Chapter 1

Introduction

Derivative estimation, a field within Monte Carlo estimation, is a set of approaches to estimate sensitivities with respect to parameters within fully specified or "white-box" parametric performance measures of discrete-event stochastic models. Derivative estimation yield estimators that represent parameter derivatives of the original model and simulation input from an underlying distribution is utilized to obtain estimates. Commonly, the same simulation input and underlying distributions are utilized to obtain these estimates. This chapter is an introduction to this area and the motivation of the subsequent work in the forthcoming chapters.

We begin with an outline of this chapter. Section 1.1 describes the concept of discrete-event stochastic models. We motivate discrete-event stochastic modelling and particularly derivative estimation of stochastic models with applications. Section 1.2 is an overview of the main gradient estimation approaches. We explicate both pathwise, Section 1.2.1, and distributional, Section 1.2.2, approaches to derivative estimation before discussing the derivative estimation simulation praxis in Section 1.2.3. We apply each of these derivative estimation methods to the normal and exponential distribution in Section 1.2.4. The results from Section 1.2.4 will be imparted in the subsequent chapters. The outline for the remainder of thesis is given in Section 1.3, providing in turn an introduction to each of the subsequent chapters.

1.1 Concepts

1.1.1 Stochastic Models

Let $X : \Omega \mapsto \mathbb{R}^J$, $X = (X_1, \ldots, X_J)$, denote a vector representing a collection of Borel measurable random variables defined on a (filtered) probability space. Let $h : \mathbb{R}^J \mapsto \mathbb{R}$ be a Borel mapping. Vector valued mappings will be considered as an elementwise extensions of the same premise. The evaluation of the map h with regard to (w.r.t.) the vector X, with Z as the outcome,

$$Z = h(X), \tag{1.1}$$

is the general construction of a discrete-event stochastic model. In a stochastic model, the random variable X is some sort of input representing a real-world

system or concern, and h represents the topology of the underlying system, a performance metric, or a convolution of both. Examples are provided in the forthcoming paragraphs. These stochastic models are either single period, static in nature, typified by Equation (1.1), or are dynamic, occurring over multiple periods or events. Common lexicon for a static discrete-event stochastic model is a sample performance function, and for a dynamic discrete-event stochastic model, a stochastically recursive sequence. In the instance of a stochastically recursive sequence, we will have a vector $\mathbf{Z} = (Z_n)$ denoting a sequence of outcomes.

Quite commonly, Monte Carlo or stochastic simulation is needed to estimate quantities for discrete-event stochastic system. Discrete-event stochastic models are inherently complex and in most cases finding analytical expressions for something mundane as a mean value or a probability of some event is either impossible or infeasible. In addition, Monte Carlo simulation is easily extendible to high-dimensional problems, which frequently occur in stochastically recursive sequences. Generating realizations from a probability distribution is made possible by analytic precepts that go with the analysis and theory within applied probability and stochastic operations research. Derivative estimation is even more dependent on Monte Carlo simulation as the derivative expressions derived from the original stochastic model are often even more complex, and perhaps with increased dimensionality. Analytic results on the sensitivities for these models are uncommonly known, and especially for stochastic recursive sequences, numerical integration methods are unwieldy. For this thesis, we will refer to Monte Carlo simulation simply as simulation.

Beginning with sample performance functions, common examples, which are also exhibited in later chapters, are the stochastic activity network (SAN), and the cash flow for the European vanilla-call option. These examples will be described presently.

A SAN graph is a directed, acyclic graph, and this form of graph represents an efficient, accurate abstraction for inter-linked processes that occur over time. We present two examples for the SAN. The first example is to utilize the SAN to ascertain the sensitivity of the overall completion time of a project under the Project Evaluation Review Technique w.r.t. completion time rates of individual tasks. This example is provided in Section 2.4.2.1 and the present is a brief account. We denote the completion times by the random vector X, encoding the six individual tasks $X_1, \ldots, X_6$ within the project. How these individual tasks are linked is by the map h, displayed in Figure 1.1, denoting the topology of the three paths of the corresponding directed, acyclic graph. As all tasks need to be finished before the project is completed, there is a task or several tasks that

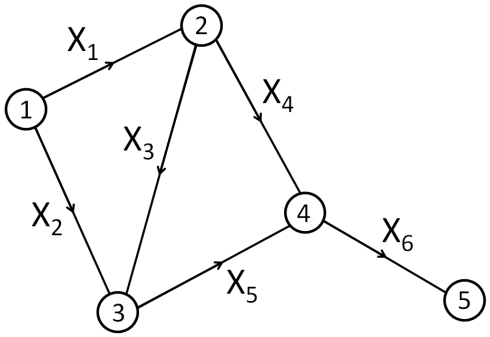

Figure 1.1: Graph of the Stochastic Activity Network (SAN).

cause the delay for the completion of the project. All delays will be observed on one of the multiple paths taken from the start to the end node. In this example, these paths are denoted by $\zeta_1,\ldots,\zeta_3$ shown in Equation (1.2) below. The completion time of this project is then the maximum completion time for each of these paths. The expression for this sample performance function is given below where ζ^* is the path with the maximal completion time:

$$h(X,\zeta^*) = \begin{cases} X_1 + X_4 + X_6 & \text{if } \zeta^* = \zeta_1 \\ X_1 + X_3 + X_5 + X_6 & \text{if } \zeta^* = \zeta_2 \\ X_2 + X_5 + X_6 & \text{if } \zeta^* = \zeta_3. \end{cases} \tag{1.2}$$

The second example in employing the SAN model is an interesting application from Garsva, [33]. In this paper, the SAN was used to characterize the incident process, modelling attacks due to external security threats and the response to these attacks on a distributed computer network. The state of the computer is characterized by five different working states, from “normal" to “non-functional" and the performance functions relate to probabilities and means amongst these working states.

A European-style vanilla call option, with an exercise price K, is a contract that permits the possibility of purchasing stocks at this price. This can only occur at the date when the contract ends, denoted by t. Let X denote the continuously compounded return of a stock over $[0,t]$, and the map h, defined Equation (1.3) below, is the discounted cash flow, with short-term interest rate r, of the specified contract:

$$h(X) = e^{-rt}\max\{e^X - K, 0\}. \tag{1.3}$$

This sample performance function is considered in detail in Section 4.6.2.1, computing sensitivities due to parameters involved in the price dynamics of the α-quantile cash flow, $\alpha \in (0,1)$, resulting from this option. We use both the Black-Scholes-Merton and Variance Gamma processes in this analysis. Even in more involved examples in the mathematical finance literature, such as path-dependent or multi-asset options, or more intricate price dynamics capturing more of the behaviour of price data, the basis is a simple sample performance function such as Equation (1.3). In Section 5.4 we consider one such complication where we estimate the sensitivity w.r.t. the barrier level of a discretely monitored "up-and-out" European-style Parisian put option. The results depend upon the notion of certain critical observed prices. Alternatively, in Cont and Deguest, [16], they propose an estimation scheme composed of a random mixture of asset pricing models based upon minimizing relative entropy to compute multi-asset European-style stock options and static-hedge ratios. Their method of discerning the price dynamics, compared to earlier works, reproduces the smile phenomenon of the implied volatility curve of the basket of stocks.

When the model is dynamic, represented by a stochastically recursive sequence, the two most common applications are queuing networks, most prominently exampled by its core recursion, the Lindley recursion, and in stochastic approximation. The most common stochastic approximation algorithm is by Robbins and Monro.

In queuing networks the fundamental element in constructing such systems is the $G/G/1/\infty$ first-in-first-out (FIFO) single-server queue. Let $X = (X_n)$, with $X_n = (S_{n-1}, A_n)$ respectively denote a sequence of independent and identically distributed (i.i.d.) random variables where X_n denotes respectively the service time needed by customer $n-1$ and the inter-arrival time of the n^{th} customer. The dynamics of this single-server sequence is modelled via a stochastically recursive sequence (Z_n) of waiting times for the customers that enter into the queue:

$$Z_n = h(Z_{n-1}, X_n) = \max\{Z_{n-1} + S_{n-1} - A_n, 0\}. \tag{1.4}$$

The above sequence is known as the Lindley recursion. Quite often the stationary behaviour, attained via Equation (1.4), of the queuing system needs to be understood. This stationary behaviour, apart from the topology of the system, depends on the traffic intensity of the constituent single-server queues, $\rho = \mathbb{E}[S_n]/\mathbb{E}[A_n]$. Estimating sensitivities provides insight into this relation via, for instance, ascertaining the sensitivity of the mean waiting time or mean queue length of a stationary queuing system w.r.t. defining parameters in the random variables (S_n), $n \geq 0$, and (A_n), $n \geq 0$. In Section 2.4.2.2 we examine a stationary feed-forward queueing system composed of three $M/M/1/\infty$ queues. We

take advantage of the feed-forward structure and the known stationary waiting time distribution of the single-server $M/M/1/\infty$ queue to determine the effect a change of an exponential service time parameter has on the overall waiting time for a customer.

Ultimately, we would like to control and optimize our stochastic representations, giving us means and policies to improve the outcomes of the physical system or concern. Stochastic approximation algorithms, such as the Robbins-Monro algorithm, and its gradient free variant, the Kiefer-Wolfowitz algorithm, are simulation based approaches that iteratively determine the optimal choice of parameters, $\theta^* = \lim_n \theta_n \in \Theta$, to aid in this decision making. These above methods are analogous to deterministic methods such as the Newton method. The text by Spall, [99], provides much insight into stochastic approximation algorithms. The unconstrained version of the Robbins-Monro algorithm is given by

$$\theta_{n+1} = h(\theta_n, X_n) = \theta_n + a_n X_n, \tag{1.5}$$

where the sequence $X = (X_n)$ denotes the parameter derivative, or sensitivity, of the mean value of a sample performance function or stochastically recursive sequence; and the deterministic sequence (a_n), $a_n > 0$, monotone decreasing towards zero, known colloquially as the gain sequence, is a collection of numbers, given certain restrictions, that ensures almost sure convergence of the parameter sequence to the true value. In the constrained case, Equation (1.5), the values of the sequence (θ_n) are projected onto the set of feasible parameter values Θ. Derivative estimation provides us the means to attain the gradient X_n. A recent example of employing the Robbins-Monro algorithm was in, [98], to estimate the parameters via the method of moments to measure the time evolution of a social network and the properties therein. The evolution of this network is modelled via a continuous time Markov Chain. Mundt, [80], uses this technique to ascertain the incidence of binge-drinking behaviour between friendships among adolescents. The results were from a stratified sample of 7th-12th graders from 132 American schools.

1.1.2 Sensitivity Analysis

Suppose that the random variable $Z(\theta) : \Theta \times \Omega \mapsto \mathbb{R}$ is a mapping of a single parameter θ. The domain Θ is an open interval within $\mathbb{R}$. A parameter within a stochastic model can either be *intrinsic* in nature, located with in one or more random variables X, influencing the dynamics, or *extrinsic*, where the influence is only induced by the mapping h. For the examples given in Section 1.1.1, these parameters were all intrinsic. For this chapter, unless specified, we assume that

θ is an intrinsic parameter. We depict the influence of the parameter on the random variable via the mapping $X(\theta) : \Theta \times \Omega \mapsto \mathbb{R}^J$.

For the moment we would like to infer the mean of the performance function $Z(\theta) = h(X(\theta))$, presuming $\mathbb{E}[Z(\theta)] < \infty$, which we remind is a general construction. Mathematically, the mean $\mathbb{E}[Z(\theta)]$ is observed by two principal methods. Firstly, $X(\theta)$, $\theta \in \Theta$, can be written as a map of a common random variable X_1 and the parameter θ. Hence, $X(\theta)$ in this representation is explicitly written as a function of θ, where $X(\theta) := g(X_1, \theta)$ for some Borel map g. The random variable X_1 is distributed with probability measure $\mathbb{P}$. The expectation $\mathbb{E}[Z(\theta)]$ is given by the Lebesgue integral

$$\mathbb{E}[Z(\theta)] = \int_{\mathbb{R}} h(g(x_1, \theta))\, \mathbb{P}(dx_1). \tag{1.6}$$

Alternatively, we can consider $X(\theta)$ to have its own separate probability measure, $X(\theta) \sim \mathbb{P}_\theta$, $\theta \in \Theta$. This implies that the parameter influence comes from the distribution, not directly exhibited:

$$\mathbb{E}_\theta[Z] = \int_{\mathbb{R}} h(x)\, \mathbb{P}_\theta(dx). \tag{1.7}$$

In this representation we have suppressed the dependence of θ as a function of the input random variables and the performance measure; i.e. $X(\theta) =: X$, and $Z(\theta) =: Z$. Correspondingly, the expectation is parameterized by θ, denoting the probability measure. The probability measures are Borel measurable, $\mathbb{P}_\theta : \Theta \times \mathscr{B}(\mathbb{R}) \mapsto \mathbb{R}$.

For this chapter, and for the thesis, we will assume that the input random variables $X(\theta)$ have a jointly continuous distribution function, implying that each element of the random vector is continuous. We will also require that the performance measure random variable $Z(\theta)$ is continuous almost everywhere. Further assumptions depend on the derivative estimation method. We denote by F_θ the cumulative distribution function (c.d.f.) of the probability measure $\mathbb{P}_\theta$ (and, respectively, F for the probability measure $\mathbb{P}$). The corresponding density function (p.d.f.) is denoted by f_θ (respectively f). Derivative estimation for discrete distributions will be discussed as a part of the overview of the main gradient estimation approaches in Section 1.2 but otherwise not be considered.

Following this explanation, derivative estimation, assuming differentiability, is then differentiating Equation (1.6) or (1.7) w.r.t. θ. If the expression for the integral is known, simulation is unnecessary as the functional behaviour w.r.t. θ is known. We therefore need to differentiate the integrand, and for the moment, assume we may interchange integral and derivative. Permission of this interchange requires some form of uniform integrability condition. Details depend

on the form of the mean performance measure, either (1.6) or (1.7), and are elucidated in Section 1.2. A good introduction into derivative estimation is Fu, [29], which focuses on the essentials: main approaches to derivative estimation, the validity of applying each approach, and implementation issues. A more detailed overview of derivative estimation is by the same author in [26] with greater emphasis on applying derivative estimation to particular discrete-event stochastic models. Some texts that discuss Monte Carlo simulation include [3], [99], and [38].

Each of the principal depictions, Equation (1.6) and (1.7), leads to its own derivative estimation method. The *pathwise* methods, or Perturbation Analysis methods, follow from Equation (1.6), in which the random variable, and performance measure, is differentiated w.r.t. the parameter. The *distributional methods* follow from Equation (1.7) in which the distributions associated with the input random variables are differentiated.

The pathwise methods all require the random variable $X(\theta) := g(X_1, \theta)$ to be an almost surely differentiable map w.r.t. θ. From this position the different perturbation analysis methods arise. For the thesis we consider two methods: *Infinitesimal Perturbation Analysis* (IPA) and *Smoothed Perturbation Analysis* (SPA). Infinitesimal Perturbation Analysis requires h to be differentiable w.r.t. θ. More specifically, IPA particularly requires the map h to be almost surely differentiable map w.r.t. to the state variables. A fuller explication where the required conditions for IPA hold are given in Section 1.2.1.1. If the conditions hold for IPA, the parameter derivative of the mean performance measure is the mean of the parameter derivative of the performance function, Equation (1.8):

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}[Z(\theta)] &= \frac{d}{d\theta}\int_{\mathbb{R}} h(x(\theta))\, f(x)\, dx \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta} h(x(\theta))\, f(x)\, dx \\ &= \mathbb{E}\left[\frac{\partial}{\partial\theta} Z(\theta)\right]. \end{aligned} \tag{1.8}$$

If, however, h is not differentiable or is infeasible to apply differentiation, SPA is applied using conditioning of random variables to integrate out portions of the integrand so that IPA can be used for the resulting conditional expectation. The "smoothed" sobriquet comes from rendering jump continuous components of maps, such as indicator mappings, to something differentiable from the random variable conditioning. SPA is further explained in Section 1.2.1.2.

Any discussion to derivative estimation should involve the Finite Difference (FD) method, the derivative estimation method that is most applied by practitioners. Equation (1.9) illustrates the forward method of the FD estimator, with

$\theta, \theta + \Delta \in \Theta$:

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta)] = \frac{\mathbb{E}[Z(\theta+\Delta)] - \mathbb{E}[Z(\theta)]}{\Delta}. \tag{1.9}$$

Observe that the IPA estimator is the sample path version of the limit in (1.9), assuming the limit exists. Conditions for FD to apply successfully are broadly similar to IPA and the simulation behaviour of the FD estimator is best when IPA is applicable. The FD method is further explained in Section 1.2.1.3.

The advantage of distributional methods is that the parameter derivative does not affect the performance function. The distribution function containing the parameter θ is at least weakly differentiable w.r.t. the parameter. These derivative estimation methods can be readily applied as an alternative to pathwise methods for intrinsic parameters, or to performance functions that are non-differentiable, such as indicator mappings or to complex performance functions in which the derivative of the distribution function can simply be inserted into the mean expression. If the probability density function f_θ is differentiable w.r.t. θ, the derivative of the mean performance function is an integral w.r.t. to a "signed" density function f'_θ:

$$\frac{d}{d\theta}\mathbb{E}_\theta[Z] = \int_{\mathbb{R}} h(x)\, f'_\theta(x)\, dx. \tag{1.10}$$

The two methods of the distributional approach differ in how they treat the signed density and convert the above integral into an expectation. The *Score Function* (SF) method, as most commonly known or the *Likelihood Ratio* method, inserts a probability distribution p and converts the signed density into a distribution specific augmentation to the performance function. This augmentation is known as the *score function* . This approach requires that f_θ is absolutely continuous w.r.t. to density function p. Commonly the inserted density function is the underlying density $p = f_\theta$. The method is known as the Likelihood Ratio method since the score function can be interpreted as the likelihood ratio function in statistics. The Score Function method is further explained in Section 1.2.2.1. From the Hahn-Jordan Theorem, the signed density in (1.10) can alternatively be written as a difference of functions where the supports of both functions are disjoint. Making use of this observation and converting these functions into density functions f^+_θ, f^-_θ is the basic idea to *Measure Valued Differentiation*. The corresponding derivative of the mean performance measure is then a scaled difference of the mean performance measure of the two modified systems with a non-unique choice of the measure-valued derivative density functions. The Measure-Valued Differentiation method is elucidated in Section 1.2.2.2.

For each of the derivative estimators the forthcoming explanations in Section 1.2, we will assume independent and identically distributed replications of the performance function $Z(\theta)$, with $Z(\theta)$ representing either a static stochastic model or a model occurring from a stochastically recursive sequence at a specified event. An example for the later performance function is the waiting time of, for instance, the 30th customer in for a FIFO single-server queue beginning from an empty state, denoted by Z_{30}. This waiting time is composed from the sequence $X = (X_n : 1 \leq n \leq 30)$ of copies $X = (A, S)$ consisting of the inter-arrival time and service time related to the customer. The random vector X_n may depend on earlier realizations of this vector. Most of the forthcoming references relate to these type of stochastically recursive sequences.

Derivative estimation is primarily used in stochastic operations research and, more prominently, in optimization and control. Derivative estimation is also applied to a lesser extent in statistical estimation. The overview [26] provides a literature review on the applications onto different types of stochastic models up to 2006. We will provide a combination of recent and/or more pertinent references.

Historically, derivative estimation was mostly applied to queueing based models and networks. Ho and Cao [47] is an early overview, see [36] for the initial application to queueing networks, and [68], [69], for imparting stochastic optimization to queueing systems. Some of the other earlier applications, such as [25], [39] and [63], were towards inventory systems. Mahajan and van Ryzin [77] applied the Robbins-Monro algorithm and IPA to optimize the initial inventory level to maximize the expected single-period profit where consumers arrive sequentially and consume according to their maximum utility. More recent accounts of derivative estimation in network-based applications are for the optimization of parameters, using SF, to model social networks in [98], and [97]; optimization of age-based threshold maintenance repair via MVD, [50]; stochastic fluid models determining IPA derivatives w.r.t. capacity thresholds [13], and control of these models in various applications such as resource contention games [106], [107].

More recently, much application of derivative estimation methods is towards finance, especially towards financial derivatives. A financial derivative is a contract in which the value of an asset at some future time is derived from prices from physical or financial assets, or physical or financial attributes. An option, Equation (1.3), is an example of a financial derivative. Evaluating parameter derivatives allows the financial agent to determine where the causes of the price occur in which the agent can hedge against these risks. In option pricing, where most of this research is based, the parameter derivatives of the option premium

are colloquially known[1] as the "Greeks"; the option premium is attained as a discounted expectation of a sample performance function or payoff function.

Other than for determining investment strategies, derivative estimation is in finance also used to find optimal drift parameter(s) for variance reduction in computing option premiums. The seminal papers on computing Greeks are Fu and Hu, [30], and Broadie and Glasserman, [10]. The first paper computed sensitivities of European and American call stock options via IPA, whilst the second paper was a comparison between the IPA and SF methods for European options. MVD was first applied to options in [49]. Computing Greeks of interest rate options has received much attention and notable papers include [41], [35], and [19]. For optimization, [102] used stochastic approximation as a part of other variance reduction techniques to compute European Asian stock options; and [2] focused on imparting stochastic approximation on different pricing models and payoff functions. Alternatively, [104] employed the Robbins-Munro algorithm to ascertain the early exercise policy and computed the price of American-style Asian options. Other applications of derivative estimation in finance include Chen and Glasserman, [15], where the authors compared the IPA, SF, and FD methods to determine the effect the choice of a hazard rate has on the pricing of credit derivatives; and Joshi and Pitt, [61], use IPA to determine the sensitivities of factors, such as a change of interest rate and rate of inflation, that influence the present value of a defined benefit pension fund.

We conclude this section with some remarks regarding derivative estimation that are tangential to the presentation. Application of derivative estimation extend to vector and matrix valued mappings of Z. As differentiation is an element-wise procedure, the derivative estimator of the vector or matrix arrives from the elementwise estimators. Derivative estimation also applies to extrinsic parameters though treatment of these problems depends on the particular model. In certain instances, such as inventory models, if the parameter is a scalar or a part of an indicator, pathwise methods can be implemented, in the later case via SPA. Alternatively, if a distributional method is used, an extrinsic parameter can be "pushed-out", a general method was proposed for SF in [88], to convert the extrinsic parameter into an intrinsic parameter by change of measure. However, applying derivative estimation to an extrinsic parameter may require additional adaptation to convert the stochastic model into either Equation (1.6) or (1.7). Chapter 5 deals with threshold parameters where conditioning or the "push-out" technique does not suffice and an analysis is required.

The emphasis on derivative estimation for discrete-event stochastic models in the above and subsequent discussion is pedagogical. To simulate a continu-

[1] The sobriquet arises as the parameter derivatives are named after Greek letters.

ous time model, such as a Stochastic Differential Equation (SDE), and to represent the evolution of the system, requires collecting a sequence of realizations, chosen at distinct times. At implementation, the continuous time model becomes a difference equation or a stochastically recursive sequence from which derivative estimation can be utilized. From [14], imparting a combination of IPA and SF to the difference equation is equivalent to discretizing the parameter derivative of the SDE attained via Malliavin calculus. More details of Malliavin Calculus is given in Section 5.1.

1.2 Approaches to Gradient Estimation

This section is an exposition to each of the derivative estimation methods. For each of the derivative estimation methods we assume that the input random vector $X(\theta)$ has continuous elements and a continuous distribution function, and that the performance function realizations $(Z_n(\theta))$, $n \geq 0$, form a sequence of at least locally continuous i.i.d. random variables. For random input vectors that contain elements with discrete distributions, we contrast the behaviour to the continuous distribution results for both gradient estimation approaches. In each section, we explain the intuition behind each of the derivative estimation methods as well as the conditions needed for each of the methods to function. For the Finite Difference method, Score Function method, and Measure Valued Differentiation there are additional aspects to consider and these will also be explored.

We first consider the pathwise methods, including FD, in Section 1.2.1 before the distributional methods in Section 1.2.2. In Section 1.2.3, we weigh the advantages and disadvantages as a technique in simulation for each of the derivative estimation methods. In Section 1.2.4 we complete the exposition by giving an illustration of each estimation method, attaining the pathwise and distributional derivatives for both the exponential and normal distributions.

Expanding on the notation from Section 1.1, let $(\Omega, \mathscr{F}, \mathbb{F}, \mathbb{P})$ be a filtered Borel measurable probability space, our basis for our discrete-event stochastic model. Let $\Theta \subset \mathbb{R}$ be an open interval denoting the domain space of the parameter θ. Let $X : \Theta \times \Omega \mapsto \mathbb{R}^J$ denote the input random vector $X(\theta)$, and let $Z : \Omega \times \Theta \mapsto \mathbb{R}$ denote the performance measure $Z(\theta)$. From Equation (1.1) these random variables are related by the map $h : \mathbb{R}^J \mapsto \mathbb{R}$. With this construction, $\tilde{\mathscr{F}} = \sigma((\cup_{n=1}^{\infty} Z_n))$, is the σ-field induced by the output random variables. For the stochastically recursive sequence the σ-field is given by $\mathscr{F} = \sigma(\tilde{\mathscr{F}} \cup \mathscr{N})$ containing the null sets, defined by $\mathscr{N}$, from our probability measure $\mathbb{P}$. The augmented filtration $\mathbb{F}$ is such that the σ-field $\mathscr{F}_n$, after $n \geq 0$ events have occurred in the stochastically

recursive sequence, is $\mathscr{F}_n = \sigma(\tilde{\mathscr{F}}_n \cup \mathscr{N})$ where $\tilde{\mathscr{F}}_n$ is the field restricted to the first n outcomes. For static models, the filtration is not needed.

For each method we provide the conditions necessary for unbiasedness. To obtain the Central Limit Theorem (CLT), as we are assuming i.i.d. random variables, we require in addition at least the second moment of the derivative estimator to be finite for all $\theta \in \Theta$. The IPA, SPA, and the distributional methods adhere to the standard rate of convergence in statistical or Monte Carlo estimation of $O(n^{-1/2})$ towards the actual value of $\partial_\theta \mathbb{E}[h(X(\theta))]$. The term n denotes the number of realizations that formed the derivative estimate. For the FD methods, the standard rate of convergence can be up to $O(n^{-1/2})$, depending on the assumptions and this is one of the additional considerations discussed.

1.2.1 Pathwise Differentiation

Given a Borel map g, we stated in Section 1.1.2 that the parameter θ is treated as an argument within the random vector $X(\theta)$, i.e., $X(\theta) = g(X_1, \theta)$. The random vector X_1 is merely a reference random variable influenced by a parameter $\theta_0 \in \Theta$ and so any two random variables induced by a parameter in Θ are related to each other. Correspondingly, for all $\theta \in \Theta$, the probability measure $\mathbb{P}_\theta$ is induced by the random variable $X(\theta)$, namely $\mathbb{P}_\theta(A) = \mathbb{P}(X_\theta^{-1}(A))$ for $A \in \mathscr{B}(\mathbb{R}^J)$.

Differentiation w.r.t. θ is considered via the sample path. For every $\omega \in \Omega$, the pathwise method considers the resulting realization, with $x := X_1(\omega)$, $[X(\theta)](\omega) := g(x, \theta)$ to be a deterministic differentiable function w.r.t. θ. The random variable $X(\theta)$ is then considered to be differentiable w.r.t. θ almost surely (with probability one) if there exists a set $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ for which $[X(\theta)](\omega)$ is a differentiable function. Reverting the $X(\theta)$ to being a vector, the gradient of the map exist if the constituent random variables induced by θ is differentiable almost surely.

The requirement of a differentiable random variable is the underlying assumption for the pathwise methods to be applicable. Frequently, the parameter of interest is either a location, $X(\theta) = X + \theta$, or scale parameter, $X(\theta) = \theta X$, and it is immediate that the random variable is differentiable w.r.t. the parameter. Pathwise differentiation is useful in many cases as the functional form of the parameter in the random variable is either a location or scale parameter. Pathwise differentiation is also advantageous for extrinsic parameters as the parameter can often be incorporated into a random variable. The differences between IPA and SPA is due to differences in properties of the map h.

This requirement of a differentiable random variable also inhibits its usefulness to discrete random variables. If the realizations to some part are a function of the parameter θ, and specifically independent of the probability of each out-

come, then differentiation via pathwise methods is permitted given the realization of the random variable. However, if θ is present in the probability distribution then pathwise differentiation cannot be applied, even if the probability of each outcome is a differentiable function w.r.t. θ.

An example of this is if Y and Z are two independent integrable random variables and X is the mixture of Y and Z where $X = Y$ with probability θ and equals Z with probability $1-\theta$. Defining as a Bernoulli random variable with parameter θ, i.e., $\eta_\theta \sim Ber(\theta) \in \{0,1\}$, in particular, $X = 1_{\eta_\theta=0} Y + 1_{\eta_\theta=1} Z$ and the parameter θ is located in the indicator mapping. However, we can condition w.r.t. the random variable η_θ and SPA can be applied.

1.2.1.1 Infinitesimal Perturbation Analysis

The main reference texts for the Infinitesimal Perturbation Analysis method are Ho and Cao, [47], and Glasserman, [36]. As mentioned in Section 1.1.2, IPA is applicable when h is a map almost everywhere differentiable w.r.t. θ. More specifically, extending from Section 1.1.2, the IPA method provides an unbiased estimator representing the parameter derivative if the following three conditions hold:

(IPA 1) There exists a subset $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that for all $\theta \in \Theta$, $X(\theta)$ is differentiable.

(IPA 2) The map h is piecewise-differentiable.

(IPA 3) The performance function $h(X(\theta))$ is Lipschitz continuous with integrable modulus of continuity for all $\theta \in \Theta$; i.e, $|h(X(\theta+\Delta)) - h(X(\theta))| \leq K(X)\Delta$ with $\mathbb{E}[K(X(\theta))] < \infty$.

A sufficient condition for Assumption **(IPA 3)** to hold is if the modulus of continuity, represented by $\sup_{\theta\in\Theta} |\partial_\theta h(X(\theta))|$, is integrable for all $\theta \in \Theta$. This observation is a consequence of the Mean Value Theorem. Hence, **(IPA 3)** may be replaced by the following stronger condition:

(IPA 4) The partial derivative $\partial_\theta h(X(\theta))$ has a uniformly bounded absolute expectation, i.e., $\mathbb{E}[\sup_{\theta\in\Theta} |\partial_\theta h(X(\theta))|] < \infty$.

For the proof of the unbiasedness of the IPA estimator, alternatively the permission of the interchange between expectation and derivative operators, Equation (1.8), we will employ Assumptions **(IPA 1)**, **(IPA 2)**, and **(IPA 3)**. From Assumption **(IPA 3)**, given that $\Delta \neq 0$:

$$\mathbb{E}\left[\frac{1}{\Delta}|h(X(\theta+\Delta)) - h(X(\theta))|\right] \leq \mathbb{E}[K(X(\theta))], \tag{1.11}$$

and from the Dominated Convergence Theorem, the limit and expectation due from (1.11) can be interchanged:

$$\lim_{\Delta\to 0}\frac{\mathbb{E}[h(X(\theta+\Delta))]-\mathbb{E}[h(X(\theta))]}{\Delta}=\mathbb{E}\left[\lim_{\Delta\to 0}\frac{h(X(\theta+\Delta)-h(X(\theta))}{\Delta}\right]. \tag{1.12}$$

As **(IPA 1)** together with the LHS of (1.12) is the definition of the parameter derivative, and, from Assumption **(IPA 2)**, the limit on the RHS of Equation (1.12) is the derivative of the performance measure, and hence Equation (1.7) is derived:

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))]=\mathbb{E}\left[\frac{\partial}{\partial\theta}h(X(\theta))\right].$$

In usage, the derivative of the performance measure, if Assumption **(IPA 1)** holds, is rewritten via the chain rule to include the pathwise derivative. As $X(\theta)$ is a J-element vector, the expression for the IPA derivative for the mean performance measure $Z(\theta)$ is given by

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta))]=\mathbb{E}\left[\sum_{i=1}^{J}\frac{\partial}{\partial x_i}\Big|_{x=X(\theta)}h(x)\,\frac{\partial}{\partial\theta}X_i(\theta)\right]. \tag{1.13}$$

In many applications, the map h is differentiable. This is readily evident for payoff functions, e.g., Equation (1.2). In addition, after some effort, many models can be written so that an IPA estimator can be attained. The shortcomings are that IPA cannot be applied to indicator mappings, and so an alternative method is needed to find the sensitivity of probability mappings. In addition, for more involved maps, such as queuing networks, IPA has to satisfy a concept known as the “commuting condition" before the method can be used.

1.2.1.2 Smoothed Perturbation Analysis

Smoothed Perturbation Analysis is a method to circumvent non-differentiable maps which contain the parameter θ. The principal book on SPA is [31]. The idea behind this method is that by conditioning an analytical expression can be attained that is differentiable w.r.t. the parameter θ.

For an explanatory purpose, we suppose that the map $h(x)=r(x_{\bar{1}})I(x_1)$ is a product of two maps, I and r, with respective state variables x_1 and $x_{\bar{1}}=(x_i:2\leq i\leq J)$. Let r be an almost surely differentiable mapping w.r.t. to $x_{\bar{1}}$ and assume that I fails to be an almost surely differentiable mapping w.r.t. x. In applications, the map I is commonly an indicator mapping. Suppose for this example, without loss of generality, that our random vector $X(\theta)$ consists of indepen-

dent elements[2]. In this example, we are conditioning w.r.t. the random vector $X_{\bar{1}}(\theta) = (X_i(\theta) : i \leq 2 \leq J)$ to attain the conditional expectation $\mathbb{E}[I(X(\theta))|X_{\bar{1}}(\theta)]$, which we assume to be almost surely differentiable. For SPA to apply in this context we require the analogous assumptions from IPA, Section 1.2.1.1, applied to the conditional expectation of the map h. Hence, we need Assumption **(IPA 1)** and the following three assumptions below:

(SPA 1) The map $\mathbb{E}[h(X(\theta))|X_{\bar{1}}(\theta)] = r(X_{\bar{1}}(\theta))\mathbb{E}[I(X(\theta))|X_{\bar{1}}(\theta)]$ is almost surely differentiable with derivative $\partial_\theta r(X_{\bar{1}}(\theta))\mathbb{E}[I(X(\theta))|X_{\bar{1}}(\theta)]$.

(SPA 2) The conditional expectation $\mathbb{E}[\mathbb{E}[h(X(\theta))|X_{\bar{1}}(\theta)]]$ is Lipschitz continuous with integrable modulus of continuity $K(X)$ with $\mathbb{E}[K(X)] < \infty$ for all $\theta \in \Theta$.

As for IPA, a sufficient condition to satisfy Assumption **(SPA 2)** is that the modulus of continuity, here represented by $\sup_{\theta\in\Theta} |\partial_\theta \mathbb{E}[h(X(\theta))|X_{\bar{1}}(\theta)]|$ is integrable. This can also be verified via the Mean Value Theorem. Hence, we may replace **(SPA 2)** by the following condition:

(SPA 3) The derivative of the conditional expectation $\partial_\theta \mathbb{E}[h(X(\theta))|X_{\bar{1}}(\theta)]$ has a uniformly bounded expectation for all $\theta \in \Theta$, i.e., $\mathbb{E}[\sup_{\theta\in\Theta} |\partial_\theta \mathbb{E}[h(X(\theta))|X_{\bar{1}}(\theta)]|]$ $< \infty$.

Because of the similarity to the IPA derivative estimator, with the conditional expectation $\mathbb{E}[h(X(\theta))|X_{\bar{1}}(\theta)]$ substituting for the performance measure, the SPA estimator is unbiased. The resulting SPA Monte Carlo derivative estimator yields

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta)] = \mathbb{E}\left[\sum_{i=2}^{J} \frac{\partial}{\partial x_i}\Big|_{x=X(\theta)} \left\{r(x_{\bar{1}})\mathbb{E}\big[I(\cdot)\big|X_{\bar{1}}(\theta) = x_{\bar{1}}\big]\right\} \frac{\partial}{\partial\theta}X_i(\theta)\right].$$

We note for the conditional expectation that the random variable $X_1(\theta)$ has been integrated from the performance measure. In SPA, the conditioned random variables may not be dependent on the parameter of interest. An example is again the parameter derivative w.r.t. the barrier level of a discretely monitored Parisian option in Section 5.4. Secondly, the map h does not need to have the product functional form. This was chosen for exposition showing that not the entire performance function needs to be conditioned.

When IPA cannot be applied or is infeasible to apply, SPA extends the scope of pathwise derivative estimation to ascertain sensitivities of discrete-event

[2] An input random vector with elements that are dependent on each other is only a complication. A common approach is to rewrite the random vector as another vector containing independent elements.

stochastic models. SPA is especially applicable to ascertain derivatives of extrinsic parameters as the distributional methods cannot be applied or are unwieldy to apply. Applying SPA to derivative estimation is an art-form, depending on the model in question.

1.2.1.3 Finite Difference Method

The popularity of the FD method lies in its intuitive simplicity: the notion of gradient as a derivative; and only little additional thought is needed to attain a derivative estimate, inserting the different parameter values to get an answer.

In the derivative estimation literature, the FD method has been much analyzed as a formal derivative estimation method. Some of the prominent papers in this topic are by L'Ecuyer and Perron [70], Glynn [42], and Glasserman and Yao [40]. The text by Asmussen and Glynn, [3], has a recent synopsis on the FD method. The major FD estimators, with $\Delta > 0$, are the forward/backward estimator, with the forward estimator given in (1.9), repeated below:

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta)] = \frac{\mathbb{E}[Z(\theta+\Delta)] - \mathbb{E}[Z(\theta)]}{\Delta}, \tag{1.14}$$

and the central estimator

$$\frac{d}{d\theta}\mathbb{E}[Z(\theta)] = \frac{\mathbb{E}[Z(\theta+\Delta)] - \mathbb{E}[Z(\theta-\Delta)]}{2\Delta}, \tag{1.15}$$

where $\theta-\Delta, \theta+\Delta \in \Theta$. A Taylor series expansion of the mean performance measures around θ proves that both the forward/backward and central estimators are biased estimates of the parameter derivative. There are a countable number of forms of the FD estimator with additional perturbations around θ that reduces the bias to an arbitrary degree, but these expressions are not as intuitively appealing.

From [70], to attain the FD estimators we require a stronger form of the differentiability conditions than the IPA derivative estimator. We require Assumptions **(IPA 2)** and **(IPA 3)** for the derivative of the performance measure as well as two similar assumptions for the second derivative:

(FD 1) The mean performance measure $\mathbb{E}[h(X(\theta))]$ is piecewise-twice differentiable with continuous derivative, and

(FD 2) It holds that $\sup_{\theta\in\Theta} \partial^2_\theta \mathbb{E}[h(X(\theta))] < \infty$; where ∂^2_θ denotes the second derivative w.r.t. the parameter.

Apart from being the derivative-free version of the IPA estimator, these like requirements are a second reason why IPA and FD are similar approaches.

Let n be the number of realizations to acquire an estimate. We will assume that estimates via the FD method are attained by common random numbers. This implies that the same generations of the random vectors are used to estimate both mean performance functions for $\partial_\theta \mathbb{E}[h(X(\theta)]$. This has the effect to increase the rate of convergence towards $\partial_\theta \mathbb{E}[h(X(\theta))]$. Given the said assumptions, from, [70], the rate of convergence to the actual value of $\partial_\theta \mathbb{E}[h(X(\theta))]$ using common random variables is $O(n^{-1/2})$ in which the step-size[3] Δ is $\Delta = O(n^{-1/2})$. To attain this rate of convergence Assumption **(FD 2)** is required, i.e., the boundedness of the derivative of the map h. This implies a Lipschitz continuity for $h(X(\theta+\Delta))-h(X(\theta))$ in which the variance, $\mathrm{Var}(h(X(\theta+\Delta))-h(X(\theta)) = O(\Delta^2)$. If the conditions for the map h do not satisfy Assumption **(FD 2)**, [42] proves a reduced rate of convergence for both types of estimators. Given a variance behaviour of $\mathrm{Var}(h(X(\theta+\Delta))-h(X(\theta)) = O(\Delta)$, the optimal order of convergence is $O(n^{-1/3})$ for the forward/backward estimator and $O(n^{-2/5})$ towards $\partial_\theta \mathbb{E}[h(X(\theta)]$ for the central FD estimator. In addition, providing separate realizations for the n input random vectors further reduces the rate of convergence. This is due to the further enlarged variance, since $\mathrm{Var}(h(X(\theta+\Delta))-h(X(\theta)) = O(1)$. For these cases, [42] provides the optimal rate for step-size as well as the optimal coefficients that belong with them.

A second reason common random numbers is preferred is that random number generation is computationally costly. For the separately estimated mean performance functions, the additional amount of time needed to attain a derivative estimate to a specified precision is compounded by generating the additional random vectors.

1.2.2 Distributional Differentiation

Distributional methods have the appeal that the map h does not need modification to evaluate a derivative. The expressions attained by the Score Function method and Measure Valued Differentiation occur from the density function, assuming continuous random variables, from a chosen distribution. Markedly, the insertion of the same density for each performance function h will automatically yield a derivative estimator. As opposed to the Finite Difference method, distributional methods can be applied without the existence of higher order derivatives.

From Section 1.1.2 we exhibit the uncertainty of the stochastic model via the family of distribution functions $\{F_\theta\}$, $\theta \in \Theta$, and hence via the density function. To emphasize, the random vector $X \sim F_\theta$ is specified via its probability measure,

[3] If the mean performance function is three times continuously differentiable, then for the central estimator the step-size is $\Delta = O(n^{-1/4})$ to attain the $O(n^{-1/2})$ convergence.

and the random vector and hence the performance function has no explicit dependency on θ. The density function of the distribution is denoted by f_θ.

We first develop Equation (1.10). A detailed discussion on the function-analytical concepts on the differentiability of measures is written in [51] and [52]. Let $\mathfrak{L}^1(f_\theta : \theta \in \Theta)$ denote the set of Borel measurable mappings h that are absolutely integrable w.r.t. the density function f_θ for $\theta \in \Theta$, and let $\mathfrak{D} \subset \mathfrak{L}^1(f_\theta : \theta \in \Theta)$. The density function f_θ is called $\mathfrak{D}$-differentiable if a finite signed density f'_θ exists such that for all $h \in \mathfrak{D}$:

$$\frac{d}{d\theta}\int_{\mathbb{R}} h(x)\, f_\theta(x)\, dx = \int_{\mathbb{R}} h(x)\, f'_\theta(x)\, dx, \tag{1.16}$$

and the density f'_θ is called the $\mathfrak{D}$-*derivative* of f_θ. By setting $h(x) = 1$, it is apparent that f'_θ is related to a signed measure

$$1 = \int_{\mathbb{R}} f_\theta(x)\, dx \tag{1.17}$$

and from (1.16)

$$0 = \int_{\mathbb{R}} f'_\theta(x)\, dx. \tag{1.18}$$

For $\mathfrak{D}$-differentiability, we require $\mathfrak{C}_b \subset \mathfrak{D}$, where $\mathfrak{C}_b$ denotes the set of continuous and bounded functions. This ensures that the derivative density f'_θ is uniquely defined. The general construction involves mappings of a family of measures $\{\mu_\theta\}$, $\theta \in \Theta$, onto a Polish space in which the equivalent condition to Equation (1.16) is in regard to a finite signed measure μ'_θ. Assuming a Borel measurable space on $\mathbb{R}^J$, $(\mathbb{R}^J, \mathscr{B}(\mathbb{R}^J))$, this construction connects to probability distributions by noting that $\mu_\theta(\mathbb{R}^J) = 1$, $\forall\, \theta \in \Theta$. The distribution function is then denoted by $F_\theta(x) = \mu_\theta((-\infty, x])$, where the interval is a condensed representation for the semi-open box $(-\infty, x] := \prod_{i=1}^{J}(-\infty, x_i]$, and the density function f_θ is the Radon-Nikodym derivative w.r.t. the Lebesgue measure λ; i.e. $f_\theta = d\mu_\theta / d\lambda$.

Let $v \geq 1$ be an integrable function, i.e. $v \in \mathfrak{L}^1(f_\theta : \theta \in \Theta)$, and let $c > 0$ be a constant. Two appropriate choices to represent the function space $\mathfrak{D}$ are the set of Borel functions $\mathfrak{B}_v$ and the subset of continuous functions $\mathfrak{C}_v$ that are bounded by cv. For the function space $\mathfrak{C}_v$, the function v is also required to be continuous. With the function space $\mathfrak{C}_b$ as a subset, the idea behind the later two spaces is to define enlarged function spaces while retaining the notion of boundedness and integrability. This equivalence is represented via the relation $|h|/cv$ in which this function is $\mathfrak{L}^1(f_\theta v)$ integrable. Specifically, if $h \in \mathfrak{B}_v$, this corresponds to $|h|/cv \in \mathfrak{B}$, the space of bounded Borel mappings for some constant $c > 0$, and if $h \in \mathfrak{C}_v$, for a continuous function v, this corresponds to $|h|/cv \in \mathfrak{C}_b$.

The purpose of the aforementioned framework is to ensure that a chosen probability distribution that describes the dynamics of the discrete-event stochastic model can be implemented for the designed performance function, attained from the underlying problem. The concept of $\mathfrak{D}$-differentiability provides an immediate verification.

Where the SF and MVD methods depart begins at Equation (1.16) and this is where we commence with the explanations. As a consequence, they have vastly different interpretations and simulation properties.

For discretely distributed random variables[4], the aforementioned explanation can be conducted w.r.t. to the counting measure. If we can find a family of distribution functions $(F_\theta), \theta \in \Theta$, in which the probability mass function is differentiable w.r.t. θ for all outcomes, both distributional methods can be applied similarly to the continuous distribution case as noted in their respective subsections. Unlike for continuous random variables, there is no mechanism to "push-in" the parameter θ from the support of the random variable to their associated probabilities. Thus distributional methods are unable to differentiate w.r.t. a parameter that is an argument of a discrete random variable.

1.2.2.1 Score Function Method

The seminal text for this derivative estimation method is by Rubinstein and Shapiro, [89].

The basic idea of the Score Function method is to rewrite Equation (1.16) for the $\mathfrak{D}$-differentiable map h by inserting the density function f_θ into the integral. The signed density f'_θ is then converted into a Radon-Nikodym[5] type expression $\mathrm{SF}_\theta(x) = (f'_\theta / f_\theta)(x)$ and this is what the score function refers to. The derivative w.r.t θ via the Score Function method is then

$$\begin{aligned} \frac{d}{d\theta}\int_{\mathbb{R}} h(x)\, f_\theta(x)\, dx &= \int_{\mathbb{R}} h(x) \left(\frac{f'_\theta}{f_\theta}\right)(x)\, f_\theta(x)\, dx \\ &= \mathbb{E}_\theta\left[h(X)\,\mathrm{SF}_\theta(X)\right]. \end{aligned} \tag{1.19}$$

If the density function $f_\theta(x) = \prod_{i=1}^{J} f_{\theta,i}(x_i)$ is a J-fold mutually independent product of the marginal densities, the derivative of the density function is attained by the product rule and we require the score function of each marginal

[4] Specifically random input vectors that contain discretely distributed elements, but this finer delineation does not provide any additional insight.

[5] Precisely the difference of two Radon-Nikodym derivatives by the Hahn-Jordan Theorem.

distribution:

$$\frac{d}{d\theta}\int_{\mathbb{R}} h(x)\, f_\theta(x)\, dx = \mathbb{E}_\theta\left[h(X) \sum_{i=1}^{J} \mathrm{SF}_\theta(X_i) \right]. \tag{1.20}$$

For many of the continuous probability distributions encountered, the score function can be written as $\mathrm{SF}_\theta(x) = \partial_\theta \ln f_\theta(x)$.

The advantages of this derivative estimation method is that the original probability distribution is used to attain the parameter derivative and that the parameter derivative is only a distribution specific augmentation of the performance function. This, together with the modularity of the approach, makes SF an attractive method.

The problem though is the high variance that is observed from simulation experiments, and in the form of the distribution given in Equation (1.20) the variance is linear in the number of times the derivative of a marginal distribution needs to be attained. In (1.20), this is a J-fold increase. This is a particular concern in terminating stochastically recursive sequences where, for instance, estimating the derivative of the 30th waiting time of a single server queue w.r.t. the inter-arrival parameter, requires 30 derivatives of the inter-arrival distribution. There is also the minor concern that the SF method cannot be applied to distributions for which the support depends on the parameter, for example, the uniform $U(0,\theta)$ distribution with support $(0,\theta)$.

Instead of inserting f_θ into the integral in (1.19), for static models any density function p can be chosen such that the family of density functions (f_θ), for $\theta \in \Theta$, is absolutely continuous w.r.t. p, i.e., the support of f_θ is a subset for the support of p for all choices of the parameter. Alternatively, if we prefer to not to simulate additional random variables, another density from the same family $f_{\theta'}$, $\theta' \in \Theta$, can be chosen, however, the support of the family of densities must not depend on θ. The consequent expectation is then written w.r.t. the chosen measure and that the score function is written as $\mathrm{SF}_p(x) = f'_\theta(x)/p(x)$. The purpose of inserting an substitute density p is to ensure that the derivative map $h(x)\mathrm{SF}_p(x)$ becomes nearly a constant function, minimizing variance. This choice of density has to be balanced against the increased computational time in calculating the derivative estimate. Nevertheless, from [3], for stochastically recursive sequences deviation from the choice $p = f_\theta$ as the inserted density function leads to much increased variance.

Reducing this explanation leads us to the conditions to implement a Score Function method, given below:

(SF 1) $f_\theta(x)$ is differentiable w.r.t. θ for all $\theta \in \Theta$ and for all x.

Let $D_\theta = \{x : f_\theta(x) > 0\}$ be the support, the pre-image of the non-zero values, of f_θ, and $D_p = \{x : p(x) > 0\}$ be the support of the density p.

(SF 2) $D_\theta \subset D_p$, i.e., for any chosen density function inserted into the integral (1.19) the density function f_θ must be absolutely continuous w.r.t. this choice.

(SF 3) The integral, $\int_{\mathbb{R}} (1 + |h(x)|) \sup_{\theta \in \Theta} \left| \partial_\theta f_\theta(x) \right| \, dx < \infty$, implying integrability of the parameter derivative.

Alternatively, if $h \in \mathfrak{B}_v$, as we will encounter is Section 1.2.2.2, Assumption **(SF 3)** is automatically satisfied for any choice of density function p.

1.2.2.2 Measure-Valued Differentiation

The principal text for the Measure Valued Differentiation estimation method is by Pflug, [83].

Rather than forming an expectation from the integral in Equation (1.16) by inserting a density function, the signed density can be written, via the Hahn-Jordan Theorem, [62], as a difference of two densities. We elucidate this idea given our Borel measurable space $(\mathbb{R}^J, \mathscr{B}(\mathbb{R}^J))$ and our family of measures $\{\mu_\theta\}$, $\theta \in \Theta$. Indeed, the space $\mathbb{R}^J$ can be partitioned into two sets, $A_\theta^\pm$, namely $\mathbb{R}^J = A_\theta^+ \cup A_\theta^-$, and $A_\theta^+ \cap A_\theta^- = \{\emptyset\}$. The sets $A_\theta^\pm$ are defined to be the largest Borel subsets of $\mathbb{R}^J$ such that for every Borel measurable subset of A_θ^+, respectively A_θ^-, the signed measure attains a positive (respectively negative) value. More formally, the sets $A_\theta^\pm = \sup\{B \subset \mathbb{R}^J : \mu'_\theta(C) >(<) 0 \ \forall \, C \subset B\}$. We denote the corresponding Hahn-Jordan measures by $\tilde{\mu}_\theta^+$, $\tilde{\mu}_\theta^-$ such that for any Borel set $C \subset \mathbb{R}^J$, the positive measures $\tilde{\mu}_\theta^\pm$ are defined by $\tilde{\mu}_\theta^\pm(C) := \pm \mu'_\theta(C \cap A_\theta^\pm)$ on $C \subset A_\theta^\pm$, and $\tilde{\mu}_\theta^\pm(C) = 0$ on $A_\theta^\mp$, the respective complementary sets. The signed measure can then be written for any Borel set C as the difference of these measures $\mu'_\theta(C) = \tilde{\mu}_\theta^+(C) - \tilde{\mu}_\theta^-(C)$.

To make use of this idea as a derivative estimation method, the measures $\tilde{\mu}_\theta^\pm$ are first converted into probability measures, before choosing the preferred form of probability measures. Then for the purposes of this explication, via the Radon-Nikdoym derivative, the measure-valued derivative densities $f_\theta^\pm$ is obtained and consequently the final expression. Due to the Hahn-Jordan Theorem, it is discernible that the derivative of the mean performance function is observed to be a scaled difference, with parameter dependent constant $c_\theta > 0$, of two mean performance functions with different distributions for the input random vector. We arrive at this destination via a few observations, beginning from Equation (1.18). Firstly, the positive measures $\mu_\theta^\pm$ are finite, and $\tilde{\mu}_\theta^+(A_\theta^+) =$

$\tilde{\mu}_\theta^-(A_\theta^-)$. In addition, the difference $\tilde{\mu}_\theta^+ - \tilde{\mu}_\theta^-$ is only unique up to a positive measure ν; i.e., for any subset $C \subset \mathbb{R}^J$:

$$\begin{aligned}\mu'_\theta(C) &= \tilde{\mu}_\theta^+(C) - \tilde{\mu}_\theta^-(C) \\ &= (\tilde{\mu}_\theta^+ + \nu)(C) - (\tilde{\mu}_\theta^- + \nu)(C) \\ &:= \mu_\theta^+(C) - \mu_\theta^-(C).\end{aligned}$$

We denote the supports for these positive measures $\mu_\theta^\pm$, $\theta \in \Theta$, by $B_\theta^\pm$. Then for any Borel set C, $\mu_\theta^+(C) = \mu_\theta^+(C \cap B_\theta^+)$, and $\mu_\theta^-(C) = \mu_\theta^-(C \cap B_\theta^-)$. From this position, given a Borel set C, the probability measures $\mathbb{P}_\theta^\pm$ is attained from the normalization of the measures $\mu_\theta^\pm$. The constant c_θ is then observed to be the total mass of $\mu_\theta^\pm$:

$$\begin{aligned}\mu'_\theta(C) &= \mu_\theta^+(C) - \mu_\theta^-(C) \\ &= \mu_\theta^+\left(B_\theta^+\right)\left(\frac{\mu_\theta^+(C)}{\mu_\theta^+(B_\theta^+)} - \frac{\mu_\theta^-(C)}{\mu_\theta^-(B_\theta^-)}\right),\end{aligned}$$

noting $\mu_\theta^+(B_\theta^+) = \mu_\theta^-(B_\theta^-)$,

$$:= c_\theta\left(\mathbb{P}_\theta^+(C) - \mathbb{P}_\theta^-(C)\right).$$

Repeating this argument for a collection of $\mathfrak{D}$-differentiable functions, and with $h \in \mathfrak{D}$, we observe how the MVD derivative estimator is attained:

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}_\theta[h(X)] &= \int_{\mathbb{R}} h(x)\left(\mu_\theta^+(dx) - \mu_\theta^-(dx)\right) \\ &= c_\theta\left(\int_{\mathbb{R}} h(x)\,\mathbb{P}_\theta^+(dx) - \int_{\mathbb{R}} h(x)\,\mathbb{P}_\theta^-(dx)\right).\end{aligned}$$

Then, from the Radon-Nikodym Theorem, $f_\theta^\pm = d\mathbb{P}_\theta^\pm / d\lambda$ are the resulting measure-valued derivative density functions for our continuous random variables, and that the MVD derivative estimator can be observed as a difference of means from two altered systems, i.e.,

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}_\theta[h(X)] &= c_\theta\left(\int_{\mathbb{R}} h(x)\,f_\theta^+(x)\,dx - \int_{\mathbb{R}} h(x)\,f_\theta^-(x)\,dx\right) \\ &:= c_\theta\left(\mathbb{E}_\theta\big[h(X^+)\big] - \mathbb{E}_\theta[h(X^-)]\right), \qquad (1.21)\end{aligned}$$

where we assume $X^\pm \sim f_\theta^\pm$. The measure-valued derivative of f_θ can be encoded by the triple $(c_\theta, f_\theta^+, f_\theta^-)$ or alternatively via the random variable representation

(c_θ, X^+, X^-). For a performance function $Z = h(X)$, the measure-valued derivative is similarly written, i.e., $Z^\pm = h(X^\pm)$.

Apart from the measure-theoretical dissection of sets, MVD can be conceived from the SF method. Given the density function f_θ, and the score function SF_θ, a measure-valued derivative pair of densities $f_\theta^\pm$, if the score function exists, can be obtained via the positive and, respectively, negative components from the score function. Specifically, the measure-valued density f_θ^+ can be written via $f_\theta^+(x) = (\mathrm{SF}_\theta(x))^+ f_\theta(x)/c_\theta$, and the definition for f_θ^- written analogously. For a function g, the function $(g(x))^+$ has the standard definition $(g(x))^+ = \max\{g(x), 0\}$, and for the function $(g(x))^-$ via $(g(x))^- = \max\{-g(x), 0\}$.

Though the intuition via SF may render that method more accessible, the set of probability distributions for which the SF method can be applied is a subset to the MVD method. For instance, a measure-valued derivative does not require the existence of the absolutely continuous measure. For example, the uniform $U(0,\theta)$ distribution with support $(0,\theta)$, the measure-valued derivative is attained via the Leibniz Theorem. The resulting expression is commonly a difference of terms, lending the measure-valued derivative type expressions.

In the multivariate case, if the density function f_θ is a J-fold product, $X \in \mathbb{R}^J$, of independently distributed marginal density functions; namely $f_\theta(x) = \prod_{i=1}^J f_{\theta,i}(x_i)$, the measure-valued derivative is also determined by the product rule. The final result, Equation (1.22), is similar to the Score Function method in this case:

$$\frac{\partial}{\partial\theta} f_\theta(x) = \sum_{i=1}^{J} d_{\theta,i} \prod_{j=1}^{i-1} f_{\theta,j}(x_j) \left(f_{\theta,i}^+(x_i) - f_{\theta,i}^-(x_i) \right) \prod_{j=i+1}^{J} f_{\theta,j}(x_j). \tag{1.22}$$

We denote the measure-valued derivative triple for each marginal density function by $(d_{\theta,i}, f_{\theta,i}^+, f_{\theta,i}^-)$, $i = 1, \ldots, J$. For products that have a null index set, $\prod_{\{\emptyset\}} f_{\theta,i} =: 1$. Let $X_{\tilde{i}}^\pm = (X_1, \ldots, X_{i-1}, X_i^\pm, X_{i+1}, \ldots, X_J)$, $1 \le i \le J$, denote the random the variable X with the i-th random variable substituted with the corresponding measure-valued derivative replacement. The corresponding MVD derivative estimator to Equation (1.21), when f_θ is a product of independent densities, reads:

$$\frac{d}{d\theta} \mathbb{E}_\theta[h(X)] = \mathbb{E}_\theta \left[\sum_{i=1}^{J} d_{\theta,i} \left(h(X_{\tilde{i}}^+) - h(X_{\tilde{i}}^-) \right) \right]. \tag{1.23}$$

The choice of the measure-valued derivative densities $f_\theta^\pm$ for this method brings up several simulation considerations. From (1.21), up to two additional random generations per appearance of a derivative are needed. In many parameter derivatives of distributions one of these densities; for instance f_θ^+, is

equal to f_θ and so we can use a realization for the random variable X for the realization X^+. Secondly, even though the sample mean of the performance function is evaluated from a collection of independently distributed random variables, there is no independence relationship implied between X_η, X_η^+, and X_η^-, for $\eta \in \{1,\ldots,n\}$. Expanding from the first point, it is advantageous to positively correlate the random variables $h(X^+)$ and $h(X^-)$ by common random number generation and, if possible, using the original realizations. This reduces both the variance and computation time. Thirdly, though any distribution can be generated via acceptance-rejection methods, some choices of f_θ^+, f_θ^-, do not require this and are more quickly generated. Fourth, independent sampling of X^+ and X^- from the Hahn-Jordan choice of density functions provides least variance. In practice, the choice of measure-valued derivative densities is evident where the policy is to either to attain f_θ for one of the density functions or to derive the Hahn-Jordan choice of densities. In many instances, correlation between X^+ and X^- is achieved.

The prime problem of MVD is the additional generation of random variables to attain the derivative estimator for a stochastically recursive sequence. From the example of the single-server queue given in Section 1.2.2.1, we would need to generate at least 30 additional random variables to find the parameter derivative w.r.t. the inter-arrival parameter for the waiting time W_{30}. The advantage for MVD for these dynamic models is that the variance does not increase due to the number of the parameter derivatives needed.

To implement the MVD derivative estimation method only one assumption is required: $\mathfrak{B}_\nu$ differentiability of f_θ for $\theta \in \Theta$:

(MVD 1) Let $\nu \geq 1$, $\nu \in \mathfrak{L}^1(f_\theta, \theta \in \Theta)$. Assume that Equation (1.21) holds for all $h \in \mathfrak{B}_\nu$.

1.2.3 Simulation Considerations

Each of the methods presented has its advantages and disadvantages, though all methods are not equally comparable in terms of efficacy. Specifically, efficacy is measured by estimation precision as well as the computational speed, and to a lesser extent broad applicability of the methods, requiring minimal assumptions. In other words, the traits a derivative estimation method would need for a practitioner would trust and implement a method as a part of the software infrastructure.

Therefore, the Finite Difference methods are considered to be an inferior estimation method as there is always a better method to attain the derivative. If map h is differentiable, IPA is available and the FD method needs the existence of higher-order derivatives of $\mathbb{E}[Z(\theta)]$ for the method to be applied. If the map

is not differentiable, the rate of convergence is lower than SPA, MVD or SF, and the variance for finite sample results is worse than SPA or MVD. Additionally, it takes considerable work to attain an optimal or near-optimal choice of step-size to minimize the variance of the FD estimator. Simplicity of the implementation method is not a main virtue as ultimately the performance of the simulation program is what matters.

Between SF and MVD, SF appears to have all of the properties that a practitioner would want in a derivative estimation method and MVD is a method method that appears to be difficult to use. However, the SF derivative estimation method does not perform well in simulations while the MVD estimation method does perform well. The advantage of the SF method is when $p = f_\theta$ where the parameter derivative can be estimated from the already generated random variables; i.e., when applied as a single-run estimation method, and the estimate is attained quickly. Even without the variance reduction considerations, MVD derivative estimators has almost always less variance than the corresponding SF estimator. Incorporating usage of common random numbers with MVD renders MVD to be a much better estimation method. The reduced computational time attaining the SF derivative estimator does not compensate for the lack of precision, in terms of work-normalized variance, [43].

When the map h is differentiable, IPA is the prudent choice of derivative estimation method. The pathwise methods are precise; it is fast to compute an estimate, as it is a single-run derivative estimation method, and only the behaviour of the parameter of interest, θ, as a function of the random variables needs to be known to attain the pathwise derivative. Attaining the derivative of the map h could be problematic due to complexity, but the derivative only has to be attained once and there are no further assumptions. When common random numbers are used MVD is either comparable to IPA in regard to variance when both methods apply. MVD suffers from computational speed due to the longer programs when computing the MVD derivative estimator and the additional random number generations needed.

When the map h is not differentiable, both MVD and SPA are promising and it becomes a decision on which method to use. MVD is a favourable estimation method when the discrete-event stochastic model has inherently much variance, and less so for the pathwise methods. For SPA, the concerns are which random variables should be conditioned in attaining the conditional expectation, whether the conditional expectation has an analytical expression, and the functional form of the integrand after the pathwise derivative is acquired. The “smoothing" increases the variance of the resulting SPA estimator to some extent. For relatively simple models, SPA requires effort in computing the condi-

tional expectation whereas for MVD the effort is in how to estimate the random variables X^+ and X^-. For relatively complex models, MVD is the sole choice as SPA is not feasibly implementable.

1.2.4 Examples

The purpose of this illustration is twofold. Firstly, this illustration provides an impression of each of the derivative estimation methods via a concrete performance function. Secondly, these attained results for the pathwise and distributional derivatives for each of the distributions presented are employed in applications in the subsequent chapters. In Section 1.2.4.1, the pathwise and distributional derivatives are determined w.r.t. the rate parameter for the exponential distribution. Each of these results are given with regard to a specific performance function, satisfying the first of our aims. In Section 1.2.4.2, the pathwise and distributional derivatives w.r.t. both the mean and standard deviation of the normal distribution are attained. In addition, parameter derivatives w.r.t. a common parameter present in both those parameters are given.

Let $X = (X_1, X_2) \in \mathbb{R}^2$ be a two element random vector and assume that the elements are independent of each other. In Section 1.2.4.1 the examples are conveyed via one of two performance functions. For the IPA, SF, and MVD derivative approaches the example performance function is, suppressing the parameter dependence,

$$h_1(X(\theta)) = X_1 e^{-X_1},$$

and for the SPA and FD methods the performance function is a little more involved, with

$$h_2(X(\theta)) = X_1 e^{-X_1} 1\{X_2 < t\},$$

for $t > 0$. For each of the estimation methods the conditions for the simulation procedures are kept in mind.

1.2.4.1 Exponential Distribution

Let $X_1(\theta) \sim E(\theta)$, $X_2(\theta) \sim E(2\theta)$, be exponentially distributed random variables with rates θ, and 2θ, $\theta > 0$; i.e., the density function $f_{\theta,1}$ for the random variable $X_1(\theta)$ is written as

$$f_{\theta,1}(x_1) = \theta e^{-\theta x_1} 1\{x_1 > 0\}, \tag{1.24}$$

and $f_{\theta,2}$ is given analogously. For all of the examples, the parameter derivative is w.r.t. θ.

IPA: Since θ is a scale parameter for X_1; that is, $X_1(\theta) = E/\theta$ where E is an exponential random variable with rate (and mean) one, the pathwise derivative for the exponential random variable yields

$$\frac{\partial}{\partial\theta}X_1(\theta) = -\frac{E}{\theta^2} = -\frac{X_1(\theta)}{\theta}. \tag{1.25}$$

The derivative of the map h_1 is given by $\partial_{x_1} h_1(x_1) = (1-x_1)\exp(-x_1)$. From Equation (1.13), following the chain rule, the IPA derivative estimator w.r.t. the parameter θ for the example is given below, implying the argument θ for random variable X_1:

$$\frac{d}{d\theta}\mathbb{E}[h_1(X(\theta))] = -\mathbb{E}\left[\frac{X_1}{\theta}(1-X_1)e^{-X_1}\right].$$

The expectation is integrable as $x_1\exp(-x_1) \le e^{-1}$ for $x \ge 0$. The map h is Lipschitz continuous w.r.t θ bounded above by a random variable which has mean $(1+\theta)/\theta^2$.

SPA: To implement the pathwise derivative for the map h_2, the random variable X_1 has to be conditioned since the indicator mapping cannot be differentiated w.r.t. the state variable x_2. The realization of the random variable $1\{X_2 < t\}$ is in $\{0,1\}$ and the θ-dependence is present in the probability of either realization occurring. The distribution function of X_2 is attained from the conditional expectation,

$$\begin{aligned}\mathbb{E}[h_2(X(\theta))] &= \mathbb{E}\left[X_1 e^{-X_1}\,\mathbb{E}[\,1\{X_2 < t\}|X_1]\right] \\ &= \mathbb{E}\left[X_1 e^{-X_1}\left(1-e^{-2\theta t}\right)\right],\end{aligned}$$

and with the derivative $\partial_{x_1} h_1$ already determined, as well as the pathwise derivative of X_1, (1.25), the SPA derivative estimator is the longer expression

$$\frac{d}{d\theta}\mathbb{E}[h_2(X(\theta))] = -\frac{1}{\theta}\left(\left(1-e^{-2\theta t}\right)\mathbb{E}\left[X_1(1-X_1)e^{-X_1}\right] - 2\theta t e^{-2\theta t}\mathbb{E}\left[X_1 e^{-X_1}\right]\right).$$

The expectation is integrable since $h_2 \le h_1$.

FD: The forward estimator, (1.14), will be used for this illustration, portraying the use of common random numbers in the derivative expression. The step-size is denoted by $\Delta > 0$. To emphasise that a single random variable is used, we

denote $X_1(\theta) = E/\theta$ and $X_1(\theta+\Delta) = E/(\theta+\Delta)$, and similarly for X_2. The exponential random variable, E, with rate one is the common random variable for this representation. Then the derivative-free FD estimator w.r.t. θ is written as

$$\frac{d}{d\theta}\mathbb{E}[h_2(X(\theta))] = \mathbb{E}\left[\frac{\frac{E}{\theta+\Delta}e^{-\frac{E}{\theta+\Delta}}\,1\{E<2(\theta+\Delta)t\} - \frac{E}{\theta}e^{-\frac{E}{\theta}}\,1\{E<2\theta t\}}{\Delta}\right].$$

The mean can be computed as $\mathbb{E}[h_2(X)] = \theta(1-\exp(-2\theta t))/(1+\theta^2)$ and so it is twice differentiable w.r.t. θ and is finite, verifying that the FD method can be applied. From the IPA Example, with $x = (x_1, x_2)$, $h_2(x)$ is also Lipschitz continuous w.r.t. θ. Consequently, if we were to use n realizations to ascertain the estimate, the rate of convergence is $O(n^{-1/2})$ and the functional form of the step-size is $\Delta(n) = O(n^{-1/2})$.

SF: Let $v(x_1) = 1 + x_1^p$, $p \in \mathbb{N}$, and depict $\mathfrak{B}_v$ for the set of Borel functions that are bounded by some finite multiple of v, such that the map $v \in \mathfrak{L}^1(f_{\theta,1} : \theta \in \Theta)$. Denoting $X_1 \sim E(\theta)$, $f_{\theta,1}$ is differentiable w.r.t. to the set of functions $\mathfrak{B}_v$. As the map $h_1 \in \mathfrak{B}_v$, Assumption **(SF 3)** is satisfied. In this example, the inserted measure $p = f_{\theta,1}$ has been chosen. This satisfies **(SF 2)**. Assumption **(SF 1)** is satisfied since $f_{\theta,1}$ is differentiable w.r.t. θ.

For the exponentially distributed random variable X_1, and for all Borel, polynomially bounded mappings $h \in \mathfrak{B}_v$, the score function is given by

$$\begin{aligned}\mathrm{SF}_\theta(x) &= -\frac{\frac{\partial}{\partial\theta}f_{\theta,1}(x)}{f_{\theta,1}(x)} \\ &= \frac{1}{\theta} - x. \end{aligned} \tag{1.26}$$

Consequently, our performance function h_1, and all $\mathfrak{B}_v$ differentiable maps, the Score Function estimator w.r.t. parameter θ becomes

$$\frac{d}{d\theta}\mathbb{E}_\theta[h_1(X_1)] = \mathbb{E}_\theta[h_1(X_1)\,\mathrm{SF}_\theta(X_1)] = \mathbb{E}_\theta\left[X_1 e^{-X_1}\left(\frac{1}{\theta} - X_1\right)\right].$$

MVD: From the SF example, let $v(x_1) = 1 + x_1^p$ for $p \in \mathbb{N}$, which is $\mathfrak{L}^1(f_{\theta,1} : \theta \in \Theta)$ integrable. Again $\mathfrak{B}_v$ denotes the set of Borel, polynomially bounded functions including h_1 in which $f_{\theta,1}$ is $\mathfrak{B}_v$-differentiable. A suitable representation, and standard choice, for the derivative of $f_{\theta,1}$ is the triple $(1/\theta, E(\theta), G(2,\theta))$: successively the exponential distribution with rate θ and the gamma distribution with

shape parameter two and scale parameter θ:

$$\begin{aligned}\frac{\partial}{\partial\theta}f_{\theta,1}(x_1) &= (1-\theta x_1)e^{-\theta x_1}\,1\{x_1>0\}\\ &= \frac{1}{\theta}\left(\theta e^{-\theta x_1}\,1\{x_1>0\} - \theta^2 x_1 e^{-\theta x_1}\,1\{x_1>0\}\right) \qquad (1.27)\\ &:= c_\theta\left(f^+_{\theta,1}(x_1) - f^-_{\theta,1}(x_1)\right).\end{aligned}$$

Hence, Assumption **(MVD 1)** is satisfied, and as the random variable $X_1^+ = X_1$, a realization can be obtained for this random variable from the original $E(\theta)$ generation. In addition, the gamma $G(2,\theta)$ distribution is equal in distribution to the sum of two independently generated exponential-θ random variables. Again the realization of X_1 can be used, and together with a independently generated copy of X_1, $Y \sim E(\theta)$, to obtain a gamma random variable, i.e. $X_1^- = X_1 + Y$. Consequently, the MVD derivative estimator w.r.t. the scale parameter for the performance function h_1, it holds that

$$\frac{d}{d\theta}\mathbb{E}_\theta[h_1(X_1)] = \frac{1}{\theta}\mathbb{E}_\theta\left[X_1 e^{-X_1} - (X_1+Y)e^{-(X_1+Y)}\right].$$

1.2.4.2 Normal Distribution

Let $X \sim N(\mu,\sigma^2)$ be a normally distributed random variable with mean μ and standard deviation $\sigma > 0$. The density function $f_{\mu,\sigma}(x)$ is written as

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right), \qquad (1.28)$$

for $x \in \mathbb{R}$. In this section only the IPA, SF, and MVD estimators for the mean and standard deviation are attained insofar as attaining the according parameter derivative of this distribution via each method. We denote the common parameter by θ and so $\mu := \mu(\theta)$, and $\sigma := \sigma(\theta)$, are mappings of θ.

IPA: Let $X(\mu,\sigma)$ denote an arbitrary normal random variable and let $N = X(0,1)$ denote the standard normal random variable with mean zero and standard deviation one. For any parameteriztion of the normal random variable $X(\mu,\sigma)$, this can be written as an affine transformation from N, i.e,

$$X(\mu,\sigma) = \mu + \sigma N,$$

and can consequently be generated from the standard normal random variable. From the above representation, the pathwise derivatives can be easily acquired:

$$\frac{\partial}{\partial\mu}X(\mu,\sigma) = 1 \quad \text{and} \quad \frac{\partial}{\partial\sigma}X(\mu,\sigma) = \frac{X(\mu,\sigma)-\mu}{\sigma}. \qquad (1.29)$$

Assuming that the conditions for the IPA estimator hold, the IPA estimator for both parameters μ and σ is provided in Equation (1.30) below. The map $h'(x)$ depicts the derivative of the performance function w.r.t. the state variable, and imply the parameter dependence for X:

$$\frac{d}{d\mu}\mathbb{E}\big[h(X(\mu,\sigma))\big] = \mathbb{E}\big[h'(X)\big] \quad \text{and} \quad \frac{d}{d\sigma}\mathbb{E}\big[h(X(\mu,\sigma))\big] = \frac{1}{\sigma}\mathbb{E}\big[h'(X)\big(X-\mu\big)\big]. \tag{1.30}$$

If both the mean $\mu := \mu(\theta)$, and standard deviation $\sigma := \sigma(\theta)$ are mappings containing the common parameter θ, then the IPA estimator w.r.t. θ is attained by standard procedure via a combination of the chain rule and both parameter derivatives, yielding

$$\frac{d}{d\theta}\mathbb{E}\big[h(X(\mu,\sigma))\big] = \mathbb{E}\left[h'(X)\left(\frac{\partial\mu}{\partial\theta} + \frac{X-\mu}{\sigma}\frac{\partial\sigma}{\partial\theta}\right)\right]. \tag{1.31}$$

Compared to (1.30), an additional assumption is needed. Equation (1.31) requires that the functions μ and σ are differentiable w.r.t. θ.

SF: To attain the Score Function estimator w.r.t. both the mean and standard deviation, the partial derivatives, respectively, $\partial_\mu f_{\mu,\sigma}$ and $\partial_\sigma f_{\mu,\sigma}$ of the density function need to be computed. For the normal distribution, these are attained following Equation (1.28):

$$\frac{\partial}{\partial\mu}f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}}\frac{(x-\mu)^2}{\sigma^3}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right), \tag{1.32}$$

for $x \in \mathbb{R}$, and

$$\frac{\partial}{\partial\sigma}f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}}\left(\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right)\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right), \quad x \in \mathbb{R}. \tag{1.33}$$

Choosing $p = f_{\mu,\sigma}$ as the inserted equivalent measure, respectively, the score functions yield

$$\begin{aligned}\mathrm{SF}_\mu(x) &= \frac{\frac{\partial}{\partial\mu}f_{\mu,\sigma}(x)}{f_{\mu,\sigma}(x)} \\ &= \frac{(x-\mu)}{\sigma^2}\end{aligned}$$

for $x \in \mathbb{R}$, and

$$\begin{aligned} \mathrm{SF}_\sigma(x) &= \frac{\frac{\partial}{\partial\sigma} f_{\mu,\sigma}(x)}{f_{\mu,\sigma}(x)} \\ &= \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}. \end{aligned}$$

If we choose $v = 1+|x|^p$, $p \in \mathbb{N}$, we observe that $v \in \mathfrak{L}^1(f_{\mu,\theta})$ for all $\mu \in \mathbb{R}$, and separately for all $\sigma > 0$. Therefore $f_{\mu,\sigma}$ is $\mathfrak{B}_v$-differentiable. Hence, the Score Function estimator both to the mean and standard deviation are written according to Equation (1.19).

If a common parameter θ is present in both the differentiable mean and standard deviation mappings; i.e., $\mu := \mu(\theta)$, $\sigma := \sigma(\theta)$, the Score Function estimator w.r.t. θ is also attained via the chain rule and the SF estimators w.r.t. μ and σ

$$\frac{d}{d\theta}\mathbb{E}_\theta[h(X)] = \mathbb{E}_\theta\left[h(X)\left(\mathrm{SF}_\mu(X)\frac{\partial\mu}{\partial\theta} + \mathrm{SF}_\sigma(X)\frac{\partial\sigma}{\partial\theta}\right)\right]. \tag{1.34}$$

No additional assumptions are needed as the derivatives w.r.t. μ, σ, are $\mathfrak{B}_v$-differentiable.

MVD: Alternatively, the parameter derivatives of the normal density function $f_{\mu,\sigma}$ can be rewritten, extracting the measure-valued derivative density functions. For the parameter density w.r.t. the mean, following Equation (1.32), two shifted-Rayleigh density functions are attained:

$$\begin{aligned} \frac{\partial}{\partial\mu} f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}\Bigg(&\frac{(x-\mu)}{\sigma^2}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right)1\{x \geq \mu\} \\ &- \frac{(\mu-x)}{\sigma^2}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right)1\{x < \mu\}\Bigg), \\ := c_\mu&\left(r^+_{\mu,\sigma}(x) - r^-_{\mu,\sigma}(x)\right), \end{aligned} \tag{1.35}$$

and the triple is denoted as $(c_\mu, r^+_{\mu,\sigma}, r^-_{\mu,\sigma})$. These Rayleigh distributions are shifted by an amount μ. The density $r^+_{\mu,\sigma}$ has support on the positive half-line $[\mu,\infty)$ while the Rayleigh density $r^-_{\mu,\sigma}$ has been reflected onto the negative half-line $(-\infty,\mu]$. For a standard Uniform distributed random variable $U \in (0,1)$, the random variable $R^+(\mu,\sigma) \sim r^+_{\mu,\sigma}$ can be generated via the inverse transform method and is given by $R^+(\mu,\sigma) = \mu + \sigma\sqrt{-2\ln(1-U)}$. Similarly, the random variable $R^-_{\mu,\sigma} \sim r^-_{\mu,\sigma}$, can be sampled from $R^-_{\mu,\sigma} = \mu - \sigma\sqrt{-2\ln(1-U)}$, using the same uniform random variable.

With regard to the standard deviation parameter, if the derivative $\partial_\sigma f_{\mu,\sigma}$ were rewritten to attain the measure-valued derivative density functions, a normal distribution function is obtained, see (1.36) below. In addition, the density function $m_{\mu,\sigma}(x)$ for the Double-Maxwell distribution is also attained:

$$\begin{aligned}
&\frac{\partial}{\partial\sigma} f_{\mu,\sigma}(x) \\
&\quad = \frac{1}{\sqrt{2\pi}}\left(\frac{(x-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right)\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right) \\
&\quad = \frac{1}{\sigma}\left(\frac{1}{\sqrt{2\pi}\sigma}\left(\frac{x-\mu}{\sigma}\right)^2\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right) - \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right)\right) \qquad (1.36) \\
&\quad := c_\sigma\left(m_{\mu,\sigma}(x) - f_{\mu,\sigma}(x)\right),
\end{aligned}$$

for $x \in \mathbb{R}$. The measure-value derivative triple is thus depicted as $(c_\sigma, m_{\mu,\sigma}, f_{\mu,\sigma})$. The Double-Maxwell random variable is denoted by $W := M(\mu,\sigma)$ and this random variable can be represented by a standard Double-Maxwell distribution, $M := M(0,1)$, with $\mu = 0$ and $\sigma = 1$, by the same affine transformation $W = \mu + \sigma M$. The standard Double-Maxwell random variable is generated via an acceptance-rejection method and is discussed in [54]. For this measure-value derivative representation, the random variable M and the standard normal random variable N can be related in distribution, increasing the correlation between the performance functions and reducing the variance. Given a standard Double-Maxwell random variable M and a uniform random variable $U \in (0,1)$ independent of M, a standard normal random variable can be generated via the product $N = UM$. In this example, a normal random variable generated by this approach is denoted by X^*. The equality in distribution $1-U = U$ provides two options for the correlation betweens these random variables.

Given $v(x) = 1 + |x|^p$, $p \in \mathbb{N}$, we make the same argument with $v \in \mathfrak{L}^1(f_{\mu,\sigma})$ for all $\mu \in \mathbb{R}$ and separately for all $\sigma > 0$ that normal density function is $\mathfrak{B}_v$-differentiable; i.e., differentiable w.r.t. with the collection of functions that are absolutely bounded by some finite multiple of v. If $h \in \mathfrak{B}_v$, satisfying Assumption **(MVD 1)**, a MVD derivative estimator w.r.t. the mean parameter μ has the form

$$\frac{d}{d\mu}\mathbb{E}_\mu[h(X)] = \frac{1}{\sqrt{2\pi}\sigma}\mathbb{E}_\mu\left[h(R^+) - h(R^-)\right],$$

where $R^\pm$ are the shifted-Rayleigh random variables defined above. Similarly, a MVD estimator w.r.t. the standard deviation parameter is written as

$$\frac{d}{d\sigma}\mathbb{E}_\sigma[h(X)] = \frac{1}{\sigma}\left(\mathbb{E}_\sigma\left[h(X^*)\right] - \mathbb{E}_\sigma[h(W)]\right),$$

where the random variable W is a Double-Maxwell random variable defined above. Alternatively, the original random variable is kept to ascertain the derivative estimator w.r.t. σ by independently sampling a Double-Maxwell random variable. The first approach is especially useful if in advance the parameter derivative w.r.t. σ is needed for the performance function.

Suppose that μ and σ are now mappings of a common parameter θ, i.e., $\mu := \mu(\theta)$ and $\sigma := \sigma(\theta)$. Provided that $\mu(\theta)$ and $\sigma(\theta)$ are differentiable mappings w.r.t. θ, applying the chain rule of differentiation with the above MVD estimators provides a MVD derivative estimator w.r.t. θ. This is depicted by

$$\frac{d}{d\theta}\mathbb{E}_\theta[h(X)] = \frac{d}{d\mu}\mathbb{E}_\mu[h(X)]\frac{\partial\mu}{\partial\theta} + \frac{d}{d\sigma}\mathbb{E}_\sigma[h(X)]\frac{\partial\sigma}{\partial\theta}. \tag{1.37}$$

and random variable generation is conducted separately for each expectation.

1.3 Outline of the Thesis

The thesis addresses two areas of derivative estimation praxis that are presently underdeveloped. The first of these topics is the paucity of results in attaining sensitivities that are not either a probability or moment-based statistic. Hong with other authors, [57], [71], [58], and, [32] provided the first results on attaining sensitivities that are based upon either the order statistic or the quantile of a distribution. They used pathwise methods, either IPA or SPA, for their results. Following from the initial work conducted in [55] and [56], in Chapters 2, 3, and 4 we provide a treatise to derivative estimation of these statistics via distributional methods. For this body of work the synopsis is given in Section 1.3.1.

The second of these topics contributes to the derivative estimation with respect to an extrinsic parameter θ that is present in a boundary of some domain. As stated in Section 1.1.2, while there are certain treatments of ascertaining extrinsic parameters via pathwise and distributional methods in certain instances, there has only been little development analyzing parameters which are in relations encoded within indicator mappings. In option pricing, important parameters in the setting of option contracts such as the exercise price exist within relations of indicator maps. Following work from Lyuu and Teng, [74]; Hong and Liu, [59]; and Wang, Fu, and Marcus, [103]; we investigate extrinsic parameter derivative estimation w.r.t. multi-asset options, barrier options, and Step/Par*asian* and Parisian options. The synopsis is given in Section 1.3.2.

1.3.1 Distributional methods into Quantile-based Sensitivity Estimation

Quantiles and quantile related performance measures are common in modelling quality of service (QoS). Indeed, in the call center industry, QoS is typically measured by the fraction of services meeting a predefined service level, which can be expressed in terms of the fraction of customers that could be helped within a pre-specified time (e.g. 90 % of customers are in contact with an agent in less than ten seconds). In public transportation networks, QoS is measured by the achieved punctuality (e.g. 95 % of trains are delayed by no more than two minutes). In risk analysis, value-at-risk and conditional value-at-risk are defined through quantiles. Finally, note that the 6-σ quality control approach in business management is another example. Here it is the goal to guarantee that 99.9996 % of produced parts are with a pre-specified range of boundary values.

To improve or optimize the quantile related performance of a system, sensitivity analysis of quantiles with respect to changes in the parameters of the underlying model are essential. Let Z have distribution function F. The quantile of Z (respectively F) at a level $\alpha \in (0,1)$, denoted by q_α, is defined as the largest value y such that the probability of attaining a value $y \leq Z$ is less than or equal to α:

$$q_\alpha = \sup\{y : F(y) \leq \alpha\}. \tag{1.38}$$

As we assume in this work that F is continuous, so that Z possesses a probability density function (p.d.f.), denoted by f, the quantile can be written more simply since F is now a bijection:

$$q_\alpha = \sup\{y : F(y) = \alpha\} = F^{-1}(\alpha). \tag{1.39}$$

Since $Z(\theta)$ is a continuous random variable, the corresponding inverse distribution $F_\theta^{-1}(x)$ is differentiable w.r.t. to its argument x, and if F_θ is differentiable w.r.t θ, so is $F_\theta^{-1}(y)$. By definition, see (1.39),

$$\alpha = F_\theta\big(q_\alpha(\theta)\big),$$

and we attain an expression for the quantile sensitivity by differentiating the previous w.r.t θ:

$$0 = \frac{\partial}{\partial\theta} F_\theta\big(q_\alpha(\theta)\big) + f_\theta\big(q_\alpha(\theta)\big)\, \partial_\theta q_\alpha(\theta),$$

or

$$\frac{\partial}{\partial\theta} q_\alpha(\theta) = -\frac{\frac{\partial}{\partial\theta} F_\theta\big(q_\alpha(\theta)\big)}{f_\theta\big(q_\alpha(\theta)\big)}. \tag{1.40}$$

Our investigations into attaining derivative estimators for quantile-based statistics of models proceeds from Equation (1.40). In Chapter 2, we derive two derivative estimators via both MVD and SF to determine the parameter derivative of ranked-data statistics for i.i.d. realizations. This applies to both the order statistic and, if the sample size is sufficiently large, the quantile. We ascertain statistical properties for these derivative estimators: strong consistency, deviation between the mean value of the estimator and the actual quantile sensitivity, and a Central Limit Theorem with the attendant confidence intervals. The results we attain for the distributional methods have analogous properties to the IPA estimator in [57]. With a choice of examples from different contexts, we compare the performance of the three derivative estimation methods. We observe that the IPA estimator is generally the most precise method and, taking computation speed into account, the best derivative method to use. If IPA is inapplicable, MVD is the best distributional approach. As a part of the investigation the consistency result for the IPA estimator in [57] is improved from “in probability" to “almost sure" while clarifying the derivation.

Chapter 3 is a theoretical analysis of the performance of the two MVD ranked-data derivative estimators. We consider the Hahn-Jordan choice of measure valued derivative densities, the effect different correlation behaviours between the original and substituted random variables have on reducing the asymptotic variance, as well as the effect importance sampling has on modifying the variance of these class of estimators. We observe that inducing correlation in the measure-valued derivative random variables means a muted effect in reducing the variance of the ranked-data statistic, and importance sampling, given the implementation, provides no performance improvement.

In Chapter 4 we develop an alternative distributional estimator for the quantile sensitivity based on i.i.d. samples, predicated on the spacing estimator, [86], for the density function of the distribution. The advantage of this estimator is that it arises only from the use of ordered realizations and the knowledge (or estimation) of the parameter derivative of the input distribution. As in Chapter 2 we establish the key theoretical properties of strong consistency, deviation between the mean value of the estimator and the actual quantile sensitivity, and establish confidence intervals. The use of the estimator is illuminated through examples from option pricing, value-at-risk evaluation of a portfolio, and queuing theory.

This work also has four appendices. Appendix A.1 contains two lemmas that mostly underpin the uniform integrability conditions needed for the derivations presented in each of three chapters. Appendix A.2 is a tabulation of results of expectations of functions of order statistics, uniformly distributed on $(0,1)$.

These are needed to determine the extent of asymptotic bias between a derivative estimator for a distributional method and the actual quantile sensitivity in Chapters 2 and 4. Appendix A.3 is the verification of an assumption for two parameter derivatives of the quantile for the cash flow of an option example in Chapter 4 where the price path is modelled via the Variance Gamma process. In Appendix B, we present measure-valued derivative density functions and their corresponding random variable generation for the parameter derivatives of the bivariate normal distribution. This follows an example computing the quantile sensitivity w.r.t. the correlation of the two stocks for the quantile of the cash flow of a spread option employing a two-dimensional Black-Scholes-Merton price process.

1.3.2 Extrinsic Greeks of Options

A financial option is a contract written by a seller that conveys to the buyer the right, but not the obligation, to buy or to sell a particular asset, shares of stock or some other financial instrument, at some maturity time t, or earlier. In return for granting this option, the seller collects a payment (the premium) from the buyer. Option pricing, which is determining the fair premium to be paid for such an option in an arbitrage-free market, along with hedging, is one of the key topics in mathematical finance. For risk-managers, however, equally important is to evaluate the sensitivity of the option premium w.r.t. various parameters such as volatility, interest rate, maturity time or strike price. Parameter derivatives of option premiums are known in the literature as the “Greeks" (they are denoted by Greek letters) and, due to their importance, they have received much attention in the mathematical finance literature.

Apart from a few well known types of options in a Black-Scholes-Merton market, option prices (and, consequently, Greeks) are very rarely obtained in closed form expressions. Since the option price is, up to a multiplicative constant, given by an expectation of a certain random variable, Monte Carlo simulation seems to be the only reasonable way to evaluate options. Numerical methods involving numerical integration and/or algorithmic PDE solving become infeasible if the dimension of the problem is large.

Chapter 5 analyzes the sensitivity of three different types of extrinsic barriers for European-style options using the Black-Scholes-Merton pricing model. Firstly, we derive derivative estimators for multi-stock options in which a payment is received if the combination of the stock prices at maturity time is above the exercise price. The parameter derivative is w.r.t. the exercise level. The estimators are attained via the Leibniz Theorem. Secondly, we determine derivative estimators for the continuously monitored Step options w.r.t. the barrier level

using properties of the quantile process of a Brownian Motion. Thirdly, we ascertain a SPA derivative estimator w.r.t. the barrier level of discretely monitored Parisian options. For the Step and Parisian options we compare our results to the finite difference methods and observe a significant improvement in performance. For the Parisian option sensitivity, we also propose a variance reduction algorithm.