optimal policy in the single-period model can be extended to a dynamic setting under reasonable conditions [19]. In general, a closed-form solution for the multi-dimensional newsvendor problem is unattainable, unlike the single-dimensional problem. We illustrate how the SAA approach can be applied to this example and present some technical details.

From linear programming theory, we can show that both the sample path function $\Pi(\cdot,\xi)$ and the expected objective function π exhibit nice structural properties. First, $\Pi(\cdot,\xi)$ is concave for any fixed ξ, and so is $\pi(\cdot) = \mathsf{E}[\Pi(\cdot,\xi)]$. If ξ has a discrete probability distribution, then both $\Pi(\cdot,\xi)$ and $\pi(\cdot)$ are piecewise linear and concave. However, we focus on random demand with a continuous probability distribution, and we would like to determine conditions under which $\pi(\cdot)$ is differentiable everywhere. Assume that ξ is finite a.s. Consider the dual problem of the linear program $\mathscr{P}_{(x,\xi)}$:

$$\begin{aligned}\mathscr{D}_{(x,\xi)}: \min_{(\mu,\lambda)\in\mathbb{R}^{p+q}_+} \; & x^{\mathsf{T}}\lambda + \xi^{\mathsf{T}}\mu \\ \text{s.t.} \quad & A^{\mathsf{T}}\lambda + \mu \geq v.\end{aligned}$$

Since ξ is finite, the primal problem has a finite optimal solution and the optimal value of the primal problem is equal to that of the dual problem. Let $\lambda(x,\xi)$ denote the optimal shadow value of the capacity constraint in the primal problem $\mathscr{P}_{(x,\xi)}$. Using duality theory, it can be shown that

$$\Pi(x,\xi) \leq \Pi(x_0,\xi) + \lambda(x_0,\xi)^{\mathsf{T}}(x - x_0), \tag{8.4}$$

and hence $\lambda(\cdot,\xi)$ is a subgradient of $\Pi(\cdot,\xi)$. Since $\Pi(\cdot,\xi)$ is concave for a fixed ξ, it is differentiable except on a set A with Lebesgue measure zero. Since ξ is a continuous random variable, A is also negligible with respect to the probability

measure. Thus, $\lambda(x,\xi)$ is unique and $\nabla_x \Pi(x,\xi) = \lambda(x,\xi)$ at a fixed x a.s. Taking the expectation in Eq. (8.4) yields that $\mathsf{E}[\lambda(\cdot,\xi)]$ is a subgradient of $\pi(\cdot)$. Therefore, $\mathsf{E}[\lambda(x,\xi)]$ is unique for all $x \in \mathbb{R}^q_+$ so that $\pi(\cdot)$ is differentiable and

$$\nabla \pi(\cdot) = \mathsf{E}[\lambda(\cdot,\xi)] = \mathsf{E}[\nabla_x \Pi(\cdot,\xi)]. \tag{8.5}$$

Note that $\Pi(\cdot,\xi)$ does not have to be differentiable everywhere, but expectation with respect to a continuous random variable ξ yields a smooth function $\pi(\cdot)$. Equation (8.5) establishes that one can interchange the expectation and differentiation operators. In Sect. 8.3 we will discuss how this interchange property basically ensures that SAA is appropriate for tackling an optimization problem.

The analysis above shows that the true function $\pi(\cdot)$ and the sample function $\Pi(\cdot,\xi)$ share the same nice structural properties; smoothness and concavity. This allows the multi-dimensional newsvendor problem to be effectively solved by the SAA method. The sample average approximation function

$$f_n(x) = \frac{1}{n}\sum_{k=1}^{n} \Pi(x,\xi_k) - c^{\mathsf{T}}x$$

is piecewise linear and concave, but not smooth everywhere. However, the sample average approximation function can be quickly smoothed out as the sample size n grows, so in practice, one can choose sufficiently large n, and then apply an algorithm for optimization of smooth concave functions to solve the sample average approximation problem using the gradient estimator $\frac{1}{n}\sum_{k=1}^{n} \lambda(x,\xi_k) - c$. If the sample average function is not smooth enough and any gradient-based algorithm is not appropriate to use, a subgradient method for convex optimization can be applied to $-f_n(\cdot)$. One can also apply two-stage stochastic linear programming algorithms to solve the sampled problem [8].

Example 8.2 (Multi-Mix Blending Problem). Consider a simple blending problem in which q products are made with p raw materials. The blend products have to satisfy certain pre-specified quality requirements. The total processing costs incurred depend on the product blending options used. Additionally, the production output has to meet minimum requirements. A complication arises when some materials are available in different quantities at different prices.

For the sake of illustration, we consider a problem with only one quality measure. For given raw material type $i = 1,\ldots,p$, and product type $j = 1,\ldots,q$, let

- Q_i be the value of the quality parameter for raw material i,
- b_j be the threshold acceptable level of quality per unit of product j,
- c_{ij} be the cost of processing one unit of raw material i for product j,
- x_{ij} be the amount of raw material i blended into product j,
- d_j be the minimum output level required of product j, and
- a_i be the available amount of raw material i.

The classical multi-mix blending problem is to determine the amount of raw material x_{ij} that minimizes the total processing cost subject to quality requirements. This problem can be formulated as a linear program. We modify this problem with the assumption that the raw material quality parameters Q_i are random with known probability distributions. In this case, the quality requirement constraints can only be satisfied with a certain probability. Thus, instead of minimizing the total processing cost, the manager chooses the amounts x of raw material to be blended in order to maximize the probability of achieving the target quality while keeping the total processing cost within a certain level. This leads to the following constrained stochastic optimization problem:

$$\min_{x\in\mathbb{R}_+^{p\times q}} f(x) = \mathsf{P}\left(\sum_{i=1}^{p}(b_j - Q_i)x_{ij} > 0, j = 1,\ldots,q\right) \tag{8.6}$$

$$\text{s.t.} \quad \sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}x_{ij} \le \tau \quad \text{(total processing cost constraints)},$$

$$\sum_{i=1}^{p} x_{ij} \ge d_j, j = 1,\ldots,q \quad \text{(demand constraints)},$$

$$\sum_{j=1}^{q} x_{ij} \le a_i, i = 1,\ldots,p \quad \text{(resource constraints)}.$$

In general, analytic evaluation of the probability objective function is intractable, particularly when the quality parameters Q_i are correlated. In applying SAA, the random element ξ is taken to be all the quality parameters Q_i. The corresponding sample function $Y(x,\xi)$ is

$$Y(x,\xi) = \mathbf{1}\left\{\sum_{i=1}^{p}(b_j - Q_i)x_{ij} > 0, j = 1,\ldots,q\right\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

Suppose that ξ is a nonnegative random vector with a continuous density function. Note that for any feasible solution x, $f(x)$ is an integral of a density function over a polyhedral set parameterized by x. By using a classical result in mathematical analysis, it can be shown that the true function f is differentiable and the gradient can be expressed as a surface integral [25]. By applying Propositions 8.2 and 8.5 in Sect. 8.3, we can show that the ULLN for f_n holds. Therefore, as long as we can obtain a solution to the sample problem, we can guarantee the convergence of the SAA optimal values. However, the sample function has a discontinuity whenever $\sum_{i=1}^{p}(b_j - Q_i)x_{ij} = 0$ for some $j = 1,\ldots,q$, and $\nabla_x Y(x,\xi) = 0$ for any x except discontinuity points, i.e., the sample average function $f_n(x)$ is differentiable except

on a set A of probability zero, and $\nabla f_n(x) = 0$ on A^c. This problem is ill-posed and any point $x \in A^c$ is a stationary point. Thus, any locally convergent algorithm that searches for a stationary point is not applicable.

One approach to this problem is to approximate the sample function by using a smooth (or piecewise linear) function. The indicator sample function has a very simple structure, only taking on values of zero or one. At any discontinuous point $x \in A$, we can obtain a smooth approximate function by smoothly connecting the sample function on an open neighborhood of x. The resulting approximate function can have a non-zero gradient that is selected to point in an uphill direction in this neighborhood. Then, we can develop an algorithm that employs the gradient to search for an approximate optimal solution. In the example above, we can use the smooth approximation $\phi(h(x,\xi),\varepsilon)$ of the sample function $Y(x,\xi)$, where $h(x,\xi) = \min\{\sum_{i=1}^{p}(b_j - Q_i)x_{ij}, j = 1,\ldots,q\}$ and $\phi : \mathbb{R}\times\mathbb{R}\backslash\{0\} \to \mathbb{R}$ is a continuously differentiable real-valued function such that for any $z \in \mathbb{R}$ and a given $\varepsilon > 0$,

(a) $\phi(z,0) = \mathbf{1}\{z \in (0,\infty)\}$, and
(b) $\mathbf{1}\{z \in (0,\infty)\} \le \phi(z,\varepsilon) \le \mathbf{1}\{z \in (-\varepsilon,\infty)\}$.

If the smoothing parameter ε goes to zero as n increases, then under a set of regularity conditions, the optimal solution of the smooth approximate problem converges to a stationary point of the true problem. This smoothing technique has been widely used, particularly for minimizing risk measures such as Value-at-Risk (VaR) and conditional Value-at-Risk (CVaR) [2, 16], as well as for handling chance constraints [23]. Xu and Zhang [63] provide simple examples of smoothing techniques and discuss the local convergence of the SAA method with a smoothed sample function.

The smoothing approach above changes the true objective function to be optimized, and while the magnitude of the change can be controlled through the parameter ε, one might wonder whether this approximation can be avoided. Sometimes a conditional expectation can be used to smooth jumps in the sample function $Y(\cdot,\xi)$. This is called smoothed perturbation analysis (SPA) [15]. SPA was developed to overcome difficulties in applying infinitesimal perturbation analysis (see Sect. 8.3) in nonsmooth settings, and has been applied to a large class of stochastic discrete event systems. To illustrate the SPA technique, we consider a company that produces only one type of product using two types of raw material. Then, for any $x = (x_1, x_2) > 0$, the corresponding objective function is

$$\begin{aligned} f(x) = \mathsf{E}[Y(x,\xi)] &= \mathsf{E}\left[\mathbf{1}\left\{b\sum_{i=1}^{2} x_i - Q_1x_1 - Q_2x_2 > 0\right\}\right] \\ &= \mathsf{E}\left[\mathsf{E}\left[\mathbf{1}\left\{b\sum_{i=1}^{2} x_i - Q_1x_1 - Q_2x_2 > 0\right\}\middle| Q_2\right]\right] \\ &= \mathsf{E}\left[F_{Q_1}\left(\frac{b\sum_{i=1}^{2} x_i - Q_2x_2}{x_1}\right)\right], \end{aligned} \tag{8.7}$$

where F_{Q_1} is the cumulative distribution function of Q_1 (under the assumption that Q_1 has a density). If the density of Q_1 is continuous, then at any fixed Q_2, the function inside the expectation in (8.7) is differentiable at x. Thus, if F_{Q_1} is known, $f(\cdot)$ can be approximated instead by taking the sample average of this smooth sample function, the expectation of which is the exact function we truly wish to minimize.

Smoothing techniques may not always be applicable, particularly when the sample function has a complex structure. However, even when not applicable, the gap between the sample function $f_n(\cdot)$ and the true function $f(\cdot)$ may still converge to 0 uniformly on the domain Θ as the sample size n grows, so provided that one can solve the SAA problem (8.2), the sequence of optimal objective function values will converge.

Example 8.3 (Bus Scheduling Problem). Passengers arrive at a bus stop according to a Poisson process with rate λ over the interval $[0,1]$. We wish to schedule the arrival times of d infinite-capacity buses over this time interval so as to minimize the expected sum of passenger wait times. We assume that an additional bus arrives at time 1 so that all waiting times are well defined and can be calculated.

Let $\tilde{x}_j$ denote the scheduled time of arrival of the jth bus, $j = 1,2,\ldots,d+1$ where $\tilde{x}_{d+1} = 1$, and let $x_j = \tilde{x}_j - \tilde{x}_{j-1}$ denote the length of the time interval between the arrivals of buses $j-1$ and j, $j = 1,2,\ldots,d+1$, where $\tilde{x}_0 = 0$. The random element ξ here may be taken to be N, the number of passengers to arrive over the time interval $[0,1]$, along with the times $T_1, T_2, \ldots, T_N$ of their arrival.

The sample function $Y(x,\xi)$ may be written

$$Y(x,\xi) = \sum_{i=1}^{N} \sum_{j=1}^{d+1} (\tilde{x}_j - T_i)\mathbf{1}\{T_i \in (\tilde{x}_{j-1}, \tilde{x}_j]\}.$$

The order-statistic property of Poisson processes may be used to show directly (e.g., [44, p. 68]) that

$$f(x) = \mathsf{E}[Y(x,\xi)] = \frac{\lambda}{2} \sum_{j=1}^{d+1} x_j^2,$$

so that $f(x)$ is convex and quadratic, and hence smooth. However, the sample function has a discontinuity whenever the arrival time of a bus coincides with the time of a passenger arrival.

Like the multi-mix blending problem, for any fixed ξ, $Y(\cdot,\xi)$ is differentiable except on a set A of probability zero. Unlike the multi-mix blending problem, the gradient of the sample function $\nabla_x Y(x,\xi)$ at $x \in A^c$ can take a non-zero value. For example, when $d = 1$, the derivative of $Y(x,\xi)$ for a fixed ξ is zero on $(0,T_1)$ and positive at any differentiable point $x \in (T_1,1)$. Note that the gradient of the sample function does not provide any useful information about the (quadratic) true function, and hence any gradient-based algorithm is highly unlikely to work well in

this setting. Approximating the sample function with smoothing techniques is not a trivial problem in this case due to the complexity of the sample function. One could try to minimize the sample average approximation problem in this example using techniques such as metamodeling. If the difference between the sample average and the true function is small enough for sufficiently large sample size n, then the true function can be well approximated with a quadratic metamodel. Indeed, by applying Propositions 8.2 and 8.5, we can show that the gap between the sample and the true function does not persist and eventually converges to 0 uniformly on $[0,1]^d$. Thus, although SAA can in principle be applied to this example, it may not be the best approach, due to the lack of useful structure in the sample functions.

In some examples, sample functions can have appealing structural properties in some variables but not in others. For example, in [14], an (s,S) inventory system is considered. When the problem is reparameterized with decision variables $\Delta = S - s$ and S, then for each fixed Δ, the sample functions are piecewise linear and convex in S. It is then natural to solve the problem by searching only over Δ, using the value of S that minimizes the sample function at each fixed Δ. More precisely, for each fixed Δ, let $S_n^*(\Delta)$ minimize $f_n(\Delta,\cdot)$. Then one can reduce the two-dimensional optimization problem of minimizing $f_n(\cdot,\cdot)$ to a one-dimensional optimization problem where one minimizes $f_n(\Delta, S_n^*(\Delta))$ over Δ. As discussed in [14], this latter problem is not unimodal in Δ, so it is difficult to solve numerically.

8.3 Detecting When SAA Is Appropriate

The key principles exemplified in Sect. 8.2 are that

1. SAA is appropriate only when the approximating functions f_n have some structure that enables the application of an efficient deterministic optimization algorithm, and
2. the limiting function f that we actually want to minimize shares that structure, so that the properties of the limiting function such as the location of local minima are similar to those of the approximating function.

The term "some structure" is intentionally vague because it can mean different things in different problems. For example, by "some structure" we might mean that all sample functions are convex, in which case convex optimization techniques can be applied to the sample functions, and we can ensure convergence to a global minimum of both the sample functions and the limiting function. Alternatively, the sample functions might not be convex, but might be differentiable, in which case gradient-based methods could be applied. A weaker property that can be exploited by numerical optimization algorithms is Lipschitz continuity.

The approximating functions f_n are observable, because we can generate them in finite time, while the limiting function f is not directly observable. Nevertheless, one can often infer structural properties of f through the corresponding properties

of the approximating functions f_n and regularity conditions that ensure that these properties persist in the limit as $n \to \infty$.

In this section, we give sufficient conditions involving only the sample functions $Y(\cdot,\cdot)$ (from which the approximating functions $f_n(\cdot)$ are built) for the *true* function $f(\cdot)$ to be continuous or differentiable at a fixed point x. If these conditions apply at each point x in the domain, then one can conclude that $f(\cdot)$ is continuous or differentiable over that domain. Perhaps surprisingly, one can often arrive at this conclusion even when the sample functions do not possess these same properties over the entire domain. Therefore, Principle 2 does not follow automatically from Principle 1.

These observations will not be surprising to those who are familiar with infinitesimal perturbation analysis (IPA), and indeed, the results presented here can be viewed as a recasting of those ideas in the SAA setting. If we take as given that SAA-appropriate problems are those for which both the approximating functions $f_n(\cdot)$ and $f(\cdot)$ are differentiable, and the derivatives of $f_n(\cdot)$ converge to those for $f(\cdot)$, then we arrive at an underlying theme of this section, which is the following meta-principle:

SAA-appropriate problems are almost exactly those in which IPA applies.

In contrast to much of the IPA literature, we explicitly treat the case where d, the dimension of the domain Θ, can be greater than one. The ideas involved are similar to the one-dimensional case, but some additional care is required. See [17, Chap. 1] for an excellent treatment of the one-dimensional case.

Our first result [27] gives sufficient conditions for $f(\cdot)$ to be continuous at a fixed point $x \in \Theta$. The result is disarmingly straightforward to state and prove. Throughout this chapter, $\|\cdot\|$ will denote the Euclidean norm. Let $B(x,\delta) = \{y : \|y-x\| \le \delta\}$ denote the closed ball of radius δ around x.

Proposition 8.1. *Fix $x \in \Theta$. Suppose that $Y(\cdot,\xi)$ is continuous at x a.s., i.e., for all ξ in a set of probability 1, $Y(x+h,\xi) \to Y(x,\xi)$ as $h \to 0$. Suppose further that the family of random variables*

$$\{Y(x+h,\xi) : x+h \in B(x,\delta)\}$$

is uniformly integrable, for some $\delta > 0$. Then $f(\cdot)$ is continuous at x.

Proof. Continuity of $Y(\cdot,\xi)$ on a set A of probability 1 ensures that

$$\begin{aligned} f(x) &= \mathsf{E}[Y(x,\xi)\mathbf{1}\{\xi \in A\}] \\ &= \mathsf{E}\left[\lim_{h\to 0} Y(x+h,\xi)\mathbf{1}\{\xi \in A\}\right] \\ &= \lim_{h\to 0} \mathsf{E}[Y(x+h,\xi)\mathbf{1}\{\xi \in A\}] \qquad (8.8) \\ &= \lim_{h\to 0} f(x+h), \end{aligned}$$

where the interchange (8.8) is justified by the uniform integrability assumption. ∎

As an immediate corollary we have the following result on the global continuity of $f(\cdot)$.

Corollary 8.1. *Suppose that the conditions of Proposition 8.1 hold at each $x \in \Theta$. Then $f(\cdot)$ is continuous on Θ.*

What is perhaps surprising about this corollary is that we may be able to establish that $f(\cdot)$ is continuous on Θ even when the sample functions $Y(\cdot,\xi)$ are discontinuous on Θ almost surely! The apparent contradiction dissolves when one realizes that the assumption of Proposition 8.1 requires continuity of $Y(\cdot,\xi)$ only locally at x. There may be discontinuities of this function at points outside a neighbourhood of x, and this neighbourhood can depend on ξ.

As an example, let us revisit the bus-scheduling problem from Sect. 8.2. The sample functions $Y(\cdot,\xi)$ have discontinuities at all points x such that a bus arrival time coincides with a passenger arrival time in the interval $(0,1)$. Consequently, the sample functions $Y(\cdot,\xi)$ are discontinuous on Θ almost surely. However, for a fixed $x \in \Theta$, $Y(\cdot,\xi)$ is continuous at x unless a passenger arrival coincides with one of the bus arrival times encoded in x, which occurs with probability 0. Furthermore, the sum of the waiting times of the N arriving passengers is bounded by N, which has finite expectation, and so $\{Y(y,\xi) : y \in B(x,\delta)\}$ is uniformly integrable. Proposition 8.1 then ensures that f is continuous at x, and Corollary 8.1 allows us to conclude that f is continuous on Θ. We already knew that f is continuous on Θ, because it is a convex quadratic. However, this same argument can be used to show continuity in other examples where the form of $f(\cdot)$ is unknown. See [27] for an example involving locating multiple ambulances. This result for the bus-scheduling example is a special case of the following general result.

Proposition 8.2. *Suppose that*

(i) for any fixed ξ, $Y(\cdot,\xi)$ is a piecewise Lipschitz continuous function, i.e., there exists a countable partition of Θ such that the restriction of $Y(\cdot,\xi)$ to the interior of each component is Lipschitz continuous,

(ii) the Lipschitz constants in all components are bounded by an integrable random variable $L(\xi)$,

(iii) the jump size at any discontinuous point $x \in \Theta$ is bounded by a random variable $J(\xi)$ with $\mathsf{E}[J^2(\xi)] < \infty$, and

(iv) for any $x \in \Theta$, $m(x,x+h,\xi) \to 0$ a.s. as $\|h\| \to 0$, where $m(x,x+h,\xi)$ is the number of discontinuity points of the sample function $Y(\cdot,\xi)$ restricted to the line segment joining x and $x+h$ and satisfies

$$\sup_{x+h \in B(x,\delta)} m(x,x+h,\xi) \le M(\xi),$$

for some $\delta > 0$ and a random variable $M(\xi)$ with $\mathsf{E}[M^2(\xi)] < \infty$.

Then the assumptions in Proposition 8.1 hold, and hence $f(\cdot)$ is continuous on Θ.

Proof. Fix $x \in \Theta$. We have

$$|Y(x+h,\xi) - Y(x,\xi)| \le \|h\| L(\xi) + J(\xi) m(x, x+h, \xi). \tag{8.9}$$

By Assumption (iv), the right hand side of (8.9) converges to zero a.s. as $h \to 0$. Thus, $Y(\cdot,\xi)$ is continuous at x a.s. Since $m(x+h,x,\xi) \le M(\xi)$ (over $x+h \in B(x,\delta)$), the right hand side of (8.9) is dominated by an integrable random variable $\|h\| L(\xi) + J(\xi) M(\xi)$. Thus, $\{|Y(x+h,\xi) - Y(x,\xi)| : x+h \in B(x,\delta)\}$ is uniformly integrable, and so is $\{Y(x+h,\xi) : x+h \in B(x,\delta)\}$. ∎

As with continuity, one can obtain differentiability results for $f(\cdot)$ based on local properties of the sample functions $Y(\cdot,\cdot)$.

Proposition 8.3. *Fix x in the interior of Θ. Suppose that $Y(\cdot,\xi)$ is differentiable at x w.p.1, and let $\nabla Y(x,\xi)$ be its gradient. Suppose further that the family of random variables*

$$\left\{ \frac{Y(x+h,\xi) - Y(x,\xi)}{\|h\|} : 0 < \|h\| \le \delta \right\} \tag{8.10}$$

is uniformly integrable, for some $\delta > 0$. Then $f(\cdot)$ is differentiable at x, and $\nabla f(x) = \mathsf{E}[\nabla Y(x,\xi)]$.

Proof. We have that for all ξ in a set A of probability 1,

$$Y(x+h,\xi) = Y(x,\xi) + h^{\mathsf{T}} \nabla Y(x,\xi) + \|h\| R(x,\xi,h), \tag{8.11}$$

where the remainder term $R(x,\xi,h) \to 0$ as $h \to 0$. For $\xi \notin A$, define $\nabla Y(x,\xi) = 0$ and $R(x,\xi,h) = 0$. Taking $h = re_i$, i.e., the ith unit vector scaled by r, for each $i = 1,2,\ldots,d$, and letting $r \to 0$, the uniform integrability assumption implies that $\mathsf{E}[|\partial Y(x,\xi)/\partial x_i|] < \infty$. Hence, all d components of $\mathsf{E}[\nabla Y(x,\xi)]$ exist and are finite. Taking expectations in (8.11), we obtain

$$f(x+h) = f(x) + h^{\mathsf{T}} \mathsf{E}[\nabla Y(x,\xi)] + \|h\| \mathsf{E}[R(x,\xi,h)],$$

so the result will follow if we show that $\mathsf{E}[R(x,\xi,h)] \to 0$ as $h \to 0$. From (8.11), we have that for $\xi \in A$,

$$\frac{Y(x+h,\xi) - Y(x,\xi)}{\|h\|} = \frac{h^{\mathsf{T}}}{\|h\|} \nabla Y(x,\xi) + R(x,\xi,h),$$

and the left-hand side is uniformly integrable (over $\|h\| \in (0,\delta]$) by assumption. But each component of $\nabla Y(x,\xi)$ is integrable, and therefore, $h^{\mathsf{T}} \nabla Y(x,\xi)/\|h\|$ is uniformly integrable for $\|h\| \in (0,\delta]$. It follows that $R(x,\xi,h)$ is uniformly integrable for $h \in (0,\delta]$, and therefore, $\mathsf{E}[R(x,\xi,h)] \to 0$ as $h \to 0$ as required. ∎

Corollary 8.2. *Suppose that the conditions of Proposition 8.3 hold at each x in the interior of Θ. Then $f(\cdot)$ is differentiable on the interior of Θ with $\nabla f(x) = \mathsf{E}[\nabla Y(x,\xi)]$.*

It is again striking that under certain verifiable conditions, one can show that $f(\cdot)$ is differentiable throughout the interior of Θ, even if the sample functions $Y(\cdot,\xi)$ are not. In fact, this is the norm in applications arising in discrete-event simulation, in that the functions $Y(\cdot,\xi)$ typically fail to be differentiable on "seams" in Θ that have measure 0.

The uniform integrability assumption is almost always verified (either locally or on the interior of Θ) by showing that $Y(\cdot,\xi)$ is Lipschitz continuous with Lipschitz constant $L(\xi)$ on the appropriate set, where $\mathsf{E}[L(\xi)] < \infty$. Indeed, the Lipschitz condition ensures that $|Y(x+h,\xi) - Y(x,\xi)| \le L(\xi)\|h\|$, and the uniform integrability requirement follows immediately. But how can this property be verified? In one dimension, one can appeal to the following result, known as the generalized mean value theorem, in which the Lipschitz constant for a sample function $Y(\cdot,\xi)$ arises from a bound on the (sample) derivative. For a proof, see [12, Sect. 8.5].

Theorem 8.1. *Let g be a continuous real-valued function on the closed interval $[a,b]$ that is differentiable everywhere except possibly on a set C of at most countably many points. Then for all x and $x+h$ in $[a,b]$ with $h \neq 0$,*

$$\left|\frac{g(x+h)-g(x)}{h}\right| \le \sup_{y\in[a,b]\setminus C} |g'(y)|.$$

In higher dimensions, we can again apply this result. One difficulty is that real-valued (sample) functions arising in discrete-event simulation often fail to be differentiable along "seams," so the set of non-differentiable points can be uncountable. Fortunately, it is sufficient for our purposes to apply the generalized mean-value theorem along certain line segments only. So long as these line segments intersect the non-differentiable set in at most countably many places, we can apply the generalized mean-value theorem. The following proposition gives sufficient conditions for the uniform integrability condition (8.10) in Proposition 8.3.

Proposition 8.4. *For some $\delta > 0$ suppose that for all ξ in a set of probability 1,*

(i) *$Y(\cdot,\xi)$ is continuous in $B(x,\delta)$;*
(ii) *$C(\xi) \cap [x,y]$ is countable for all $y \in B(x,\delta)$, where $C(\xi)$ denotes the points of non-differentiability of $Y(\cdot,\xi)$ in $B(x,\delta)$ and $[x,y]$ denotes the line segment joining x and y; and*
(iii) $\displaystyle\sup_{y\in B(x,\delta)\setminus C(\xi)} \|\nabla Y(y,\xi)\| \le L(\xi) < \infty.$

If $\mathsf{E}[L(\xi)] < \infty$, then the uniform integrability condition (8.10) holds.

Proof. For $\|h\| \le \delta$ and ξ in the set of probability 1,

$$|Y(x+h,\xi) - Y(x,\xi)| \le \sup_{y\in[x,x+h]\setminus C(\xi)} |h^{\mathsf{T}}\nabla Y(y,\xi)| \tag{8.12}$$

$$\le \|h\| \sup_{y\in[x,x+h]\setminus C(\xi)} \|\nabla Y(y,\xi)\| \tag{8.13}$$

$$\le \|h\| L(\xi),$$

where (8.12) and (8.13) follow from the generalized mean-value theorem and the Cauchy–Schwarz inequality, respectively. The result follows since $L(\xi)$ is integrable. ∎

Sometimes one can verify the Lipschitz property directly, as in the following example.

Example 8.4. A depot is to be located in the unit square $[0,1]^2$. Each night a set of N requests for pickups the following day is made, where N has finite mean. Conditional on $N \geq 1$, the N pickup locations are independent and identically distributed with density $p(\cdot)$ on the unit square. The pickups are completed in a single tour by a van that travels in a straight line from pickup to pickup (Euclidean distance), visiting all pickups before returning to the base. The sequence of pickups is chosen so as to minimize the total travel distance of the van, i.e., the sequence of pickups is the solution to a traveling salesperson problem, starting and finishing at the depot. In this case, the random element ξ consists of the number and locations of pickups, and x gives the Cartesian coordinates of the depot. The goal is to select the depot location to minimize the expected distance traveled by the van.

Here, $Y(x,\xi)$ gives a realization of the distance traveled by the van. We can write

$$Y(x,\xi) = \min_{\pi} Y(x,\xi,\pi), \tag{8.14}$$

where π is a permutation specifying the order in which pickups are visited and $Y(x,\xi,\pi)$ is the resulting distance traveled. (We exclude duplicate permutations that are the reverse of each other in this pointwise minimum.) Each function $Y(\cdot,\xi,\pi)$ is differentiable and in fact has partial derivatives bounded by 2. (To see why, notice that $Y(x,\xi,\pi)$ gives the sum of the distance from x to the first pickup, the distance from the last pickup to x, and the sum of the "internal" distances between the pickups of the permutation. The internal distances do not change as x varies.) Hence, $Y(\cdot,\xi,\pi)$ is Lipschitz with Lipschitz constant 2 for all ξ and π. It then follows from (8.14) that $Y(\cdot,\xi)$ is Lipschitz with Lipschitz constant 2. Furthermore, for fixed x, the set of ξ for which $Y(\cdot,\xi)$ fails to be differentiable at x are such that multiple permutations attain the minimum in (8.14). This set has probability 0 since pickup locations have a density. It follows from our previous discussion that $f(\cdot)$ is differentiable at x and $\nabla f(x) = \mathsf{E}[\nabla Y(x,\xi)]$.

8.4 Known Properties

In this section, we discuss some known properties for well-structured unconstrained optimization problems. Here, "well-structured" means that the sample function enjoys some structural property such as continuity or differentiability. We first investigate under which conditions the optimal solution and value of the SAA problem approach those of the true problem as the sample size n grows. Then,

we discuss how quickly this convergence occurs via the Central Limit Theorem (CLT). We also briefly present the convergence of local solutions for both smooth and nonsmooth problems.

8.4.1 Almost Sure Convergence

As we briefly discussed in Sect. 8.2, uniform convergence of the sample average functions is the key condition for establishing convergence of optimal objective values and solutions in SAA. Indeed, one immediate consequence is the consistency of the SAA optimal values, i.e., $v_n^* \to v^*$ a.s. as $n \to \infty$, where v_n^* and v^* are the optimal objective values of the SAA problem (8.2) and the true problem (8.1), respectively. To see why, note that for a fixed sequence $\{\xi_n : n \geq 1\}$, $\{f_n : n \geq 1\}$ can be viewed as a sequence of deterministic functions. Suppose that f_n converges to the true function f uniformly on Θ. Then, for any sequence $\{x_n\} \subset \Theta$ converging to $x \in \Theta$, $f_n(x_n)$ converges to $f(x)$. Many problems, including those with discontinuous sample functions in Sect. 8.2, satisfy this uniform convergence. When the sample function $Y(\cdot, \xi)$ is convex a.s., the pathwise LLN is equivalent to the ULLN on a compact set [55, Corollary 3]. In a problem with non-convex functions, the following result shows that the conditions for the continuity of the true function $f(\cdot)$ discussed in Sect. 8.3 are, in fact, sufficient to ensure the uniform convergence of the approximating functions on a compact set.

Proposition 8.5. *Let Θ be a nonempty compact set. For any fixed $x \in \Theta$, suppose that $Y(\cdot, \xi)$ is continuous at x a.s., and there exists $\delta > 0$ such that the family of random variables $\{Y(y, \xi) : y \in B(x, \delta)\}$ is uniformly integrable. Then $\{f_n(x)\}$ converges to $f(x)$ uniformly on Θ a.s. as $n \to \infty$.*

Proof. The proof can be carried out by adapting the proof of Proposition 7 in [55]. Choose $\tilde{x} \in \Theta$. Let $\{\delta_k \leq \delta(\tilde{x}) : k = 1, 2, \ldots\}$ be a sequence of positive numbers decreasing to 0, and define

$$\alpha_k(\xi) = \sup_{x \in B(\tilde{x}, \delta_k)} |Y(x, \xi) - Y(\tilde{x}, \xi)|.$$

By the continuity of $Y(\cdot, \xi)$ at $\tilde{x}$, $\alpha_k(\xi)$ goes to zero a.s. as k increases. The uniform integrability assumption ensures that $\{\alpha_k(\xi) : k = 1, 2 \ldots\}$ is uniformly integrable, and hence

$$\lim_{k \to \infty} \mathsf{E}[\alpha_k(\xi)] = \mathsf{E}\left[\lim_{k \to \infty} \alpha_k(\xi)\right] = 0.$$

Note that

$$\sup_{x\in B(\tilde{x},\delta_k)} |f_n(x) - f_n(\tilde{x})| \le \frac{1}{n}\sum_{i=1}^{n} \alpha_k(\xi_i). \tag{8.15}$$

By the LLN, the right-hand side of (8.15) converges to $\mathsf{E}[\alpha_k(\xi_i)]$ a.s. as $n \to \infty$. Thus, for given $\varepsilon > 0$, there exists a neighborhood V of $\tilde{x}$ such that a.s. for sufficiently large n,

$$\sup_{x\in V\cap\Theta} |f_n(x) - f_n(\tilde{x})| < \varepsilon.$$

Since Θ is compact, there exists a finite number of points $x_1, \ldots, x_m \in \Theta$ and corresponding neighborhoods $V_1, \ldots, V_m$ covering Θ such that a.s. for sufficiently large n,

$$\sup_{x\in V_j\cap\Theta} |f_n(x) - f_n(x_j)| < \varepsilon, j = 1, \ldots, m. \tag{8.16}$$

By Proposition 8.1, $f(\cdot)$ is continuous. Thus, we can choose the neighborhoods $V_1, \ldots, V_m$ in such a way that

$$\sup_{x\in V_j\cap\Theta} |f(x) - f(x_j)| < \varepsilon, j = 1, \ldots, m. \tag{8.17}$$

By the LLN, with probability 1 (w.p.1) for sufficiently large n,

$$|f_n(x_j) - f(x_j)| < \varepsilon, j = 1, \ldots, m. \tag{8.18}$$

Combining (8.16)–(8.18), w.p.1 for sufficiently large n, we have

$$\sup_{x\in\Theta} |f_n(x) - f(x)| < 3\varepsilon. \qquad \blacksquare$$

When the sample function is continuous, the ULLN implies the continuity of the true function f. However, in general, the continuity of the true function is not a necessary condition for uniform convergence of the approximating function f_n. For example, consider a cumulative distribution function (cdf) $f(x) = P(\xi \le x)$ and the empirical cdf $f_n(x)$. By the Glivenko–Cantelli Theorem [7, p. 269], f_n converges to f uniformly on $\mathbb{R}$ even if f is discontinuous. Optimizing a discontinuous function is in general a difficult problem, and many practical problems naturally exhibit continuity properties. In this chapter, we therefore focus on problems where f is continuous, unless the domain Θ is a discrete set.

We introduce some notation to proceed to the convergence results below. Let Π_n^* and π^* denote the set of optimal solutions of the SAA and the true problems, respectively. We define the Euclidean distance from a point x to a set B to be

$d(x,B) = \inf_{y\in B}\|x-y\|$, and the distance between two sets $A,B \subset \mathbb{R}^q$ to be $\mathbb{D}(A,B) = \sup\{d(x,B) : x \in A\}$. In the next theorem, we give convergence results based on the continuity of the true function and uniform convergence.

Theorem 8.2 (Theorem 5.3, [57]). *Suppose there exists a compact subset $C \subset \mathbb{R}^d$ such that*

(i) π^* *is non-empty and contained in* C,
(ii) $\{f_n(x)\}$ *converges to* $f(x)$ *uniformly on* C *a.s. as* $n \to \infty$, *and*
(iii) for sufficiently large n, Π_n^* *is non-empty and contained in* C *a.s.*

Then $v_n^* \to v^*$. *Furthermore, if the true function* $f(\cdot)$ *is continuous on* C, *then* $\mathbb{D}(\Pi_n^*, \pi^*) \to 0$ *a.s. as* $n \to \infty$.

Proof. Fix $\varepsilon > 0$. Uniform convergence of f_n to f on C ensures that

$$f_n(x) \geq f(x) - \varepsilon$$

for all $x \in C$, for sufficiently large n a.s. The assumption that $\Pi_n^* \subseteq C$ ensures that v_n^* is attained on C for sufficiently large n a.s., so $v_n^* \geq v^* - \varepsilon$ for sufficiently large n a.s. Since ε was arbitrary, $\liminf_n v_n^* \geq v^*$ a.s. Also, since there exists $x^* \in \pi^* \subseteq C$, $v_n^* \leq f_n(x^*) \to v^*$ as $n \to \infty$ a.s. Thus, $v_n^* \to v^*$ as $n \to \infty$ a.s. Turning to convergence of the solution set, suppose that $\mathbb{D}(\Pi_n^*, \pi^*) \nrightarrow 0$. Then there exists $X_n \in \Pi_n^*$ such that for some $\varepsilon > 0$, $d(X_n, \pi^*) \geq \varepsilon$ for all $n \geq 1$. Since C is compact, by passing to a subsequence if necessary, X_n converges to a point $x^* \in C$, and $f(x^*) > v^*$. On the other hand,

$$f_n(X_n^*) - f(x^*) = [f_n(X_n^*) - f(X_n^*)] + [f(X_n^*) - f(x^*)] \tag{8.19}$$

Both the first term and the second term in the right hand side of (8.19) converge to zero by the uniform convergence assumption and continuity of f, respectively. Thus, $v_n^* \to f(x^*) > v^*$, which contradicts the fact that $v_n^* \to v^*$. ∎

Theorem 8.2 ensures that, if X_n^* solves the SAA problem exactly, then $d(X_n^*, \pi^*) \to 0$ a.s. as $n \to \infty$. Moreover, if the true problem has a unique optimal solution x^*, then $X_n^* \to x^*$. When the sample functions are convex, the set of regularity conditions in Theorem 8.2 can be relaxed by using the theory of epi-convergence [57, Theorem 5.4].

Now we consider the case where Θ is a finite set and discuss the convergence of ε-optimal solutions in the SAA method. We first introduce some notation. For $\varepsilon \geq 0$, let

$$\pi^*(\varepsilon) := \{x \in \Theta : f(x) \leq v^* + \varepsilon\}, \quad \Pi_n^*(\varepsilon) := \{x \in \Theta : f_n(x) \leq v_n^* + \varepsilon\} \tag{8.20}$$

denote the ε-optimal solutions for the true and the SAA problems, respectively. Since Θ is finite, the pathwise LLN implies the ULLN. Thus, the a.s. convergence of v_n^* to v^* is guaranteed. Furthermore, asymptotic normality of v_n^* follows under

moment conditions if the optimal solution is unique. Also, it can be shown that for any $\varepsilon \geq 0$, $\Pi_n^*(\varepsilon) \subset \pi^*(\varepsilon)$ w.p.1 for n sufficiently large [28]. This means that any ε-optimal solution of the SAA problem is an ε-optimal solution of the true problem for large enough n. In particular, if the true problem has a unique solution x^*, then $\Pi_n^* = \{x^*\}$ w.p.1 for n large enough. But, how quickly does the probability of $\{\Pi_n^*(\varepsilon) \subset \pi^*(\varepsilon)\}$ approach 1 as n increases? Large deviation analysis shows that under a mild regularity condition (essentially finiteness of the moment generating function of $Y(x,\xi)$ at each fixed x), the probability $\mathsf{P}\{\Pi_n^*(\delta) \nsubseteq \pi^*(\varepsilon)\}$ for $0 \leq \delta < \varepsilon$, converges to zero at an exponential rate. We discuss this further in Sect. 8.5.

8.4.2 Convergence Rates for the SAA Method

There exists a well-developed statistical inference for estimators obtained from the SAA approach. From this inference, we can obtain error bounds for obtained solutions and select the sample size n to obtain a desired level of accuracy. The first result below by [32] states that the estimator v_n^* for v^* is negatively biased and the expected value of v_n^* monotonically increases. This monotonicity property of $\mathsf{E}[v_n^*]$ is desirable in the sense that we can expect a tighter lower bound as n increases.

Proposition 8.6. *For all* $n \geq 1$, $\mathsf{E}[v_n^*] \leq \mathsf{E}[v_{n+1}^*]$, *and* $\mathsf{E}[v_n^*] \leq v^*$.

Proof. Since $\xi_1, \xi_2, \ldots$ are i.i.d.,

$$\begin{aligned}
\mathsf{E}[v_{n+1}^*] = \mathsf{E}\left[\min_{x\in\Theta} \frac{1}{n+1}\sum_{i=1}^{n+1} Y(x,\xi_i)\right] &= \mathsf{E}\left[\min_{x\in\Theta} \frac{1}{n+1}\sum_{i=1}^{n+1}\left(\frac{1}{n}\sum_{j\neq i} Y(x,\xi_j)\right)\right] \\
&\geq \frac{1}{n+1}\sum_{i=1}^{n+1} \mathsf{E}\left[\min_{x\in\Theta}\left(\frac{1}{n}\sum_{j\neq i} Y(x,\xi_j)\right)\right] \\
&= \mathsf{E}[v_n^*].
\end{aligned}$$

For any $\tilde{x} \in \Theta$, $f_n(\tilde{x}) \geq \min_{x\in\Theta} f_n(x)$. By taking expectation on both sides, we have

$$v^* = \min_{x\in\Theta} f(x) = \min_{x\in\Theta} \mathsf{E}\,[f_n(x)] \geq \mathsf{E}\,[\min_{x\in\Theta} f_n(x)] = \mathsf{E}[v_n^*]. \qquad \blacksquare$$

Next, we discuss the asymptotic behavior of the SAA optimal objective value v_n^*. For a sequence of random variables $\{X_n\}$ and deterministic constants β_n, we say that $X_n = o_p(\beta_n)$, if $X_n/\beta_n \to 0$ in probability. We also say that $X_n = O_p(\beta_n)$, if $\{X_n/\beta_n\}$ is bounded in probability (tight), i.e., for any $\varepsilon > 0$, there exists $M > 0$ such that $\mathsf{P}(|X_n/\beta_n| > M) < \varepsilon$, for all n.

First, assuming that $\mathsf{E}[Y^2(x,\xi)] < \infty$, we have the CLT for any fixed $x \in \Theta$,

$$\sqrt{n}(f_n(x) - f(x)) \xrightarrow{d} Z(x)$$

as $n \to \infty$, where $\xrightarrow{d}$ signifies convergence in distribution, and $Z(x) \sim \mathcal{N}(0, \sigma^2(x))$, $\sigma^2(x) \equiv \mathrm{Var}[Y(x,\xi)]$. The CLT implies that the error $f_n(x) - f(x)$ is of order $O_p(n^{-1/2})$. Under a set of mild regularity conditions, the same canonical convergence rate of v_n^* can be obtained by applying a multidimensional version of the CLT to f.

Theorem 8.3 (Theorem 5.7, [57]). *We suppose that*

(i) Θ *is compact,*
(ii) $\mathsf{E}[Y^2(x,\xi)] < \infty$, *for some* $x \in \Theta$,
(iii) $Y(\cdot,\xi)$ *is Lipschitz on* Θ *with Lipschitz constant* $L(\xi)$ *a.s., and* $\mathsf{E}\left[L^2(\xi)\right] < \infty$.

Then,

$$v_n^* = \inf_{x \in \pi^*} f_n(x) + o_p(n^{-1/2})$$

and

$$\sqrt{n}(v_n^* - v^*) \Rightarrow \inf_{x \in \pi^*} Z(x) \tag{8.21}$$

as $n \to \infty$, *where* Z *is a Gaussian process on* Θ *with* $\mathsf{E}[Z(x)] = 0$ *and* $\mathrm{Cov}(Z(x),Z(y)) = \mathrm{Cov}(Y(x,\xi),Y(y,\xi))$, *for all* $x, y \in \Theta$.

Proof. The essential idea of the proof is to employ a functional CLT for f_n and the Delta method [7] to $V(f)$, where V is the real-valued functional given by $V(g) = \min_{x \in \Theta} g(x)$, for any continuous function g on Θ. ∎

When the true problem has a unique optimal solution x^*, (8.21) implies that v_n^* is asymptotically normally distributed. It again follows from (8.21) that under some uniform integrability conditions, the bias $\mathsf{E}[v_n^*] - v^*$ is $O(n^{-1/2})$. If the true problem has a unique solution, $\mathsf{E}[v_n^*] - v^*$ is $o(n^{-1/2})$, and with additional moments and second order conditions on f, $\mathsf{E}[v_n^*] - v^*$ is, in fact, $O(n^{-1})$.

The SAA optimal solution X_n^* requires a stronger set of conditions to achieve the same asymptotic properties as v_n^*. When f is smooth and has a unique solution, under some regularity conditions, the solution X_n^* of the SAA problem converges to the unique solution x^* of the true problem at the canonical rate $n^{-1/2}$. One of the essential regularity conditions is that the true function f increases quadratically near the unique solution x^*. (It may converge at a faster rate if the function f increases linearly near x^*, as may happen if x^* lies on the boundary of Θ.) We say that the quadratic growth condition is satisfied at $\tilde{x}$ if there exists $\alpha > 0$ and a neighborhood V of $\tilde{x}$ such that for all $x \in \Theta \cap V$,

$$f(x) \geq f(\tilde{x}) + \alpha \|x - \tilde{x}\|^2.$$

If Θ is a convex, full dimensional set and $\tilde{x}$ lies in the interior of Θ, then the quadratic growth condition is equivalent to the second order sufficient optimality condition,

i.e., the Hessian matrix $\nabla^2 f(\tilde{x})$ is positive definite. In the convergence result below, we provide conditions that are relatively easy to understand and a sketch of the proof. The readers are referred to [53,57] for the proof under more general regularity conditions.

We say $g : \mathbb{R}^d \to \mathbb{R}$ is Fréchet differentiable at x if there exists a bounded linear operator $D_x g : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\lim_{\|u\| \downarrow 0} \frac{|g(x+u) - g(x) - D_x g(u)|}{\|u\|} = 0.$$

Theorem 8.4. *Assume that the following hold:*

(i) *The true function f has a unique minimizer $x^* \in \Theta$.*
(ii) *$Y(\cdot, \xi)$ is Lipschitz with Lipschitz constant $L(\xi)$ on Θ a.s., and $\mathsf{E}[L(\xi)] < \infty$.*
(iii) *$Y(\cdot, \xi)$ is continuously differentiable at any x in a neighborhood of x^* a.s.*
(iv) *$\mathsf{E}[\|\nabla_x Y(x, \xi)\|^2] < \infty$, for some $x \in \Theta$.*
(v) *$\nabla_x Y(\cdot, \xi)$ is Lipschitz with Lipschitz constant $K(\xi)$ in a neighborhood of x^* a.s., and $\mathsf{E}\left[K^2(\xi)\right] < \infty$.*
(vi) *f satisfies the quadratic growth condition at x^*.*

Then, $\|X_n^* - x^*\| = O_p(n^{-1/2})$. Furthermore, assume that

(vii) *There exists a neighborhood U of x^* and $\alpha > 0$ such that for every u in a neighborhood of zero, the following problem*

$$\min_{x \in \Theta} f(x) + u \cdot x \tag{8.22}$$

has an optimal solution $x^*(u) \in U$ and the quadratic growth condition holds at $x^*(u)$.

If $x^*(u)$ is Fréchet differentiable at $u = 0$ and $D_0 x^*(\cdot)$ is continuous,

$$\sqrt{n}(X_n^* - x^*) \Rightarrow D_0 x^*(Z) \tag{8.23}$$

as $n \to \infty$, where Z is a multivariate normal vector with mean 0 and covariance matrix

$$\Sigma = \mathsf{E}\left[(\nabla f(x^*, \xi) - \nabla f(x^*))^{\mathsf{T}} (\nabla f(x^*, \xi) - \nabla f(x^*))\right].$$

Moreover, if $D_0 x^*(\cdot)$ is linear, then $\sqrt{n}(X_n^* - x^*)$ is asymptotically normally distributed.

Proof. By Assumptions *(i)*–*(iii)*, $X_n^* \to x^*$ a.s. as $n \to \infty$, and $f(\cdot)$ is Lipschitz continuous and continuously differentiable at x^*. Let $\delta_n(x) = f_n(x) - f(x)$.

By the quadratic growth condition (vi) and the generalized mean-value theorem, we have

$$\|X_n^* - x^*\| \leq \frac{\sup_{x \in B(x^*, \|X_n^* - x^*\|)} \|\nabla \delta_n(x)\|}{\alpha}.$$

With Assumptions (iv)–(v), by applying the functional CLT and the continuous mapping theorem [7] to $\nabla \delta_n(\cdot)$, we have $\sup_{x \in B(x^*, \|X_n^* - x^*\|)} \|\nabla \delta_n(x)\| = O_p(n^{-1/2})$, and hence $\|X_n^* - x^*\| = O_p(n^{-1/2})$ follows.

By applying the quadratic growth condition (vii) to $x^*(\nabla \delta_n(x^*))$ and using the Lipschitz continuity of $\nabla f(\cdot)$, it can be shown that

$$X_n^* = x^*(\nabla \delta_n(x^*)) + o_p(n^{-1/2}).$$

Since $x^*(u)$ is Fréchet differentiable at $u = 0$,

$$x^*(\nabla \delta_n(x^*)) - x^* = D_0 x^*(\nabla \delta_n(x^*)) + o_p(n^{-1/2}).$$

Thus, it follows from the CLT on $\nabla \delta_n(x^*)$ and the continuous mapping theorem that

$$\sqrt{n}(X_n^* - x^*) = D_0 x^*(\sqrt{n} \nabla \delta_n(x^*)) + o_p(1) \Rightarrow D_0 x^*(Z). \qquad \blacksquare$$

A second-order condition can ensure the quadratic growth condition (vii) for the parameterized objective function $f(x) + u \cdot x$. For example, if Θ is convex, f is twice continuously differentiable, and $\nabla^2 f(x^*)$ is positive definite, Assumption (vii) holds by setting α as the lower bound of the smallest eigenvalue of $\nabla^2 f(x)$ in a neighborhood of x^*. If $x^*(\cdot)$ is Lipschitz, the Fréchet derivative $D_0 x^*(\cdot)$ is continuous and linear, and thus the asymptotic normality of X_n^* can be ensured.

8.4.3 The SAA Method in the Nonconvex Case

Thus far, the convergence theory for SAA methods that we have presented has been derived under the assumption that we can produce a global minimum of the SAA problem in Θ, and hence the theory can be applied primarily to convex problems. In the nonconvex case, the best that we can hope for from a computational point of view is that we can generate local minimizers of the SAA problems. When the sample function is differentiable almost surely on Θ, the validity of IPA ensures the convergence of the first order points of the SAA problem to those of the true problem. This reaffirms the key principle observed in Sect. 8.3: when IPA is valid, the SAA method is appropriate.

For $x \in \Theta$, $\mathcal{N}(x)$ denotes the normal cone to Θ at x. For x in the interior of Θ, $\mathcal{N}(x) = \{0\}$. For x on the boundary of Θ, $\mathcal{N}(x)$ is the convex cone generated by the outward normals of the faces on which x lies. When Θ is convex,

$$\mathcal{N}(x) = \{y \in \mathbb{R}^d : y^\mathsf{T}(x' - x) \leq 0, \text{for all } x' \in \Theta\}.$$

A first-order critical point x of a smooth function f satisfies

$$-\nabla f(x) = z \text{ for some } z \in \mathcal{N}(x),$$

i.e., the direction of most rapid descent lies in the normal cone (of directions we cannot move in without leaving Θ). Let S_n^* and S^* be the set of first-order critical points of the approximating function f_n and the true function f in Θ, respectively. The following theorem states first-order convergence results of the SAA method, and is an immediate result of [55, Proposition 19].

Theorem 8.5. *Suppose there exists a compact subset $C \subset \mathbb{R}^d$ such that*

(i) *S^* is non-empty and contained in C,*
(ii) *the true function $f(\cdot)$ is continuously differentiable on an open set containing C,*
(iii) *$\{\nabla f_n(x)\}$ converges to $\nabla f(x)$ uniformly on C, a.s. as $n \to \infty$, and*
(iv) *for sufficiently large n, S_n^* is non-empty and contained in C w.p.1.*

Then $\mathbb{D}(S_n^*, S^*) \to 0$ a.s. as $n \to \infty$.

Proof. The proof can be derived from stochastic generalized equations. We do not introduce them here; rather, we present a relatively easier version of the proof with the added assumption that the domain Θ is compact and convex. Suppose that $\mathbb{D}(S_n^*, S^*) \nrightarrow 0$. Since Θ is compact, by passing to a subsequence if necessary, we can assume that there exists a convergent sequence of solutions $\{X_n^* \in S_n^*\}$ such that for some $\varepsilon > 0$, $d(X_n^*, S^*) \geq \varepsilon$ for all $n \geq 1$. Let x^* be a limit point of $\{X_n^*\}$, and then $x^* \notin S^*$. On the other hand, since Θ is convex and each X_n^* satisfies the first order criticality condition, for any $u \in \Theta$

$$\nabla f_n(X_n^*)^\mathsf{T}(u - X_n^*) \geq 0 \text{ w.p.1.}$$

By (ii) and (iii),

$$\nabla f_n(X_n^*)^\mathsf{T}(u - X_n^*) \to \nabla f(x^*)^\mathsf{T}(u - x^*)$$

a.s. as $n \to \infty$. Thus, $\nabla f(x^*)^\mathsf{T}(u - x^*) \geq 0$ for all $x \in \Theta$. But $x^* \notin S^*$ implies that for some $u \in X$ $\nabla f(x^*)^\mathsf{T}(u - x^*) < 0$, which is a contradiction. ∎

The assumptions (ii) and (iii) above are satisfied under the sufficient conditions for a valid IPA gradient estimator presented in Sect. 8.3. When the sample path function is continuously differentiable a.s. at any $x \in \Theta$, $\nabla Y(\cdot, \xi)$ is uniformly integrable under the assumptions in Proposition 8.4. By applying Proposition 8.5 to each component of $\nabla Y(\cdot, \xi)$, we can show the continuity of $\nabla f(\cdot)$ and the uniform convergence of $\{\nabla f_n(\cdot)\}$.

Theorem 8.5 implies that the limit point of any solution sequence $\{X_n^* \in S_n^*\}$ must lie in S^*. This does not guarantee that $\{X_n^*\}$ converges almost surely. When there are multiple critical points, the particular critical point chosen from S_n^* depends, among other things, on the optimization algorithm that is used. The existence of a unique

first-order critical point can ensure convergence. However, this condition tends to be difficult to verify in practice.

The second order convergence of the SAA method can be obtained by further strengthening the assumptions in Theorem 8.5. Now, we select X_n^* from a set of local minimizers of the SAA problem. By passing to a subsequence if necessary, we assume that $\{X_n^*\}$ converges to some random point $x^* \in \Theta$ a.s. as $n \to \infty$. The additional condition required is there must exist a neighborhood of X_n^* in which X_n^* is a local minimizer and this neighborhood does not shrink to a singleton when $n \to \infty$.

Theorem 8.6 (Theorem 4.1, [5]). *Suppose that the assumptions in Theorem 8.5 hold. Furthermore, assume that for any fixed sample path $\xi = \{\xi_1, \xi_2, \ldots\}$, there exist $n_0 > 0$ and $\delta > 0$ such that for all $n \geq n_0$ and $x \in B(X_n^*, \delta) \cap \Theta$, $f_n(X_n^*) \leq f_n(x)$. Then x^* is a local minimum of $f(\cdot)$ w.p.1.*

Nonsmooth objective functions arise in a number of interesting stochastic optimization problems such as stochastic programs with recourse and stochastic min-max problems [50]. To close this section, we briefly discuss the local convergence of the SAA method in the nonsmooth setting. When the true and sample functions are continuous and nonsmooth, we can derive convergence results based on the Clarke generalized gradient [10]. For a locally Lipschitz function f, the generalized gradient ∂f can be defined as the convex hull of all the limit points of $\nabla f(x_k)$, where $\{x_k\}$ is any sequence which converges to x while avoiding the points where $\nabla f(x_k)$ does not exist. With some technical definitions, the expectation of the generalized gradient of the sample function can be well-defined [21, 62].

Essentially, the same principle from the smooth case can hold for the nonsmooth problem. When IPA is valid, i.e., $\partial \mathsf{E}[Y(x,\xi)] = \mathsf{E}[\partial_x Y(x,\xi)]$, SAA can be appropriate and the first order convergence can be achieved. A sufficient condition for the validity of IPA is that the sample function is locally Lipschitz continuous with integrable Lipschitz constant. This condition is a fairly general condition in the Lipschitz continuous setting, just as it is in the smooth case.

8.5 SAA Implementation

Implementing the SAA method is conceptually straightforward since only two choices need to be made: the sample size with which to generate the sample-path problem, and the numerical procedure with which to solve the generated sample-path problem. Assuming a numerical procedure is (somehow) chosen using cues laid out in Sects. 8.2 and 8.3, the only remaining task then is choosing an appropriate sample size. Towards making this decision, a reasonable question might be to ask what minimum sample size ensures that the solution resulting from the generated sample-path problem is of a stipulated quality, with a specified probability. In what follows, we present a "minimum sample size" result that answers this question. This is followed by a discussion of certain refined versions of SAA that are aimed at enhancing the implementability of the SAA method.

8.5.1 Sample Size Choice

Recall that for $\varepsilon \geq 0$, $\pi^*(\varepsilon)$ and $\Pi_n^*(\varepsilon)$ denote the ε-optimal solutions for the true and the sample-path problems, respectively. Theorem 8.7 presents an expression for the sample size n that guarantees that $\mathsf{P}\{\Pi_n^*(\delta) \not\subseteq \pi^*(\varepsilon)\} \leq \alpha$, for given $\alpha > 0, \varepsilon > 0$, and a chosen constant $\delta < \varepsilon$. The implication is that when an SAA problem is generated with a sample size exceeding the expression provided, the resulting solution is guaranteed to be ε-optimal with probability exceeding $1 - \alpha$. To guide intuition, we present the result only for the setting where Θ is finite. The corresponding expression for the general case follows in a straightforward fashion after making additional assumptions that help to approximate $\{f(x) : x \in \Theta\}$ with $\{f(x) : x \in \tilde{\Theta}\}$, where $\tilde{\Theta}$ is an appropriately chosen finite set that in a certain precise sense "approximates" Θ.

Theorem 8.7 (Theorem 5.18, [51]). *Suppose there exists a constant $\sigma > 0$ such that for any $x \in \Theta \setminus \pi^*(\varepsilon)$, the moment generating function $M_x(t)$ of the random variable $Y(x,\xi) - f(x)$ satisfies $M_x(t) \leq \exp(\sigma^2 t^2/2), \forall t \in \mathbb{R}$. Then, for $\varepsilon > 0, 0 \leq \delta < \varepsilon$, and $\alpha \in (0,1)$, any n satisfying*

$$n \geq \frac{2\sigma^2 \ln(\frac{|\Theta|}{\alpha})}{(\varepsilon - \delta)^2} \tag{8.24}$$

guarantees that $\mathsf{P}\{\Pi_n^*(\delta) \not\subseteq \pi^*(\varepsilon)\} \leq \alpha$.

The proof of Theorem 8.7 proceeds by using the crude bound

$$\begin{aligned} \mathsf{P}\{\Pi_n^*(\delta) \not\subseteq \pi^*(\varepsilon)\} &\leq \sum_{x \in \Theta \setminus \pi^*(\varepsilon)} \mathsf{P}\{f_n(x) \leq v^* + \varepsilon\} \\ &\leq |\Theta| \exp\{-n\eta(\delta,\varepsilon)\} \\ &\leq |\Theta| \exp\{-n(\varepsilon - \delta)^2/2\sigma^2\}, \end{aligned} \tag{8.25}$$

where $\eta(\delta,\varepsilon) = \min_{x \in \Theta \setminus \pi^*(\varepsilon)} I_x(-\delta)$, and $I_x(\cdot)$ is the large deviations rate function of $Y(x,\xi) - f(x)$. The expression in (8.24) then follows upon replacing the left-hand side of (8.25) with α and then solving for the sample size. Note that in choosing a sample size through (8.24), the tolerance δ to which the SAA problem is solved still needs to be chosen by the user. It can also be seen from the expression in (8.24) that the dependence of the minimum sample size on the error probability α is logarithmic, and hence weak.

The sample size directive given by Theorem 8.7, while useful in some SAA settings, can be overly conservative [31, 51], often resulting in a loss in computational efficiency. This is unsurprising considering the crude bound leading to (8.25), and the existence of unknown constants, e.g., σ^2 in the case of finite Θ and several others in the case of continuous Θ, that nevertheless need to be chosen by the user. Such loss in efficiency resulting from the sometimes impractical sample size directives has been one of the primary impediments to SAA's implementability.

8.5.2 *Refined SAA Methods*

With a view towards easier implementation, various refined versions [11, 22, 35, 36, 45] of the SAA method have recently been proposed. In what follows, we discuss one of these paradigms, Retrospective Approximation (RA), in further detail. (The similarly named "Retrospective Optimization" technique was introduced by [20], but to describe the SAA method.)

Recall the efficiency issue associated with the SAA method. SAA dictates that a single sample-path problem be generated with a large enough sample size and solved to adequate tolerance. However, the minimum sample size required to ensure that the resulting solution is of stipulated quality may be so large as to render the procedure not viable. To thwart this difficulty, RA proposes a slight refinement of the SAA paradigm. Instead of solving a single sample-path problem generated with a large enough sample size, RA proposes to generate and solve a sequence of sample-path problems. The sequence of sample-path problems are generated using a nondecreasing sequence of sample sizes $\{m_k\}$, that are then solved to increasing stringency using a sequence of error-tolerances $\{\varepsilon_k\}$ that converge to zero. When the paradigm works as intended, the resulting sequence of solutions approaches the true solution asymptotically. More importantly, the paradigm is constructed to preserve efficiency. The early iterations are efficient because they involve sample-path problems generated with small sample sizes. The later iterations are efficient, at least in principle, due to the use of "warm starts," where solutions from previous iterations are used as initial guesses to the subsequent problems.

Towards further clarification, we now list RA as a nonterminating algorithm.

RA Components:

(i) A procedure for solving a generated sample-path problem to specified tolerance vector ε_k.
(ii) A sequence $\{m_k\}$ of sample sizes tending to infinity.
(iii) A sequence $\{\varepsilon_k\}$ of error-tolerances tending to zero.
(iv) A sequence of weights $\{w_{kj} : j = 1, 2, \ldots, k\}$ for each iteration.

RA Logic:

0. Initialize the retrospective iteration number $k = 1$.
1. Generate a sample-path problem with sample size m_k. Use RA component (i) with a "warm start," i.e., with $\overline{X}_{k-1}$ as the initial guess, to solve the generated problem to within error-tolerance ε_k. Obtain a retrospective solution X_k.
2. Use component (iv) to calculate the solution $\overline{X}_k$ as the weighted sum of retrospective solutions $\{X_i\}_{i=1}^k$:

$$\overline{X}_k = \sum_{j=1}^{k} w_{kj} X_j.$$

3. Set $k \leftarrow k + 1$ and go to 1.

Step 1 of the RA listing is deliberately left ambiguous, and is to be made precise depending on the problem context. For example, in the context of using RA within global SO problems, "solving a sample-path problem to within tolerance ε_k" can mean identifying a point X_k whose optimality gap as measured with respect to the objective function $f_{m_k}(x)$ is at most ε_k.

The iterates resulting from the RA paradigm, for the context of global SO, are strongly consistent under conditions similar to those imposed within the SAA method. The proof follows in a rather straightforward fashion from the corresponding theorem [51, Theorem 5.3] in the SAA context in combination with some standard results on M-estimators [52].

Theorem 8.8. *Assume*

A_1. *The feasible region* Θ *is compact, and the set of global minima* $\pi^* \subset \Theta$ *of the function* f *is nonempty.*

A_2. *The sequence of sample functions* $\{f_n(x)\}$ *is such that the set of global minima* Π_n^* *of the function* f_n *is nonempty for large enough* n *w.p.1.*

A_3. *The functional sequence* $\{f_n(x)\} \to f(x)$ *uniformly as* $n \to \infty$ *w.p.1.*

A_4. *The function* f *is continuous on* Θ.

A_5. *The sequence of sample sizes* $\{m_k\}$ *and the sequence of error-tolerances* $\{\varepsilon_k\}$ *in the RA paradigm are chosen to satisfy* $\{m_k\} \to \infty$ *and* $\varepsilon_k \to 0$ *as* $k \to \infty$.

A_6. *Given* $s > 0$, *define the* i*th sum of the first* s *weights* $w_i(s) = \sum_{j=1}^{s} w_{ij}$ *for each* $i \geq s$. *The weights* $\{w_{ij}\}$ *are chosen so that* $w_i(s) \to 0$ *as* $i \to \infty$.

A_7. *The sample-path problems are solved to obtain a retrospective solution* X_k *satisfying* $\|f_{m_k}(X_k) - v_{m_k}^*\| \leq \varepsilon_k$ *when* $\Pi_{m_k}^* \neq \phi$, *with* $v_{m_k} = \inf\{f_{m_k}(x) : x \in \Theta\}$.

Then the sequences $\{f(X_k) - v^*\}$, $\{d(X_k, \pi^*)\} \to 0$ *w.p.1. (Assume* $d(X_k, \pi^*) = \infty$ *if* $\Pi_{m_k}^* = \emptyset$.*)*

The RA method above was presented as a nonterminating algorithm where a sequence of sample sizes $\{m_k\}$ for problem generation and a sequence of error-tolerances $\{\varepsilon_k\}$ relevant during problem solution need to be chosen by the user. This raises the natural question of how these sequences should be chosen to ensure efficiency. Pasupathy [35] partially addresses this question and presents guidelines on choosing these sequences as a function of the convergence rate of the numerical procedure in use. For example, it is shown that for efficiency, it may be best to choose $\varepsilon_k = O(1/\sqrt{m_k})$ when the numerical procedure in use converges at a linear rate. (*Convergence rates* are defined rigorously in Sect. 8.6.2.) Furthermore, when using linearly convergent numerical procedures, efficiency dictates that it is best to choose sample sizes $\{m_k\}$ such that $\limsup m_k/m_{k-1}^p = 0$ for all $p > 1$. Likewise, when using numerical procedures that have superlinear convergence rates, efficiency dictates that it is best to choose $\{m_k\}$ such that $\limsup m_k/m_{k-1}^p < \infty$ for all $p > 1$. We discuss these results in more detail in Sect. 8.6.

More recently, [45] addresses the obvious drawback that the directives provided in [35] are at best asymptotic. In other words, while the results in [35] recommend the rates at which the sample size sequence $\{m_k\}$ and the error-tolerance sequence

$\{\varepsilon_k\}$ should converge to zero, these recommendations still leave a large family of sequences from which to choose. Royset [45] remedies this in the specific context of solving smooth stochastic programs (e.g., when derivatives of the function $Y(x,\xi)$ are observable and $Y(x,\xi)$ is Lipschitz with the Lipschitz constant having finite expectation) with a numerical solver that is linearly convergent (e.g., projected gradient method using Armijo step sizes as detailed in [38]). Using a model that approximates the progress made by the linearly convergent numerical procedure in use, [45] formulates a dynamic program to identify generation-effort/solution-effort trade-off at the beginning of each iteration within RA. The output of the dynamic program includes the sample size that should be used for each generated problem and the computational effort that should be expended toward solving each generated problem.

8.6 Asymptotic Efficiency Calculation

As noted earlier, refined SAA methods like RA, are constructed with a view towards implementation. Does this construction result in any real computational savings? In other words, does RA enjoy provable efficiency gains over the SAA paradigm? In this section, we answer this question in some detail. Towards first providing a benchmark for an asymptotic rate calculation, we present a very concise overview and analysis of stochastic approximation (SA) which, alongside the SAA method, is a standard technique for solving SO problems. This is followed by Sect. 8.6.2 where we discuss the maximum achievable convergence rate by the SAA method. Section 8.6.3 presents the analogous calculation for the RA method.

Towards setting up the problem of identifying asymptotic efficiency, suppose that the optimal solution to the true problem, x^*, is unique, and suppose that we want to obtain a solution that is within a prescribed distance ε from x^*. (All distances are Euclidean unless otherwise noted.) Suppose also that we measure computational effort in terms of the number of simulation replications required, i.e., the number of times $Y(x,\xi)$ is computed, for various x and ξ. Here we take the function $Y(\cdot,\cdot)$ as fixed, rather than allowing it to come from a class of functions as in many varieties of complexity theory; see, e.g., [33]. Moreover, this measure ignores the effort required to compute, e.g., gradients. However, our results will be restricted to *rates* of convergence that ignore proportionality constants, so as long as gradients are obtained through schemes that only proportionally increase the work, e.g., finite differences and infinitesimal perturbation analysis, then our results will be unaffected. Finally, we also ignore the internal computations of an optimization algorithm beyond the simulation effort. Such computations often heavily depend on the dimension of the problem, but since we are fixing f, the dimension is also fixed.

8.6.1 Asymptotic Rates for Stochastic Approximation

A well-understood, general-purpose method for solving stochastic optimization problems, alternative to using the SAA principle, is stochastic approximation [26, 30, 42]. For unconstrained problems in $\mathbb{R}^d$, the classical stochastic approximation algorithm is a simple recursion that produces a sequence of points $\{\tilde{X}_n : n \geq 0\}$, each of which lies in $\mathbb{R}^d$. The recursion requires an initial point $\tilde{X}_0$, a positive gain sequence $\{a_n : n \geq 0\}$, and a sequence of vectors $\{\hat{\nabla} f(\tilde{X}_n) : n \geq 0\}$ in $\mathbb{R}^d$, where $\hat{\nabla} f(\tilde{X}_n)$ is an estimate of $\nabla f(\tilde{X}_n)$. A simple version of a stochastic-approximation recursion for a minimization problem is then

$$\tilde{X}_{n+1} = \tilde{X}_n - a_n \hat{\nabla} f(\tilde{X}_n). \tag{8.26}$$

For the problems we consider here, the gradient estimator can usually be taken to be $\nabla Y(\tilde{X}_n, \xi_n)$, i.e., the gradient of $Y(\cdot, \xi_n)$ evaluated at $\tilde{X}_n$, where $(\xi_n : n \geq 0)$ are i.i.d., since under fairly general conditions (Sect. 8.3), this gradient estimator is unbiased and has other desirable qualities like bounded variance. In that case, if $f(\cdot)$ is smooth, has a unique global minimizer x^*, and $a_n = a/n$ with $a > 0$ sufficiently large, then under additional nonrestrictive conditions,

$$\sqrt{n}(\tilde{X}_n - x^*) \Rightarrow N(0, \Lambda), \tag{8.27}$$

as $n \to \infty$, for a certain $d \times d$ matrix Λ. See [4, Chap. VIII] for an overview of this result and a sketch of how it can be established using the "Ordinary Differential Equation" approach.

The CLT in (8.27) is striking in that the recursion (8.26) is trivial to implement, involves almost no computation beyond the calculation of a sample gradient at each iteration, and is very generally applicable. If the number of iterations of (8.26) is completed in c units of computer time, $n(c)$ grows roughly linearly in c (as would be the case if, e.g., sample gradients are computed in constant time), then a time-changed version of the CLT (8.27) establishes that the resulting SA estimator has an error $\tilde{X}_{n(c)} - x^* = O_p(c^{-1/2})$. Equivalently, the computational effort required to obtain an error of order ε with SA is $O_p(\varepsilon^{-2})$.

It is generally known that the performance of the recursion in (8.26) is highly dependent on the gain sequence $\{a_n\}$. (In fact, even when the gradient estimator $\nabla Y(\tilde{X}_n, \xi_n)$ is directly observable and $a_n = a/n$ $(a > 0)$, convergence to the root fails if the constant a falls below a certain threshold, akin to the parallel-chord method for nonlinear root-finding [34, p. 181].) Accordingly, the last three decades have seen enormous attention given to the question of choosing the gain sequence $\{a_n\}$; see [3, 9, 30, 40] and Chap. 6. While we do not go into any further detail on this question, two key facts stand out. First, within the context of the iteration (8.26), the fastest achievable convergence rate is $O_p(c^{-1/2})$ [40]. Second, a remarkably simple scheme independently developed by [39, 49], and surveyed under the moniker "Polyak–Ruppert averaging" in [4, Chap. VIII], achieves this

maximum rate. The scheme involves using the step-size sequence $a_n = a/n^\gamma$ for some $\gamma \in (0,1)$, and then estimating the root x^* via the direct average

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i.$$

Under mild conditions, the Polyak–Ruppert averaging scheme enjoys a CLT of the same form as (8.27), although with a different covariance matrix Λ. Furthermore, this happens *irrespective* of the value of the constant $a > 0$. (The small-sample performance is, however, seriously affected by the choice of the constant a.) The Polyak–Ruppert averaging scheme also has other optimality properties related to the matrix Λ that appears in the limit; see [4, Chap. VIII].

8.6.2 Asymptotic Rates for the SAA Method

As noted in Sect. 8.6.1, the Polyak–Ruppert averaging scheme achieves the maximum possible convergence rate of $O_p(c^{-1/2})$ within the context of stochastic approximation. Loosely speaking, this amounts to requiring $O_p(\varepsilon^{-2})$ computational effort if one wants to obtain ε accuracy. How does the SAA method perform in comparison? Towards setting up this question rigorously, recall the SAA method again—a single sample-path problem is generated with sample size n and solved using a chosen numerical solver. Furthermore, since solving the generated problem to infinite precision is usually impossible in practice, suppose we execute k iterations of the numerical procedure on the generated problem to obtain a solution $X_n(k)$. The total budget expended in the process is then simply $c = n \times k$. It seems clear that, under certain conditions (e.g., numerical procedure cannot solve to infinite precision in a finite number of steps), as the available budget $c \to \infty$, the sample size n and the number of steps k should satisfy $n, k \to \infty$ to ensure that the optimality gap of $X_n(k)$ converges to zero in any reasonable sense. However, what relationship between n and k (for given c) ensures that such convergence happens at the fastest possible rate? Moreover, what is the corresponding maximal rate?

In a recent paper, [46] provide an answer to these questions. Before we summarize their results, let us introduce definitions relating to the convergence rates of the numerical solver in use. These definitions appear in more restrictive form in [46].

Denote the numerical procedure acting on the sample function $f_n(x)$ by the map $A(x) : \Theta \to \Theta$. Let $A^k(x)$ denote the iterate obtained after k successive applications of the map $A(\cdot)$ on the initial iterate x. In all three definitions that follow, we assume that the function $f_n(x)$ attains its infimum $v_n^* := \inf\{f_n(x) : x \in \Theta\}$ and that $f_n(A^k(x)) \to v_n^*$ as $k \to \infty$ for all $x \in \Theta$. Also, to avoid trivialities, assume that $f_n(A^{k+1}(x))$ is different from v_n^* for all k. Denote $Q_t = \limsup_{k\to\infty} |f_n(A^{k+1}(x)) - v_n^*|/|f_n(A^k(x)) - v_n^*|^t$.

Definition 8.1. The numerical procedure $A(x) : \Theta \to \Theta$ is said to exhibit pth-order sublinear convergence if $Q_1 \geq 1$, and there exist constants $p, s > 0$ such that $p = \sup\{r : f_n(A^k(x)) - v_n^* \leq s/k^r \text{ for all } x \in \Theta\}$.

When $f_n(x)$ is convex and Θ is a closed convex set, the subgradient method [33, Sect. 3.2] for nonsmooth convex optimization exhibits sublinear convergence with $p = 1/2$. Similarly, when $f_n(x)$ is strongly convex with $\Theta := \mathbb{R}^d$, the optimal gradient method [33, Sect. 2.2] is sublinear with $p = 2$.

Definition 8.2. The numerical procedure $A(x) : \Theta \to \Theta$ is said to exhibit linear convergence if $Q_1 \in (0, 1)$ for all $x \in \Theta$.

The definition of linear convergence implies that there exists a constant θ satisfying $f_n(A(x)) - v_n^* \leq \theta(f_n(x) - v_n^*)$ for all $x \in \Theta$. The projected gradient method with Armijo steps [38] when executed on certain smooth problems exhibits a linear convergence rate.

Definition 8.3. The numerical procedure $A(x) : \Theta \to \Theta$ is said to exhibit superlinear convergence if $Q_1 = 0$ for all $x \in \Theta$. The convergence is said to be pth-order superlinear if $Q_1 = 0$ and $\sup\{t : Q_t = 0\} = p < \infty$ for all $x \in \Theta$.

When $f_n(x)$ is strongly convex and twice Lipschitz continuously differentiable with observable derivatives, Newton's method is second-order superlinear. For settings where the derivative is unobservable, there is a slight degradation in the convergence rate, but Newton's method remains superlinear [6, p. 338].

We are now ready to summarize the main results of [46] through Theorem 8.9, which is in essence a characterization of the maximum achievable convergence rate when using a sublinearly convergent algorithm within the SAA method, and should be juxtaposed with the $O_p(c^{-1/2})$ rate achievable using stochastic approximation as discussed in Sect. 8.6.1.

Theorem 8.9 (Convergence Rate for the SAA Method). *Let the following assumptions hold.*

A_1. *The expectation* $\mathsf{E}[Y^2(x, \xi)] < \infty$ *for all* $x \in \Theta$.
A_2. *The function* $Y(x, \xi)$ *is Lipschitz w.p.1, and has Lipschitz constant* $K(\xi)$ *having finite expectation.*
A_3. *The function* $f_n(x)$ *attains its infimum on* Θ *for each n w.p.1.*

Also, let $c = n \times k$ *and* $n/c^{1/(2p+1)} \to a$ *as* $c \to \infty$, *with* $a \in (0, \infty)$. *Then, if the numerical procedure exhibits pth-order sublinear convergence,*

$$c^{p/(2p+1)}(f_n(A^k(x)) - v^*) = O_p(1) \text{ as } c \to \infty.$$

The crucial message given by Theorem 8.9 is that in the context of the SAA method, the maximum achievable convergence rate is $O_p(c^{-p/(2p+1)})$ when the numerical procedure in use exhibits p-th order sublinear convergence. (While Theorem 8.9 does not directly assert that $O_p(c^{-p/(2p+1)})$ is the maximum achievable rate, [46] show this rigorously.) [46] also demonstrate that the corresponding rates when

using linearly convergent and pth-order superlinearly convergent procedures are $O_p((c/\log c)^{-1/2})$ and $O_p((c/\log\log c)^{-1/2})$, respectively.

Two observations relating to the assertions in [46] are noteworthy. First, the fastest achievable convergence rate within the SAA method depends on the numerical procedure in use, with faster numerical procedures affording a faster rate. This is not so surprising when one sees that the SAA method splits the available budget ($c = n \times k$) between sampling and solving. Since faster numerical procedures incur a smaller cost to solving, they facilitate attainment of a faster convergence rate. Second, none of the families of numerical procedures considered are capable of attaining the canonical convergence rate $O_p(c^{-1/2})$ that is seen in stochastic approximation. Such degradation from the canonical convergence rate can be explained as the "price" of using a numerical procedure. In other words, unless the numerical procedure used within SAA is capable of infinite precision with only a finite amount of computing effort, there is always a degradation in the convergence rate due to the fact that a non-negligible portion of the budget is expended towards solving the generated problem.

8.6.3 Asymptotic Rates for the RA Method

In this section, we present an analogous analysis for the maximum achievable convergence rates within the RA method. Recall that in the RA method, instead of generating and solving a single sample-path problem as in the SAA method, a sequence of sample-path problems are generated with sample sizes $\{m_k\}$ and solved to corresponding error-tolerances $\{\varepsilon_k\}$. In analyzing the achievable convergence rates within the RA method, we then seek an asymptotic relationship between the error $\|X_k - x^*\|$ incurred at the end of k iterations, and the corresponding total work done C_k. The following result, adapted from [35], captures this relationship as a function of the convergence rate of the numerical procedure in use, but with strict stipulations on the sample-path structure and the ability to observe their derivatives.

Theorem 8.10. *Assume that Assumptions (i)–(vi) of Theorem 8.4 hold. In addition, let the following assumptions hold:*

- A_1. *The sample function $f_n(x)$ has a unique minimum X_n^* w.p.1.*
- A_2. *When $f_n(x)$ attains a unique minimum X_n^*, $f_n(x)$ is twice differentiable at X_n^*. Furthermore, the matrix of second-order partial derivatives (Hessian) of $f_n(x)$ at X_n^* is positive definite with smallest eigenvalue uniformly bounded away from 0 w.p.1.*
- A_3. *The solution X_k obtained from the kth iteration of RA satisfies $\|\nabla f_{m_k}(X_k)\| \leq \varepsilon_k$.*
- A_4. *The numerical procedure used to solve the sample-path problems in RA exhibits p-th order sublinear convergence or pth-order linear convergence with respect to the observed derivatives.*
- A_5. *The sample sizes are increased linearly, i.e., $m_k/m_{k-1} = c > 1$ for all k.*
- A_6. *The error-tolerances are chosen so that $\varepsilon_k = O(1/\sqrt{m_k})$.*

Then the sequence of solutions obtained using the RA procedure satisfies $C_k\|X_k - x^*\|^2 = O_p(1)$ *as* $k \to \infty$*, where* C_k *is the total amount of computational work done until the kth iteration and is given by* $C_k = \sum_{i=1}^{k} N_i m_i$*. Here* N_i *is the number of points visited by the numerical procedure during the ith iteration.*

Proof. The proof proceeds along lines very similar to the proof of Theorem 5 in [35]. In what follows, we provide only a proof sketch, and only for the case where the numerical procedure in use exhibits linear convergence. The corresponding proof for the sublinear convergence case follows almost directly after appropriately changing the expression in (8.28).

We first see, since the numerical procedure is assumed to exhibit linear convergence, that

$$N_i = O_p\left(1 + \frac{1}{\log r}\left(\log \frac{\varepsilon_i}{\|\nabla f_{m_i}(X_{i-1})\|}\right)\right), \tag{8.28}$$

for some $r \in (0,1)$. Using the Delta method [7] and Assumptions A_1, A_2, A_3, we write

$$\|\nabla f_{m_i}(X_{i-1})\| = O_p(\|X_i^* - X_{i-1}^*\|) + \varepsilon_{i-1}. \tag{8.29}$$

Since Assumptions (i)-(vi) of Theorem 8.4 hold, $\|X_i^* - x^*\| = O_p(1/\sqrt{m_i})$ and hence $\|X_i^* - X_{i-1}^*\| = O_p(1/\sqrt{m_i} + 1/\sqrt{m_{i-1}})$. Combining this with (8.28) and (8.29) yields

$$N_i = O_p\left(1 + \frac{1}{\log r}\left(\log \frac{\varepsilon_i}{1/\sqrt{m_i} + 1/\sqrt{m_{i-1}} + \varepsilon_{i-1}}\right)\right). \tag{8.30}$$

Now use Assumption A_6 to obtain

$$C_k\|X_k - x^*\|^2 = O_p\left((\sum_{i=1}^{k} m_i)(1/\sqrt{m_k} + \varepsilon_k)^2\right). \tag{8.31}$$

Finally, use (8.31) and Assumption A_5 to conclude that the assertion holds. ∎

Theorem 8.10 asserts that, as long as the sample size and error-tolerance sequences are chosen strategically, the error in the obtained solution converges to zero at the canonical rate. This assertion is interesting, since we will recall from Sect. 8.6.2 that the canonical convergence rate is unachievable in the context of the SAA method, barring unlikely contexts where the numerical procedure exhibited exceptionally fast convergence rates. It is also noteworthy that Theorem 8.10 assumes that the derivatives of the sample path are observable. This is to help with terminating the individual iterations of the RA algorithm, and could probably be relaxed further by assuming instead that the derivative is estimated using a consistent estimator appropriately constructed from function observations. The assumption

about the numerical procedure exhibiting at least linear convergence is easily satisfied, e.g., projected gradient method [38] for certain smooth problems; and Newton's method [34, Chap. 9] when used on smooth convex programs with observable derivatives.

8.7 Conclusions

We have provided a guide to the principle of SAA for simulation optimization, with a discussion on when SAA might be an appropriate solution method, how the potential for such applicability can be detected, and an appraisal of SAA's implementation and efficiency characteristics. An interesting observation on SAA's applicability is that it can be applied whenever infinitesimal perturbation analysis [17] for gradient estimation can be applied. Loosely speaking, both of these methods become applicable when the sample problems and true problem share characteristics that are important for numerical optimization software, chief among which are continuity, differentiability, and the approximate location of optimal solutions. SAA has a well-developed large-sample theory, both within the global and the local optimality contexts. The latter context seems especially useful within application settings.

On the question of asymptotic efficiency, recent results have established that a straightforward implementation of SAA is inferior to stochastic approximation. This difference in efficiency stems entirely from the fact that SAA, by construction, stipulates that the optimization software being employed uses a fixed sample size irrespective of how close the current solution is to an optimal solution. Towards remedying this, a refinement of SAA called retrospective approximation has been developed. The refinement increases sample sizes at a carefully controlled rate as the numerical optimization proceeds, and in the process recovers the same rate of convergence (up to a multiplicative constant) as stochastic approximation.

Throughout this chapter, we have made the assumption that samples are i.i.d., but that is not essential to the established theory. Indeed, one can apply any of several variance reduction methodologies that induce dependence, and for the most part the theory remains relatively unchanged; see [57]. One can also generate the samples using techniques such as quasi-Monte Carlo and randomized versions thereof [29]. Some of these topics, along with stochastic constraints, are treated in detail in the following chapter.

Acknowledgements This work was partially supported by National Science Foundation grants CMMI-0800688 and CMMI-1200315, and by Singapore MOE Academic Research Fund grant WBS R-266-000-049-133.

References

1. S. Ahmed and A. Shapiro. Solving chance-constrained stochastic programs via sampling and integer programming. In *Tutorials in Operations Research*, pages 261–269. INFORMS, 2008.
2. S. Alexander, T. F. Coleman, and Y. Li. Minimizing CVaR and VaR for a portfolio of derivatives. *Journal of Banknig and Finance*, 30(2):584–605, 2006.
3. S. Andradóttir. A scaled stochastic approximation algorithm. *Management Science*, 42:475–498, 1996.
4. S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modeling and Applied Probability*. Springer, New York, 2007.
5. F. Bastin, C. Cirillo, and P. L. Toint. Convergence theory for nonconvex stochastic programming with an application to mixed logit. *Mathematical Programming B*, 108:207–234, 2006.
6. M. S. Bazaara, H. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York, NY., 2006.
7. P. Billingsley. *Probability and Measure*. Wiley, New York, NY., 1995.
8. J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, 1997.
9. M. Broadie, D. M. Cicek, and A. Zeevi. General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. *Operations Research*, 59(5):1211–1224, 2011.
10. F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley, New York, 1983.
11. G. Deng and M. C. Ferris. Variable-number sample-path optimization. *Mathematical Programming*, 117:81–109, 2009.
12. J. A. Dieudonné. *Foundations of Modern Analysis*. Academic Press, New York, 1960.
13. M. C. Fu. Gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, pages 575–616. Elsevier, Amsterdam, 2006.
14. M. C. Fu and K. Healy. Techniques for optimization via simulation: an experimental study on an (s, S) inventory system. *IIE Transactions*, 29(3):191–199, 1993.
15. M. C. Fu and J. Q. Hu. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer, Boston, 1997.
16. A. A. Gaivoronski and G. Pflug. Value-at-Risk in portfolio optimization: Properties and computational approach. *Journal of Risk*, 7(2):1–31, 2005.
17. P. Glasserman. *Gradient Estimation Via Perturbation Analysis*. Kluwer, The Netherlands, 1991.
18. G. Gürkan, A. Y. Özge, and S. M. Robinson. Sample-path solution of stochastic variational inequalities. *Mathematical Programming*, 84:313–333, 1999.
19. J. M. Harrison and J. A. Van Mieghem. Multi-resource investment stragies: Operational hedging under demand uncertainty. *European Journal of Operational Research*, 113:17–29, 1999.
20. K. Healy and L. W. Schruben. Retrospective simulation response optimization. In B. L. Nelson, D. W. Kelton, and G. M. Clark, editors, *Proceedings of the 1991 Winter Simulation Conference*, pages 954–957. IEEE, Piscataway, NJ, 1991.
21. T. Homem-de-Mello. Estimation of derivatives of nonsmooth performance measures in regenerative systems. *Mathematics of Operations Research*, 26:741–768, 2001.
22. T. Homem-de-Mello. Variable-sample methods for stochastic optimization. *ACM Transactions on Modeling and Computer Simulation*, 13:108–133, 2003.
23. L. J. Hong, Y. Yang, and L. Zhang. Sequential convex approximations to joint chance constrained programs: A Monte Carlo approach. *Operations Research*, 59(3):617–630, 2011.
24. H. Hu, T. Homem-de-Mello, and S. Mehrotra. Sample average approximation of stochastic dominance constrained programs. *Mathematical Programming*, 133(1–2):171–201, 2012.
25. A. I. Kibzun and Y. S. Kan. *Stochastic Programming Problems with Probability and Quantile Functions*. Wiley, New York, 1996.

26. J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.
27. S. Kim and S. G. Henderson. The mathematics of continuous-variable simulation optimization. In S. J. Mason, R. R. Hill, L. Moench, and O. Rose, editors, *Proceedings of the 2008 Winter Simulation Conference*, pages 122–132. IEEE, Piscataway, NJ, 2008.
28. A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.
29. M. Koivu. Variance reduction in sample approximations of stochastic programs. *Mathematical Programming*, 103(3):463–485, 2005.
30. H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, New York, 2nd edition, 2003.
31. J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19:674–699, 2008.
32. W. K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.
33. Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Norwell, MA, 2004.
34. J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, NY., 1970.
35. R. Pasupathy. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58:889–901, 2010.
36. R. Pasupathy and S. Kim. The stochastic root-finding problem: overview, solutions, and open questions. *ACM Transactions on Modeling and Computer Simulation*, 21(3):23, 2011.
37. E. L. Plambeck, B.-R. Fu, S. M. Robinson, and R. Suri. Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, 75:137–176, 1996.
38. E. Polak. *Optimization: Algorithms and Consistent Approximations*. Springer, New York, NY, 1997.
39. B. T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh.*, 7:98–107, 1990.
40. B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.
41. A. Ravindran, D. T. Phillips, and J. J. Solberg. *Operations Research: Principles and Practice*. Wiley, New York, NY, 2nd ed. edition, 1987.
42. H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
43. S. M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21:513–528, 1996.
44. S. M. Ross. *Stochastic Processes*. Wiley, New York, 2nd edition, 1996.
45. J. Royset. On sample size control in sample average approximations for solving smooth stochastic programs. *Computational Optimization and Applications*, 55(2):265–309, 2013.
46. J. Royset and R. Szechtman. Optimal budget allocation for sample average approximation. *Operations Research*, 61(3):762–776, 2013.
47. R. Y. Rubinstein and A. Shapiro. Optimization of static simulation models by the score function method. *Mathematics and Computers in Simulation*, 32:373–392, 1990.
48. R. Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, Chichester, 1993.
49. D. Ruppert. Stochastic approximation. In B. K. Ghosh and P. K. Sen, editors, *Handbook of Sequential Analysis*, pages 503–529. Marcel Dekker, New York, 1991.
50. A. Ruszczyński. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12:32–49, 1987.
51. A. Ruszczyński and A. Shapiro, editors. *Stochastic Programming. Handbook in Operations Research and Management Science*. Elsevier, New York, NY, 2003.
52. R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., New York, NY, 1980.

53. A. Shapiro. Asymtotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18:829–845, 1993.
54. A. Shapiro. Simulation-based optimization – convergence analysis and statistical inference. *Stochastic Models*, 12(3):425–454, 1996.
55. A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, Handbooks in Operations Research and Management Science. Elsevier, 2003.
56. A. Shapiro. Sample average approximation. In S. I. Gass and M. C. Fu, editors, *Encyclopedia of Operations Research and Management Science*, pages 1350–1355. Springer, 3rd edition, 2013.
57. A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. MPS-SIAM Series on Optimization. SIAM-MPS, Philadelphia, PA, 2009.
58. Stochastic Programming Society. Stochastic programming introduction. www.stoprog.org, 2014.
59. J. A. Van Mieghem and N. Rudi. Newsvendor networks: Inventory management and capacity investment with discretionarty activities. *Manufacturing and Service Operations Management*, 4(4):313–335, 2002.
60. S. Vogel. Stability results for stochastic programming problems. *Optimization*, 19(2):269–288, 1988.
61. S. Vogel. A stochastic approach to stability in stochastic programming. *Journal of Computational and Applied Mathematics*, 45:65–96, 1994.
62. R. Wets. Stochastic programming. In G. L. Nemhauser, A. H. G.Rinnooy Kan, and M. J. Todd, editors, *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*, pages 573–629. Elsevier, 1989.
63. H. Xu and D. Zhang. Smooth sample average approximation of stationary point in nonsmooth stochastic optimization and application. *Mathematical Programming*, 119:371–401, 2009.