Chapter 5
Stochastic Gradient Estimation

Michael C. Fu

Abstract This chapter reviews simulation-based methods for estimating gradients, which are central to gradient-based simulation optimization algorithms such as stochastic approximation and sample average approximation. We begin by describing approaches based on finite differences, including the simultaneous perturbation method. The remainder of the chapter then focuses on the direct gradient estimation techniques of perturbation analysis, the likelihood ratio/score function method, and the use of weak derivatives (also known as measure-valued differentiation). Various examples are provided to illustrate the different estimators—for a single random variable, a stochastic activity network, and a single-server queue. Recent work on quantile sensitivity estimation is summarized, and several newly proposed approaches for using stochastic gradients in simulation optimization are discussed.

5.1 Introduction

For optimization problems with *continuous-valued* decision variables, the availability of gradients can dramatically improve the effectiveness of solution algorithms, but in the stochastic setting, since the outputs are themselves random, finding or deriving *stochastic* gradient estimators can itself be a challenging problem, which constitutes the subject of this chapter. The three succeeding chapters on stochastic approximation and sample average approximation—Chaps. 6, 7, and 8—highlight the central role that stochastic gradients play in simulation optimization. In addition to their use in gradient-based simulation optimization, these estimators have other important applications in simulation, most notably *sensitivity analysis*, e.g., factor screening to decide which factors are the most critical, and hedging of financial instruments and portfolios.

M.C. Fu (✉)
University of Maryland, College Park, MD, USA
e-mail: mfu@umd.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_5

Consider the general optimization problem

$$\min_{x \in \Theta} f(x), \tag{5.1}$$

where $x \in \Theta \subseteq \mathbb{R}^d$. In the context of simulation optimization considered here, f is not directly available but instead the simulation model returns a noisy output $Y(x, \xi)$, where ξ represents the randomness. We consider two forms of the objective function, the commonly used expected value performance

$$f(x) = \mathsf{E}[Y(x, \xi)], \tag{5.2}$$

and the quantile

$$f(x) = q_\alpha(x) = \sup\{y : \mathsf{P}(Y(x, \xi) \leq y) \leq \alpha, \ 0 < \alpha < 1, \tag{5.3}$$

where $\alpha = \mathsf{P}(Y(x, \xi) \leq q_\alpha(x))$ when Y is a continuous random variable.

We introduce two examples that will also be used later to illustrate the various direct gradient estimators:

- A *stochastic activity network* is a directed acyclic graph where the arcs have random activity times. The decision variables influence the distribution of these activity times. The output performance to be considered is the total time to go from a designated source to a designated sink in the network. We will specifically consider the longest path performance where the decision variables (input parameters) are in the individual activity time probability distributions.
- A first-come, first-served (FCFS) *single-server queue*, where the customer arrival process and the customer service times are both stochastic and independent of each other. The output performance to be considered is the average time spent in the system by a customer, denoted by T, and the input parameters will be in the service time distribution(s). When the arrival process is renewal, and the service times are independent and identically distributed (i.i.d.), this is known as a $G/G/1$ (or sometimes written $GI/GI/1$) queue. A simple optimization problem could be to choose the mean service time $x > 0$ to minimize

  $$f(x) = \mathsf{E}[T(x, \xi)] + c/x,$$

  where c can be viewed as the cost of having a faster server.

To solve (5.1) for either setting (5.2) or (5.3), a natural adaptation of steepest descent in deterministic nonlinear optimization is stochastic approximation (SA), which is an iterative update scheme on the parameter that takes the following general form for finding a zero of the objective function gradient:

$$x_{n+1} = \Pi_\Theta \left(x_n - a_n \hat{\nabla} f(x_n) \right), \tag{5.4}$$

where $\hat{\nabla} f$ is an estimate of the gradient ∇f, $\{a_n\}$ is the so-called gain (also known as step-size) sequence, and Π_Θ denotes a projection back into the feasible region Θ when the update (5.4) would otherwise take x_{n+1} out of Θ. Guaranteeing with probability 1 (w.p.1) convergence of x_n requires $a_n \to 0$, but at a rate that cannot be too quick, with a common set of conditions being

$$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty.$$

In practice, a_n is often adjusted to a constant value after some number of iterations, which theoretically only guarantees weak convergence (in distribution). The gain sequence need not be deterministic, i.e., it could depend on the output that is generated, e.g., Kesten's rule [31], which decreases a_n only when the sign of the derivative estimate changes. Under appropriate conditions, there are also Central Limit Theorem results that characterize the asymptotic behavior of x_n (cf. [34]).

When $\hat{\nabla} f$ is an unbiased estimator of ∇f, the SA algorithm is generally referred to as being of the Robbins–Monro (RM) [38] type, whereas if $\hat{\nabla} f$ is only asymptotically unbiased, e.g., using a finite difference estimate with the difference going to zero at an appropriate rate, then the algorithm is referred to being of the Kiefer–Wolfowitz (KW) [32] type; see Chap. 6 for details. The Robbins–Monro SA algorithm generally has a canonical asymptotic convergence rate of $n^{-1/2}$, in contrast to $n^{-1/3}$ for the Kiefer–Wolfowitz SA algorithm.

A key challenge for using an SA algorithm for simulation optimization, is the sensitivity of the early transient finite-time behavior of (5.4) to the sequence $\{a_n\}$; for KW-type algorithms, there is the additional choice of the difference sequence. For example, the behavior of SA for the commonly used sequence $a_n = a/n$ $(a > 0)$ is very sensitive to the choice of a. If a is too small, then the algorithm will "crawl" towards the optimum, even at the $1/\sqrt{n}$ asymptotic rate. On the other hand, if a is chosen too large, then extreme oscillations may occur, resulting in an "unstable" progression. Iterate averaging, whereby the estimated optimum is not the latest value of x_n but an average of a window of most recent values, can reduce the sensitivity. Robust SA is a further generalization involving a weighted (based on $\{a_n\}$) average. Addressing the choice of $\{a_n\}$—as well as other issues such as how to project onto the feasible region Θ, which might be specified indirectly (e.g., in a mathematical programming formulation) and possibly involve "noisy" constraints that also have to be estimated along with the objective function—is one of the main topics of Chap. 6. Robust SA and other generalizations and extensions of iterate averaging, along with finite-time analysis of the resulting algorithms, are described in more detail in Chap. 7.

The rest of this chapter is organized as follows. Section 5.2 summarizes the finite difference approaches, including the simultaneous perturbation method that is especially useful in high-dimensional problems. Section 5.3 describes the direct gradient estimation techniques of perturbation analysis, the likelihood ratio/score function method, and the weak derivatives method (also known as measure-valued differentiation) in detail, including illustrative examples (Sects. 5.3.3 and 5.3.4), a summary of some basic theoretical tools (Sect. 5.3.5), and guidelines for the practitioner (Sect. 5.3.6). Section 5.4 treats the more recent work on quantile sensitivity estimation. Section 5.5 describes some new developments in using direct stochastic gradients in simulation optimization. Section 5.6 concludes by briefly describing the main application areas in historical context and providing some primary reference material for further reading. The content in Sects. 5.2 and 5.3 draws heavily from Fu [14], whereas the exposition on quantile sensitivities (Sect. 5.4) and recently proposed approaches to using stochastic gradients in simulation optimization (Sect. 5.5) is new.

An important note on notation: f will later be used to denote a probability density function (p.d.f.) rather than the objective function, so it will be replaced by J henceforth; also, what has been referred to as the decision variable(s) x in the optimization problem (5.1) and in the general SA recursion (5.4) will become the parameter (vector) θ. Specifically, the goal of the rest of the chapter will be to estimate either

$$\nabla_\theta J(\theta) = \nabla_\theta \mathsf{E}[Y(\theta,\xi)],$$

or in Sect. 5.4

$$q'_\alpha(\theta),$$

where q_α is defined by (5.3) and θ is scalar.

5.2 Indirect Gradient Estimators

We divide the approaches to stochastic gradient estimation into two main categories—indirect and direct—which we now more specifically define. An indirect gradient estimator usually has two characteristics: (a) it only estimates an *approximation* of the true gradient value, e.g., via a secant approximation in the scalar case; and (b) it uses only function evaluations (performance measure output samples) from the original (unmodified) system of interest. When used in SA, the resulting algorithms are commonly referred to as *gradient-free* or *stochastic zeroth-order* methods. A direct gradient estimator tries to estimate the true gradient using some additional analysis of the underlying stochastics of the model. More specifically, we will refer to the indirect gradient estimation approach as one in which the simulation output is treated as coming out of a given black box, by which we mean it satisfies two assumptions: (a) no knowledge of the underlying mechanics of the simulation model is used in deriving the estimators, such as knowing the input probability distributions; and (b) no changes are made in the execution of the simulation model itself, such as changing the input distribution for importance sampling. Note that this entails satisfying *both* assumptions; many of the direct gradient estimation techniques can be *implemented* without changing anything in the underlying simulation, but they may require some knowledge of the simulation model, such as the input distributions or some of the system dynamics. In the case of stochastic simulation, as opposed to online estimation based on an actual system, it could be argued that to carry out the simulation most of these mechanics need to be known, i.e., one cannot carry out a stochastic simulation without specifying the input distributions. Here, we simply use the two assumptions to distinguish between the two categories of approaches and not to debate whether an estimator is "model" dependent or not. In terms of stochastic approximation algorithms, indirect and direct gradient estimators generally correspond to Kiefer–Wolfowitz and Robbins–Monro algorithms, respectively.

We describe two indirect gradient estimators: finite differences and simultaneous perturbation. Following our definition, these approaches require no knowledge of the workings of the simulation model, which is treated as a black box.

5.2.1 Finite Differences

The straightforward brute-force method for estimating a gradient is simply to use finite differences, i.e., perturbing the value of each component of θ separately while holding the other components at the nominal value. If the value of the perturbation is too small, the resulting difference estimator could be extremely noisy, because the output is stochastic; hence, there is a trade-off between bias and variance in making this selection, and unless all components of the parameter vector are suitably "standardized" a priori, this choice must be done for each component separately, which could be a burdensome task for high-dimensional problems.

The simplest finite difference estimator is the one-sided forward difference gradient estimator, with ith component given by

$$\frac{Y(\theta + c_i e_i, \xi_{2,i}) - Y(\theta, \xi_{1,i})}{c_i}, \tag{5.5}$$

where c is the vector of differences (c_i the perturbation in the ith direction) and e_i denotes the unit vector in the ith direction.

A more accurate estimator is the two-sided symmetric (or central) difference gradient estimator, with ith component given by

$$\frac{Y(\theta + c_i e_i, \xi_{2,i}) - Y(\theta - c_i e_i, \xi_{1,i})}{2c_i}, \tag{5.6}$$

which corresponds to the estimator used in the original Kiefer–Wolfowitz SA algorithm. The variance reduction technique of common random numbers (CRN) can be thought of being the case where $\xi_{1,i} = \xi_{2,i} = \xi_i$. In stochastic simulation, using CRN can reduce the variance of the gradient estimators substantially, although in practice synchronization is an important issue, since merely using the same random number seeds is typically not effective. The symmetric difference estimator given by (5.6) is more accurate, but it requires $2d$ objective function estimates (simulation replications) per gradient estimate, as opposed to $d + 1$ function estimates (simulation replications) for the one-sided estimator given by (5.5).

5.2.2 Simultaneous Perturbation

Introduced by Spall in 1992 [41], simultaneous perturbation stochastic approximation (SPSA) is targeted at high-dimensional problems, due to the property that the number of simulation replications needed to form an estimator of the gradient is independent of the dimension of the parameter vector. Specifically, the ith component of the simultaneous perturbation (SP) gradient estimator is given by

$$\frac{Y(\theta+\mathbf{C}\Delta,\xi_2)-Y(\theta-\mathbf{C}\Delta,\xi_1)}{2c_i\Delta_i}, \tag{5.7}$$

where $\Delta=(\Delta_1,\ldots,\Delta_d)$ is a d-dimensional vector of perturbations, which are generally assumed i.i.d. as a function of iteration and independent across components. In this case, $\mathbf{C}$ contains the set of differences for each component as a diagonal matrix with the differences $\{c_i\}$ on the diagonal. The key difference between this estimator and a finite difference estimator is that the *numerator* of (5.7)—corresponding to a difference in the function estimates—is the *same* for all components (i.e., independent of i), whereas the numerator in the symmetric difference estimator given by (5.6) involves a different pair of function estimates for each component (i.e., is a function of i). Thus, the full gradient estimator requires only *two* function estimates, *regardless of the size of the dimension d*. On the other hand, since d random numbers must be generated to produce the perturbation sequence Δ at each iteration, if generating function estimates is relatively inexpensive in terms of computation, then this procedure may not be computationally superior to the previous finite difference approaches. In most simulation optimization settings, however, generating simulation output responses $Y(\theta,\xi)$ is relatively quite expensive. SPSA has also been applied in situations where the output $J(\theta)$ is actually deterministic (no random Y) but expensive to generate, e.g., requires computationally intensive finite-element method calculations.

The key requirement on the perturbation sequence to guarantee w.p.1 convergence of SPSA is that each term have mean zero and finite inverse second moments. Thus, the normal (Gaussian) distribution is prohibited, and the most commonly used distribution is the symmetric Bernoulli, whereby the perturbation takes the positive and negative (equal in magnitude, e.g., ± 1) value w.p. 0.5. Intuitively, convergence comes about from the averaging property of the random directions selected at each iteration, i.e., in the long-run, each component will converge to the correct gradient even if at any particular iteration the estimator may appear odd. Thus, an interesting alternative to using random perturbation sequences $\{\Delta\}$ is to use deterministic sequences [3,46], analogous to the use of quasi-Monte Carlo.

A very similar gradient estimator for use in SA algorithms is the random directions gradient estimator [33], whose ith component is given by

$$\frac{[Y(\theta+c\Delta,\xi_2)-Y(\theta-c\Delta,\xi_1)]\,\Delta_i}{2c_i}. \tag{5.8}$$

Instead of dividing by the perturbation component, the difference term multiplies the component. Thus, normal distributions can be used for the perturbation sequence, and convergence requirements translate the moment condition to a bound on the second moment, as well as zero mean. Of course, a correspondence to the SP estimator can be made by simply taking the componentwise inverse, but in practice the performance of the two resulting SA algorithms differs substantially.

An extensive and frequently updated annotated bibliography for SPSA can be found on the World Wide Web at http://www.jhuapl.edu/SPSA/.

5.3 Direct Gradient Estimators

When available, direct gradient estimators offer the following advantages:

- They are generally unbiased, which results in faster convergence rates when implemented in a simulation optimization algorithm, whether stochastic approximation, sample average approximation, or response surface methodology.
- They eliminate the need to determine appropriate values for the finite difference sequences, which influence the accuracy of the estimator. Smaller values of c in (5.5)–(5.8) lead to lower bias but usually at the cost of increased variance, to the point of possibly giving the wrong sign for small enough values.
- They are generally more computationally efficient.

We begin with the case where the output is an expectation, and write the output Y in terms of all the input random variables $\mathbf{X} \equiv \{X_i\}$:

$$J(\theta) = \mathsf{E}[Y(\mathbf{X})] = \mathsf{E}[Y(X_1,\ldots,X_N)], \tag{5.9}$$

where N is a fixed finite number, and for notational brevity, the display of the randomness ξ and the dependence on the parameter θ will often be suppressed in the following derivations. The various direct gradient estimation techniques are distinguished by their treatment of the dependence on θ in (5.9):

sample (pathwise) vs. *measure (distributional).*

As illustrated in the examples that follow, many settings allow either dependence, leading to different gradient estimators.

To derive direct gradient estimators, we write the expectation using what is sometimes called the law of the unconscious statistician:

$$\mathsf{E}[Y(\mathbf{X})] = \int y dF_Y(y) = \int Y(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}), \tag{5.10}$$

where F_Y and $F_{\mathbf{X}}$ denote the distributions of Y and $\mathbf{X}$, respectively. In fact, when estimating expected value performance, stochastic simulation can be viewed as a way of implicitly carrying out this relationship, i.e., the simulation model is given input random variables with known distributions, and produces samples of output random variables, for which we would like to characterize the distributions.

For simplicity in discussion, we will assume henceforth that the parameter θ is scalar, because the vector case can be handled by taking each component individually. In view of the right-hand side of (5.10), we revisit the question as to the location of the parameter in a stochastic setting. Putting it in the sample performance $Y(\cdot;\theta)$ corresponds to the view of perturbation analysis (PA), whereas if it is absorbed in the distribution $F(\cdot;\theta)$, then the approach follows that of the likelihood ratio (LR) method (also known as the score function (SF) method) or weak derivatives (WD) (also known as measure-valued differentiation). In the general setting where the parameter is a vector, it is possible that some of the components would be most naturally located in the sample performance, while others would be easily retained in the distributions, giving rise to a mixed approach. For example, in an (s,S) inventory control system, it might be most effective to use PA for the control parameters (decision variables) s and S, and WD or LR/SF for the demand distribution parameters.

Let f_X denote the p.d.f. of all of the input random variables (not to be confused with the original objective function f as defined by (5.1)). Differentiating (5.10), and assuming an interchange of integration and differentiation is permissible, we write two cases:

$$\frac{d\mathsf{E}[Y(X)]}{d\theta} = \begin{cases} \displaystyle\int_{-\infty}^{\infty} Y(x)\frac{df_X(x;\theta)}{d\theta}dx & (5.11) \\[2ex] \displaystyle\int_0^1 \frac{dY(X(\theta;u))}{d\theta}du, & (5.12) \end{cases}$$

where x and u, as well as the integrals, are N-dimensional. For notational simplicity, these N-dimensional multiple integrals are written as a single integral throughout, and we also assume one random number u produces one random variate x. In (5.11), the parameter appears in the distribution directly, whereas in (5.12), the underlying uncertainty is considered the uniform random numbers; this dichotomy corresponds to the respective distributional (measure) and pathwise (sample) dependencies.

For expositional ease in introducing the approaches, we begin by assuming that the parameter only appears in X_1, which is generated *independently* of the other input random variables. So for the case of (5.12), we use the chain rule to write

$$\begin{aligned} \frac{d\mathsf{E}[Y(X)]}{d\theta} &= \int_0^1 \frac{dY(X_1(\theta;u_1),X_2,\ldots)}{d\theta}du \\ &= \int_0^1 \frac{\partial Y}{\partial X_1}\frac{dX_1(\theta;u_1)}{d\theta}du. \end{aligned} \tag{5.13}$$

In other words, the estimator takes the form

$$\frac{\partial Y(X)}{\partial X_1}\frac{dX_1}{d\theta}, \tag{5.14}$$

where the parameter appears in the transformation from random number to random variate, and the derivative is expressed as the product of a sample path derivative and derivative of a random variable. The issue of what constitutes the latter will be taken up shortly, but this approach is called infinitesimal perturbation analysis (IPA). For the $M/M/1$ queue, the sample path derivative could be derived using Lindley's equation, relating the time in system of a customer to the service times (and interarrival times, which are not a function of the parameter).

Assume that X_1 has marginal p.d.f. $f_1(\cdot;\theta)$ and that the joint p.d.f. for the remaining input random variables $(X_2,\ldots)$ is given by f_{-1}, which has no (functional) dependence on θ. Then the assumed independence gives $f_X = f_1 f_{-1}$, and the expression (5.11) involving differentiation of a density (measure) can be further manipulated using the product rule of differentiation to yield the following:

$$\frac{d\mathsf{E}[Y(X)]}{d\theta} = \int_{-\infty}^{\infty} Y(x)\frac{\partial f_1(x_1;\theta)}{\partial\theta} f_{-1}(x_2,\ldots)dx \tag{5.15}$$

$$= \int_{-\infty}^{\infty} Y(x)\frac{\partial \ln f_1(x_1;\theta)}{\partial\theta} f_X(x)dx. \tag{5.16}$$

In other words, the estimator takes the form

$$Y(X)\frac{\partial \ln f_1(X_1;\theta)}{\partial\theta}. \tag{5.17}$$

Since the term $\frac{\partial \ln f_1(\cdot;\theta)}{\partial\theta}$ is the well-known (efficient) score function in statistics, this approach has been called the score function (SF) method. The other name given to this approach—the likelihood ratio (LR) method—comes from the closely related likelihood ratio function given by

$$\frac{f_1(\cdot;\theta)}{f_1(\cdot;\theta_0)},$$

which when differentiated with respect to θ gives

$$\frac{\partial f_1(\cdot;\theta)/\partial\theta}{f_1(\cdot;\theta_0)},$$

which is equal to the score function upon setting $\theta_0 = \theta$.

On the other hand, if instead of expressing the right-hand side of (5.15) as (5.16), the density derivative is written as

$$\frac{\partial f_1(x_1;\theta)}{\partial\theta} = c_1(\theta)\left(f_1^{(+)}(x_1;\theta) - f_1^{(-)}(x_1;\theta)\right),$$

it leads to the following relationship:

$$\frac{d\mathsf{E}[Y(X)]}{d\theta} = \int_{-\infty}^{\infty} Y(x)\frac{\partial f_X(x;\theta)}{\partial\theta}dx$$

$$= c_1(\theta)\left(\int_{-\infty}^{\infty} Y(x)f_1^{(+)}(x_1;\theta)f_{-1}(x_2,\ldots)dx - \int_{-\infty}^{\infty} Y(x)f_1^{(-)}(x_1;\theta)f_{-1}(x_2,\ldots)dx\right).$$

The triple $\left(c_1(\theta), f_1^{(+)}, f_1^{(-)}\right)$ constitutes a weak derivative (WD) for f_1, which is in general not unique. The corresponding WD estimator is of the form

$$c_1(\theta)\left(Y(X_1^{(+)}, X_2, \ldots) - Y(X_1^{(-)}, X_2, \ldots)\right), \tag{5.18}$$

where $X_1^{(-)} \sim f_1^{(-)}$ and $X_1^{(+)} \sim f_1^{(+)}$, henceforth often abbreviated $X^{(\pm)} \sim f^{(\pm)}$. The term *weak derivative* comes about from the possibility that $\frac{\partial f_1(\cdot;\theta)}{\partial\theta}$ in (5.15) may not be proper, but its *integral* may be well-defined, e.g., it might involve delta-functions (impulses), corresponding to probability mass functions (p.m.f.s) of discrete distributions. Note that even for a given WD representation, only the *marginal* distributions for the two random variables $X^{(\pm)}$ are specified, i.e., their joint distribution is not constrained, so the "estimator" given by (5.18) is not really completely specified. Since (5.18) is a difference of two terms that appear similar, one might expect that generating the two random variables using CRN rather than independently would be beneficial, and it is indeed true in many situations, but such a conclusion is problem dependent. However, for the Hahn–Jordan WD representation (to be described later in Sect. 5.3.2), independent generation turns out to be the method that minimizes the variance of the WD estimator [44].

If in the expression (5.12), the interchange of expectation and differentiation does not hold (e.g., if Y is an indicator function), then as long as there is more than one input random variable, appropriate conditioning will often allow the interchange as follows:

$$\frac{d\mathsf{E}[Y(X)]}{d\theta} = \int_0^1 \frac{d\mathsf{E}[Y(X(\theta;u))|Z(u)]}{d\theta}du, \tag{5.19}$$

where $Z \subseteq \{X_1, \ldots, X_N\}$. This conditioning is known as smoothed perturbation analysis (SPA), because it seeks to "smooth" out a discontinuous function. SPA leads to an estimator of the following form:

$$\frac{\partial\mathsf{E}[Y(X)|Z]}{\partial X_1}\frac{dX_1}{d\theta}. \tag{5.20}$$

Note that taking Z as the entire set leads back to (5.14).

Remark. For SPA, the conditioning in (5.19) was done with respect to a subset of the input random variables only. Further conditioning can be done on events in the system, which leads to an estimator of the following general form:

$$\frac{dY}{d\theta} + \mathsf{E}_Z[\delta Y|\mathscr{B}]\frac{d\mathsf{P}_Z(\mathscr{B})}{d\theta}, \tag{5.21}$$

where the subscript indicates a corresponding conditional expectation/probability, $\mathscr{B}$ is an appropriately selected event, and δY represents a change in the performance measure under the conditioned (usually called "critical") event. In this case, if the probability rate $\frac{d\mathsf{P}_Z(\mathscr{B})}{d\theta}$ is 0, the estimator (5.21) also reduces to IPA. On the other hand, if the IPA term $\frac{dY}{d\theta}$ is zero, the estimator may coincide with the WD estimator in certain cases, with correspondences between $c(\theta)$ and the probability rate, and between the difference term in (5.18) and the conditional expectation in (5.21).

5.3.1 Derivatives of Random Variables

PA estimators—e.g., those shown in (5.14), (5.20), (5.21)—require the notion of derivatives of random variables. The mathematical problem for defining such derivatives consists of constructing a family of random variables parameterized by θ on a *common* probability space, with the point of departure being a set of parameterized distribution functions $\{F(\cdot;\theta)\}$. We wish to construct $X(\theta) \sim F(\cdot;\theta)$ s.t. $\forall\theta \in \Theta$, $X(\theta)$ is differentiable w.p.1. The sample derivative is then defined in the intuitive manner as

$$\frac{dX(\theta,\omega)}{d\theta} = \lim_{\Delta\theta\to 0}\frac{X(\theta+\Delta\theta,\omega)-X(\theta,\omega)}{\Delta\theta},$$

where ω denotes a sample point in the underlying probability space. If the distribution of X is known, we have [21,42]

$$\frac{dX(\theta)}{d\theta} = -\frac{\partial F(X;\theta)/\partial\theta}{\partial F(X;\theta)/\partial X}, \tag{5.22}$$

where we use the (slightly abusive) notation $\dfrac{\partial F(X;\theta)}{\partial X} = \dfrac{\partial F(x;\theta)}{\partial x}\bigg|_{x=X}$.

Definition. For a distribution function $F(x;\theta)$, θ is said to be a *location* parameter if $F(x+\theta;\theta)$ does not depend on θ; θ is said to be a *scale* parameter if $F(x\theta;\theta)$ does not depend on θ; and θ is said to be a *generalized scale* parameter if $F(\bar{\theta}+x\theta;\theta)$ does not depend on θ, for some fixed $\bar{\theta}$ (usually a location parameter) not dependent on θ.

In these special cases, one can use the following sample derivatives for the three respective cases (location, scale, generalized scale):

$$\frac{dX}{d\theta}=1,\quad \frac{dX}{d\theta}=\frac{X}{\theta},\quad \frac{dX}{d\theta}=\frac{X-\bar{\theta}}{\theta}.$$

The most well-known example is the normal distribution, with the mean being a location parameter and the standard deviation a generalized scale parameter. Similarly, the two parameters in the Cauchy, Gumbel (extreme value), and logistic distributions also correspond to location and generalized scale parameters. Other examples include the mean of the exponential being a scale parameter; and the mean of the uniform distribution being a location parameter, with the half-width being a generalized scale parameter. In the special case $U(0,\theta)$, the single parameter is an ordinary scale parameter. Also, for $\mathscr{N}(\theta,(\theta\sigma)^2)$, θ is an ordinary scale parameter. See Table 5.1 in Sect. 5.3.3 for more examples.

Lastly, note that for a given distribution, there may be multiple ways to generate a random variate, which leads to different derivatives, some of which may be unbiased and some of which may not. This is called the role of representations, and is illustrated with a simple example (Example 5.5) in Sect. 5.3.3.

5.3.2 Derivatives of Measures

As we have seen already, both the LR/SF and WD estimators rely on differentiation of the underlying measure, so the parameters of interest should appear in the underlying (input) distributions. If this is not the case, then one approach is to try to "push" the parameter out of the performance measure into the distribution, so that the usual differentiation can be carried out. This is achieved by a change of variables, which is problem dependent.

Recall that we introduced the idea of a weak derivative by expressing the derivative of a density (p.d.f.) as an appropriately normalized difference of two p.d.f.s, i.e., the triple $(c(\theta),f^{(+)},f^{(-)})$ satisfying

$$\frac{\partial f(x;\theta)}{\partial\theta}=c(\theta)\left(f^{(+)}(x;\theta)-f^{(-)}(x;\theta)\right).$$

This idea can be generalized without the need for a differentiable density, as long as the integral exists with respect to a certain set of (integrable) "test" functions, say $\mathscr{L}$, e.g., the set of continuous bounded functions.

Definition. The triple $(c(\theta),F^{(+)},F^{(-)})$ is called a *weak derivative* for distribution (c.d.f.) F if for all functions $g\in\mathscr{L}$ (not a function of θ),

$$\frac{d}{d\theta}\int g(x)dF(x;\theta)=c(\theta)\left(\int g(x)dF^{(+)}(x;\theta)-\int g(x)dF^{(-)}(x;\theta)\right).$$

Remark. As mentioned earlier, the derivative is "weak" in the sense that the density derivative may not be defined in the usual sense, but in terms of generalized functions integrable with respect to the functions in $\mathscr{L}$, as in the "definition" of a delta function in terms of its integral. The concept of a weak derivative need not be restricted to probability measures, but any finite signed measures. Lastly, note that a WD gradient estimate may require as many as $2d$ additional simulations for the vector case (a pair for each component), unlike LR/SF and IPA estimators, which will always require just a single simulation.

One choice for the weak derivative (density) that is readily available is

$$\frac{\partial f}{\partial \theta} = c\left(f^{(+)} - f^{(-)}\right), \tag{5.23}$$

where

$$f^{(-)} = \frac{1}{c}\left(\frac{\partial f}{\partial \theta}\right)^{-}, \quad f^{(+)} = \frac{1}{c}\left(\frac{\partial f}{\partial \theta}\right)^{+}, \tag{5.24}$$

$(x)^{+} \equiv \max\{x,0\}$, $(x)^{-} \equiv \max\{-x,0\}$, and $c = \int\left(\frac{\partial f}{\partial \theta}\right)^{+} dx = \int\left(\frac{\partial f}{\partial \theta}\right)^{-} dx$, using the fact that

$$\int f(x)dx = 1 \quad \Longrightarrow \quad \int \frac{\partial f}{\partial \theta} dx = 0.$$

The representation given by (5.23) and (5.24) is the Hahn–Jordan decomposition, which will always exist for probability measures, and results in a decomposition involving two measures with complementary support. It can be shown in this case that generating the two random variables according to $f^{(+)}$ and $f^{(-)}$ independently minimizes variance for the WD estimator [44].

Remark. The representation is clearly not unique. In fact, for any non-negative integrable function h, we have

$$\frac{\partial f}{\partial \theta} = c\Big([f^{(-)} + h] - [f^{(+)} + h]\Big) = \tilde{c}\Big([f^{(-)} + h]/(1 + \int h) - [f^{(+)} + h]/(1 + \int h)\Big),$$

where $\tilde{c} = c(1 + \int h)$. Thus, one way to obtain the estimator using the original simulation is to choose a representation in which both measures have the same support as the original measure. Then importance sampling can be applied, so that the original simulation can be used to generate the estimator without the need for simulating the system under alternative input distributions. Perhaps the most direct way to achieve this is to add the original measure itself to both $f^{(-)}$ and $f^{(+)}$ and renormalize appropriately, i.e., choose $h = f$ above:

$$\frac{\partial f}{\partial \theta} = 2c\left([f^{(-)} + f]/2 - [f^{(+)} + f]/2\right).$$

5.3.3 Input Distribution Examples

We now demonstrate some of these concepts on a single random variable. Section 5.3.4 then considers the two examples introduced at the beginning of the chapter (stochastic activity network and single-server queue).

Example 5.1. Let $X \sim \exp(\theta)$, an exponential random variable with mean θ and p.d.f. given by

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \mathbf{1}\{x > 0\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The usual construction of the random variable is

$$X(\theta;u) = -\theta \ln u,$$

where u represents a random number. Differentiating both expressions, we get

$$\begin{aligned}
\frac{\partial f(x;\theta)}{\partial\theta} &= \left[\frac{x}{\theta^2}\frac{1}{\theta}e^{-x/\theta} - \frac{1}{\theta^2}e^{-x/\theta}\right]\mathbf{1}\{x > 0\} \\
&= f(x;\theta)\left[\frac{x}{\theta^2} - \frac{1}{\theta}\right] \\
&= \frac{1}{\theta}\left[\frac{x}{\theta^2}e^{-x/\theta}\mathbf{1}\{x > 0\} - f(x;\theta)\right] \\
&= \frac{1}{\theta e}\left[\frac{e}{\theta}\left(\frac{x}{\theta} - 1\right)e^{-x/\theta}\mathbf{1}\{x > \theta\} - \frac{e}{\theta}\left(1 - \frac{x}{\theta}\right)e^{-x/\theta}\mathbf{1}\{0 < x \le \theta\}\right], \\
\frac{dX(\theta;u)}{d\theta} &= -\ln u \;=\; \frac{X(\theta;u)}{\theta}.
\end{aligned}$$

In the third and fourth lines above, the density derivative (which is itself *not* a density) has been expressed as the difference of two densities multiplied by a constant. This demonstrates that the weak derivative representation is not unique, with the last decomposition being the Hahn–Jordan decomposition, noting that $x = \theta$ is the point at which $df(x;\theta)/d\theta$ changes sign. The following correspond to the LR/SF, WD (a) & (b), and IPA estimators, respectively:

$$\begin{aligned}
&Y(X)\frac{1}{\theta}\left(\frac{X_1}{\theta} - 1\right), \\
&\frac{1}{\theta}\left[Y(X_{1a}^{(+)},\ldots) - Y(X_{1a}^{(-)},\ldots)\right], \quad \frac{1}{\theta e}\left[Y(X_{1b}^{(+)},\ldots) - Y(X_{1b}^{(-)},\ldots)\right], \\
&\frac{dY}{dX_1}\frac{X_1}{\theta},
\end{aligned}$$

where $X_{1a}^{(-)} \sim exp(\theta)$ and $X_{1a}^{(+)} \sim Erl(2,\theta)$, where "$Erl$" is an Erlang distribution (see Table 5.1), and $X_{1b}^{(-)} \sim \theta - \text{trunc}(Erl(2,\theta),[0,\theta])$ and $X_{1b}^{(+)} \sim \theta + Erl(2,\theta)$, where "trunc$(F,[a,b])$" represents a distribution (c.d.f.) F truncated to the range $[a,b]$. Since an $Erl(2,\theta)$ distribution can be generated by the sum of two i.i.d. exponentially distributed random variables, one way to realize the first WD estimator would be to use $X_{1a}^{(-)} = X_1$ and then generate another $\tilde{X}_1 \sim exp(\theta)$ independent of the original X_1, giving the WD estimator

$$\frac{1}{\theta}\left[Y(X_1+\tilde{X}_1,\ldots) - Y(X_1,\ldots)\right].$$

The following is a simple example that demonstrates that the WD estimator is more broadly applicable than the LR/SF estimator.

Example 5.2. Let $X \sim U(0,\theta)$. Then we have the following:

$$\begin{aligned}
f(x;\theta) &= \frac{1}{\theta}\mathbf{1}\{0<x<\theta\},\\
X(\theta;u) &= u\theta,\\
\frac{\partial f(x;\theta)}{\partial\theta} &= \frac{1}{\theta}\left[\delta(\theta-x) - \frac{1}{\theta}\mathbf{1}\{0<x<\theta\}\right] \qquad (5.25)\\
&= \frac{1}{\theta}\left[\delta(\theta-x) - f(x;\theta)\right],\\
\frac{dX(\theta;u)}{d\theta} &= u \;=\; \frac{X(\theta;u)}{\theta},
\end{aligned}$$

where we define the Dirac δ-function as the "derivative" of a step function by

$$\mathbf{1}\{x \geq \theta\} = \int_{-\infty}^{x} \delta(y-\theta)dy. \qquad (5.26)$$

On the right-hand side of Eq. (5.25), we have the difference of densities for a mass at θ and the original $U(0,\theta)$ distribution, respectively, i.e., the weak derivative representation $(1/\theta,\theta,F)$, where θ indicates a deterministic distribution with mass at θ. So, for example, the estimator in (5.18) would be given by

$$\frac{1}{\theta}\left(Y(\theta,X_2,\ldots) - Y(X_1,X_2,\ldots)\right).$$

This is a case where the LR/SF estimator is ill-defined, due to the δ-function. Another example is the following.

Example 5.3. Let $X \sim Par(\alpha,\theta)$, which represents the Pareto distribution with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$, and p.d.f. given by

$$f(x) = \theta^{\alpha}\alpha x^{-(\alpha+1)}\mathbf{1}\{x \geq \theta\}.$$

Once again the LR/SF estimator does not exist (for θ), due to the appearance of the parameter in the indicator function that controls the support of the distribution, whereas WD estimators can be derived (see Table 5.1 at the end of the section).

However, the very general exponential family of distributions leads to a nice form for the LR/SF estimator.

Example 5.4. Let θ denote the vector of parameters in a p.d.f. that can be written in the following form:

$$f(x;\theta) = k(\theta)\exp\left(\sum_i v_i(\theta)t_i(x)\right)h(x),$$

where the functions h and $\{t_i\}$ are independent of θ, and the functions k and $\{v_i\}$ do not involve the argument. Then it is straightforward to derive

$$\frac{\partial \ln f(x;\theta)}{\partial\theta} = \frac{\nabla k(\theta)}{k(\theta)} + \sum_i \nabla v_i(\theta)t_i(x).$$

Examples include the normal, gamma, Weibull, and exponential, for the continuous case, and the binomial, Poisson, and geometric for the discrete case.

As mentioned in Sect. 5.3.1, the application of PA (both IPA or SPA) depends on the way the stochastic processes in the system are represented. We illustrate this through a simple random variable example. In terms of simulation, this means that a different representation used to generate the random variable could lead to a different PA estimator. For instance, in Example 5.1, an alternative equivalent representation is $X = -\theta\ln(1-u)$, which in this case leads to the same IPA estimator X/θ. Since the underlying distribution is identical for the different representations, the LR/SF and WD estimators are not dependent on the process representation, but as noted earlier, the same distribution has infinitely many possible WD estimators.

Example 5.5. For $\theta \in (0,1)$, let

$$X \sim \begin{cases} U(0,1) \text{ w.p. } \theta, \\ U(1,2) \text{ w.p. } 1-\theta, \end{cases}$$

a mixture of two uniform distributions, with $\mathsf{E}[X] = 1.5-\theta$ and $d\mathsf{E}[X]/d\theta = -1$. A straightforward construction/representation using two random numbers is

$$X = \mathbf{1}\{U_1 \le \theta\}U_2 + \mathbf{1}\{U_1 > \theta\}(1+U_2), \tag{5.27}$$

where $U_1,\ U_2 \sim U(0,1)$ are *independent*. However, since

$$\frac{dX}{d\theta} = 0 \text{ w.p.1,}$$

this clearly leads to a biased estimator. Note that viewed as a function of θ, X jumps from $1+U_2$ down to U_2 at $\theta = U_1$. However, an unbiased estimator can be obtained by using the following construction in which the “coin flipping” and uniform generation are correlated:

$$X = \mathbf{1}\{U \le \theta\}\frac{U}{\theta} + \mathbf{1}\{U > \theta\}\left(1 + \frac{U-\theta}{1-\theta}\right), \text{ where } U \sim U(0,1),$$

$$\Longrightarrow \frac{dX}{d\theta} = -\frac{U}{\theta^2}\mathbf{1}\{U \le \theta\} + \frac{U-1}{(1-\theta)^2}\mathbf{1}\{U > \theta\},$$

which is unbiased (has expectation equal to $d\mathsf{E}[X]/d\theta = -1$). This construction is based on the property that the distributions of the random variable U/θ under the condition $\{U < \theta\}$ and the random variable $(U-\theta)/(1-\theta)$ under the condition $\{U \ge \theta\}$ are both $U(0,1)$. From a simulation perspective, this representation has the additional advantage of requiring only a single random number to generate X instead of two as in the previous construction. In this case, the construction also corresponds to the inverse transform representation. In terms of the derivative, the crucial property of the representation is that X is continuous across $\theta = U$. One can easily construct other single random number representations that do not have this desirable characteristic, e.g.,

$$X = \mathbf{1}\{U \le \theta\}(1 - \frac{U}{\theta}) + \mathbf{1}\{U > \theta\}\left(1 + \frac{1-U}{1-\theta}\right), \text{ where } U \sim U(0,1),$$

$$\Longrightarrow \frac{dX}{d\theta} = \mathbf{1}\{U \le \theta\}\frac{U}{\theta^2} + \mathbf{1}\{U > \theta\}\frac{1-U}{(1-\theta)^2},$$

which is biased (has expectation +1), the intuitive reason being the discontinuity of X at $U = \theta$, where it jumps from 0 to 2.

For the first representation given by (5.27), which used two random numbers and led to a biased IPA estimator, SPA can be applied by conditioning on U_2 as follows:

$$X = \mathsf{E}[X_1|U_2] = U_2\theta + (1+U_2)(1-\theta) = 1 + U_2 - \theta,$$

leading to the trivially unbiased “estimator” $dX/d\theta = -1$.

To derive the WD and LR/SF estimators, the p.d.f. is given by

$$f(x;\theta) = \theta\mathbf{1}\{x \in (0,1]\} + (1-\theta)\mathbf{1}\{x \in (1,2]\},$$

so

$$\frac{\partial f(x;\theta)}{\partial\theta} = \mathbf{1}\{x\in(0,1]\} - \mathbf{1}\{x\in(1,2]\}, \tag{5.28}$$

$$\frac{\partial \ln f(x;\theta)}{\partial\theta} = \frac{1}{\theta}\mathbf{1}\{x\in(0,1]\} - \frac{1}{1-\theta}\mathbf{1}\{x\in(1,2]\}, \tag{5.29}$$

and the obvious WD from (5.28) is simply $(1, U(0,1), U(1,2))$, corresponding to the Hahn–Jordan decomposition, whereas the LR/SF estimator from (5.29) is given by

$$Y(X)\cdot\begin{cases}\frac{1}{\theta} & \text{if } X\le 1,\\ -\frac{1}{1-\theta} & \text{otherwise.}\end{cases}$$

However, as noted in the remark at the end of Sect. 5.3.2, the WD representation is not unique, so for example, one can add and subtract a $U(0,1)$ density in (5.28) to get

$$\mathbf{1}\{x\in(0,1]\} + \mathbf{1}\{x\in(0,1]\} - \mathbf{1}\{x\in(1,2]\} - \mathbf{1}\{x\in(0,1]\}$$

$$= 2\left[\mathbf{1}\{x\in(0,1]\} - \frac{1}{2}\mathbf{1}\{x\in(0,2]\}\right],$$

yielding the alternative WD representation $(2, U(0,1), U(0,2))$.

Discrete distributions present separate challenges for the different approaches. Basically, when the parameter appears in the support *probabilities*, then LR/SF and WD can be easily applied, whereas IPA is in general not applicable. The reverse is true, however, if the parameter appears instead in the support *values*. The next two examples demonstrate this dichotomy, where we work directly with the probability mass function (p.m.f.) $p(x;\theta) = \mathsf{P}(X = x)$, instead of densities with δ-functions. Let $Ber(p;a,b)$ denote a Bernoulli distribution that takes value a w.p. p and b w.p. $1-p$. We start with an example where the parameter θ is the Bernoulli probability.

Example 5.6. Let $X \sim Ber(\theta;a,b)$, $a\neq b$, which has p.m.f.

$$p(x;\theta) = \theta\mathbf{1}\{x=a\} + (1-\theta)\mathbf{1}\{x=b\},$$

so

$$\frac{\partial p}{\partial\theta} = \mathbf{1}\{x=a\} - \mathbf{1}\{x=b\},$$

which can be viewed as the difference of two (deterministic) masses at a and b (with $c(\theta)=1$), and is the Hahn–Jordan decomposition in this case. For the LR/SF estimator, we have

$$\frac{\partial \ln p}{\partial\theta} = \frac{\mathbf{1}\{x=a\} - \mathbf{1}\{x=b\}}{\theta\mathbf{1}\{x=a\} + (1-\theta)\mathbf{1}\{x=b\}} = \frac{1}{\theta}\mathbf{1}\{x=a\} - \frac{1}{1-\theta}\mathbf{1}\{x=b\}.$$

Note the similarity of both the WD and LR/SF estimators to the previous example. In this case, there is no way to construct X such that it will be differentiable w.p.1. For example, the natural construction/representation

$$X = a\mathbf{1}\{U \leq \theta\} + b\mathbf{1}\{U > \theta\}$$

yields $dX/d\theta = 0$ w.p.1, so IPA is not applicable.

In contrast, now consider an example where the parameter θ is one of the support values.

Example 5.7. Let $X \sim Ber(p;\theta;0),\ \theta \neq 0,\ E[X] = p\theta,\ dE[X]/d\theta = p$, which has p.m.f.

$$p(x;\theta) = p\mathbf{1}\{x = \theta\} + (1-p)\mathbf{1}\{x = 0\},$$

which is not differentiable with respect to θ, so LR/SF and WD estimators cannot be derived. The natural random variable construction

$$X = \theta\mathbf{1}\{U \leq p\}$$

leads to the unbiased

$$\frac{dX}{d\theta} = \mathbf{1}\{U \leq p\} = \mathbf{1}\{X = \theta\} = \frac{X}{\theta}.$$

The IPA estimator $dX/d\theta = \mathbf{1}\{X = \theta\}$ holds even if additional values are added to the underlying support, as long as the additional values do not involve θ. If θ enters into them, then the estimator can be easily modified to reflect the additional dependence.

For many common input distributions, Table 5.1 provides the necessary derivatives needed to implement each of the three methods (IPA, LR/SF, WD). Recall also that the two parameters in the Cauchy, Gumbel, and logistic distributions (not given in the table) are location and (generalized) scale parameters, so the IPA expressions would be the same as for the normal distribution. The entry for the mean of the normal has an interesting implementation for the WD estimator, based on the observation that a normally distributed random variable $\mathcal{N}(\mu, \sigma^2)$ can be generated via the product of a uniform $U(0,1)$ random number and a double-sided Maxwell $Mxw(\mu, \sigma^2)$ random variate (generated independently of each other, cf. [25], which also provides a method for generating from this distribution). Implementation using such pairs of independent $U(0,1)$ and $Mxw(\mu, \sigma^2)$ distributed random variates results in a WD derivative estimator with provably the lowest variance for any polynomial output function. Of course, in most settings the output is not polynomial; furthermore, the WD estimator requires an additional simulation replication per partial derivative.

Table 5.1 Derivatives for common input distributions (NA = not applicable)

Distribution	Parameterization	p.d.f./p.m.f. $f(x)/p(x)$	Support
Bernoulli	$Ber(p;a,b)$	$p\mathbf{1}\{x=a\}+(1-p)\mathbf{1}\{x=b\}$	$\{a,b\}$
geometric	$geo(p)$	$(1-p)^{x-1}p$	$\mathbb{Z}^+$
negative binomial	$negbin(n,p)$	$\begin{pmatrix} x-1 \\ n-1 \end{pmatrix}(1-p)^{x-n}p^n$	$\{n,n+1,\ldots\}$
binomial	$bin(n,p)$	$\begin{pmatrix} n \\ x \end{pmatrix}p^x(1-p)^{n-x}$	$\mathbb{N}$
Poisson	$Poi(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	$\mathbb{N}$
normal (Gaussian)	$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mathbb{R}$
Maxwell (2-sided)	$Mxw(\mu,\sigma^2)$	$\frac{(x-\mu)^2}{\sqrt{2\pi}\sigma^3}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mathbb{R}$
uniform	$U(a,b)$	$\frac{1}{b-a}$	$[a,b]$
exponential	$exp(\beta)$	$\beta^{-1}e^{-x/\beta}$	$x \geq 0$
Weibull	$Wei(\alpha,\beta)$	$\alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}$	$x \geq 0$
gamma	$gam(\alpha,\beta)$	$\frac{\beta^{-\alpha}x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)}$	$x \geq 0$
Erlang	$Erl(\alpha,\beta)$	$\frac{\beta^{-n}x^{n-1}e^{-x/\beta}}{(n-1)!}$	$x \geq 0$
Pareto	$Par(\alpha,\beta)$	$\alpha\beta^{\alpha}x^{-(\alpha+1)}$	$x \geq \beta$

$\alpha > 0$ is generally the shape parameter, μ is a location parameter, and $\beta > 0$ is a scale parameter; $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$; $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$; $geo(p) = negbin(1,p)$, $Erl(n,\beta) = gam(n,\beta)$ for $n \in \mathbb{Z}^+$; $Wei(1,\beta) = gam(1,\beta) = exp(\beta)$. The definition of the parameter β used here in the exponential, Weibull, and gamma distributions is the inverse of what is often found in the literature ([35] being a notable exception), but makes β a scale parameter. This is why the WD expressions below for the exponential, Weibull, and gamma distributions differ slightly from the table in [24].

input dist $X \sim F$	IPA $\frac{dX}{d\theta}$	LR/SF $\frac{\partial \ln f(X;\theta)}{\partial\theta}$	WD $(c(\theta),F^{(+)},F^{(-)})$
$Ber(\theta;a,b)$	NA	$\frac{1}{\theta}\mathbf{1}\{X=a\}$ $-\frac{1}{1-\theta}\mathbf{1}\{X=b\}$	$(1,a,b)$
$Ber(p;\theta,b)$	$\mathbf{1}\{X=\theta\}$	NA	NA
$geo(\theta)$	NA	$\frac{1}{\theta}+\frac{1-X}{1-\theta}$	$(\frac{1}{\theta},geo(\theta),negbin(2,\theta))$
$bin(n,\theta)$	NA	$\frac{X}{\theta}-\frac{n-X}{1-\theta}$	$(n,1+bin(n-1,\theta),bin(n-1,\theta))$
$Poi(\theta)$	NA	$\frac{X}{\theta}-1$	$(1,1+Poi(\theta),Poi(\theta))$
$\mathcal{N}(\theta,\sigma^2)$	1	$\frac{X-\theta}{\sigma^2}$	$\left(\frac{1}{\sqrt{2\pi}\sigma},\theta+Wei(2,\frac{1}{2\sigma^2}),\ \theta-Wei(2,\frac{1}{2\sigma^2})\right)$
$\mathcal{N}(\mu,\theta^2)$	$\frac{X-\mu}{\theta}$	$\frac{1}{\theta}\left[\left(\frac{x-\mu}{\theta}\right)^2-1\right]$	$(\frac{1}{\theta},Mxw(\mu,\theta^2),\mathcal{N}(\mu,\theta^2))$
$U(0,\theta)$	$\frac{X}{\theta}$	NA	$(\frac{1}{\theta},\theta,U(0,\theta))$
$U(\theta-\gamma,\theta+\gamma)$	1	NA	$(\frac{1}{2\gamma},\theta+\gamma,\theta-\gamma)$
$U(\mu-\theta,\mu+\theta)$	$\frac{X-\mu}{\theta}$	NA	$(\frac{1}{\theta},Ber(0.5;\mu-\theta,\mu+\theta),U(\mu-\theta,\mu+\theta))$
$exp(\theta)$	$\frac{X}{\theta}$	$\frac{1}{\theta}\left(\frac{X}{\theta}-1\right)$	$(\frac{1}{\theta},Erl(2,\theta),exp(\theta))$
$Wei(\alpha,\theta)$	$\frac{X}{\theta}$	$\frac{1}{\theta}\left[\left(\frac{X}{\theta}\right)^{\alpha}-\alpha\right]$	$(\frac{\alpha}{\theta},[Erl(2,\theta^{\alpha})]^{1/\alpha},Wei(\alpha,\theta))$
$gam(\alpha,\theta)$	$\frac{X}{\theta}$	$\frac{1}{\theta}\left(\frac{X}{\theta}-\alpha\right)$	$(\frac{\alpha}{\theta},gam(\alpha+1,\theta),gam(\alpha,\theta))$
$Par(\alpha,\theta)$	$\frac{X}{\theta}$	NA	$(\frac{\alpha}{\theta},Par(\alpha,\theta),\theta)$

5.3.4 Output Examples

We consider the two examples introduced at the beginning of the chapter: stochastic activity network and single-server queue.

Stochastic Activity Network

A stochastic activity network will be given by a directed acyclic graph, defined by M nodes and N directed arcs representing activities. The activity times are given by random variables $X_i,\ i = 1,\dots,N$. Without loss of generality, we take node 1 as the source (origin) and node M as the sink (destination). A path P is a set of directed arcs going from source to sink. Let $\mathscr{P}$ denote the set of all paths from source to sink, and P^* denote the set of arcs on the optimal path corresponding to the project duration given by Y (e.g., shortest or longest path, depending on the problem), i.e.,

$$Y = \sum_{j \in P^*} X_j,$$

where P^* itself is a random variable. We wish to estimate $d\mathsf{E}[Y]/d\theta$, where θ is a parameter in the distribution(s) of the activity times $\{X_i\}$.

Example 5.8. An example of a five-node network with six arcs is shown in Fig. 5.1, in which there are three paths: $\mathscr{P} = \{(1,4,6),(1,3,5,6),(2,5,6)\}$. If the longest path is the performance measure of interest, then

$$\begin{aligned} Y &= \max\{X_1+X_4+X_6, X_1+X_3+X_5+X_6, X_2+X_5+X_6\} \\ &= X_6 + \max\{X_1+X_4, X_1+X_3+X_5, X_2+X_5\}. \end{aligned}$$

For a specific realization, $\{X_1 = 9,\ X_2 = 15,\ X_3 = 8,\ X_4 = 16,\ X_5 = 11,\ X_6 = 12\}$, $Y = 12 + \max\{9+16, 9+8+11, 15+11\} = 40$ and $P^* = (1,3,5,6)$.

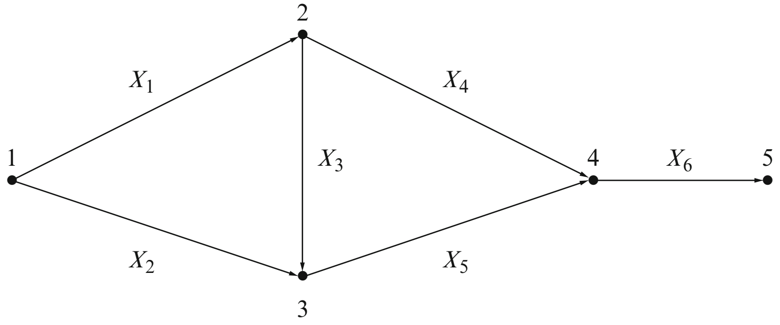

Fig. 5.1 Example stochastic activity network

Denote the c.d.f. and p.d.f. of X_i by F_i and f_i, respectively. For simplicity, assume all of the activity times are *independent*. Even so, it should be clear that duration of paths in $\mathscr{P}$ will not in general be independent, e.g., Example 5.8, where all three of the paths include arc 6, so clearly the durations are not independent.

Let θ be a parameter in the distribution of a single X_i, i.e., in f_i and F_i only. Then the IPA estimator is given by

$$\frac{dY}{d\theta} = \frac{dX_i}{d\theta}\mathbf{1}\{i \in P^*\}.$$

The LR/SF estimator is given by

$$Y\frac{\partial \ln f_i(X_i;\theta)}{\partial \theta}.$$

The WD estimator is of the form

$$c(\theta)\left(Y(X_1,\ldots,X_i^{(+)},\ldots,X_N) - Y(X_1,\ldots,X_i^{(-)},\ldots,X_N)\right)$$

where $X_i^{(\pm)} \sim F_i^{(\pm)}$, and $(c(\theta),F_i^{(+)},F_i^{(-)})$ is a weak derivative for F_i.

If we allow the parameter to possibly appear in all of the distributions, then the IPA estimator is found by applying the chain rule:

$$\frac{dY}{d\theta} = \sum_{i \in P^*} \frac{dX_i}{d\theta},$$

whereas the LR/SF and WD estimators are derived by applying the product rule of differentiation to the underlying input distribution, applying the independence that allows the joint distribution to be expressed as a product of marginals. In particular, the LR/SF estimator is given by

$$Y(X)\left(\sum_{i=1}^{N} \frac{\partial \ln f_i(X_i;\theta)}{\partial \theta}\right).$$

The WD estimator is of the form

$$\sum_{i=1}^{N} c_i(\theta)\left(Y(X_1,\ldots,X_i^{(+)},\ldots,X_N) - Y(X_1,\ldots,X_i^{(-)},\ldots,X_N)\right),$$

where $X_i^{(\pm)} \sim F_i^{(\pm)}, i = 1,\ldots,N$, and $(c_i(\theta),F_i^{(+)},F_i^{(-)})$ is a weak derivative for F_i.

Example 5.9. We illustrate several cases for Example 5.8 when $\theta = 10$ is the mean of the exponential distribution for one or all of the activity times. For the WD estimator, assume that the WD used is the entry from Table 5.1, i.e., $(1/\theta, Erl(2,\theta), exp(\theta))$, so that for the distribution(s) in which θ enters, $X_i^{(+)} \sim$

$Erl(2,10)$ and $X_i^{(-)} = X_i$. Assume the same values for X_i as in Example 5.8, and the following outputs for $X_i^{(+)}$: $X_1^{(+)} = 17$, $X_2^{(+)} = 33$, $X_3^{(+)} = 15$, $X_4^{(+)} = 40$, $X_5^{(+)} = 20$, $X_6^{(+)} = 25$.

Case 1: θ is the mean of the first activity time, i.e., $X_1 \sim exp(\theta)$.
The IPA estimate is simply $X_1/\theta = 9/10 = 0.9$, since X_1 is on the critical path in Example 5.8. The LR/SF estimate is given by $(40)(1/10)(9/10-1) = -0.4$. For WD, in the "+" network, the longest path remains the same as in the original network, and the longest path length simply increases by the difference in X_1, so the WD estimate is given by (1/10)(48-40) = 0.8.

Case 2: θ is the mean of the second activity time, i.e., $X_2 \sim exp(\theta)$.
The IPA estimate is 0, since X_2 is not on the critical path for Example 5.8. The LR/SF estimate is given by $(40)(1/10)(15/10-1) = 2$. For WD, in the "+" network, the longest path becomes (2,5,6), and the WD estimate is given by (1/10)(56-40) = 1.6.

Case 3: θ is the mean of the sixth activity time, i.e., $X_6 \sim exp(\theta)$.
The IPA estimate is $X_6/\theta = 12/10 = 1.2$, since X_6 is always on the critical path. The LR/SF estimate is given by $(40)(1/10)(12/10-1) = 0.8$, and the WD estimate is given by (1/10)(53-40) = 1.3.

Case 4: θ is the mean of *all* of the activity times, i.e., $X_i \sim exp(\theta)$ i.i.d.
The IPA estimate is $(X_1 + X_3 + X_5 + X_6)/\theta = 40/10 = 4.0$. The LR/SF estimate is given by $(40)(1/10)(-0.1+0.5-0.2+0.6+0.1+0.2) = 4.4$. For WD, the longest path has to be computed separately for six different network realizations; and the WD estimate is the sum of the six differences: (1/10)(8+16+7+21+9+13) = 7.4.

If instead we were interested in estimating $\mathsf{P}(Y > y)$ for some fixed y, the WD and LR/SF estimators would simply replace Y with the indicator function $\mathbf{1}\{Y > y\}$. For example, in Case 1 of Example 5.9, for any $y < 40$, the LR/SF estimate is given by (1/10)(9/10-1) = -0.01, and the WD estimate is (1/10)(1-1) = 0; for $y \geq 40$, the LR/SF estimate is 0; for $y \geq 48$, the WD estimate is (1/10)(0-0) =0, whereas for $40 \leq y < 48$, the WD estimate is (1/10)(1-0)=0.1. However, IPA would not apply, since the indicator function is inherently discontinuous, so an extension of IPA such as SPA is required. On the other hand, if the performance measure were $\mathsf{P}(Y > \theta)$, then since the parameter does not appear in the distribution of the input random variables, WD and LR/SF estimators cannot be derived without first carrying out an appropriate change of variables. These cases are addressed in [15].

Single-Server Queue

We illustrate each of the three direct gradient estimators for the FCFS $G/G/1$ queue. Let A_i be the interarrival time between the $(i-1)$st and ith customer, X_i the service time of the ith customer, and T_i the system time (in queue plus in service) of the

ith customer. The sample performance of interest is the average system time over the first N customers $\overline{T}_N = \frac{1}{N}\sum_{i=1}^{N} T_i$, and we take θ as a parameter in the service time distribution(s). Assume that the system starts empty, so that the times of the first N service completions are completely determined by the first N interarrival times and first N service times. Also assume that the arrival process is independent of the service times, which are also independent of each other but not necessarily identically distributed, with the p.d.f. and c.d.f. for X_i given by f_i and F_i, respectively.

The system time of a customer for a FCFS single-server queue satisfies the recursive Lindley equation:

$$T_{i+1} = X_{i+1} + (T_i - A_{i+1})^+. \tag{5.30}$$

The IPA estimator is obtained by differentiating (5.30):

$$\frac{dT_{i+1}}{d\theta} = \frac{dX_{i+1}}{d\theta} + \frac{dT_i}{d\theta}\mathbf{1}\{T_i \geq A_{i+1}\}, \tag{5.31}$$

so that the IPA estimator for the derivative of average system time is given by

$$\frac{d\overline{T}_N}{d\theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{dT_i}{d\theta} = \frac{1}{N}\sum_{m=1}^{M}\sum_{i=n_{m-1}+1}^{n_m}\sum_{j=n_{m-1}+1}^{i}\frac{dX_j}{d\theta}, \tag{5.32}$$

where M is the number of busy periods observed and n_m is the index of the last customer served in the mth busy period ($n_0 = 0$ and $n_M = N$ for M complete busy periods), and expressions for $dX/d\theta$ for many input distributions can be found in Table 5.1. Implementation of the estimator involves keeping track of two running quantities, one for (5.31) and another for the summation in (5.32); thus, the additional computational overhead is minimal, and *no alteration of the underlying simulation is required.*

To derive an LR/SF estimator, we use the fact that the interarrival times and service times are independently generated, so the joint p.d.f. on the input random variables will simply be the product of the p.d.f.s of the joint interarrival time distribution and the individual service time distributions given by

$$g(A_1,\ldots,A_N)\prod_{i=1}^{N} f_i(X_i;\theta),$$

where g denotes the joint p.d.f. of the interarrival times. Thus, the straightforward LR/SF estimator would be given by

$$\left(\frac{d\overline{T}_N}{d\theta}\right)_{LR} = \overline{T}_N\sum_{i=1}^{N}\frac{\partial \ln f_i(X_i;\theta)}{\partial\theta}, \tag{5.33}$$

where expressions for some common input distributions can be found in Table 5.1.

The WD estimator is also relatively straightforward, just incorporating the product rule of differentiation as before:

$$\left(\frac{d\overline{T}_N}{d\theta}\right)_{WD} = \sum_{i=1}^{N} c_i(\theta)\Big[\overline{T}_N(A_1,\ldots,A_N,\ldots,X_i^{(+)},\ldots) - \overline{T}_N(A_1,\ldots,A_N,\ldots,X_i^{(-)},\ldots)\Big],$$

where $X_i^{(\pm)} \sim F_i^{(\pm)}$, $i = 1,\ldots,N$ for $(c_i(\theta), F_i^{(+)}, F_i^{(-)})$ a weak derivative of F_i (again, see Table 5.1). Note that in general, implementation of the estimator requires $2N$ separate sample paths and resulting sample performance estimates whenever the parameter appears in N input random variables.

Example 5.10. We illustrate the numerical calculation for the three estimators when $\theta = 10$ is the mean of the exponential distribution for two cases: the first service time only or all of the service times. Again, for the WD estimator, assume that the WD used is the entry from Table 5.1, i.e., $(1/\theta, Erl(2,\theta), exp(\theta))$, so that for the distribution(s) in which θ enters, $X_i^{(+)} \sim Erl(2,10)$ and $X_i^{(-)} = X_i$. Take $N = 5$, with the first five arrivals occurring at $t = 10, 20, 30, 40, 50$, i.e., $A_i = 10$, $i = 1,2,3,4,5$, and the following service times generated:

$$X_1 = 15,\; X_2 = 7,\; X_3 = 11,\; X_4 = 9,\; X_5 = 6.$$

For these values, it turns out that all five customers are in the same busy period, i.e., all except the first customer have to wait, and we get the following outputs:

$$T_1 = 15,\; T_2 = 12,\; T_3 = 13,\; T_4 = 12,\; T_5 = 8;\; \overline{T}_5 = 12.0.$$

For the WD estimate, we also need the following (only first entry for the 1st case):

$$X_1^{(+)} = 25,\; X_2^{(+)} = 12,\; X_3^{(+)} = 21,\; X_4^{(+)} = 19,\; X_5^{(+)} = 11.$$

Letting $T_j^{(+i)} \equiv T_j(\ldots,X_i^{(+)},\ldots)$ and $\overline{T}_N^{(+i)} \equiv \overline{T}_N(\ldots,X_i^{(+)},\ldots) = \frac{1}{N}\sum_{j=1}^{N} T_j^{(+i)}$, we compute the following values for $T_j^{(+i)}$ and $\overline{T}_5^{(+i)}$:

i \ j	1	2	3	4	5	$\overline{T}_5^{(+i)}$
1	25	22	23	22	18	22.0
2	15	17	18	17	13	16.0
3	15	12	23	22	18	18.0
4	15	12	13	22	18	16.0
5	15	12	13	12	13	13.0

Note that since all the service times $X^{(+)}$ are longer than the original service times, all five customers remained in a single busy period on the "+" path.

Case 1: θ is the mean of the first service time only, i.e., $X_1 \sim exp(\theta)$.
The IPA estimate is simply $[5(X_1/\theta)]/5 = 15/10 = 1.5$; the LR/SF estimate is $(12)(1/10)(15/10-1) = 0.6$; and the WD estimate is (1/10)(22−12) = 1.0.

Case 2: θ is the mean of *all* of the service times, i.e., $X_i \sim exp(\theta)$ i.i.d.
The IPA estimate is $[(5X_1+4X_2+3X_3+2X_4+X_5)/\theta]/5 = 3.2$; the LR/SF estimate is $(12)(1/10)(0.5-0.3+0.1-0.1-0.4) = -0.24$; and the WD estimate is $(1/10)(10+4+6+4+1) = 2.5$.

Variance Reduction

Both the LR/SF and WD estimators may have variance problems if the parameter appears in all of the distributions, e.g., if it is the common mean when the service times are i.i.d. The variance of the LR/SF estimator given by (5.33) increases linearly with N, so some sort of truncation is generally necessary. For the single-server queue example, the regenerative structure provides such a mechanism, to be described shortly. For the WD estimator, although the variance of the estimator may not increase with N, implementation may not be practical for large N. However, in many cases, the expression can be simplified, making the computation more acceptable. As discussed earlier, the variance properties of a WD estimator depend heavily on the particular weak derivative(s) used and the coupling (correlation) between $X_i^{(+)}$ and $X_i^{(-)}$.

Using regenerative theory, the mean steady-state system time can be written as a ratio of expectations:

$$\mathsf{E}[T] = \frac{\mathsf{E}[Q]}{\mathsf{E}[\eta]},$$

where η is the number of customers served in a busy period and Q is the sum of the system times of customers served in a busy period. Differentiation yields

$$\frac{d\mathsf{E}[T]}{d\theta} = \frac{d\mathsf{E}[Q]/d\theta}{\mathsf{E}[\eta]} - \frac{d\mathsf{E}[\eta]/d\theta}{\mathsf{E}[\eta]}\mathsf{E}[T].$$

Applying the natural LR/SF estimators for each of the terms separately leads to the following regenerative estimator over M busy periods, for the i.i.d. case where θ appears in the common service time p.d.f. f_X:

$$\begin{aligned}\left(\frac{d\overline{T}_N}{d\theta}\right)_{LR} &= \frac{1}{N}\sum_{m=1}^{M}\left\{\left(\sum_{i=n_{m-1}+1}^{n_m} T_i\right)\sum_{i=n_{m-1}+1}^{n_m}\frac{\partial \ln f_X(X_i;\theta)}{\partial\theta}\right\}\\ &\quad -\frac{1}{N}\sum_{m=1}^{M}\left\{(n_m-n_{m-1})\sum_{i=n_{m-1}+1}^{n_m}\frac{\partial \ln f_X(X_i;\theta)}{\partial\theta}\right\}\overline{T}_N.\end{aligned}$$

The advantage of this estimator is that the summations are bounded by the length of the busy periods, so provided the busy periods are relatively short, the variance of the estimators should be tolerable.

Higher Derivatives

For the WD estimator, a second derivative estimator would take exactly the same form as before, the only difference being that $(c_i(\theta), F_i^{(+)}, F_i^{(-)})$ should be a weak *second* derivative of F_i.

Using the regenerative method as before, the second derivative LR/SF estimator is also relatively easy to derive:

$$\begin{aligned}\left(\frac{d^2\overline{T}_N}{d\theta^2}\right)_{LR} &= \frac{1}{N}\sum_{m=1}^{M}\left\{\left(\sum_{i=1}^{n_m} T_i\right)\sum_{i=n_{m-1}+1}^{n_m}\left[\frac{\partial^2 \ln f_X(X_i;\theta)}{\partial\theta^2}+\left(\frac{\partial \ln f_X(X_i;\theta)}{\partial\theta}\right)^2\right]\right\}\\ &\quad -\frac{1}{N}\sum_{m=1}^{M}\left\{(n_m-n_{m-1})\sum_{i=n_{m-1}+1}^{n_m}\left[\frac{\partial^2 \ln f_X(X_i;\theta)}{\partial\theta^2}+\left(\frac{\partial \ln f_X(X_i;\theta)}{\partial\theta}\right)^2\right]\right\}\frac{1}{N}\sum_{j=1}^{N}T_j.\end{aligned}$$

On the other hand, IPA will not work for higher derivatives for the single-server queue example. The implicit assumption used in deriving an IPA estimator is that small changes in the parameter results in small changes in the sample performance, which translates to the boundary condition in (5.31) being unchanged by differentiation. In general, the interchange (5.11) will typically hold if the sample performance is continuous with respect to the parameter. For the Lindley equation, although T_{n+1} in (5.30) has a "kink" at $T_n = A_{n+1}$, it is still continuous at that point, which is the intuition behind why IPA works. Unfortunately, the "kink" means that the derivative given by (5.31) has a discontinuity at $T_n = A_{n+1}$, so that IPA will fail for the second derivative.

An unbiased SPA second derivative estimator can be derived under the additional assumption that the arrival process has independent interarrival times, by conditioning on all *previous* interarrival and service times at each departure, which determines the system time, say T_n, with the corresponding next interarrival time, A_{n+1}, unconditioned. We provide a brief informal derivation based on sample path intuition (refer to Fig. 5.2). For the right-hand estimator, in which we assume $\Delta T_n > 0$ (technically it should refer to $\Delta\theta$), the only "critical" events are those departures that terminate a busy period, with the possibility that two busy periods coalesce (idle period disappears) due to a perturbation. Letting g_n and G_n denote the

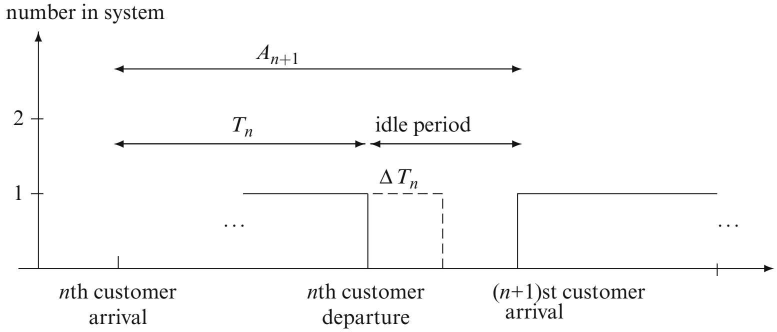

Fig. 5.2 Quantities used in deriving FCFS single-server queue SPA estimator

respective p.d.f. and c.d.f. of A_n, the corresponding probability rate (conditional on T_n) is then calculated as follows:

$$\lim_{\Delta\theta\to 0} \frac{\mathsf{P}(T_n+\Delta T_n \geq A_{n+1}|T_n < A_{n+1})}{\Delta T_n} = \frac{g_{n+1}(T_n)}{1-G_{n+1}(T_n)}\frac{dT_n}{d\theta},$$

and the corresponding effect would be that the ΔT_n perturbation would be propagated to the next busy period. The complete SPA estimator is given by

$$\left(\frac{d^2\overline{T}_N}{d\theta^2}\right)_{SPA} = \frac{1}{N}\sum_{m=1}^{M}\sum_{i=n_{m-1}+1}^{n_m}\sum_{j=n_{m-1}+1}^{i}\frac{d^2X_j}{d\theta^2} + \frac{1}{M}\sum_{m=1}^{M}\frac{g_{n_m+1}(T_{n_m})}{1-G_{n_m+1}(T_{n_m})}\left(\frac{dT_{n_m}}{d\theta}\right)^2,$$

where $\frac{d^2X}{d\theta^2}$ is well-defined when $F_X(X;\theta)$ is twice differentiable, and, in particular, $\frac{d^2X}{d\theta^2}=0$ for location, scale, and generalized scale parameters.

5.3.5 *Rudimentary Theory*

A basic requirement for the stochastic gradient estimator is that it be unbiased.

Definition. The gradient estimator $\hat{\nabla}_\theta J(\theta)$ is *unbiased* if $\mathsf{E}[\hat{\nabla} J(\theta)] = \nabla_\theta J(\theta)$.

Basically, unbiasedness requires the exchange of the operations of differentiation (limit) and integration (expectation), as was assumed in deriving (5.11) and (5.12). Although in theory *uniform integrability* is both a necessary and sufficient condition allowing the desired exchange of limit and expectation operators, in practice the key result used in the theoretical proofs of unbiasedness is the (Lebesgue) dominated convergence theorem. In the case of PA, the bounding involves properties of the performance measure, whereas in LR/SF and WD, the bounding involves the distribution functions (probability measures).

Theorem 5.1 (Dominated Convergence Theorem). If $\lim_{n\to\infty} g_n = g$ w.p.1 and $|g_n| \leq M\ \forall n$ w.p.1 with $\mathsf{E}[M] < \infty$, then $\lim_{n\to\infty} \mathsf{E}[g_n] = \mathsf{E}[g]$.

Take $\Delta\theta \to 0$ instead of $n \to \infty$, and g is the gradient estimator, so $g_{\Delta\theta}$ is the limiting sequence that defines the sample (path) gradient. Verifying that an actual bound exists is often a non-trivial task in applications, especially in the case of perturbation analysis.

Considering the two equations in (5.11), we translate these conditions to

$$g_{\Delta\theta} = \frac{Y(\theta+\Delta\theta)-Y(\theta)}{\Delta\theta}, \tag{5.34}$$

$$g_{\Delta\theta} = Y(x)\frac{f(x;\theta+\Delta\theta)-f(x;\theta)}{\Delta\theta}, \tag{5.35}$$

for IPA and LR/SF, respectively.

For IPA, the dominated convergence theorem bound implied by (5.34) corresponds to Lipschitz continuity on the sample performance function Y, so that the usual conditions required are piecewise differentiability and Lipschitz continuity of Y, where the Lipschitz modulus is integrable, i.e., $\exists M > 0$ with $\mathsf{E}[M] < \infty$ s.t.

$$|Y(\theta+\Delta\theta)-Y(\theta)| \leq M|\Delta\theta|.$$

In practice, the following generalization of the mean value theorem is useful.

Theorem 5.2 (Generalized Mean Value Theorem). Let Y be a continuous function that is differentiable on a compact set $\tilde{\Theta} = \Theta\backslash\tilde{D}$, where $\tilde{D}$ is a set of countably many points. Then, $\forall\theta, \theta+\Delta\theta \in \Theta$,

$$\left|\frac{Y(\theta+\Delta\theta)-Y(\theta)}{\Delta\theta}\right| \leq \sup_{\theta\in\tilde{\Theta}}\left|\frac{dY}{d\theta}\right|.$$

If $Y(\theta)$ can be shown to be continuous and piecewise differentiable on Θ w.p.1, then it just remains to show

$$\mathsf{E}\left[\sup_{\theta\in\tilde{\Theta}}\left|\frac{dY}{d\theta}\right|\right] < \infty,$$

to satisfy the conditions required for unbiasedness via the dominated convergence theorem. Basically, in order for the chain rule to be applicable, the sample performance function needs to be continuous with respect to the underlying random variable(s). This translates into requirements on the form of the performance measure and on the dynamics of the underlying stochastic system. The applicability of IPA may depend on how the input processes are constructed/generated, as was illustrated in Example 5.5, where one representation led to a biased estimator while another led to an unbiased estimator. In applying SPA, there is the choice of conditioning quantities (cf. (5.19)/(5.20)), which affects how easily the resulting conditional expectation can be estimated from sample paths. In Example 5.5, the representation that led to a biased IPA estimator only had two random variables, so there was a limited choice on what to condition, and the obvious choice led immediately to an unbiased SPA estimator.

For the LR/SF method, the bound is applied to the (joint) p.d.f. (or p.m.f.). *Note that the bound on $f(x;\theta)$ is with respect to the parameter θ and not its usual argument.* For WD, the required interchange is guaranteed by the definition of the weak derivative, but the sample performance must be in the set of "test" functions $\mathscr{L}$ in the definition, which again generally requires the dominated convergence theorem.

The previous examples can be used to show in very simple cases where difficulties arise. Consider the p.d.f.

$$f(x;\theta) = \frac{1}{\theta}\mathbf{1}\{0 < x < \theta\},$$

where the LR/SF method does not apply. In this case, f viewed as a function of θ for fixed x has a discontinuity at $\theta = x$. Similarly, consider the function

$$\mathsf{P}(Y > y) = \mathsf{E}[\mathbf{1}\{Y > y\}],$$

for which IPA will not work. In this case, the performance measure is an indicator function, which is discontinuous in its argument. In both of these simple examples, the dominated convergence theorem cannot be applied, because the required quantity cannot be bounded. However, since the dominated convergence theorem provides only *sufficient* conditions, it is possible in some very special situations (neither of which these two examples satisfy), unbiasedness may still hold.

In addition to the basic requirement for an unbiased estimator, it is important for the estimators to have low variance. There are also a multitude of choices of WD triples for a given input distribution, and this determines both the amount of additional simulation required and the variance of the resulting WD output gradient estimator. For LR/SF estimators, the variance of the estimator could also be a problem if care is not taken in implementation, e.g., a naïve estimator may lead to a linear increase in variance with respect to the simulation horizon, as in the single-server queue example.

5.3.6 Guidelines for the Practitioner

Here we summarize some key considerations in applying the three direct gradient estimation methods (PA, LR/SF, WD) (cf. [14]):

- IPA is generally inapplicable if there is a *discontinuity* in the sample performance or the underlying system dynamics; the *commuting condition* (see [21]) can be used to check the latter by considering possible event sequences in the system. Smoothness may depend on system representation, as mentioned in Sect. 5.3.1 and illustrated in Example 5.5.
- SPA uses *conditional Monte Carlo*, so just as in its use for variance reduction, the chief challenges of applying this approach include choosing what to condition on and being able to compute (or estimate) the resulting conditional expectation. The derived estimator may require additional simulations; see [18] for a comprehensive treatment of SPA.
- LR/SF and WD are more difficult to apply when the parameter does not appear explicitly in a probability distribution (so-called "structural" parameters), in which case an appropriate change of variables needs to be found.
- When the parameter of interest is known to be a *location* or (generalized) *scale* parameter of the input distribution, then IPA is particularly easy to apply, regardless of how complicated the actual distribution may be.
- For the LR/SF method, if the parameter appears in an input distribution that is reused frequently such as in an i.i.d. sequence of random variables, e.g., interarrival and service times in a queueing system, truncation of some sort will usually be required to mitigate the linear increase in variance.
- Application of the WD method generally requires two selections to be made: (a) which (non-unique) weak derivative $(c, F^{(+)}, F^{(-)})$ representation to use; and (b) how to correlate (or couple) the random variables generated from $F^{(+)}$ and $F^{(-)}$. Table 5.1 provides recommendations for many common distributions, and the Hahn–Jordan decomposition given by (5.23) and (5.24) always provides a fallback option. For the continuous distribution WD representations in Table 5.1, the use of common random numbers can often reduce variance, whereas for the Hahn–Jordan WD representation, it is best to generate the random variables independently [44]. High-dimensional vectors may require many additional simulations.
- For discrete distributions, IPA can usually be applied if the parameter occurs in the possible *values* of the input random variable, whereas LR/SF and WD can be applied if the parameter occurs in the *probabilities*.
- *Higher derivative* estimators are generally easy to derive using the LR/SF or WD method, but the former often leads to estimators with large variance and the latter may require a large number of additional simulations.

5.4 Quantile Sensitivity Estimation

We begin by considering the sensitivities of the order statistics, which naturally leads to quantile sensitivity estimation. This is the approach taken in [44]; see [16, 29, 30] for alternative approaches.

For a sample size n of random variables $Y_j,\ j = 1,\dots,n$, we consider the ith order statistic $Y_{(i)}$, where the order statistics are defined by

$$Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(i)} \le \cdots \le Y_{(n)}.$$

For simplicity, assume the $\{Y_j\}$ are all continuous random variables, so that equality occurs w.p.0. The order-statistics definition assumes neither independence nor identical distributions for the underlying $\{Y_j\}$. When $i = \lceil \alpha n \rceil$, where $\lceil x \rceil$ denotes the ceiling function that returns the next integer greater than or equal to x, $Y_{(i)}$ will correspond to the quantile estimator for q_α, which we denote by $\hat{q}^n_\alpha(\theta) \equiv Y_{(\lceil \alpha n \rceil)}$. We write $Y_{(i)}(Y_1,\dots,Y_n)$ as necessary to show explicit dependence on the $\{Y_j\}$, which will be the case for the WD estimator.

Under the setting that Y_i are i.i.d. and θ is a (scalar) parameter in the (common) distribution of $\{Y_j\}$, the first objective will be to estimate

$$J'(\theta) \equiv \frac{d\mathsf{E}[Y_{(i)}]}{d\theta}.$$

Then the respective IPA, LR/SF, and WD estimators are given by

$$\frac{dY_{(i)}}{d\theta}, \tag{5.36}$$

$$Y_{(i)} \sum_{j=1}^{n} \frac{\partial \ln f_{Y_j}(Y_j;\theta)}{\partial\theta}, \tag{5.37}$$

$$c(\theta) \sum_{j=1}^{n} \left[Y_{(i)}(Y_1,\dots,Y_j^{(+)},\dots,Y_n) - Y_{(i)}(Y_1,\dots,Y_j^{(-)},\dots,Y_n) \right], \tag{5.38}$$

where f_{Y_j} denotes the p.d.f. of Y_j. Note that for the IPA estimator, $dY_{(i)}/d\theta$ corresponds to the $dY/d\theta$ for $Y_{(i)}$ and NOT the ith order statistic of $\{dY_i/d\theta\}$, i.e., if we write the order statistics for the IPA estimators of $\{dY_i/d\theta\}$ as

$$\left[\frac{dY}{d\theta}\right]_{(1)} \le \cdots \left[\frac{dY}{d\theta}\right]_{(j)} \le \cdots \left[\frac{dY}{d\theta}\right]_{(n)},$$

then in general,

$$\frac{dY_{(j)}}{d\theta} \ne \left[\frac{dY}{d\theta}\right]_{(j)}.$$

For the setting where Y_i are i.i.d., since the dependence of Y on its arguments doesn't depend on the order,

$$\mathsf{E}\left[Y_{(j)}\frac{\partial \ln f_Y(Y_j)}{\partial\theta}\right]=\mathsf{E}\left[Y_{(k)}\frac{\partial \ln f_Y(Y_k)}{\partial\theta}\right]\ \forall j,k,$$

where f_Y denotes the common p.d.f., so that the following LR/SF estimator

$$nY_{(j)}\frac{\partial \ln f_Y(Y_j)}{\partial\theta}\ \forall j,$$

has the same expectation as the original LR/SF estimator given by (5.37), but now the linearly (with n) increasing variance becomes quadratic in n. The other problem with these estimators, whether this one or the original (5.37), is that they depend on f_Y, which is in general unknown in the simulation setting, where the Y_i denote the output of i.i.d. simulation replications, a function of input random variables, say $X_1,\dots,X_n$, whose distributions are known.

Similarly, we can eliminate the linearly (with n) increasing number of replications for the WD estimator by noting that if Y^* is independent of all Y_i (which are i.i.d.) but not necessarily having the same distribution, then the following is true: $\forall j,k\in\{1,\dots,n\}$,

$$Y_{(i)}(Y_1,\dots,Y_{j-1},Y^*,Y_{j+1},\dots,Y_n)\overset{d}{=}Y_{(i)}(Y_1,\dots,Y_{k-1},Y^*,Y_{k+1},\dots,Y_n).$$

As a result, the following estimator with the same expectation and order $n-1$ less pairs of simulations can be used:

$$nc(\theta)\left[Y_{(i)}(Y_1,\dots,Y_j^{(+)},\dots,Y_n)-Y_{(i)}(Y_1,\dots,Y_j^{(-)},\dots,Y_n)\right]$$

for any j. Again, as in the LR/SF case, the $Y_j^{(+)}$ and $Y_j^{(-)}$ need to be derived as a function of the (common) distribution of the $\{Y_i\}$.

It turns out that although all of the order statistics sensitivity estimators are unbiased, going to the quantile estimation setting by increasing the sample size only leads to asymptotic unbiasedness and not consistency in the general case, so that batching is required to obtain a consistent estimator of the quantile sensitivity q'_α, where q_α is defined by (5.3). Specifically, although the usual quantile estimator $\hat{q}^n_\alpha\equiv Y_{(\lceil\alpha n\rceil)}$ is strongly consistent, i.e.,

$$\lim_{n\to\infty}\hat{q}^n_\alpha=q_\alpha \text{ w.p.1},$$

for the quantile sensitivity estimator $\hat{q}'^n_\alpha(\theta)\equiv dY_{(\lceil\alpha n\rceil)}/d\theta$, *it does* **not** *follow in general that*

$$\lim_{n\to\infty}\hat{q}'^n_\alpha(\theta)=q'_\alpha(\theta)$$

in any sense (strong or weak), except that the mean converges correctly, i.e.,

$$\lim_{n\to\infty} \mathsf{E}[\hat{q}'^{n}_{\alpha}(\theta)] = q'_{\alpha}(\theta).$$

To obtain consistency requires batching, i.e., for k batches each of sample size n, the estimator is

$$\bar{q}'^{n,k}_{\alpha}(\theta) = \frac{1}{k}\sum_{i=1}^{k} \hat{q}'^{n,i}_{\alpha}(\theta),$$

where $\hat{q}'^{n,i}_{\alpha}(\theta)$ is the ith estimate out of k batches for whichever estimator is used—IPA, LR/SF, or WD, given by (5.36), (5.37), or (5.38), respectively—for $\hat{q}'^{n,i}_{\alpha}(\theta)$. Then it can be established that

$$\lim_{\substack{k\to\infty\\ n\to\infty}} \hat{q}'^{n,k}_{\alpha}(\theta) = q'_{\alpha}(\theta).$$

The unbatched IPA quantile sensitivity estimator (5.36) does turn out to be provably consistent if the following (very restrictive) condition is satisfied [30]:

There exists a function ϕ s.t. $\frac{dY}{d\theta} = \phi(Y)$.

We illustrate this condition with two simple examples, the first of which satisfies this condition, and the second of which does not. In both of these toy examples, the distribution of the output Y is known, so the LR/SF and WD quantile sensitivity estimators can also be written down explicitly.

Example 5.11. Take Example 5.1 with $Y = X\theta$ where X is exponentially distributed with mean 1, i.e., $X \sim exp(1)$, so $Y \sim exp(\theta)$. Since $dY/d\theta = X = Y/\theta$, the condition is satisfied with $\phi(y) = y/\theta$, and the unbatched IPA estimator $Y_{\lceil \alpha n \rceil}/\theta$ is consistent.

Example 5.12. Take $Y = \theta X_1 + X_2$, where X_1 and X_2 are both $\mathscr{N}(0,1)$ and independent, so $Y \sim \mathscr{N}(0, \theta^2 + 1)$. Then $dY/d\theta = X_1 = (Y - X_2)/\theta$, which still involves the input random variable X_2, and there is no way to write it in terms of Y only.

However, $\mathsf{E}[dY/d\theta|Y] = \mathsf{E}[X_1|Y] = \theta Y/(\theta^2+1) = -\partial_2 F_Y(Y;\theta)/\partial_1 F_Y(Y;\theta)$, where ∂_i denotes the partial derivative with respect to the ith argument ($i = 1,2$), and the relationship

$$\mathsf{E}\left[\frac{dY}{d\theta}\middle| Y\right] = -\frac{\partial_2 F_Y(Y;\theta)}{\partial_1 F_Y(Y;\theta)}$$

can be shown to hold in general [30].

5.5 New Approaches for Using Direct Stochastic Gradients in Simulation Optimization

This section provides an overview of several new developments in using direct stochastic gradient estimators for simulation optimization: three approaches to enhancing metamodels for response surface methodology (RSM), and combining indirect and direct estimators when used in stochastic approximation (SA), a simulation optimization approach introduced earlier in the introduction and treated in-depth in Chaps. 6 and 7. The new estimator is called a Secant-Tangents AveRaged (STAR) gradient, because it averages two direct (tangent) gradient estimators and one indirect (finite-difference secant) gradient estimator.

Before describing the other three approaches, we briefly summarize RSM; see Chap. 4 for details. Like SA, RSM is a sequential search procedure. The central component of RSM is the fitting of the local response surface using a metamodel, and the most common procedure used is regression. Specifically, after the preliminary scaling and screening of the input variables (called factors in the experimental design terminology), there are two main phases to RSM. In Phase I, which is iterative, a linear regression model is generally used to estimate a search direction to explore. Once a relatively flat area is found, RSM proceeds to Phase II where a higher-order—usually quadratic—model is fitted, which is used to estimate the optimum.

Because regression analysis arose from physically observed processes, it assumes that the only data points generated are measurements of the value of the dependent variable for each combination of independent variable values. In the simulation setting, the availability of direct gradient estimates opens up new possibilities that have just recently begun to be exploited. We begin by discussing a promising new approach that generalizes traditional regression, which is called Direct Gradient Augmented Regression (DiGAR).

Another metamodeling technique that can be used for RSM is kriging, which also arose from physical measurements. In the simulation setting, a generalization called stochastic kriging, is apropos. We discuss two enhancements to stochastic kriging that exploit the availability of direct gradient estimates: Stochastic Kriging with Gradients (SKG), which is analogous to DiGAR, and Gradient Extrapolated Stochastic Kriging (GESK), which uses the gradients in a totally different manner by generating new output data.

For these three approaches, we follow the notation of statisticians in using y for the output and x for the input.

5.5.1 Direct Gradient Augmented Regression (DiGAR)

Consider the usual regression setting with independent variable x and dependent variable y, where $n > 1$ data points $(x_1, y_1), \ldots, (x_n, y_n)$ are given. Both independent and dependent variables take values from a continuous domain. For expositional ease, here we describe only the one-dimensional setting, for which the basic DiGAR model is the following:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{5.39}$$

$$g_i = \beta_1 + \varepsilon_i', \tag{5.40}$$

where $g_i, i = 1, 2, \ldots, n$ are the gradient estimates with residuals $\{\varepsilon_i'\}$, and $\{\varepsilon_i\}$ denote the residuals of the outputs. The first line corresponds to traditional regression, whereas the second line adds the direct gradient estimates in the natural way. A weighted least-squares objective function is defined by

$$L = \alpha \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + (1-\alpha) \sum_{i=1}^{n} (g_i - \beta_1)^2, \tag{5.41}$$

where $\alpha \in [0,1]$. Note that $\alpha = 1$ corresponds to standard regression, whereas $\alpha = 0.5$ corresponds to the ordinary least squares (OLS) DiGAR model where the function estimates and derivative estimates are equally weighted, and $\alpha = 0$ corresponds to using only the gradient information. Solving the least-squares problem by minimizing (5.41) yields the following $\alpha-$DiGAR estimators for the parameters in the regression model given by (5.39)/(5.40):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \;\; \hat{\beta}_1 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \frac{1-\alpha}{\alpha}\bar{g}}{\frac{1}{n}\sum\limits_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1-\alpha}{\alpha}}, \tag{5.42}$$

where the bars indicate the sample means of the respective quantities. If the estimators for the gradients are also unbiased, then the $\alpha-$DiGAR estimators are also unbiased. In the homogeneous setting where the variances of the function and gradient estimates are given by σ^2 and σ_g^2, respectively, the following result (Proposition 4 in [19]) provides conditions under which the new DiGAR estimator guarantees variance reduction for the slope estimator in the regression model, assuming unbiased estimators in the uncorrelated setting.

Theorem 5.3. *For* $\mathsf{E}[\varepsilon_i] = \mathsf{E}[\varepsilon_i'] = 0 \;\; \forall i,$ $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{Cov}(\varepsilon_i', \varepsilon_j') = 0, i \neq j,$ $\mathrm{Cov}(\varepsilon_i, \varepsilon_j') = 0 \; \forall \; i, j,$

$$\frac{\sigma_g^2}{\sigma^2} \le \frac{2\alpha}{1-\alpha} + \frac{1}{\frac{1}{n}\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \Longleftrightarrow \mathrm{Var}(\hat{\beta}_1^{DiGAR}) \le \mathrm{Var}(\hat{\beta}_1^{standard}),$$

where $\hat{\beta}_1^{DiGAR}$ *and* $\hat{\beta}_1^{standard}$ *denote the* α*-DiGAR slope estimator (5.42) and the standard slope estimator, respectively.*

Even stronger guarantees are available in the maximum likelihood estimator (MLE) setting (Proposition 7 in [19]). However, all of these results are for the one-dimensional uncorrelated setting. For the multivariate setting $\mathbf{x}_i = \{x_{ij}, j = 1,\dots,d\}$ with corresponding output $\{y_i\}$ and partial derivatives $\{g_{ij}\}$, the extension of the least-squares objective function (5.41) for the α-DiGAR model

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i, \\ g_{ij} &= \beta_j + \varepsilon'_{ij}, \end{aligned}$$

with non-negative weights that sum to 1, $\{\alpha_j, j = 0,1,\dots,d\}$, is

$$\alpha_0 \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{d} \beta_j x_{ij})^2 + \sum_{j=1}^{d} \alpha_j \sum_{i=1}^{n} (g_{ij} - \beta_j)^2,$$

which when minimized yields the following slope estimators:

$$\hat{\beta}_j = \frac{\sum\limits_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y}) - \sum_{k \neq j} \beta_k \sum\limits_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j) + n \frac{\alpha_j}{\alpha_0} \bar{g}_j}{\sum\limits_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 + n \frac{\alpha_j}{\alpha_0}}, \quad j = 1,\dots,d,$$

which reduces to the previous expression (5.42) with $\alpha_0 = \alpha, \alpha_1 = 1 - \alpha$, when there is just a single input.

In the simulation optimization setting, simulation replications are often computationally expensive, so it is desirable to use as few of them as possible for each value of the input variable. When applying RSM for sequential search, the direction of improved performance is perhaps the most critical output of the fitted metamodel in Phase I. In several numerical experiments reported in [19] for an $M/M/1$ queue with relatively small number of simulation replications per design point, the slope of the standard linear regression model often gave the wrong sign, whereas the DiGAR model always estimated the sign correctly. Thus, where applicable, DiGAR should provide a better metamodel for simulation optimization using RSM.

5.5.2 Stochastic Kriging with Gradients (SKG)

The stochastic kriging (SK) model, introduced by [1], takes multivariate input $\{(\mathbf{x}_i, n_i)\}$, $i = 1,2,\dots,k$, which generates $y_j(\mathbf{x}_i)$ as the simulation output from replication j at design point $\mathbf{x}_i$, where $\mathbf{x} = (x_1, x_2, \dots, x_d)^T \in \mathbb{R}^d$. Stochastic kriging models $y_j(\mathbf{x}_i)$ as

$$y_j(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)^T \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \varepsilon_j(\mathbf{x}_i), \tag{5.43}$$

where $\mathbf{f}(\mathbf{x}_i) \in \mathbb{R}^p$ is a vector with known functions of $\mathbf{x}_i$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector with unknown parameters to be estimated. It is assumed that M is a realization of a mean zero stationary random process (or random field). The simulation noise for replication j taken at $\mathbf{x}_i$ is denoted as $\varepsilon_j(\mathbf{x}_i)$. The trend term $\mathbf{f}(\mathbf{x}_i)^T\boldsymbol{\beta}$ represents the overall surface mean. The stochastic nature in M is sometimes referred to as *extrinsic* uncertainty. The uncertainty in ε_j comes from the nature of stochastic simulation, and it is sometimes referred to as *intrinsic* uncertainty.

Given the simulation response outputs $\left\{y_j(\mathbf{x}_i)\right\}_{j=1}^{n_i}$, $i = 1,2,\dots,k$, denote the sample mean of response output and simulation noise at $\mathbf{x}_i$ as

$$\bar{y}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j(\mathbf{x}_i), \quad \bar{\varepsilon}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \varepsilon_j(\mathbf{x}_i), \tag{5.44}$$

and model the averaged response output as

$$\bar{y}(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)^T\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i).$$

The SKG framework, introduced in [10, 11], parallels DiGAR in the stochastic kriging setting by modeling the added gradient information analogously:

$$\bar{g}^r(\mathbf{x}_i) = \left(\frac{\partial \mathbf{f}(\mathbf{x}_i)}{\partial x_r}\right)^T \boldsymbol{\beta} + \frac{\partial \mathsf{M}(\mathbf{x}_i)}{\partial x_r} + \bar{\delta}^r(\mathbf{x}_i).$$

5.5.3 Gradient Extrapolated Stochastic Kriging (GESK)

Rather than modeling the gradient directly as in DiGAR and SKG, Gradient Extrapolated Stochastic Kriging (GESK) *extrapolates* in the neighborhood of the original design points $\{\mathbf{x}_i\}$, $i = 1,2,\dots,k$, i.e., additional response data are generated via linear extrapolations using the gradient estimates as follows:

$$\mathbf{x}_i^+ = \mathbf{x}_i + \Delta\mathbf{x}_i, \quad y_j(\mathbf{x}_i^+) = y_j(\mathbf{x}_i) + \mathbf{g}_j(\mathbf{x}_i)^T\Delta\mathbf{x}_i, \tag{5.45}$$

where $\Delta\mathbf{x} = (\Delta x_1, \Delta x_2, \dots, \Delta x_d)^T$, and $\bar{y}(\mathbf{x}_i^+)$ is defined similarly as $\bar{y}(\mathbf{x}_i)$ as in (5.44). Different extrapolation techniques can be applied in (5.45), and multiple points can also be added to the neighborhood of $\mathbf{x}_i$. For simplicity, here we assume that the same step size is used for all design points, i.e., $\Delta\mathbf{x}_i = \Delta\mathbf{x}$, $i = 1,2,\dots,k$, and that only a single additional point is added in the neighborhood of each point. Figure 5.3 depicts the idea where the gradient is indicated by the arrow and the extrapolated point by the cross.

The GESK model requires the choice of step sizes for the extrapolated points; large step sizes allow better coverage but at the cost of additional bias since the linearity is less likely to hold further from the original point. Thus, there is a basic

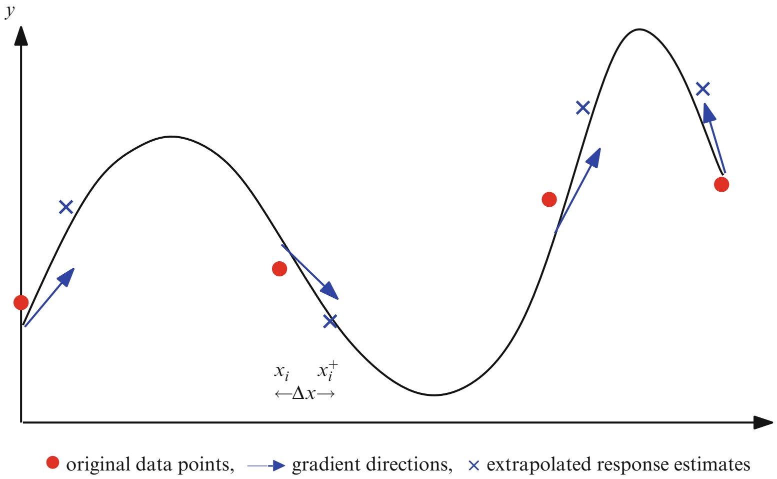
y
x_i
x_i^+
$\leftarrow \Delta x \rightarrow$
● original data points, → gradient directions, × extrapolated response estimates

Fig. 5.3 Illustration of gradient-extrapolated response points

bias-variance tradeoff to consider. In [37], this tradeoff is analyzed in a simplified setting, leading to conditions under which improvement can be guaranteed. To illustrate the potential improvements in performance that SKG and GESK offer over ordinary stochastic kriging, we present a simple stylized numerical example from [37] for a highly nonlinear function with added noise.

Numerical Example

The output is $y_j(x) = f(x) + \varepsilon_j(x)$, where $f(x) = \exp(-1.4x)\cos(7\pi x/2)$ and $\varepsilon_j(x) \sim \mathscr{N}(0,1)$, and the gradient estimate is given by $g_j(x) = f'(x) + \delta_j(x)$, where $\delta_j(x) \sim \mathscr{N}(0,25)$. Note that the variance of the direct gradient estimates are higher than those of the response outputs, generally the situation found in stochastic simulation settings. The Gaussian correlation function $R_{\mathsf{M}}(x,x') = \exp\{-\theta(x-x')^2\}$ is used for the stochastic kriging models. The number of design points is six, and the number of replications per design point is 50. Predictions are made at $N = 200$ equally spaced points in $[-2,0]$. Figure 5.4 shows the results in the form of graphs for a typical macro-replication, where both SKG and GESK are considerably better than SK as a result of incorporating gradient estimates, and both SKG and GESK better capture the trend of the response surface.

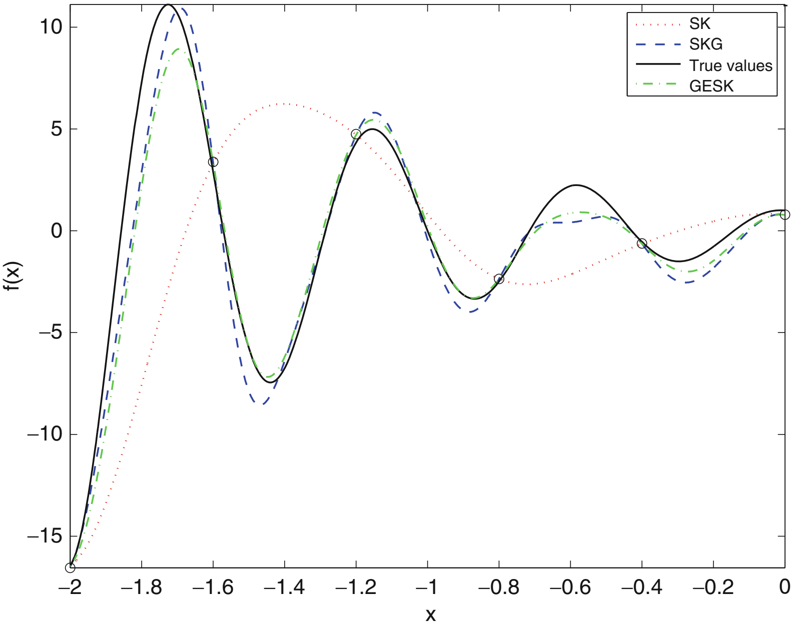

Fig. 5.4 Fitted curves for a representative macro-replication (6 points, 50 replications per point)

5.5.4 Secant-Tangents AveRaged Stochastic Approximation (STAR-SA)

Another proposal for the use of direct stochastic gradients is combining them with indirect gradients obtained from function estimates. In the deterministic optimization setting, there is nothing gained by the use of the function values themselves if exact gradients are available, but in the simulation setting the direct gradient estimates are noisy, so averaging them with indirect (finite difference) gradient estimates could potentially reduce the variance at the cost of adding some bias. This is the idea of the Secant-Tangents AveRaged (STAR) gradient estimator used in the STAR-SA algorithm introduced in [8,9]:

$$g^{\mathrm{STAR}}(x,\xi) = \alpha\frac{Y(x+c,\xi)-Y(x-c,\xi)}{2c} + (1-\alpha)\frac{g(x+c,\xi)+g(x-c,\xi)}{2}, \tag{5.46}$$

where $\alpha \in [0,1]$, and for notational convenience, the same noise is assumed for both points, e.g., through common random numbers, which is a convex combination of a symmetric finite difference (secant) and an average of two direct gradient (tangent) estimators. More details are provided in the following chapter on stochastic approximation, Chap. 6, including the extension to higher dimensions in the form of STAR-SPSA [8].

5.6 Concluding Remarks

The direct estimation techniques PA, LR/SF, and WD have been applied to a wide variety of application domains, the dominant ones being queueing, inventory, and finance. Historically, the early research was all on queueing systems, motivated originally by problems in manufacturing and communication networks. The first work in the inventory control setting was [13, 23]. The first work in finance was [5, 17], which considered IPA, SPA, and LR/SF estimators (cf. the book [22]). More recent research in the finance setting includes the first work on quantile sensitivity estimation [29]; the first combined IPA/LR estimator for options pricing [45]; and the development of various WD estimators, as described in the dissertation [44], which also includes an extensive treatment of IPA and LR/SF. Other areas of application include preventive maintenance, statistical process control, and traffic light signal control. Stochastic gradient estimation approaches not covered in this chapter include frequency domain experimentation and Malliavin calculus, the latter primarily used in continuous-time finance settings; see [14] for references on these various applications.

More details on IPA can be found in the books by Glasserman [21], Ho and Cao [26], and Cao [6], whereas a comprehensive treatment of SPA can be found in the book by Fu and Hu [18]. Although IPA and SPA are the best known forms of perturbation analysis, other versions include rare perturbation analysis [4], structural IPA [12], discontinuous perturbation analysis [40], and augmented IPA [20]. This chapter has treated gradient estimation, hence the focus on *infinitesimal* PA, but perturbation analysis originally arose from investigating the effects of a finite perturbation [27]; see also [7,28,43]. LR/SF is discussed in the books by Rubinstein and Shapiro [39], Glasserman [22], and Asmussen and Glynn [2], all of which also include discussion of IPA. The weak derivative method was introduced by Pflug [36], and many of the WD entry derivations in Table 5.1 can be found in [24].

Acknowledgements This work was supported in part by the National Science Foundation under Grants CMMI-0856256 and ECCS-0901543, and by the Air Force Office of Scientific Research under Grant FA9550-10-10340.

References

1. B. Ankenman, B. L. Nelson, and J. Staum. Stochastic kriging for simulation metamodeling. *Operations Research*, 58(2):371–382, 2010.
2. S. Asmussen and P. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York, 2007.
3. S. Bhatnagar, M. C. Fu, S. I. Marcus, and I. J. Wang. Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modeling and Computer Simulation*, 13(2):180–209, 2003.
4. P. Brémaud and F. J. Vázquez-Abad. On the pathwise computation of derivatives with respect to the rate of a point process: The phantom RPA method. *Queueing Systems: Theory and Applications*, 10:249–270, 1992.

5. M. Broadie and P. Glasserman. Estimating security price derivatives using simulation. *Management Science*, 42(2):269–285, 1996.
6. X.-R. Cao. *Realization Probabilities: The Dynamics of Queuing Systems*. Springer-Verlag, Boston, Massachusetts, 1994.
7. C. G. Cassandras and S. G. Strickland. On-line sensitivity analysis of Markov chains. *IEEE Transactions on Automatic Control*, AC-34:76–86, 1989.
8. M. Chau, M. C. Fu, and H. Qu. Multivariate stochastic approximation using a Secant-Tangents AveRaged (STAR) gradient estimator. Technical report, Working paper, University of Maryland, College Park, 2014.
9. M. Chau, H. Qu, and M. C. Fu. A new hybrid stochastic approximation algorithm. In *Proceedings of the 12th International Workshop on Discrete Event Systems*, 2014.
10. X. Chen. *Enhancing Stochastic Kriging Metamodels for Computer Simulation*. PhD thesis, Northwestern University, 2012.
11. X. Chen, B. Ankenman, and B. L. Nelson. Enhancing stochastic kriging metamodels with gradient estimators. *Operations Research*, 61(2):512–528, 2013.
12. L. Y. Dai and Y. C. Ho. Structural infinitesimal perturbation analysis for derivative estimation in discrete event dynamic systems. *IEEE Transactions on Automatic Control*, 40:1154–1166, 1995.
13. M. C. Fu. Sample path derivatives for (s,S) inventory systems. *Operations Research*, 42(2):351–364, 1994.
14. M. C. Fu. Gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, chapter 19, pages 575–616. Elsevier, 2006.
15. M. C. Fu. Sensitivity analysis in Monte Carlo simulation of stochastic activity networks. In F. B. Alt, M. C. Fu, and B. L. Golden, editors, *Perspectives in Operations Research: Papers in Honor of Saul Gass' 80th Birthday*, pages 351–366. Springer, 2006.
16. M. C. Fu, L. J. Hong, and J. Q. Hu. Conditional Monte Carlo estimation of quantile sensitivities. *Management Science*, 55(12):2019–2027, 2009.
17. M. C. Fu and J. Q. Hu. Sensitivity analysis for Monte Carlo simulation of option pricing. *Probability in the Engineering and Informational Sciences*, 9(3):417–446, 1995.
18. M. C. Fu and J. Q. Hu. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic Publishers, 1997.
19. M. C. Fu and H. Qu. Regression models augmented with direct stochastic gradient estimators. *INFORMS Journal on Computing*, 26(3):484–499, 2014.
20. A. Gaivoronski, L. Y. Shi, and R. S. Sreenivas. Augmented infinitesimal perturbation analysis: An alternate explanation. *Discrete Event Dynamic Systems: Theory and Applications*, 2:121–138, 1992.
21. P. Glasserman. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Boston, Massachusetts, 1991.
22. P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004.
23. P. Glasserman and S. Tayur. Sensitivity analysis for base-stock levels in multi-echelon production-inventory systems. *Management Science*, 41:263–281, 1995.
24. B. Heidergott, F. Vazquez-Abad, G. Pflug, and T. Farenhorst-Yuan. Gradient estimation for discrete-event systems by measure-valued differentiation. *ACM Transactions on Modeling and Computer Simulation*, 20(1):5.1–5.28, 2010.
25. B. Heidergott, F. Vazquez-Abad, and W. Volk-Makarewicz. Sensitivity estimation for Gaussian systems. *European Journal of Operational Research*, 187(1):193–207, 2008.
26. Y. C. Ho and X. R. Cao. *Perturbation Analysis and Discrete Event Dynamic Systems*. Kluwer Academic, 1991.
27. Y. C. Ho, M. A. Eyler, and T. T. Chien. A gradient technique for general buffer storage design in a serial production line. *International Journal of Production Research*, 17:557–580, 1979.
28. Y. C. Ho and S. Li. Extensions of infinitesimal perturbation analysis. *IEEE Transactions on Automatic Control*, AC-33:827–838, 1988.
29. L. J. Hong. Estimating quantile sensitivities. *Operations Research*, 57(1):118–130, 2009.

30. G. Jiang and M. C. Fu. On estimating quantile sensitivities via infinitesimal perturbation analysis. *Operations Research*, under review, 2014.
31. H. Kesten. Accelerated stochastic approximation. *The Annals of Mathematical Statistics*, 29(1):41–59, 1958.
32. J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–266, 1952.
33. H. J. Kushner and D. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.
34. H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 2nd edition, 2003.
35. A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, 3rd edition, 2000.
36. G. C. Pflug. Sampling derivatives of probabilities. *Computing*, 42:315–328, 1989.
37. H. Qu and M. C. Fu. Gradient extrapolated stochastic kriging. *ACM Transactions on Modeling and Computer Simulation*, 24(4): forthcoming, 2014.
38. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.
39. R. Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, 1993.
40. L. Y. Shi. Discontinuous perturbation analysis of discrete event dynamic systems. *IEEE Transactions on Automatic Control*, 41:1676–1681, 1996.
41. J. C. Spall. Multivariate stochastic approximation using simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:332–341, 1992.
42. R. Suri and M. A. Zazanis. Perturbation analysis gives strongly consistent sensitivity estimates for the $M/G/1$ queue. *Management Science*, 34:39–64, 1988.
43. P. Vakili. Using a standard clock technique for efficient simulation. *Operations Research Letters*, 10(8):445–452, 1991.
44. W. Volk-Makarewicz. *Advances in Derivative Estimation: Ranked Data, Quantiles, and Options*. PhD thesis, Vrije (Free) University of Amsterdam, 2014.
45. Y. Wang, M. C. Fu, and S. I. Marcus. A new stochastic derivative estimator for discontinuous payoff functions with application to financial derivatives. *Operations Research*, 60(2):447–460, 2012.
46. X. Xiong, I. J. Wang, and M. C. Fu. An asymptotic analysis of stochastic approximation with deterministic perturbation sequences. In *Proceedings of the 2002 Winter Simulation Conference*, pages 285–291. IEEE, Piscataway, NJ, 2002.