Chapter 6

A Primer for Gradient Estimation

6.1 Motivation

EXAMPLE 6.1. In the mining industry, an investment of θ dollars is used to extract the ore (or "raw material") from the ground. The actual amount of valuable minerals is a random variable X_θ that depends on the mining strategies (excavation, exploration, purification, cut-off grades, etc) as well as on the quality of the soil. For a given geographical area, say an open pit mine for example, X_θ is non decreasing in θ. Following economic theory, the demand curve determines the market price as a function $p(q)$, given that the total quantity supplied is q. Let S be the aggregate supply from all other competitors.

The net cost of the mining is thus $\theta - X_\theta\, p(X_\theta + S)$, which is a random variable (if negative, it means the company has a net profit). Strategic planning for large investments is a decision problem: large investments may yield too much supply with prices lowering below production cost, and investing too little may yield too small a revenue. Suppose that a mining company has a forecast $\hat{S}$ on the quantity supplied by the competitors, then the net cost of mining is given by

$$J(\theta) = \theta - \mathbb{E}[X_\theta\, p(X_\theta + \hat{S})].$$

Finding the investment θ that minimizes the net cost leads to

$$\min_\theta J(\theta).$$

The stochastic version of the gradient search method of Chapter 1 can be implemented via our results in Chapters 2 and 3, using

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n,$$

where Y_n is an estimator of the negative of the derivative of the objective function, that is, we seek

$$\mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}] = \frac{d}{d\theta}\mathbb{E}[g(X_\theta)] - 1, \quad g(x) \overset{\text{def}}{=} x\, p(x + \hat{S}),$$

and we assume that $\hat{S}$ is a random variable independent of x. Notice that the function g is in general a non-linear function of x. Suppose that for a fixed investment amount θ we can simulate the output variables $X_\theta, \hat{S}$. Then the sample average of $g(X_\theta)$ is an unbiased estimator of the profit. However, for the purpose of optimization, we need also to *estimate the gradient* of the profit.

⁂

As we have seen in part I, many problems in stochastic optimization can be stated as non-linear optimization problems, like the above mining example, where measurements of performance functions, constraints and their derivatives may be noisy. In particular, we established the conditions under which numerical gradient-based algorithms can be used to approximate the solution to such problems, if we can build appropriate estimators of the gradients.

While finding appropriate gradient estimators can be done in an ad hoc manner, it is preferable to have a calculus of gradient estimation at hand that allows to build gradient estimates for complex problems in a systematic way. Like in calculus, we will first study gradient estimation for simple problems. To transcend these results to problems that are of relevance in practice, we will identify typical stochastic models encountered in applications and then show how the elementary gradient results are related to these problem settings. The analysis of gradient estimators for more elaborate situations will be given in the subsequent chapters.

To simply the presentation, we consider the one-dimensional case only, that is, we assume $\theta \in \mathbb{R}$ and "derivative estimation" might be a more appropriate title of this part. However, as a gradient is easily obtained through its partial derivatives the restriction to the one-dimensional case comes at no loss of generality and we follow the standard literature in referring to the material presented in this part of the monograph as gradient estimation.

6.2 One Dimensional Distributions

In this section we explain the key issues of gradient estimation by means of simple examples. Let Θ be some non-empty connected subset of $\mathbb{R}$. For $\theta \in \Theta$, let $X(\theta)$ denote some random variable defined on an underlying probability space, and let h be some real-valued measurable mapping such that $\mathbb{E}[h(X(\theta))]$ is defined for any $\theta \in \Theta$. Estimating the derivative of $\mathbb{E}[h(X(\theta))]$ with respect to θ is of key importance in optimizing $\mathbb{E}[h(X(\theta))]$. In other words, we ask how to estimate

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

Since optimization with unbiased derivative information is preferable to biased estimators, we are seeking for a measurable (possibly random) mapping ψ such that

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \mathbb{E}[\psi(h, X(\theta), \theta)]. \tag{6.1}$$

6.2.1 Infinitesimal Perturbation Analysis

The above estimation problem can be dealt with in a straightforward way provided that the random variable $h(X(\theta))$ is differentiable and interchanging expectation and differentiation is justified; for details see Example 6.2 below. Indeed, in this case one obtains

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \mathbb{E}\left[\frac{d}{d\theta}h(X(\theta))\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}X(\theta)\, h'(X(\theta))\right], \tag{6.2}$$

where $h'(x)$ denotes the derivative of $h(x)$ with respect to x. It is worth noting that sample path derivatives are measurable, see Exercise 6.2.

For many distributions that are of importance in applications, the sample path derivative of $X(\theta)$ with respect to θ can be obtained in a simple form, yielding an efficient derivative estimator. This approach of using sample path derivatives as estimator for the derivative of an expected value is called *infinitesimal perturbation analysis* (IPA).

EXAMPLE 6.2. Consider an exponential random variable $X(\theta)$ with mean value $\theta > 0$:

$$\mathbb{P}(X(\theta) \leq x) = 1 - e^{-x/\theta},$$

for $x \geq 0$. The exponential distribution is used for modeling life times of items, think of the life time of a light bulb, or service times in queueing systems. For U uniformly distributed on $[0, 1]$ and independent of everything else, let

$$\tilde{X}(\theta) = -\theta \ln(1 - U), \tag{6.3}$$

then $\tilde{X}(\theta)$ is a version on $X(\theta)$, i.e., $\tilde{X}(\theta)$ and $X(\theta)$ are equal in distribution (refer to Exercise 6.1). Without loss of generality we will in the following identify $\tilde{X}(\theta)$ and $X(\theta)$. Taking derivatives yields

$$\frac{d}{d\theta} X(\theta) = -\ln(1 - U) = \frac{1}{\theta} X(\theta).$$

Inserting this expression for the derivative into (6.2) yields

$$\frac{d}{d\theta} \mathbb{E}[h(X(\theta))] = \mathbb{E}\left[\frac{1}{\theta} X(\theta)\, h'(X(\theta))\right]. \tag{6.4}$$

Letting

$$\psi(h, X(\theta), \theta) = \frac{1}{\theta} X(\theta)\, h'(X(\theta))$$

the expression for the derivative in (6.4) is of the general form (6.1).

Observe that above we have paramterized the exponential distribution via the mean θ. While this a natural parameterization from the point of control theory as θ is a simple scaling of the realizations, in probability theory the exponential distribution is typically parameterized via its rate or inverse scale. More formally, we denote by Exponential (λ) the exponential distribution with rate λ:

$$\mathbb{P}(X(\lambda) \leq x) = 1 - e^{-\lambda x},$$

for $x \geq 0$. Let $X(\theta)$ follow an Exponential (θ) distribution, then the IPA derivative is given by

$$\frac{d}{d\theta} X(\theta) = \frac{d}{d\theta}\left(-\frac{1}{\theta} \ln(1 - U)\right) = -\frac{1}{\theta} X(\theta).$$

Note that the sample path derivatives of the mean and rate parameterization differ only by sign.

⁂

EXAMPLE 6.3. Refer to Example 6.1. Assume that the price function is of the form $p(x) = x^{-1/d}$. To find the optimal investment level θ, one may use

$$\theta_{n+1} = \theta_n - \epsilon\left(1 - \frac{d}{d\theta} \mathbb{E}\left[X(\theta_n)\,(X(\theta_n) + \hat{S})^{-1/d}\right]\right).$$

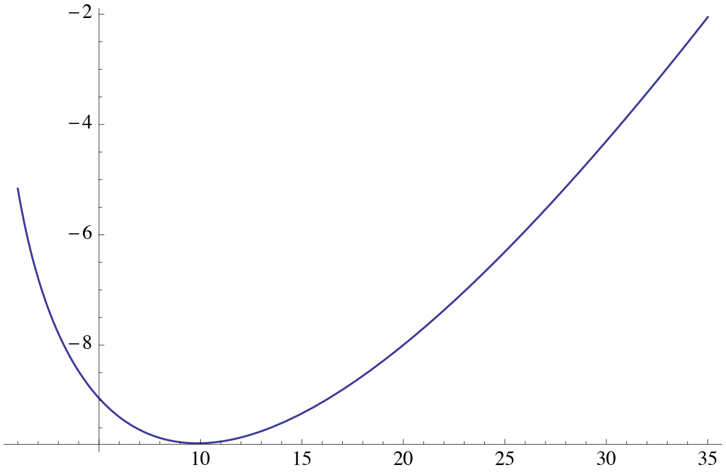


Figure 6.1: Objective function $J(\theta)$

Assume that $X(\theta)$ is exponential with mean 50θ. Then in general, it is necessary to evaluate the expectation numerically, that is, using $d = 2, S = 1$ the integral

$$\mathbb{E}\left[\frac{X(\theta)}{\sqrt{X(\theta)+1}}\right] = 50\theta \int_0^\infty \frac{xe^{-x/(50\theta)}}{\sqrt{x+1}}\,dx$$

is not in closed form. Figure 6.1 shows the form of the function. We obtained this plot using Mathematica, and it required 190 seconds to execute. Clearly the function is convex and has a unique minimum θ^* (around 10).

It is possible to calculate the derivative of this integral expression, but then each step in the recursion will take approximately 200 seconds of execution time, rendering the iterative method very inefficient. We explore now the alternative using the stochastic approximation method and generating consecutive independent samples of $X(\theta_n)$. But, for that we need to use estimation of the derivative of $J(\theta)$. Following Example 6.2,

$$X(\theta) \overset{\mathcal{L}}{=} -50\,\theta\ln(1-U) \implies \frac{d}{d\theta}X(\theta) = \frac{X(\theta)}{\theta},$$

and $h(x) = x/\sqrt{1+x}$, so that

$$h'(X(\theta)) = \frac{1}{(1+X(\theta))^{1/2}} - \frac{X(\theta)}{2(1+X(\theta))^{3/2}}.$$

Putting this together, use

$$\theta_{n+1} = \theta_n - \epsilon\left(1 - \frac{X(\theta_n)}{\theta_n}h'(X(\theta_n))\right).$$

A resulting trajectory of this stochastic approximation is shown in Figure 6.2. It took Mathematica 0.0238 seconds to run 1000 iterations of the algorithm. If more accuracy is desired, one can follow the directives in Chapter 5 to choose appropriate step size sequences to achieve better precision. Comparing the execution times it should be apparent that it is beneficial to use derivative estimation rather than numerical integration in cases like this.

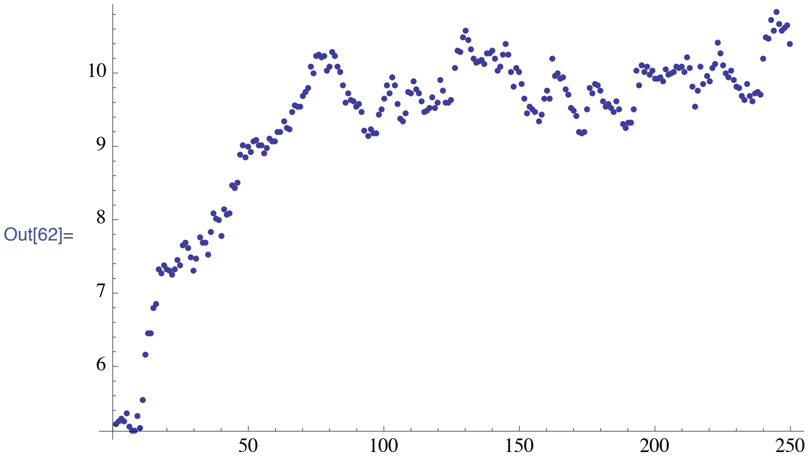


Figure 6.2: A trajectory of the stochastic approximation $\{\theta_n\}$.

Example 6.4. This example provides an overview of IPA derivatives for standard distributions. The derivation of the expressions is left as an exercise to the reader. Let $X(\mu, \sigma)$ have normal distribution with mean μ and standard deviation σ, for $\sigma > 0$. The density is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbb{R}.$$

If Z is a standard normal variable, i.e., $Z \sim$ Normal (0 ,1), then $X(\mu, \sigma) \overset{\text{def}}{=} \mu + \sigma Z$ is a sample of Normal (μ, σ^2), which gives the following IPA derivatives

$$\frac{\partial}{\partial \mu} X(\mu, \sigma) = 1,$$

and

$$\frac{\partial}{\partial \sigma} X(\mu, \sigma) = Z = \frac{X(\mu, \sigma) - \mu}{\sigma}.$$

We denote by Weibull (α,λ) the Weibull distribution with shape parameter α and scale parameter λ and cdf

$$1 - e^{-(\frac{x}{\lambda})^\alpha}, \quad x \geq 0.$$

Samples from the Weibull (α,λ) can be obtained from

$$X_{\alpha,\lambda} = \lambda \left(-\ln(1-U)\right)^{\frac{1}{\alpha}},$$

where U is uniformly distributed on $[0, 1]$. IPA derivative with respect to the scale λ is

$$\frac{\partial}{\partial \lambda} X_{\alpha,\lambda} = \left(-\ln(1-U)\right)^{\frac{1}{\alpha}} = \frac{1}{\lambda} X_{\alpha,\lambda}.$$

Differentiating with respect to α,

$$\frac{\partial}{\partial \alpha} X_{\alpha,\lambda} = \frac{\lambda}{\alpha} \left(-\ln(1-U)\right)^{\frac{1-\alpha}{\alpha}}.$$

Note that the above derivative cannot be written in a straightforward way as a transformation of $X_{\alpha,\lambda}$. Later on we will discuss a method for deriving an IPA derivative, see Lemma 7.1.

Let Pareto-(α,λ) denote the Pareto type II distribution (also known as Lomax distribution) with scale parameter λ and shape parameter α having cdf

$$1-\left(1+\frac{x}{\lambda}\right)^{-\alpha}, \quad x \geq 0.$$

Note that for $\alpha > 1$ the mean value is given by $\lambda/(\alpha-1)$. Samples of the Parteo-(α,λ)-distribution can be obtained from

$$X_{\alpha,\lambda} = \lambda((1-U)^{\alpha}-1).$$

The IPA derivative with respect to the scale parameter is given by

$$\frac{\partial}{\partial\lambda}X_{\alpha,\lambda} = (1-U)^{\alpha}-1 = \frac{1}{\lambda}X_{\alpha,\lambda},$$

whereas the derivative with respect to α leads to

$$\frac{\partial}{\partial\alpha}X_{\alpha,\lambda} = \lambda(1-U)^{\alpha}\ln(1-U).$$

The above shape derivative cannot be written in a straightforward way as a transformation of $X_{\alpha,\lambda}$. Lemma 7.1 in the following chapter will provide a method for obtaining an IPA derivative.

⁂

6.2.2 Score Function

It is not always possible to differentiate sample paths or to interchange expectation and differentiation without harming unbiasedness of the estimator. Examples will be provided later on in the text. In the following we discuss alternative methods for obtaining the derivative of $\mathbb{E}[h(X(\theta))]$. To this end let f_θ denote the Lebesgue density of $X(\theta)$. Rewriting the expected value as integral over f_θ the derivative estimation problem becomes

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \frac{d}{d\theta}\int h(x)f_\theta(x)\,dx.$$

Assuming that f_θ is differentiable with respect to θ and that interchanging differentiation and integration is justified, one obtains

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)\,dx. \tag{6.5}$$

The problem with this manipulation is that $\frac{\partial}{\partial\theta}f_\theta(x)$ fails to be a density and we thus cannot sample from it. To see this, note that for a density f_θ it holds that $\int f_\theta(x)dx = 1$ and taking derivatives yields

$$0 = \frac{d}{d\theta}\int f_\theta(x)dx = \int \frac{\partial}{\partial\theta}f_\theta(x)\,dx.$$

Hence, $\partial f_\theta(x)/\partial\theta$ integrates out to zero and is therefore not a density. Often a density can be introduced by means of a simple analytical transformation. This is done as follows

$$\begin{aligned}\int h(x)\frac{\partial}{\partial\theta}f_\theta(x)\,dx &= \int h(x)\left(\frac{\partial}{\partial\theta}f_\theta(x)\right)\frac{f_\theta(x)}{f_\theta(x)}\,dx \\ &= \int h(x)\left(\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}\right)f_\theta(x)\,dx. \qquad (6.6)\end{aligned}$$

Note that

$$\frac{\partial}{\partial\theta}\log(f_\theta(x)) = \frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(x)}$$

and letting

$$S(\theta,x) = \frac{\partial}{\partial\theta}\log(f_\theta(x)), \qquad (6.7)$$

the derivative in (6.6) can be written in random-variable-language as

$$\int h(x)\frac{\partial}{\partial\theta}f_\theta(x)\,dx = \mathbb{E}[h(X(\theta))S(\theta,X(\theta))]$$

yielding

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \mathbb{E}[h(X(\theta))S(\theta,X(\theta))],$$

or, in the notation of (6.1),

$$\psi(h,X(\theta),\theta) = h(X(\theta),S(\theta,X(\theta))).$$

The mapping $S(\theta,\cdot)$ is called the *score function* and the estimation approach is called the *score function method* (SF).

EXAMPLE 6.5. Revisit the exponential distribution with mean $\theta\in(0,\infty)=\Theta$ in Example 6.2. Note that the distribution of $X(\theta)$ has Lebesgue density $f_\theta(x)=\exp(-x/\theta)/\theta$, for $x\geq 0$ and $\theta>0$. The score function can be computed as

$$S(\theta,x) = \frac{1}{\theta}\left(\frac{x}{\theta}-1\right)$$

and the score function estimator becomes

$$\psi(h,X(\theta),\theta) = h(X(\theta))\frac{1}{\theta}\left(\frac{X(\theta)}{\theta}-1\right). \qquad (6.8)$$

EXAMPLE 6.6. This example provides an overview of the Score Functions for standard distributions. The derivation of the expressions is left as an exercise to the reader. Let Bernoulli (θ) denote the Bernoulli distribution on $\{0,1\}$ assigning probability θ to 1 and $1-\theta$ to 0 with density

$$f_\theta(n) = \theta^n(1-\theta)^{1-n}, \quad n\in\{0,1\}$$

and note that $X(\theta)=\mathbf{1}_{\{U\leq\theta\}}$ yields a Bernoulli (θ) sample, for U uniform on $[0,1]$. Then the Score Function reads

$$S(\theta,n) = \frac{n}{\theta}+\frac{n-1}{1-\theta}, \quad n\in\{0,1\}.$$

For the Poisson (θ) distribution with cdf $\theta^x e^{-\theta}/x!$, for $x \in \mathbb{N}$, the Score Function can be computed to be

$$S(\theta, n) = \frac{x}{\theta} - 1, \quad n \geq 0.$$

For the Normal (μ, σ^2) distribution we obtain for the Score Function with respect to $\theta = \mu$ by

$$S_\mu(\theta, x) = \frac{x - \mu}{\sigma^2}, \quad x \in \mathbb{R},$$

and the Score Function with respect to $\theta = \sigma$ by

$$S_\sigma(\theta, x) = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(x - \mu)^2, \quad x \in \mathbb{R}.$$

For the Weibull (α,β) distribution we obtain as Score Functions for $\theta = \alpha$

$$S_\alpha(\theta, x) = \alpha + \left(1 + \frac{x}{\lambda}\right) \ln\left(\frac{x}{\lambda}\right)$$

and for $\theta = \beta$

$$S_\lambda(\theta, x) = \frac{\alpha^2}{\lambda^2}\left(\left(\frac{x}{\lambda}\right)^\alpha - 1\right), \quad x \in \mathbb{R}.$$

Consider the Gamma (α, β) with shape parameter α and rate parameter β the pdf is given by

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\beta x}, x > 0,$$

for $\alpha, \beta > 0$, where $\Gamma(\alpha)$ denotes the gamma function. For $\alpha \in \mathbb{N}$, Gamma (α, β) is the distribution of the sum of α independent exponential variables with rate β. For $\theta = \alpha$ we obtain

$$S_\alpha(\theta, x) = \ln(\beta x) - \psi(\alpha), \quad x \in \mathbb{R},$$

where $\psi(\alpha)$ denotes the digamma function, and for $\theta = \beta$ it holds that

$$S_\beta(\theta, x) = \frac{\alpha}{\beta} - x, \quad x \in \mathbb{R}.$$

⁂

6.2.3 Measured Valued Differentiation

The score function was introduced to deal with the fact that $\frac{\partial}{\partial \theta} f_\theta(x)$ in (6.5) fails to be a density. An alternative way of dealing with this problem stems from measure theory. There it is shown that under quite general conditions it is possible to write $\frac{\partial}{\partial \theta} f_\theta(x)$ as re-scaled difference between two densities. To see this, let

$$c_\theta = \int \max\left(\frac{\partial}{\partial \theta} f_\theta(x), 0\right) dx = \int \max\left(-\frac{\partial}{\partial \theta} f_\theta(x), 0\right) dx, \tag{6.9}$$

and introduce new densities

$$f_\theta^+(x) = \frac{1}{c_\theta} \max\left(\frac{\partial}{\partial \theta} f_\theta(x), 0\right) \tag{6.10}$$

and

$$f_\theta^-(x) = \frac{1}{c_\theta} \max\left(-\frac{\partial}{\partial \theta} f_\theta(x), 0\right). \tag{6.11}$$

Inserting these densities into the right-hand side of (7.18) yields

$$\frac{d}{d\theta}\int h(x) f_\theta(x) dx = c_\theta \left(\int h(x) f_\theta^+(x)\, dx - \int h(x) f_\theta^-(x)\, dx \right), \tag{6.12}$$

for any h such that interchanging differentiation and integration on the left-hand side of (7.19) is justified. The above approach for obtaining the derivative is called *measure-valued differentiation* (MVD). As we will see later on, MVD applies under more general conditions.

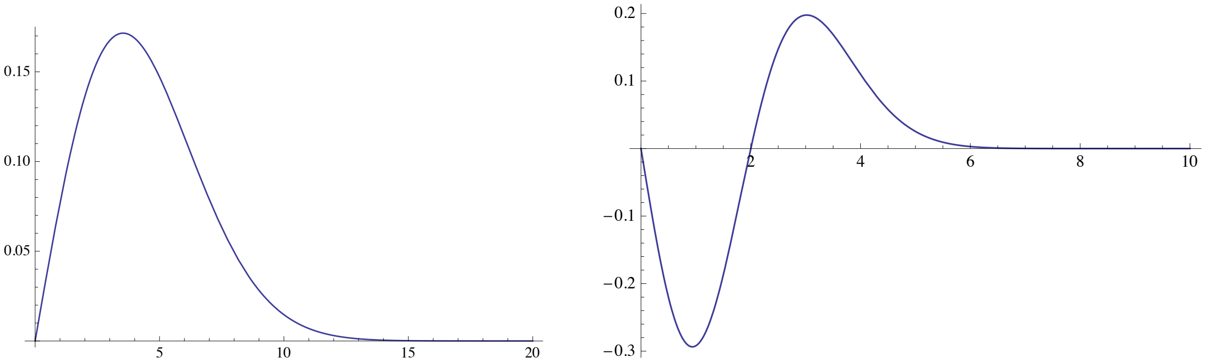


Figure 6.3: Left: Weibull$(2, \theta)$ density f_θ. Right: $\partial f_\theta / \partial\theta$. Plots show the functions at $\theta = 5$.

Figure 6.3 shows a plot of f_θ and the corresponding $\partial f_\theta/\partial\theta$ for the Weibull$(2,\theta)$ distribution. The constant c_θ corresponds to the area under the positive part of the plot to the right, which equals the area under the negative part, because a density f_θ integrates to 1 for every value of θ. f_θ^+ (f_θ^-) is built as the normalized positive (negative) parts of this derivative and thus they are true densities. Observe that this construction implies that the supports of the two densities are disjoint.

Let $X(\theta) \sim f_\theta$ and $X^{\pm}(\theta) \sim f_\theta^{\pm}$. Then in the notation of (6.1) the MVD estimator is the measurable (random) function

$$\psi(h, X(\theta), \theta) = c_\theta(h(X^+(\theta)) - h(X^-(\theta))).$$

Note that ψ is a random mapping. Compare with the IPA and SF examples where ψ is a deterministic mapping.

EXAMPLE 6.7. In the case of the exponential distribution with mean θ, see Example 6.2 and Example 6.5, the θ-derivative of the density can be written as follows

$$\frac{\partial}{\partial\theta} f_\theta(x) = \frac{1}{\theta^3}(x-\theta)e^{-\frac{x}{\theta}} = \frac{1}{\theta}\left(\frac{x}{\theta^2}e^{-\frac{x}{\theta}} - \frac{1}{\theta}e^{-\frac{x}{\theta}}\right). \tag{6.13}$$

Noting that

$$f_\theta^e(x) = \frac{x}{\theta^2}e^{-\frac{x}{\theta}}$$

is the density of the distribution of the sum of two independent exponential random variables with mean θ, known as Gamma-(2,1/θ)-distribution, the derivative expression becomes

$$\frac{d}{d\theta}\int h(x) f_\theta(x)\, dx = \int h(x) \frac{\partial}{\partial\theta} f_\theta(x)\, dx = \frac{1}{\theta}\left(\int h(x) f_\theta^e(x)\, dx - \int h(x) f_\theta(x)\, dx\right). \tag{6.14}$$

Letting $Y(\theta)$ be an exponential random variable with mean θ and independent of $X(\theta)$, the sum $Y(\theta) + X(\theta)$ is Gamma-(2,1/θ)-distributed and the estimator reads in random variable language

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \frac{1}{\theta}\mathbb{E}[h(X(\theta) + Y(\theta)) - h(X(\theta))].$$

From (6.13) the Hahn-Jordan decomposition , i.e., seperating the positive and negative part, would result into the densities

$$\begin{aligned} f_\theta^+(x) &= \frac{(x-\theta)e^{1-x/\theta}}{\theta^2} & x > \theta, \\ f_\theta^-(x) &= \frac{(\theta-x)e^{1-x/\theta}}{\theta^2} & x \le \theta \end{aligned}$$

with $c_\theta = e^{-1}/\theta$, which do not belong to an identifiable family of distributions. Although the MVD is also unbiased using $X^\pm(\theta) \sim f_\theta^\pm$, in this case we fid the problem that generating random variables with these distributions is not efficient (inverse function method must be evaluated numerically). This example also shows that in practice, one seeks a representation of the "plus" and "minus" measures that is easy to implement.

⁂

An overview on MVD representations of standard distributions, such as provided in Example 6.4 for IPA and in Example 6.6 for the Score Functions, is postponed to the next chapter.

We have presented a brief motivation of the main techniques for finding unbiased gradient estimators. We summarize the findings for the exponential distribution in the following example.

Example 6.8. For the case of $X(\theta)$ following an exponential distribution with mean θ we have derived the following IPA, SF and MVD estimator for $d\mathbb{E}[h(X(\theta))]/\theta$:

$$\text{IPA} \qquad \frac{1}{\theta}\mathbb{E}\big[X(\theta)h'(X(\theta))\big] = \frac{d}{d\theta}\mathbb{E}\big[h(X(\theta))\big] = \begin{cases} \frac{1}{\theta}\mathbb{E}\big[h(X(\theta))(X(\theta)/\theta - 1)\theta\big] & \text{SF} \\ \\ \frac{1}{\theta}\mathbb{E}\big[h(X(\theta) + Y(\theta)) - h(X(\theta))\big] & \text{MVD,} \end{cases} \tag{6.15}$$

with $X(\theta)$ and $Y(\theta)$ are i.i.d. Note that the estimator on the left-hand side of (6.15) is based on sample-path analysis whereas the estimators on the right-hand side of (6.15) are based on a distributional analysis. The estimators differ in functional form, complexity and variance.

⁂

6.3 A Taxonomy of Gradient Estimation

In this section we introduce the basic gradient estimation problems. The classification is based on the time horizon of the stochastic experiment, which can be either static (i.e., finite and deterministic) or random (and almost surely finite). Moreover, gradients of steady-state characteristics are of interest, and, for Markov processes, gradients of stationary characteristics. In the following we will introduce these estimation problems and we will introduce key examples that will serve as benchmark problems for gradient estimation.

6.3.1 The Static Problem

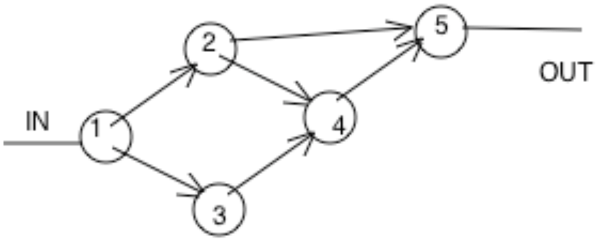
In a *static model* the expected value is evaluated over a finite (and non-random) collection of random variables. This gives rise to the following definition.

Definition 6.1 *For $\theta \in \Theta$, let $X_i(\theta)$, $1 \leq i \leq N$, be a collection of random variables defined on some underlying probability space, such that $X_i(\theta)$ is a measurable mapping on some measurable space $(S, \mathcal{S})$. For a measurable mapping $L : S^N \to \mathbb{R}$, the static gradient estimation problem is to find an unbiased estimator for*

$$\frac{d}{d\theta}\mathbb{E}[L_N(X_1(\theta), \ldots, X_N(\theta))],$$

provided the expression exists.

Example 6.9. [Reliability] Consider the reliability network of the figure below. The system has 5 components whose logical interconnections are depicted in terms of the arcs joining the nodes. Individual components are either working (on) or not (down). For the system to operate, there must be a path between "IN" and "OUT" with components that are all working.



The *life* L of the system is defined as the amount of time that it operates from the moment that all components are new. Each component $i = 1, \ldots, 5$ works for a random time $T_i > 0$ with distribution G_i, after which it brakes down. Suppose that the component lifetimes $\{T_i\}$ are mutually independent with finite expected values. Clearly, $L = T_{i^*}$, where i^* is the index of the last component to brake down that causes the system to brake down as well, and it is a random index.

There are three possible paths: $R_1 = \{1, 2, 5\}$, $R_2 = \{1, 2, 4, 5\}$ and $R_3 = \{1, 3, 4, 5\}$. Route i fails at time $L_i = \min(T_j, j \in R_i)$, therefore the life of the system is of the form $L = \max(L_1, L_2, L_3)$. Let I_i be the index of the component that causes the failure of route R_i, that is, $I_i = \arg\min(T_j, j \in R_i)$. Then we can also represent L by the expression $L_3 = \max(T_{I_1}, T_{I_2}, T_{I_3})$, or $L = T_{i^*}$, where $i^* = \arg\max(T_{I_1}, T_{I_2}, T_{I_3})$.

Suppose that $T_3 \sim$ Exponential $(1/\theta)$, so that $\theta = \mathbb{E}[T_3]$ is the parameter of interest. In order to determine the sensitivity of L_3 with respect to θ, or to find the value of θ that minimizes $\mathbb{E}[L_3]$ the derivative of $\mathbb{E}[L_3]$ with respect to θ has to be determined, i.e.,

$$\frac{d}{d\theta}\mathbb{E}\big[\max(T_{I_1}, T_{I_2}, T_{I_3})\big].$$

⁂

Example 6.10. [Sojourn Times (static)] Customers arrive at a service station according to a renewal point process. The inter-arrival times $\{A_n : n \in \mathbb{N}\}$ are iid, with $\mathbb{E}[A_n] < \infty$ and $\mathbb{P}(A_n = 0) = 0$. Customers are served in order of arrival, and consecutive service times are iid random variables

$\{S_n(\theta) : n \in \mathbb{N}\}$. Interarrival times and service times are assumed to be mutually independent. The common distribution of the service times G_θ depends on the parameter $\theta = \mathbb{E}[S_n(\theta)]$. This is what is known as the GI/GI/1 queueing model with FCFS ("first-come first-served") service discipline. See Example 4.6 for an earlier treatment of this model.

Consider the process of consecutive sojourn (or system) times $\{X_n(\theta)\}$, denoting the total time that the corresponding customer is in the system (from arrival to end of service). The arrival process starts at $T_0 = 0$, and the time of arrival of customer n is $T_n = \sum_{i=1}^n A_i$. Denote the departure time of the n-th customer by $D_n(\theta)$ and let

$$X_n(\theta) = D_n(\theta) - T_n, \quad n \geq 1. \tag{6.16}$$

Then $X_n(\theta)$ denotes the total time that the n-th customer is in the system (from arrival to end of service), also called *sojourn (or system) time*. As we will explain in the following, the process of consecutive sojourn times $\{X_n(\theta)\}$ forms a Markov chain. To see this, note that if the n-th customer leaves the system prior to the $(n + 1)$-st arrival, i.e., $D_n(\theta) - T_{n+1} \leq 0$, then customer $n + 1$ has no wait and enters service immediately in which case the sojourn is equal to the customer's service time:

$$X_{n+1}(\theta) = S_{n+1}(\theta) \, .$$

If, on the other hand, the $(n+1)$-st arrival takes place when the n-th customer is still at the server, i.e., $D_n(\theta) - T_{n+1} > 0$, then customer $n + 1$ has to wait until the previous departs, that is the total wait is

$$D_n(\theta) - T_{n+1} = D_n(\theta) - T_n - A_{n+1} \overset{(6.16)}{=} X_n(\theta) - A_{n+1}$$

before service can commence and the sojourn time becomes

$$X_{n+1}(\theta) = S_{n+1}(\theta) + X_n(\theta) - A_{n+1}.$$

To summarize, consecutive sojourn times follow the recursive relation:

$$X_{n+1}(\theta) = \max(0, X_n(\theta) - A_{n+1}) + S_{n+1}(\theta), \; n \geq 0, \tag{6.17}$$

where we assume that the system starts empty and formally set $X_0(\theta) = 0$. The above recursive relation is called *Lindley* recursion and shows that $\{X_n(\theta)\}$ is a Markov chain.

Let $W_n(\theta)$ denote the waiting time of the n-th customer (the time the customer spends in the system until start of service), then we obtain by (7.23) a similar recursion for the waiting times

$$W_{n+1}(\theta) = \max(0, W_n(\theta) + S_n(\theta) - A_{n+1}), \tag{6.18}$$

for $n \geq 0$, and $W_0(\theta) = S_0(\theta) = 0$.

The sojourn process $\{X_n(\theta)\}$ is adapted to the filtration $\{\mathfrak{F}_n\}$, where $\mathfrak{F}_n$ is the σ-algebra generated by $\{A_1, \ldots, A_n; S_1(\theta), \ldots S_n(\theta)\}$. Suppose that we are interested in evaluating the derivative of the average sojourn times of the first N customers, that is,

$$L_N(X_1(\theta), \ldots, X_N(\theta)) = \frac{1}{N} \sum_{n=1}^{N} X_n(\theta). \tag{6.19}$$

This is a static gradient estimation problem.

A typical application of the above static gradient in an optimization problem is the following. Let $c(\theta)$ denote the cost/energy needed for operating the server at speed θ. Then a stochastic approximation can be applied to minimize

$$J(\theta) = \alpha_1 c(\theta) + \frac{\alpha_2}{N} \sum_{n=1}^{N} \mathbb{E}[X_n(\theta)]$$

the weighted sum of cost and average sojourn time, for weights $\alpha_1, \alpha_2 > 0$. Note that $J(\theta)$ has no closed-form solution and, with the exception of rather small values of N such as $1, 2$ or 3, $J(\theta)$ cannot be solved numerically. Thus, simulation is the only available approach to solving $\min_\theta J(\theta)$. In order to implement stochastic approximation it is required to use an estimator of the derivative of $J(\theta)$. More complex problems involve networks of queues, each with their own service parameter, and in those cases the gradient may have large dimension.

⁂

6.3.2 The Random Horizon Problem

In the static setting the performance L is evaluated over a fixed finite number of observations. An extension of this setup is to let L depend on a finite but random number of observations. The precise definition is given in the following.

Definition 6.2 *For $\theta \in \Theta$, let $\{X_i(\theta) : i \in \mathbb{N}\}$ be stochastic process defined on some underlying probability space, such that $X_i(\theta)$ is a measurable mapping on some measurable space $(S, \mathcal{S})$. Moreover, let τ_θ be a stopping time adapted to the natural filtration of $\{X_i(\theta) : i \in \mathbb{N}\}$. For $n \in \mathbb{N}$, let $L_n : S^n \to \mathbb{R}$. The random horizon gradient estimation problem is to find an unbiased estimator for*

$$\frac{d}{d\theta}\mathbb{E}[L_{\tau_{\tau_\theta}}(X_1(\theta), \ldots, X_{\tau_\theta}(\theta))],$$

provided the expected values exists.

EXAMPLE 6.11. [Sojourn Times (cycle)] We illustrate the random horizon gradient estimation problem with a classical example from queuing theory, using the sojourn time model in Example 6.10. Since service times and interarrival times are assumed to be mutually independent sequences of independent random variables it holds that whenever the system empties, the sequence of sojourn times starts anew independent of the past. Suppose that the queue starts initially empty, which implies that the first arriving customer does not have to wait and the first sojourn time equals the service time, i.e., $X_1(\theta) = S_1(\theta)$. Let τ_θ be the first time instance after $n = 1$ such that $X_n(\theta) = S_n(\theta)$, i.e.,

$$\tau_\theta = \inf\{n > 1 : X_n(\theta) = S_n(\theta)\},$$

or, equivalently,

$$\tau_\theta = \inf\{n > 1 : W_n(\theta) = 0\}, \tag{6.20}$$

and we set $\tau_\theta = \infty$ if either set on the above right hand sides is empty. In words, τ_θ is the first time an arriving customer (other than the first one) experiences no waiting time. Then $(X_1(\theta), \ldots, X_{\tau_\theta - 1}(\theta))$ is called a *cycle* of the sojourn time sequence, and the accumulated sum of sojourn times over a cycle is given by

$$\mathbb{E}\Big[L_{\tau_\theta - 1}(X_1(\theta), \ldots, X_{\tau_\theta - 1}(\theta))\Big]$$

with

$$L_N(X_1(\theta), \dots, X_N(\theta)) = \sum_{n=1}^{N} X_n(\theta).$$

Note that in the definition of a cycle we only add the first $\tau_\theta - 1$ sojourn times. The reason for this is that the index of the last sojourn before the end of a cycle is not a stopping time, whereas the first sojourn time of a cycle is one. The random horizon problem is that of estimating

$$\frac{d}{d\theta}\mathbb{E}\left[\sum_{n=0}^{\tau_\theta - 1} X_n(\theta)\right] = \frac{d}{d\theta}\mathbb{E}\Big[L_{\tau_\theta - 1}(X_1(\theta), \dots, X_{\tau_\theta - 1}(\theta))\Big].$$

⁂

The random horizon problem is related to the static problem. Suppose that

$$\psi(N; X_1(\theta), \dots, X_N(\theta)),$$

for $N \geq 1$, solves the static gradient estimation problem, i.e.,

$$\frac{d}{d\theta}\mathbb{E}[L_N(X_1(\theta), \dots, X_N(\theta)] = \mathbb{E}[\psi(N; X_1(\theta), \dots, X_n(\theta))],$$

for all N, then, as we will show in Section 8.2.1 later on in the next text, under appropriate smoothness conditions the following, somewhat surprisingly simple, equation holds

$$\frac{d}{d\theta}\mathbb{E}[L_{\tau_\theta}(X_1(\theta), \dots, X_{\tau(\theta)}(\theta))] = \mathbb{E}[\psi(\tau(\theta); X_1(\theta), \dots, X_{\tau(\theta)}(\theta))].$$

Hence, evaluating the static gradient estimator over the random horizon provides an unbiased gradient estimator.

6.3.3 The Steady-State Problem

Studying the long term behavior of a system, leads to the following gradient estimation problem.

Definition 6.3 *For $\theta \in \Theta$, let $\{X_i(\theta) : i \in \mathbb{N}\}$ be stochastic process defined on some underlying probability space, such that $X_i(\theta)$ is a measurable mapping on some measurable space $(S, \mathcal{S})$. Let $L : S \to \mathbb{R}$, the steady-state gradient estimation problem is to find an unbiased estimator for*

$$\frac{d}{d\theta}\left(\lim_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} L(X_n(\theta))\right]\right),$$

provided the expression exists.

There is a close relationship between the steady-state costs and the random horizon cost. To see this, suppose that $X_n(\theta)$ is regenerative processes with associated sequence of renewal times $\{\eta_n(\theta)\}$, see Section B.7 in the Appendix for details. From renewal theory it follows that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} L(X_n(\theta)) = \frac{\mathbb{E}\left[\sum_{n=\eta_k(\theta)}^{\eta_{k+1}(\theta)-1} L(X_n(\theta))\right]}{\mathbb{E}[\eta_2(\theta) - \eta_1(\theta)]},$$

w.p.1, for any $k \geq 1$, provided that the expected values are finite. When $\{X_n(\theta)\}$ is ergodic, it holds that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} L(X_n(\theta)) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[L(X_n(\theta))]$$

with probability one, and for the steady-state problem we have

$$\begin{aligned} \frac{d}{d\theta} L_\infty(\theta) &= \frac{d}{d\theta} \left(\frac{\mathbb{E}\left[\sum_{n=\eta_k(\theta)}^{\eta_{k+1}(\theta)-1} L(X_n(\theta))\right]}{\mathbb{E}[\eta_2(\theta)-\eta_1(\theta)]} \right) \\ &= \frac{\frac{d}{d\theta}\mathbb{E}\left[\sum_{n=\eta_k(\theta)}^{\eta_{k+1}(\theta)-1} L(X_n(\theta))\right]}{\mathbb{E}[\eta_2(\theta)-\eta_1(\theta)]} + L_\infty(\theta) \frac{\frac{d}{d\theta}\mathbb{E}\left[\eta_2(\theta)-\eta_1(\theta)\right]}{\mathbb{E}[\eta_2(\theta)-\eta_1(\theta)]}, \end{aligned} \tag{6.21}$$

for any $k \geq 1$.

A more detailed discussion will be provided in the following section. It is worth noting that the expression on the above right hand side is very challenging for simulation, as it contains fractions of expected values. Indeed, estimators for fractions of this type are studied in the simulation literature and, generally speaking, only asymptotic unbiasedness can be established (as the number of observed cycles tends to infinity). The same holds true for estimating the corresponding confidence intervals. These drawbacks may render simulations based on the above right hand side impractical.

6.3.4 Markov Processes: The Stationary Problem

Markov processes are a common tool for modeling and analyzing complex systems. Under appropriate conditions, a Markov process has a unique stationary distribution which is also the unique limiting distribution. In this case the steady-state gradient estimation problem is related to the stationary gradient estimation problem to be introduced in the following.

Consider a family of general homogeneous Markov process $\mathbf{X}(\theta) = \{X_n(\theta)\}$, where for each $\theta \in \Theta$, $X_n(\theta) \in S$, and the measurable state space $(S, \mathcal{S})$ is a general space (not necessarily discrete or countable). Let $P_{\theta,n}$ denote the Markov kernel of $\mathbf{X}(\theta)$, which is given by

$$P_{\theta,n}(s, A) = \mathbb{P}\big(X_{n+1}(\theta) \in A \,|\, X_n(\theta) = s\big), \quad n \in \mathbb{N},$$

for $s \in S$ and $A \in \mathcal{S}$. The Markov process, respectively, the Markov kernel is called *homogenous* if $P_{\theta,n}$ is independent of n.

Definition 6.4 *Let P_θ be a homogenous Markov kernel and let π_θ be a probability distribution on the state space $(S, \mathcal{S})$ such that*

$$\forall A \in \mathcal{S}: \quad \pi_\theta(A) = \int_S P_\theta(s, A)\pi_\theta(s)ds,$$

then π_θ is called a stationary distribution *for P_θ.*

Generally speaking, a Markov process is called *ergodic* if the stationary distribution is the unique limiting distribution. This implies that for an ergodic Markov process the distribution of $X_n(\theta)$ converges independently of the initial distribution to π_θ. Hence, for ergodic Markov processes, the steady-state gradient estimation problem can alternatively be phrased in terms of the stationary distribution.

Definition 6.5 *For $\theta \in \Theta$, let $\mathbf{X}(\theta) = \{X_i(\theta) : i \in \mathbb{N}\}$ be a homogenous Markov process with unique stationary distribution π_θ. Let L a real-valued mapping such that $\int_S |L(s)|\pi_\theta(s)ds$ is finite for $\theta \in \Theta$. The stationary gradient estimation problem is to find an unbiased estimator for*

$$\frac{d}{d\theta}\mathbb{E}\left[L(\tilde{X}(\theta))\right],$$

where $\tilde{X}(\theta)$ is distributed according to π_θ.

EXAMPLE 6.12. Consider the single-server queue put forward in Example 6.10. Provided that service times and interarrival times are i.i.d. and mutually independent, it is well known that the *stability condition* $\mathbb{E}[S_1(\theta)] < \mathbb{E}[A_1(\theta)]$ implies that there exists a stationary waiting time. More specifically, there exists a random variable $W(\theta)$ such that $W(\theta)$ and $\max(0, W(\theta) + S_n(\theta) - A_{n+1})$ are equal in distribution, for $n \geq 1$, see (6.18). This result holds more general for max-plus linear queueing systems, see [4, 20]. In the case of an $M/M/1/\infty$ queue with mean interarrival time $1/\lambda$ and mean service time θ it holds that :

$$\mathbb{P}(W(\theta) \leq x) = 1 - \lambda\theta e^{-((1/\theta)-\lambda)x}, \quad x > 0,$$

and $\mathbb{P}(W(\theta) = 0) = 1 - \lambda\theta$. Suppose that operating the server with mean service time θ consumes energy $c(\theta)$. Then solving

$$\min_\theta \bigg(\alpha_1\mathbb{E}[W(\theta)] + \alpha_2 c(\theta)\bigg),$$

with $\alpha_i > 0$, $i = 1, 2$, will find the speed of the server that balances the wait of a job in equilibrium and the energy consumption of the server in an optimal way.

Under ergodicty of $\{X_n(\theta)\}$ the steady-state performance equals the expected stationary performance. As explained in Example 6.11, the sequence of $\{X_n(\theta)\}$ starts independent anew whenever $X_n(\theta) = S_n(\theta)$. Hence, $\{X_n(\theta)\}$ is a regenerative process. Call $\{\eta_k(\theta)\}$ the corresponding renewal times. If the system starts empty, it holds that $\eta_1(\theta) = 1$ and the next regeneration epoch occurs at time $\eta_2(\theta) = \tau_\theta$, where we use the notation introduced in Example 6.20. Following (6.21), we obtain

$$\frac{d}{d\theta}\lim_{N\to\infty}\sum_{n=1}^{N} L(X_n(\theta)) = \frac{d}{d\theta}\mathbb{E}\left[L(\tilde{X}(\theta))\right] = \frac{\frac{d}{d\theta}\mathbb{E}\left[\sum_{n=1}^{\tau_\theta - 1}(X_n(\theta))\right]}{\mathbb{E}[\tau_\theta - 1]} + \mathbb{E}\left[L(\tilde{X}(\theta))\right]\frac{\frac{d}{d\theta}\mathbb{E}\left[\tau_\theta\right]}{\mathbb{E}[\tau_\theta - 1]},$$

with probability one, where existence of the derivative and finiteness of expected values is assumed. ⁂

The above extends to the more general case of Harris recurrent Markov chains. As already detailed in Section 6.3.3, the expression on the above right hand side does not lead to unbiased gradient estimation and may not be efficient. To overcome this drawback, approaches to directly differentiating the stationary distribution have been developed, which will be discussed in the next chapter.

6.4 Exercises

EXERCISE 6.1. Consider the distribution function $F_\theta(x)$ of some real-valued random variable $X(\theta)$ with parameter θ. The image of the distribution is by definition $[0, 1]$. Suppose that the random variable has a continuous density so that $F_\theta(\cdot)$ is strictly increasing. Let U be a uniform-$[0,1]$-random variable. Show that $X(\theta, U) = F_\theta^{-1}(U)$ is a random variable with distribution F_θ on $(\Omega, \mathbb{P})$. This is sometimes called the canonical representation and it is a means for generating general distributions from a uniform (pseudo) random number generator.

EXERCISE 6.2. For Θ a non-empty connected subset of $\mathbb{R}$, let $X(\theta)$, for $\theta \in \Theta$, be a real-valued random variable and let h be a mapping from $\mathbb{R}$ to $\mathbb{R}$ such that $h(X(\theta))$ is measurable for all $\theta \in \Theta$. Show that if h is differentiable and if $X(\theta)$ is differentiable with respect to θ at some point $\theta_0 \in \Theta$, then $dh(X(\theta))/d\theta$ at θ_0 is measurable with respect to $\mathcal{F}$. (Recall that $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the underlying probability space.)

EXERCISE 6.3. Compute the Score Function for Pareto II distribution with respect to the scale and shape parameter.

EXERCISE 6.4. Consider a random variable $X(\theta)$. Assume that $\theta = \mathbb{E}[X(\theta)]$ is a scale parameter of the distribution F_θ; in other words, using the inverse function representation, $X(\theta) = F_\theta^{-1}(U) = \theta F_1^{-1}(U)$. Explain which of the following is a sufficient condition for:

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

(a) $\mathbb{E}[|X(1)|] < \infty$.

(b) $\mathbb{E}[X(\theta)^2] < \infty$.

(c) $\mathbb{E}[|\theta X(\theta)|] < \infty$.

EXERCISE 6.5. Repeat the stochastic approximation method for the mining investment problem of Example 6.3 using IPA, SF and MVD for derivative estimation.

(a) Using simulation, estimate the corresponding confidence intervals and CPU times for the three derivative estimation methods and compare.

(b) Apply an appropriate theorem from Part I to establish convergence of the stochastic approximation to the true optimal value θ^*. Specify your choice of step size sequence.

(c) Run the stochastic approximations and discuss.

Chapter 7

Gradient Estimation for the Static Problem

In Part I we discussed utilization of finite differences to estimate required gradients for the feedback of stochastic approximation procedures. It is clear from the results in Chapter 5 that such methods suffer from two sources: they require two evaluations (or simulations) at different values of θ (so they take more CPU time per iteration) and they introduce a bias that slows down the convergence rate. It is for these reasons that we study *unbiased* gradient estimation.

Chapter 6 introduced the formulas for the three broad approaches to gradient estimation, and a taxonomy of problems under study. The reader is referred to Examples 6.6 and Example 6.4 for a summary of known results that can become handy when solving problems.

In this chapter we present the main technical tools that help establish the conditions under which such gradient estimation methods yield unbiased estimators. We will discuss IPA, SF and MVD. Moreover, we will present an extension of IPA called smoothed perturbation analysis (SPA).

7.1 Perturbation Analysis: IPA and SPA

The strength of IPA methodology is its ease of use for the static problem. This section presents the mathematical methodology and proofs for unbiasedness of IPA. As well, we discuss the conditions for unbiasedness of IPA and present examples for verification.

7.1.1 Basic Results and Techniques

In this section we present the main technical tools for establishing unbiasedness of the IPA estimator. Lipschitz continuity is a key condition for unbiasedness of the IPA estimator.

Definition 7.1 *Let $\Theta \subset \mathbb{R}$ be an open connected set, such that $X(\theta)$, for $\theta \in \Theta$, is a real-valued random variable defined on a common underlying probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Let θ be an interior point of Θ. We say that $h(X_\theta)$ is almost surely (locally) Lipschitz continuous on Θ if the set $\mathcal{N} \subset \Omega$ of realizations such that*

$$|h(X_{\theta_1}(\omega) - h(X_{\theta_0}(\omega))| \leq |\theta_1 - \theta_0| K(\omega)$$

has probability one, i.e., $\mathbb{P}(\mathcal{N}) = 1$, and $\mathbb{E}[K] < \infty$, where measurability of $\mathcal{N}$ is assumed.

In the following we often write $\frac{d}{d\theta} f(\theta_0)$ for the derivative of a differentiable mapping f at a point θ, i.e., we let

$$\left.\frac{d}{d\theta}\right|_{\theta=\theta_0} f(\theta) = \frac{d}{d\theta} f(\theta_0)$$

to simplify notation when this causes no confusion. First, we state the key IPA theorem.

Theorem 7.1 *Let $\Theta \subset \mathbb{R}$ be an open connected set, such that $X(\theta)$ is a measurable mapping on a common underlying probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ Let $\theta_0 \in \Theta$. If,*

(i) the sample path derivative $dX(\theta)/d\theta$ exists with probability one at θ_0,

(ii) the mapping $h : S \to \mathbb{R}$ is differentiable,

(iii) the mapping $h(X(\theta))$ is Lipschitz continuous on Θ with probability one, then

then

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))]\bigg|_{\theta=\theta_0} = \mathbb{E}\left[\frac{d}{d\theta}X(\theta)\, h'(X(\theta))\bigg|_{\theta=\theta_0}\right],$$

where $h'(x)$ denotes the derivative of $h(x)$ with respect to x.

Proof: Let $\{\Delta_n\}$ be sequence such that Δ_n tends to 0 as n tends to infinity. By condition (iii), the finite difference expressions are path-wise bounded by an integrable mapping K, for formally,

$$\frac{1}{|\Delta_n|}|h(X(\theta+\Delta_n)) - h(X(\theta))| \leq K,$$

with $\mathbb{E}[K] < \infty$. By conditions (i) and (ii), the expression on the above left-hand side converges to

$$h'(\theta)\,\frac{d}{d\theta}X(\theta)$$

with probability one. By the Dominated Convergence Theorem (Theorem B.9 in the Appendix) it then follows that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{\Delta_n}\mathbb{E}[h(X(\theta+\Delta_n) - h(X(\theta))] &= \mathbb{E}\left[\lim_{n\to\infty}\frac{1}{\Delta_n}\mathbb{E}[h(X(\theta+\Delta_n) - h(X(\theta))]\right]\\ &= \mathbb{E}\left[\frac{d}{d\theta}X(\theta)\,h'(X(\theta))\right],\end{aligned}$$

which proves the claim.

QED

The existence of the sample path derivative of $X(\theta)$ in condition (i) above can be deduced with the following result in a general way, see [30, 16, 14].

Lemma 7.1 *Let F_θ denote the cumulative distribution function of $X(\theta)$ for $\theta \in \Theta$ having support S. If $F_\theta(x)$ is continuously differentiable with respect to θ on Θ, and continuously differentiable with respect to x on S, then it holds for any interior point θ of Θ with probability one that*

$$\frac{d}{d\theta}X(\theta) = -\frac{\frac{\partial}{\partial\theta}F_\theta(X(\theta))}{\frac{\partial}{\partial x}F_\theta(X(\theta))}.$$

Proof: Denote the (generalized) inverse of F_θ by F_θ^{-1}, i.e.,

$$X(\theta, u) = F_\theta^{-1}(u),$$

for $u \in [0,1]$. Using the fact that $F_\theta(X(\theta))$ is uniformly distributed on $[0,1]$, we arrive at $X(\theta, u) = F_\theta^{-1}(F_\theta(X(\theta)))$ for $u = F_\theta(X(\theta))$. In particular it holds for Δ such that $\theta + \Delta \in \Theta$:

$$F_{\theta+\Delta}(F_{\theta+\Delta}^{-1}(u)) = F_\theta(F_\theta^{-1}(u)), \tag{7.1}$$

for $u \in [0,1]$. Applying Taylor's theorem, yields

$$F_{\theta+\Delta}(F_{\theta+\Delta}^{-1}(u)) = F_\theta(F_\theta^{-1}(u)) + \frac{\partial}{\partial\theta} F_\xi(\eta)\Delta + \frac{\partial}{\partial x} F_\xi(\eta)(F_{\theta+\Delta}^{-1}(u) - F_\theta^{-1}(u))$$

for some point (ξ, η) on the line segment joining $(\theta, F_\theta^{-1}(u))$ and $(\theta + \Delta, F_{\theta+\Delta}^{-1}(u))$. By (7.1), this yields

$$\frac{\partial}{\partial\theta} F_\xi(\eta)\Delta + \frac{\partial}{\partial x} F_\xi(\eta)(F_{\theta+\Delta}^{-1}(u) - F_\theta^{-1}(u)) = 0.$$

Hence

$$-\frac{\frac{\partial}{\partial\theta} F_\xi(\eta)}{\frac{\partial}{\partial x} F_\xi(\eta)} = \frac{(F_{\theta+\Delta}^{-1}(u) - F_\theta^{-1}(u))}{\Delta} = \frac{X(\theta+\Delta, u) - X(\theta, u))}{\Delta}$$

Letting Δ tend to zero the claim follows from the continuity of the partial derivatives of F_θ.

QED

The following examples shows how IPA derivatives can be obtained by Lemma 7.1 for the case of the shape of the Weibull and Pareto II distribution, see Example 6.4 where for these instances sample path differentiation of $X(\theta)$ didn't yield an expression for the derivative as a function of $X(\theta)$ in a straightforward way.

EXAMPLE 7.1. Consider the Weibull (α, λ) distribution. Then, the IPA derivative with respect to $\alpha = \theta$ can be obtained by

$$\frac{\partial}{\partial\alpha} X(\theta) = \frac{\left(\frac{X(\theta)}{\lambda}\right)^{\alpha} \ln\left(\frac{X(\theta)}{\lambda}\right) e^{-\left(\frac{X(\theta)}{\lambda}\right)^{\alpha}}}{\frac{\alpha}{\lambda}\left(\frac{X(\theta)}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{X(\theta)}{\lambda}\right)^{\alpha}}} = \frac{X(\theta)}{\alpha} \ln\left(\frac{X(\theta)}{\lambda}\right),$$

compare with Example 6.4. In the same vein we find IPA derivative with respect to the shape parameter of the Pareto II distribution by

$$\frac{\partial}{\partial\alpha} X(\theta) = \frac{\left(1 + \frac{X(\theta)}{\lambda}\right)^{-\alpha} \ln\left(1 + \frac{X(\theta)}{\lambda}\right)}{\left(1 + \frac{X(\theta)}{\lambda}\right)^{-\alpha-1}} = \left(1 + \frac{X(\theta)}{\lambda}\right) \ln\left(1 + \frac{X(\theta)}{\lambda}\right).$$

⁂

The next example shows an application of IPA the estimation of quantile sensitivity.

EXAMPLE 7.2. Let F_θ be continuously differentiable with respect to both θ and x. We denote the α-quantile of F_θ by

$$q^\alpha = F_\theta^{-1}(\alpha) = \inf_u \{u \in \mathbb{R} : F_\theta(u) \geq \alpha\},$$

for $\alpha \in (0,1)$. Provided that $f_\theta(x)$ is larger than zero on a neighbourhood of q^α, differentiation the implicit equation $\alpha = F_\theta(q^\alpha)$ with respect to θ yields

$$\frac{d}{d\theta} q^\alpha = -\frac{\frac{\partial}{\partial\theta} F_\theta(q^\alpha)}{f_\theta(q^\alpha)}. \tag{7.2}$$

By Lemma 7.1, we can read the above equation for $X(\theta)$ with cdf F_θ as

$$\frac{d}{d\theta} q^\alpha = \mathbb{E}[X'(\theta) | X(\theta) = q^\alpha].$$

Denote for $(X_\theta(i) : 1 \leq i \leq m)$ the order statistic by $X_\theta(i:m), 1 \leq i \leq m$, and observe that due to the continuity of F_θ it holds

$$X_\theta(1{:}m) < X_\theta(2{:}m) < \cdots < X_\theta(m:m)$$

with probability one. Then, it is known that

$$\lim_{m\to\infty} X_\theta(\lceil m\alpha \rceil : m) = q^\alpha \quad a.s.,$$

where $\lceil m\alpha \rceil$ is the smallest integer greater than or equal to $m\alpha$; see [13]. By continuity we arrive at

$$\lim_{m\to\infty} \mathbb{E}[X'(\theta) | X(\theta) = X_\theta(\lceil m\alpha \rceil : m)] = \frac{d}{d\theta} q^\alpha \quad a.s.,$$

which shows that

$$\frac{\partial}{\partial\theta} X_\theta(\lceil m\alpha \rceil : m)$$

is an asymptotically unbiased estimator for $dq^\alpha/d\theta$.

***

For particular families of distributions, differentiability of $X(\theta)$ can be established directly without making use of Lemma 7.1. In fact, often we can compute the sample path derivatives without having explicit knowledge of the distribution, which establishes an interesting robustness property of IPA.

Definition 7.2 *A parameter $\theta \in \Theta \subset \mathbb{R}$ of a family of probability distributions $\{F_\theta, \theta \in \Theta\}$ is called*

- *a* location parameter *if $F_\theta(x) = F_0(x - \theta)$, and*
- *a* scale parameter *if $F_\theta(x) = F_1(x/\theta)$.*

Examples of location parameters are the mean of a normal distribution and the mean of the distribution of the random variable $X(\theta) = \theta + U$, where $U \sim U(-1,1)$. Examples of scale parameters are the standard deviation of the normal distribution,the mean of the exponential distribution, and the mean of the random variable $X(\theta) = U\theta$, where $U \sim U(-1,1)$.

Proposition 7.1 *Let θ be the* location parameter *of the distribution function of $X(\theta)$. Then,*

$$\frac{d}{d\theta} X(\theta) = 1.$$

Let θ be the scale parameter *of distribution function of $X(\theta)$. Then,*

$$\frac{d}{d\theta} X(\theta) = X(1) = \frac{1}{\theta} X(\theta).$$

Proof: The proof follows from the fact that if θ is a location parameter, then $X(\theta) \overset{\mathcal{L}}{=} \theta + X(0)$. If it is a scale parameter, then $X(\theta) = \theta X(1)$.

QED

We now turn to the discussion of the third condition in Theorem 7.1. We have the following result from analysis, which we prove here, as the proof provides insight into the relation of differentiability and continuity to Lipschitz continuity.

Lemma 7.2 *Let f be a real-valued mapping such that f is continuous on $[\theta_a, \theta_b]$ and differentiable on (θ_a, θ_b), for $-\infty < \theta_a < \theta_b < \infty$, except for an at most denumerable set of points $\Theta_0 \subset [\theta_a, \theta_b]$. Then, f is Lipschitz continuous on $[\theta_a, \theta_b]$ with modulus*

$$\sup\left\{ \left| \frac{d}{d\theta} f(\theta) \right| : \theta \in \Theta \setminus \Theta_0 \right\},$$

provided the above expression is finite.

Proof: Let $\theta_a \leq \theta_0 < \theta_1 \leq \theta_b$. Then, using a telescopic sum

$$|f(\theta_0) - f(\theta_1)| \leq \sum_i |f(\theta^i) - f(\theta^{i+1})|,$$

where $\{\theta^i\}$ are the points of non-differentiability in Θ_0 falling into $[\theta_0, \theta_1]$ such that $\theta^i < \theta^{i+1}$. By the mean value theorem

$$|f(\theta^i) - f(\theta^{i+1})| \leq |\theta^i - \theta^{i+1}| K_i,$$

where

$$K_i = \sup_{\theta \in (\theta^i, \theta^{i+1})} \left| \frac{d}{d\theta} f(\theta) \right|.$$

Hence,

$$|f(\theta_0) - f(\theta_1)| \leq \sum_i |f(\theta^i) - f(\theta^{i+1})| \leq \sum_i |\theta^i - \theta^{i+1}| \sup_i K_i = |\theta_0 - \theta_1| \sup_i K_i,$$

which establishes the result.

QED

We now arrive at the following sufficient condition for the third condition in Theorem 7.1 that is typically easily applicable in applications.

Lemma 7.3 *Let $\Theta = [\theta_a, \theta_b] \subset \mathbb{R}$, with $-\infty < \theta_a < \theta_b < \infty$. Suppose that*

(i) with probability one, the function $h(X(\theta))$ is continuous on Θ, and the set Θ_0 of points of non-differentiability of $h(X(\theta))$ has no accumulation point in (θ_a, θ_b),

(ii) there exists a random variable K such that $\mathbb{E}[K] < \infty$ and

$$\sup\left\{ \left| \frac{d}{d\theta} h(X(\theta)) \right| : \theta \in \Theta \setminus \Theta_0 \right\} \leq K,$$

then $h(X(\theta))$ is with probability one Lipschitz continuous on Θ and the Lipschitz modulus has finite first moment.

Proof: By Lemma 7.2, for $\theta_1 > \theta_0$, where $\theta_1, \theta_0 \in \Theta$, it holds almost surely that

$$|h(X(\theta_1) - h(X_{\theta_0})| \le |\theta_1 - \theta_0| \sup_{\theta \in [\theta_a, \theta_b] \setminus \Theta_0} \left| \frac{d}{d\theta} h(X(\theta)) \right| .$$

The expression on the above right-hand side serves a.s. as a Lipschitz modulus for $h(X(\theta))$ on Θ. Hence, any K such that

$$\sup \left\{ \left| \frac{d}{d\theta} h(X(\theta)) \right| : \theta \in \Theta \setminus \Theta_0 \right\} \le K$$

is a Lipschitz modulus as well and the proof follows from the assumption that $\mathbb{E}[K] < \infty$.

QED

By virtue of Lemma 7.2 we may apply the Dominate Convergence Theorem to mappings that have points of non-differentiability.

Example 7.3. Let $X(\theta)$ be exponentially distributed with mean $\theta > 0$. Since θ is a scale parameter of the exponential distribution, it follows that

$$\frac{d}{d\theta} h(X(\theta)) = \frac{1}{\theta} X(\theta)\, h'(X(\theta)),$$

for any differentiable mapping h with derivative h', which establishes condition (i) in Theorem 7.1. Let $\hat{\Theta} = (\theta_l, \theta_r) \subset \Theta$ such that $\theta \in \hat{\Theta}$. Using the representation put forward in (6.3) it follows for any differentiable mapping h with monotone derivative that

$$\sup_{\theta \in \hat{\Theta}} \frac{d}{d\theta} h(X(\theta)) \le \frac{1}{\theta_l} X_{\theta_r}\, h'(X_{\theta_r}).$$

Hence, provided that $\mathbb{E}[X_{\theta_r}\, h'(X_{\theta_r})]$ is finite, Lemma 7.3 yields that condition (ii) in Theorem 7.1 is satisfied. Applying Theorem 7.1 then yields for any differentiable mapping with monotone derivative

$$\frac{d}{d\theta} \mathbb{E}[h(X(\theta))] = \frac{1}{\theta} \mathbb{E}[X(\theta)\, h'(X(\theta))],$$

provided that $\mathbb{E}[|X_{\theta_r}\, h'(X_{\theta_r})|] < \infty$ for some $\theta_r > \theta$. Note that the condition on h can be simplified in some cases. Indeed, if h' is bounded, then $\mathbb{E}[X_{\theta_r}] < \infty$ is sufficient for unbiasedness which is trivially satisfied. This is the case for the identity mapping $h(x) = x$. Applying this result directly for the problem in Example 6.3 yields that the feedback

$$Y_n = 1 - \frac{X(\theta_n)}{\theta_n} \left(\frac{1}{(1 + X(\theta_n))^{1/2}} - \frac{X(\theta_n)}{2(1 + X(\theta_n))^{3/2}} \right)$$

is an unbiased estimator of the desired derivative, which shows why the stochastic approximation converges to the optimal solution.

***

Example 7.4. We revisit Example 6.9 and establish unbiasedness of the IPA estimator for the sensitivity of the system's life to this particular component. Use the Skorohod representation (see Example 6.2) to establish that T_3 has the same distribution as $-\theta \ln(1 - U(\omega))$ where $U \sim U(0, 1)$ and use this representation in what follows. Fix ω and rewrite the life of the system as a function of θ

as follows. To simplify notation, call $h(\theta,\omega) = L(T_1(\omega), T_2(\omega), T_3(\theta,\omega), T_4(\omega), T_5(\omega))$. We will next show that

$$\begin{aligned} h(\theta,\omega) &= \begin{cases} T_3(\theta,\omega) & \text{if } T_2(\omega) < T_3(\theta,\omega) < \bar{T}(\omega) \overset{\text{def}}{=} \min(T_1(\omega), T_4(\omega), T_5(\omega)) \\ T_{i^*}(\omega)\ (i^* \neq 3) & \text{otherwise} \end{cases} \\ &= \begin{cases} \theta(-\ln(1-U) & \text{if } T_2(\omega) < T_3(\theta,\omega) < \bar{T}(\omega) \\ T_{i^*}(\omega)\ (i^* \neq 3) & \text{otherwise} \end{cases} \end{aligned}$$

so that for each ω, h is a piece-wise linear function of θ.

To find the expression above, notice first that if $T_2 \geq \bar{T}$ then route R_3 breaks down at time $\min(T_3(\theta), \bar{T})$. If $T_3(\theta) < \bar{T}$ then both routes R_1, R_2 are working until time $\bar{T}$, and otherwise, the system also breaks down at $\bar{T}$, so in this case $h(\theta) = \bar{T}$ is independent of θ. Now suppose that $T_2 < \bar{T}$. If $T_3(\theta) \leq T_2$ then route R_3 breaks at time $T_3(\theta)$ but the other routes are still working, and the system breaks down at T_2. If $T_3(\theta) \geq \bar{T}$ then routes R_1 and R_2 break first, and then route R_3 works until time $\bar{T}$, when the system fails (although component 3 is still working), so the system life is $\bar{T}$. Finally, if $T_2 < T_3(\theta) < \bar{T}$ then routes R_1 and R_2 break at time T_2, and route R_3 breaks exactly at time $T_3(\theta)$, which is then the value of the system's life.

if $T_2 < \bar{T}$, define $\underline{\theta} = -T_2(\omega)/\ln(1-U(\omega)), \bar{\theta} = -\bar{T}(\omega)/\ln(1-U(\omega))$. Then

$$h(\theta,\omega) = \begin{cases} T_2(\omega) & \theta \leq \underline{\theta} \\ \theta(-\ln(1-U(\omega))) & \underline{\theta} \leq \theta \leq \bar{\theta} \\ \bar{T}(\omega) & \theta \geq \bar{\theta} \end{cases}$$

and notice that this function is continuous. Therefore the Lipschitz constant is bounded in absolute value by the largest slope, namely $-\ln(1-U)$. Because this is a random variable with distribution $\exp(1)$, it has bounded expectation, which implies that we can interchange derivative and expectation for h. The IPA estimator is the resulting derivative, expressed in terms of the original random variables $(T_i; i = 1, \ldots, 5)$, which is:

$$\widehat{L}^{\mathrm{IPA}}(\theta,\omega) = \begin{cases} \frac{T_3(\theta,\omega)}{\theta} & T_2(\omega) < T_3(\theta,\omega) < \bar{T}(\omega) = \min(T_1(\omega), T_4(\omega), T_5(\omega)) \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

It follows from Theorem 7.2 that, although the derivative is a discontinuous random variable, it is unbiased for $dL(\theta)/d\theta$. It is important to realise that the expectation of $L^{\mathrm{IPA}}(\theta)$ is independent of the "true" representation of the random variable $T_3(\theta)$ as a function of ω, because $T_3(\theta) \overset{\mathcal{L}}{=} -\theta\ln(1-U)$ is sufficient to establish the result.

$\ast\ast\ast$

So far, we have considered the case of perturbing a single random variable. Many models that are of importance in applications are driven by a finite collection of random variables, and in the remainder of this section we explain how IPA can be used in these models. The following theorem states sufficient conditions for unbiasedness of the IPA estimator for static gradient estimation problem.

Theorem 7.2 *Let $X_i(\theta)$, $1 \leq i \leq N$, be (not necessarily statistically independent) random variables.*

Suppose that

(i) with probability one, the function $L_N(X_1(\theta), \ldots, X_N(\theta))$ is continuous on Θ, and the set Θ_0 of points of non-differentiability of $L_N(X_1(\theta), \ldots, X_N(\theta))$ has no accumulation point in (θ_a, θ_b),

(ii) the cost function $L_N(X_1(\theta), \ldots, X_N(\theta))$ is Lipschitz continuous in θ with integrable Lipschitz modulus,

then

$$\frac{d}{d\theta}\mathbb{E}[L_N(X_1(\theta), \ldots, X_N(\theta))] = \mathbb{E}\left[\sum_{i=1}^{N} \frac{d}{d\theta}X_i(\theta)\, \frac{\partial}{\partial x_i} L_N(X_1(\theta), \ldots, X_N(\theta))\right].$$

EXAMPLE 7.5. Let $\{X_n(\theta)\}$ be an i.i.d. sequence of exponentially distributed random variables with mean value θ. Moreover, for $n \geq 1$, let

$$T_\theta(n) = \sum_{k=1}^{n} X_k(\theta) = L_n(X_1(\theta), \ldots, X_n(\theta)).$$

By definition, the sequence $\{T_\theta(n)\}$ corresponds to the arrival times of a Poisson process with rate $1/\theta$. It is easily seen that L_n satisfies the condition put forward in Theorem 7.2. In particular, it holds that $\partial L_n / \partial x_i = 1$, which yields

$$\frac{d}{d\theta}T_\theta(n) = \sum_{k=1}^{N} \frac{\partial}{\partial x_i} L_n(X_1(\theta), \ldots, X_n(\theta)) \frac{d}{d\theta}X_i(\theta) = \frac{1}{\theta}T_\theta(n)$$

for the IPA estimator. That the estimator is indeed unbiased can be verified here from the fact that $\mathbb{E}[T_\theta(n)] = n\theta$ and

$$\frac{d}{d\theta}\mathbb{E}[T_\theta(n)] = n = \frac{1}{\theta}\mathbb{E}[T_\theta(n)]$$

together with Example 6.2. For this academic example, one could also use the fact that $T_\theta(n)$ has a Gamma distribution with scale parameter θ. The number of jumps of the Poisson counting process between time zero and T is given by

$$N_\theta(T) = \sum_{n=0}^{\infty} \mathbf{1}_{\{T_\theta(n) \leq T < T_\theta(n+1)\}} = \sup\{n : T_\theta(n) \leq T\}.$$

Since $N_\theta(T)$ is a piecewise constant mapping in θ it follows that

$$\frac{d}{d\theta}N_\theta(T) = 0$$

with probability one. Hence, the IPA estimator would yield the erroneous outcome 0, where it is well known that $\mathbb{E}[N_\theta(T)] = T/\theta$ and the true derivative is thus given by $-T/\theta^2$. To see why IPA fails for this performance function, note that $N_\theta(T)$ fails to be Lipschitz continuous in θ on a non-random neighborhood of θ. An important observation is that the results above hold true for other models, if θ is a scale parameter of the distribution of $X(\theta)$. This becomes relevant when the actual distribution of inter-arrival times is not known, but it can be argued that the variable θ is a scale parameter.

⁂

Next, we provide an IPA analysis for the more demanding case of sojourn times in a single server queue.

EXAMPLE 7.6. We revisit the sojourn time example as introduced in Example 6.10 with notations and basic assumptions as detailed in this example. In addition we assume that θ is a scaling parameter of the service time distribution. Suppose that we are interested in solving the static gradient estimation problem with IPA for

$$L_N(X_1(\theta), \ldots, X_N(\theta)) = \frac{1}{N}\sum_{n=1}^{N} X_n(\theta). \tag{7.4}$$

In words, L_N is the average sojourn time over the first N customers in a single server queue.

Because L_N is an average, it suffices to show that each $X_n(\theta)$ is Lipschitz continuous w.p.1. We use induction on (7.23). Using the fact that θ is a scaling parameter, we obtain, for $n = 1$, $X_1(\theta) = S_1(\theta) = \theta S_1(1)$ is almost surely Lipschitz continuous with modulus $l_1 = S_1(1)$. We now proof by induction that $X_n(\theta)$ is a.s. Lipschitz continuous with Lipschitz modulus

$$l_n = \sum_{k=1}^{n} S_k(1).$$

For the induction, note that

$$X_n(\theta) < A_{n+1} \quad \Rightarrow \quad X_n(\theta+\delta) - A_{n+1} < X_n(\theta+\delta) - X_n(\theta). \tag{7.5}$$

Use (7.23) and the fact that $X_n(\theta+\delta) \geq X_n(\theta)$ to show that

$$\begin{aligned} X_{n+1}(\theta+\delta) - X_{n+1}(\theta) &= \\ &= S_{n+1}(\theta+\delta) - S_{n+1}(\theta) + \begin{cases} X_n(\theta+\delta) - X_n(\theta) & \text{if } A_{n+1} < X_n(\theta) \\ X_n(\theta+\delta) - A_{n+1} & \text{if } X_n(\theta) < A_{n+1} < X_n(\theta+\delta) \\ 0 & \text{otherwise,} \end{cases} \\ &\overset{(7.5)}{\leq} S_{n+1}(\theta+\delta) - S_{n+1}(\theta) + \begin{cases} X_n(\theta+\delta) - X_n(\theta) & \text{if } A_{n+1} < X_n(\theta+\delta) \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

which yields that

$$|X_{n+1}(\theta+\delta) - X_{n+1}(\theta)| \leq \delta(S_{n+1}(1) + l_n) = \delta l_{n+1},$$

and completes the induction argument. Since $\mathbb{E}[l_n] \leq n < \infty$, we have shown that L_N is a.s. Lipschtiz continuous. The resulting IPA derivative is:

$$\frac{d}{d\theta} L_N(X_1(\theta), \ldots, X_N(\theta)) = \frac{1}{N}\sum_{n=1}^{N} \frac{dX_n(\theta)}{d\theta}, \tag{7.6}$$

and it can be computed recursively using an auxiliary *derivative process*:

$$Z_n(\theta) \overset{\text{def}}{=} \frac{dX_n(\theta)}{d\theta} = \frac{dS_n(\theta)}{d\theta} + \frac{dX_{n-1}(\theta)}{d\theta}\mathbf{1}_{\{X_{n-1}(\theta) > A_n\}},$$

which can be calculated online, as the process $\{X_n(\theta)\}$ is being simulated or observed. In the original formulation of (7.6), identify $\tilde{X}_n(\theta) = Z_n(\theta)$, and set $G(Z_1, \ldots, Z_N) = (1/N)\sum_n Z_n$. In this case no augmentation of the filtration is needed.

An alternative and popular computation procedure is provided now. Let $\alpha(n)$ be the index of the last customer prior to i that encounters the server idle, that is: $\alpha(n) = \max(k \leq n\colon\ X_{k-1}(\theta) < A_k)$, which is $\mathfrak{F}_n(\theta)$-measurable and therefore a stopping time. Then

$$X_n(\theta) = \sum_{i=\alpha(n)}^{n} S_i(\theta), \tag{7.7}$$

and therefore we can write the derivative as:

$$Z_n(\theta) = \frac{dX_n(\theta)}{d\theta} = \sum_{i=\alpha(n)}^{n} \frac{dS_i(\theta)}{d\theta},$$

(where the sum from n to n contains the one single term). This gives the final computation:

$$\frac{d}{d\theta} L_N(X_1(\theta), \ldots, X_N(\theta)) = \frac{1}{N} \sum_{n=1}^{N} Z_n(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=\alpha(n)}^{n} \frac{dS_i(\theta)}{d\theta}. \tag{7.8}$$

***

7.1.2 Smoothed Perturbation Analysis

The key limitation of the sample path approach is that the combined random variable $h(X(\theta))$ has to be Lipschtiz continuous on Θ with probability one, for some open interval Θ not depending on the sample path. The following example illustrates the point for a very simple model.

EXAMPLE 7.7. Suppose that $X(\theta) \sim \text{Bernoulli}(\theta)$ is a random variable with a Bernoulli distribution. If we use the Skorohod representation, then $X(\theta, U) \overset{\mathcal{L}}{=} \mathbf{1}_{\{U \leq \theta\}}$, for U a uniform random variable on $(\Omega, \mathbb{P})$. Here $\mathbb{E}(X(\theta)) = \mathbb{P}(U \leq \theta) = \theta$ so that the derivative satisfies $\frac{d}{d\theta}\mathbb{E}(X(\theta)) = 1$. However, for every U, $X(\theta, U)$ is piecewise constant with a jump at $U = \theta$, so that for every $\omega\colon U(\omega) \neq \theta$, $X'(\theta, U) = 0$. Because $\mathbb{P}(U(\omega) = \theta) = 0$, we have here that $X'(\theta, U) = 0$ a.s., and consequently $\mathbb{E}(X'(\theta, U)) = 0 \neq \frac{d}{d\theta}\mathbb{E}(X(\theta)) = 1$.

***

Sometimes it is possible to overcome this obstacle using a conditioning approach, see the following example.

EXAMPLE 7.8. We revisit the sojourn time example in Example 6.10 and Example 7.6. Consider the probability that the nth sojourn time is smaller or equal to some finite positive value β given by

$$\mathbb{P}(X_n(\theta) \leq \beta) = \mathbb{E}\big[\mathbf{1}_{\{X_n(\theta) \leq \beta\}}\big]. \tag{7.9}$$

Note that even though $X_n(\theta)$ is almost surely Lipschitz continuous in θ, see Example 7.6, the indicator function $\mathbf{1}_{\{X_n(\theta) \leq \beta\}}$ fails to be Lipschitz continuous. To see this note that for $\Delta > 0$ it holds that

$$\mathbf{1}_{\{X_n(\theta+\Delta) \leq \beta\}} - \mathbf{1}_{\{X_n(\theta) \leq \beta\}}$$

equals one if and only if $X_n(\theta) \leq \beta < X_n(\theta + \Delta)$ and is otherwise zero. Hence, Lipschitz continuity fails.

Recall the basic recursion for sojourn times given in (7.23) and let F_θ denote the distribution function of $S_n(\theta)$. By calculation,

$$\begin{aligned}\mathbb{P}(X_n(\theta) \leq \beta) &= \mathbb{E}\big[\mathbb{P}(X_n(\theta) \leq \beta \,|\, X_{n-1}(\theta))\big] \\ &= \mathbb{E}\big[\mathbb{P}(\max(0, X_{n-1}(\theta) - A_n) + S_n(\theta) \leq \beta)\big] \\ &= \mathbb{E}\big[\mathbb{P}(S_n(\theta) \leq \max(\beta - \max(0, X_{n-1}(\theta) - A_n), 0)\big] \\ &= \mathbb{E}\big[F_\theta(\max(\beta - \max(0, X_{n-1}(\theta) - A_n), 0))\big].\end{aligned}$$

Since θ is a scaling parameter of F_θ it follows that F_θ is Lipschitz continuous with respect to θ and differentiable. It thus holds that

$$\frac{d}{d\theta}\mathbb{P}(X_n(\theta) \leq \beta) = \mathbb{E}\left[\frac{d}{d\theta}F_\theta(\max(\beta - \max(0, X_{n-1}(\theta) - A_n), 0))\right]$$

Let $\{X_n(\theta)\}$ denote the consecutive sojourn times (Markovian) for a fixed value of θ and define

$$Y_n(\theta) = \frac{d}{d\theta}F_\theta(\max(\beta - \max(0, X_{n-1}(\theta) - A_n), 0)).$$

Then by construction, the estimator is unbiased for (7.9).

⁂

The example above shows how appropriate conditioning can establish an unbiased gradient estimator in case IPA fails. The basic principle can be described as follows. Consider $L(\theta) \stackrel{\text{def}}{=} L(X_\theta)$ such that at θ_0, for every ω

$$\lim_{\Delta \downarrow 0} L(\theta_0 + \Delta) = L^+(\theta_0) \text{ and } \lim_{\Delta \uparrow 0} L(\theta_0 + \Delta) = L^-(\theta_0)$$

exists a.s. In case $L(X_\theta)$ is a.s. continuous at θ_0 we obtain $L^-(\theta_0) = L^+(\theta_0)$. Otherwise $L^+(\theta_0) - L^-(\theta_0)$ expresses the height of the jump of $L(\theta)$ at θ_0, which was in the Example 7.8 an indicator mapping with height 1.

For $\Delta \neq 0$, let $\Omega_{\theta_0}(\Delta)$ be the event that $L(X_\theta)$ is differentiabe at θ_0 and Lipschtiz contnuous on $[\theta_0 - |\Delta|, \theta_0 + |\Delta|]$. Then

$$\begin{aligned}\frac{1}{\Delta}\mathbb{E}[L(X_{\theta_0+\Delta}) - L(X_{\theta_0})] &= \frac{1}{\Delta}\mathbb{E}[L(X_{\theta_0+\Delta}) - L(X_{\theta_0})|\Omega_{\theta_0}(\Delta)]\mathbb{P}(\Omega_{\theta_0}(\Delta)) \\ &\quad + \mathbb{E}[L(X_{\theta_0+\Delta}) - L(X_{\theta_0})|\Omega^c_{\theta_0}(\Delta)]\frac{\mathbb{P}(\Omega^c_{\theta_0}(\Delta))}{\Delta}\end{aligned}$$

Typically, the probabilities on the above right hand side are smooth and it holds that

$$\lim_{\Delta \to 0} \mathbb{P}(\Omega_{\theta_0}(\Delta)) \stackrel{\text{def}}{=} p_{\theta_0} \tag{7.10}$$

and

$$\lim_{\Delta \to 0} \frac{1}{\Delta}\mathbb{P}(\Omega^c_{\theta_0}(\Delta)) \stackrel{\text{def}}{=} \hat{p}_{\theta_0}. \tag{7.11}$$

Let

$$\Omega_{\theta_0} = \bigcap_{\Delta \neq 0} \Omega_{\theta_0}(\Delta).$$

Provided the limits (7.10) and (7.11) exist, we obtain

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\mathbb{E}[L(X_{\theta_0+\Delta})-L(X_{\theta_0})] \;=\; \mathbb{E}\left[\frac{d}{d\theta}L(X_{\theta_0})\middle|\,\Omega_{\theta_0}\right]p_{\theta_0}+\mathbb{E}\big[L^+(\theta_0)-L^-(\theta_0)\big|\Omega^c_{\theta_0}\big]\hat{p}_{\theta_0}.$$

The first part on the above right hand side is noticeably the IPA contribution to the derivative, and the second part is a rate weight of the jump size at θ_0. The above conditioning approach is known as *smoothed perturbation analysis* (SPA). While SPA has been successfully applied in many situations, it provides only ad hoc solutions as the conditioning depends on the structure of the sample path function and the type of dependence of $L(X(\theta))$ on θ. For more details on SPA we refer to [17, 12]. Observe that in Example 7.8 we discussed an application of SPA where the IPA contribution was zero and $L^+(\theta_0)-L^-(\theta_0)=1$ so that analysis reduced to the computation of an unbiased estimator for the rate $\hat{p}_{\theta_0}$.

⋆7.1.3 An Indirect Approach To Establishing Unbiasedness

When presenting IPA we have chosen for a constructive approach. First we constructed the sample path derivatives, and then we asked whether they provide an unbiased estimator. This is the usual approach in the simulation literature. In this section we discuss a different approach to establishing the existence of a sample path derivative and the proof of unbiasedness. The approach is based on a basic result from measure theory for absolutely continuous mappings. The resulting estimators are called “stochastic derivatives” in the applied mathematics literature, rather than IPA derivatives.

Let (X,d) be a metric space and let $I\subset\mathbb{R}$ be an interval. A mapping f from I to X is called *absolutely continuous* on I if for every $\epsilon>0$, there exists $\delta>0$ such that whenever a (finite or infinite) sequence of pairwise disjoint sub-intervals $[x_k,y_k]$ of i satisfies

$$\sum_k |y_k-x_k|<\delta$$

then

$$\sum_k d\left(f(y_k),f(x_k)\right)<\epsilon.$$

If f is absolutely continuous, then f has a derivative almost everywhere, the derivative is Lebesgue integrable, and its integral is equal to the increment of f, i.e., if f is absolutely continuous, then the set of $x\in I$ such that f' fails to exit has Lebesgue measure zero and it holds for $x,y\in$ with $y>x$ that

$$f(y)-f(x)=\int_x^y f'(z)\lambda(dz),$$

where $\lambda(\cdot)$ denotes the Lebesgue measure. The usefulness of the concept of absolute continuity for gradient estimation stems from the fact that a Lipschitz-continuous function is absolutely continuous.

Let $X(\theta)$ be Lipschitz continuous on some finite interval Θ. In order to make use of arguments based on absolute continuity, we have to consider $X(\theta)$ as a random mapping on an extended sample space. More specifically, let, for all θ , $X(\theta)$ be defined on some underlying probability space $(\Omega,\mathcal{F},P)$. Let $\Theta=[\theta_a,\theta_b]\subset\mathbb{R}$. Equip $\mathbb{R}$ with the usual topology and equip Θ with its Borel field, denoted by $\mathcal{B}$. For the following consider $X(\theta)$ as random variable on the product space

$(\Omega \times \Theta, \mathcal{F} \otimes \mathcal{B}, P \otimes \lambda)$, where $\mathcal{F} \otimes \mathcal{B}$ denotes the product field of $\mathcal{F}$ and $\mathcal{B}$ and $P \otimes \lambda$ denotes the product probability measure of P and the Lebesgue measure. By absolute continuity of $X(\theta)$ it holds

$$X(\theta_b) - X(\theta_a) = \int_{\theta_a}^{\theta_b} X'(r)\, \lambda(dr),$$

and we conclude that $X(\theta)$ is for Lebesgue almost all θ in Θ differentiable. Note that since $X(\theta)$ is measurable as a mapping of θ, it follows that $X'(\theta)$ is measurable. Fubini's theorem now gives

$$\mathbb{E}[X(\theta_b)] - \mathbb{E}[X(\theta_a] = \int_{\theta_a}^{\theta_b} \mathbb{E}[X'(r)]\, \lambda(dr),$$

from which we conclude that $\mathbb{E}[X(\theta)]$ is absolutely continuous and that $\mathbb{E}[X'(\theta)]$ is for almost all θ the derivative of $\mathbb{E}[X(\theta)]$ on Θ:

$$\frac{d}{d\theta}\mathbb{E}[X(\theta)] = \mathbb{E}[X'(\theta)], \tag{7.12}$$

for almost all θ in Θ.

To summarize, the above approach is rather elegant in that interchanging differentiation and integration can be replaced by a simple application of Fubini's theorem. However, the fact that we only have Lebesgue almost everywhere differentiability on Θ makes the result in (7.12) somewhat esoteric, as for any given θ_0 we cannot say whether (7.12) holds or not.

7.2 The Score-Function Method (SF)

7.2.1 Basic Results and Techniques

Definition 7.3 *Let ν and μ be measures on a common measurable space $(\Omega, \mathfrak{F})$. A measurable mapping f is called a μ-*density *of ν if for all $A \in \mathfrak{F}$ it holds that*

$$\nu(A) = \int_A f(x)\mu(dx).$$

The μ-density of ν is sometimes denoted by $\left[\frac{d\nu}{d\mu}\right]$ and it is also called the Radon-Nikodym derivative*.*

A sufficient condition for the Radon-Nikodym derivative to exists is that ν is absolutely continuous with respect to μ (written $\nu << \mu$), *i.e.*, for any measurable set A, $\nu(A) = 0 \implies \mu(A) = 0$. This is also expressed by saying that ν is *dominated* by μ.

REMARK. For continuous distributions, the "probability density" commonly used is the μ-density of the given distribution with respect to the Lebesgue measure.

EXAMPLE 7.9. Let $X(\alpha)$ have Beta distribution with parameters $\alpha > 0, \beta > 0$, that is, the Lebesgue density of $X(\alpha)$ is

$$g_\alpha(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \text{ where } \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

and $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$ the Gamma function. Let $Y \sim g_1(x)$ be another Beta random variable. Then for any interval $A \subset [0,1]$,

$$\begin{aligned}\mathbb{P}(X(\alpha) \in A) &= \int_A g_\alpha(x)\,dx \\ &= \int_A \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,dx \\ &= \int_A \left(\frac{x^{\alpha-1}B(1,\beta)}{B(\alpha,\beta)}\right)\frac{(1-x)^{\beta-1}}{B(1,\beta)}\,dx \\ &= \mathbb{P}\left(\frac{g_\alpha(Y)}{g_1(Y)}Y \in A\right).\end{aligned}$$

Here we identify $\mu(dx)$ with the pdf $f_1(x)\,dx$, and $\nu_\alpha(dx)$ with $g_\alpha(x)\,dx$. The μ-density of ν_α can be obtained by scaling the μ-density of ν_1 by the fraction of the densities

$$f_\alpha(x) = \frac{g_\alpha(x)}{g_1(x)}.$$

Now suppose that $X(p)$ has a Binomial distribution Bin(n,p) so

$$\mathbb{P}(X(p) = k) = \binom{n}{k}p^k(1-\theta)^{n-k},$$

and let $Y \sim$ Bin$(n, 0.1)$. Then for any subset A of $\{1,\ldots,n\}$

$$\mathbb{P}(X(p) \in A) = \sum_{k\in A}\left(\frac{p^k(1-p)^{n-k}}{(0.1)^k(0.9)^{n-k}}\right)\binom{n}{k}(0.1)^k(0.9)^{n-k}.$$

Here, we identify ν_p as the Binomial Bin(n,p) distribution and μ with the Bin$(n,0.1)$ distribution, so the μ-density of ν_p can obtained from via the μ-density of $\nu_{0.1}$ via ratio of the likelihoods

$$f_p(k) = \frac{p^k(1-p)^{n-k}}{(0.1)^k(0.9)^{n-k}}.$$

⁂

REMARK. When X is a random variable with distribution μ, the Radon Nykodim derivative $f(X)$ is a random variable. In statistics it is common to use this ratio evaluated at sample observations of the random variable, and in that context $f_\theta(x)$ is called the *Likelihood ratio*.

If the μ-density of a measure ν exists, then for any integrable function h and any $A \in \mathfrak{F}$

$$\int_A h(x)\,\nu(dx) = \int_A h(x)\left[\frac{d\nu}{d\mu}\right]\mu(dx).$$

This transformation is called a "change of measure". In particular, it allows to "push out" the dependency on θ of the random variables $X(\theta)$ to the density f_θ, while the measure μ is independent of θ.

As already explained in Section 6.2, the score function approach considers the θ-dependency to be given via the density and, therefore, the key step in applying the score function is to interchange integration with respect to a θ-dependent density and to be able to differentiate that density with respect to θ. The precise statement is given in the following theorem.

Theorem 7.3 *For $\theta \in \Theta$ let ν_θ be a probability measure with μ-density f_θ, for some measure μ and assume that the family of μ-densities f_θ has common support. If*

(i) f_θ is μ-almost everywhere differentiable with respect to θ,

(ii) for a measurable mapping h it holds that

$$|h(x)f_{\theta+\Delta}(x) - h(x)f_\theta(x)| \leq |\Delta| k(x),$$

for all Δ such that $|\Delta| < \epsilon$, where ϵ is some small number independent of x, and

$$\int k(x)\mu(dx) < \infty,$$

then

$$\frac{d}{d\theta}\int h(x)\nu_\theta(dx) = \int h(x)S(\theta,x)\nu_\theta(dx),$$

where

$$S(\theta, x) = \frac{\partial}{\partial\theta}\log(f_\theta(x)) \quad \mu\text{-a.s.}$$

Proof: Under the conditions of the theorem, for any $\theta, \theta_0 \in \Theta$, the measure ν_θ has ν_{θ_0}-density f_θ/f_{θ_0} (see Exercise 7.7). By the Dominated Convergence Theorem it holds that interchanging differentiation and integration is justified. This yields

$$\begin{aligned}
\lim_{\Delta\to 0}\frac{1}{\Delta}\left(\int h(x)\nu_{\theta+\Delta}(dx) - \int h(x)\nu_\theta(dx)\right) &= \lim_{\Delta\to 0}\frac{1}{\Delta}\left(\int h(x)\frac{f_{\theta+\Delta}(x)}{f_{\theta_0}(x)}\nu_{\theta_0}(dx) - \int h(x)\frac{f_\theta(x)}{f_{\theta_0}(x)}\nu_{\theta_0}(dx)\right)\\
&= \int h(x)\lim_{\Delta\to 0}\frac{1}{\Delta}\left(\frac{f_{\theta+\Delta}(x)}{f_{\theta_0}(x)} - \frac{f_\theta(x)}{f_{\theta_0}(x)}\right)\nu_{\theta_0}(dx)\\
&= \int h(x)\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_{\theta_0}(x)}\nu_{\theta_0}(dx)
\end{aligned}$$

Letting $\theta = \theta_0$ completes the proof.

QED

The statement in Theorem 7.3 reads in random variable language as follows. Let $X(\theta)$ be distributed according to μ_θ such that for all $\theta \in \Theta$ it holds that μ_θ has a density with respect to some measure ν. Under the condition of Theorem 7.3 it then holds that

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta))] = \mathbb{E}[h(X(\theta))S(\theta, X(\theta))].$$

Following the Mean-Value Theorem, a sufficient condition for (ii) in Theorem 7.3 is that

$$\int |h(x)| \sup_{\theta\in\Theta_0}\left|\frac{\partial}{\partial\theta}f_\theta(x)\right|\mu(dx) < \infty,$$

where Θ_0 is some neighborhood of θ.

The following example illustrated this.

EXAMPLE 7.10. Consider $\mathbb{E}[h(X_\theta)]$ for X_θ exponential with mean θ. In the following we establish the conditions in Theorem 7.3. Condition (i) is satisfied, in particular,

$$\frac{\partial}{\partial\theta} f_\theta(x) = \left(\frac{x}{\theta^2} - 1\right) \frac{1}{\theta^2} \exp(-x/\theta),\ x \geq 0,$$

where f_θ denotes the exponential density with mean θ. Given $\theta \in [\theta_0, \theta_1]$, it holds for all θ that

$$\left|\frac{\partial}{\partial\theta} f_\theta(x)\right| \leq \left(\frac{x}{\theta_0^2} - 1\right) \frac{1}{\theta_0^2} \exp(-x/\theta_1) = k(x). \tag{7.13}$$

Note that f_θ is a Lebesgue density and therefore $\mu(dx) = dx$ in condition (ii) in Theorem 7.3. Condition (ii) becomes

$$\int |h(x)|\, k(x) dx < \infty,$$

with $k(x)$ as defined in (7.17). Alternatively, condition (ii) can be expressed via a random variable representation as follows

$$\int |h(x)|\, k(x) dx = \int |h(x)| \left(\frac{x}{\theta_0^2} - 1\right) \frac{\theta_1}{\theta_0^2} f_{\theta_1}(x) dx = \frac{\theta_1}{\theta_0^2} \mathbb{E}\left[|h(X_{\theta_1})| \left(\frac{X_{\theta_1}}{\theta_0^2} - 1\right)\right] < \infty.$$

Since all moments of the exponential distribution are finite, the above inequality holds for any polynomially bounded cost function h. Observe that the expression on the above right hand site resembles in part the Score Function (compare with Example 6.5) but due to that fact that it contains two different parameters it cannot be written as Score Function. For any polynomially bounded h, Theorem 7.3 thus yields

$$\frac{d}{d\theta} \mathbb{E}[h(X_\theta)] = \mathbb{E}\left[h(X_\theta) \frac{1}{\theta} \left(\frac{X_\theta}{\theta} - 1\right)\right],$$

see Example 6.5.

***

EXAMPLE 7.11. We revisit the normal distribution from Example 6.4. The normal distribution with mean μ and standard deviation σ, for $\sigma > 0$, has density

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbb{R}.$$

Denote by $\theta = \mu_0$ the point at which we want to take a derivative with respect to μ. For $\epsilon > 0$, choose $[\mu_0 - \epsilon, \mu_0 + \epsilon]$. Since μ is a shift parameter of $f(x; \mu, \sigma^2)$, we have for $\mu \in [\mu_0 - \epsilon, \mu_0 + \epsilon]$

$$|f(x; \mu, \sigma^2) - f(x; \mu_0, \sigma^2)| \leq \begin{cases} f(x; \mu_0 - \epsilon, \sigma^2) & x \leq \mu_0 - \epsilon, \\ \frac{1}{\sigma\sqrt{2\pi}} & \mu_0 - \epsilon \leq x \leq \mu_0 + \epsilon, \\ f(x; \mu_0 + \epsilon, \sigma^2) & \mu_0 + \epsilon \leq x, \end{cases}$$

which yields a bound $k(x)$. Note that $f(x; \mu, \sigma^2)$ is a Lebesgue density and therefore $\mu(dx) = dx$ in condition (ii) in Theorem 7.3. Condition (ii) becomes

$$\int |h(x)|\, k(x) dx = \int_{-\infty}^{\mu_0 - \epsilon} |h(x)| f(x; \mu_0 - \epsilon, \sigma^2) dx + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu_0 - \epsilon}^{\mu_0 + \epsilon} |h(x)| dx + \int_{\mu_0 + \epsilon}^{\infty} |h(x)| f(x; \mu_0 + \epsilon, \sigma^2) dx.$$

Since the middle integral over a compact range is finite by definition, condition (ii) can be checked via

$$\int |h(x)| f(x; \mu_0 - \epsilon, \sigma^2) dx < \infty \quad \text{and} \quad \int |h(x)| f(x; \mu_0 + \epsilon, \sigma^2) dx < \infty.$$

As all moments of the normal distribution are finite, we see that condition (ii) holds for all polynomially bounded mappings h.

We now turn to the Score Function applied to the standard deviation. For μ fixed, note that the fact that σ is a scale parameter implies for $\sigma \leq \sigma_0 + \epsilon$, for $\epsilon > 0$, that there exist $M > 0$ such that

$$|f(x; \mu, \sigma^2) - f(x; \mu, (\sigma_0 + \epsilon)^2)| \leq \begin{cases} f(x; \mu, (\sigma_0 + \epsilon)^2)| & x \leq -M + \mu \text{ and } M + \mu \leq x, \\ \frac{1}{(\sigma_0 + \epsilon)\sqrt{2\pi}} & -M + \mu \leq x \leq M + \mu. \end{cases}$$

Following the line of argument for the mean μ, we see that condition (ii) for differentiation with respect to σ at σ_0 can be checked via

$$\int |h(x)| f(x; \mu, (\sigma_0 + \epsilon^2)) dx < \infty.$$

$\divideontimes\divideontimes\divideontimes$

In the following we present a worked out example of a more demanding application.

EXAMPLE 7.12. This is an example of what is known as the *optimal stopping* problem. Suppose you are selling a property. Offers come at random times. If you accept the offer then you sell the property. Otherwise you keep the property and wait for future offers. During the time that you keep the property, there are expenses incurred at a rate of c dollars per unit time. The amounts of the offers are iid positive random variables $\{Q_i\}$ with a bounded density $f(q)$, so that $\mathbb{E}(Q) < \infty$. We will assume that the intervals between offers are given by $\{T_i\}$, a sequence of iid random variables with unknown distribution.

You wish to determine a policy for stopping as follows. Let θ be your decision variable, and consider the rule :

$$X_\theta = \min\{n \colon Q_n \geq \theta\},$$

then you sell your property at this time, with a corresponding cost of:

$$J(\theta) = c \sum_{i=1}^{X_\theta} T_i - Q_{X_\theta}.$$

We wish to find the optimal threshold policy that has best expected cost, that is,

$$\min_\theta \mathbb{E}\left[c \sum_{i=1}^{X_\theta} T_i - Q_{X_\theta} \right].$$

This is a complex problem to solve, particularly when various distributions are not known analytically, and statistics or consecutive observations of the random variables must be used concurrently with the optimization. We are interested in estimating the derivative:

$$\frac{d}{d\theta} J(\theta) = c \frac{d}{d\theta} \mathbb{E}\left[c \sum_{i=1}^{X_\theta} T_i) \right] - \mathbb{E}[Q_{X_\theta}] \stackrel{\text{def}}{=} c g(X_\theta) - \mathbb{E}[Q_{X_\theta}].$$

The problem can be greatly simplified by noticing that X_θ has a geometric distribution. To see this, let $\xi_n = \mathbf{1}_{\{Q_i \geq \theta\}}$, then $\{\xi_n\}$ are iid Bernoulli random variables with

$$p_\theta \overset{\text{def}}{=} \mathbb{P}(\xi = 1) = \int_\theta^\infty f(q)\, dq,$$

where $f(\cdot)$ is the density (assumed to exist) of the offer amount. Then X_θ is the index of the first "success" in the sequence $\{\xi_n\}$, which has a Geometric distribution. On the other hand, by construction, Q_{X_θ} has a density which is the conditional density $f(q)/p_\theta$, on $[\theta, \infty)$. That is, the distribution of Q_{X_θ} is independent of X_θ. This helps to solve the problem estimating separately the derivatives of $g(\theta)$ and of $\mathbb{E}(Q_{X_\theta})$, which we now do.

The density function $f(q)$ is assumed to be known, so it may be possible to evaluate its expectation $\mathbb{E}(Q_{X_\theta})$ as a function of θ and then take the derivative. However for complicated distributions the integral may have to be calculated numerically, and derivatives may be difficult to evaluate. Alternatively, we can use:

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}(Q_{X_\theta}) &= \frac{d}{d\theta}\left(\frac{1}{p_\theta}\int_\theta^\infty q\, f(q)\, dq\right) = -\frac{p'_\theta}{p_\theta^2}\int_\theta^\infty q\, f(q)\, dq + \frac{1}{p_\theta}\theta f(\theta)\\ &= \frac{1}{p_\theta}\left(\theta\, f(\theta) - p'_\theta\, \mathbb{E}(Q_{X_\theta})\right),\end{aligned}$$

so that $(\theta\, f(\theta) - p'_\theta\, Q_{X_\theta})/p_\theta$ is an unbiased estimator for the derivative. See Exercise 7.4 for an instance of this problem where Q_i has Weibull distribution. In that example, $p_\theta = 1 - F(\theta)$ is known analytically, but the moments of Q_{X_θ} are not available in closed form.

We now turn to the problem of estimating the derivative of $g(X_\theta)$. The Geometric distribution is given by:

$$\mathbb{P}(X = n) = (1 - p_\theta)^{n-1}\, p_\theta.$$

To apply Theorem 7.3, let μ be the measure on the integers corresponding to a Geometric distribution with parameter $\mu \in (0, 1)$. To see that that distribution of X_θ has a μ-density, notice that for any set $A \subset \mathbb{N}$ we have:

$$\mathbb{P}(X_n \in A) = \sum_{n \in A}\left(\frac{(1-p_\theta)^{n-1}\, p_\theta}{(1-\mu)^{n-1}\, \mu}\right)(1-\mu)^{n-1}\, \mu,$$

so that the density is:

$$f_\theta(n) = \frac{(1-p_\theta)^{n-1}\, p_\theta}{(1-\mu)^{n-1}\, \mu},$$

which is differentiable in $\Theta = (0, \infty)$ for each integer n. Weak Lipschitz continuity holds ffor any measurable mapping g on the integers, such that for any α in a small neighborhood of $\theta \in \Theta$ it holds:

$$\mathbb{E}[X_\alpha\, |g(X_\alpha)|] = \sum_{n \in \mathbb{N}} n|g(n)|\, p_\alpha(n) < \infty. \tag{7.14}$$

Indeed, if this is the case then using the Mean Value Theorem as is Lemma 7.3 we can bound:

$$|g(n)(f_{\theta+\delta}(n) - f_\theta(n)| \leq |g(n)\delta f'_\alpha(n),$$

for $\theta \leq \alpha \leq \theta + \delta$. The derivative of the geometric density is:

$$f'_\theta(n) \overset{\text{def}}{=} \frac{\partial}{\partial\theta}f_\theta(n) = -p'_\theta\frac{(n-1)(1-p_\theta)^{n-2} + (1-p_\theta)^{n-1}}{(1-\mu)^{n-1}\, \mu} \tag{7.15}$$

We now show that $-|g|f'_\alpha$ has a bounded expectation w.r.t. μ, which is condition (ii) of Theorem 7.3. Indeed,

$$\begin{aligned} -\sum_{n\in\mathbb{N}} |g(n)|f'(\alpha)\mu(n) &= p'_\alpha \left(\sum_{n\in\mathbb{N}} (n-1)(1-p_\alpha)^{n-2}|g(n) + \sum_{n\in\mathbb{N}} (1-p_\alpha)|g(n)| \right) \\ &= \frac{p'_\alpha}{p_\alpha(1-p_\alpha)} \mathbb{E}[(X_\alpha - 1)|g(X_\alpha)|] + \frac{p'_\alpha}{p_\alpha}\mathbb{E}[|g(X_\alpha)|] < \infty. \end{aligned}$$

To apply Theorem 7.3 for the function $g(n) = \sum_{i=1}^n T_i$, we only need to verify condition (7.14). Using conditional expectations and independence of $\{T_i\}$ from X_θ,

$$\mathbb{E}\left[X_\theta \left(\sum_{i=1}^{X_\theta} T_i \right) \right] = \mathbb{E}\left[X_\theta \, \mathbb{E}\left[\sum_{i=1}^{X_\theta} T_i \,|\, X_\theta \right]\right] = \mathbb{E}\left[X_\theta \, \mathbb{E}[X_\theta \bar{T}] \right] = \bar{T}\, \mathbb{E}[X_\theta^2] < \infty,$$

where $\bar{T} = \mathbb{E}[T_i]$. A Geometric random variable has finite moments, which is why $\mathbb{E}[X_\theta^2] < \infty$ for all $\theta \in \Theta$. It is straightforward to calculate the Score Function:

$$S(\theta, X_\theta) = \frac{\partial}{\partial\theta} \ln f_\theta(X_\theta) = \frac{\partial}{\partial\theta} \left((X_\theta - 1)\ln(1-p_\theta) + \ln p_\theta \right) = -p'_\theta \left(\frac{(X_\theta - 1)}{1-p_\theta} + \frac{1}{p_\theta} \right)$$

To finalize, the unbiased derivative estimator using SF is

$$J'_{\mathrm{SF}}(\theta) = -p'_\theta \left(\frac{(X_\theta - 1)}{1-p_\theta} + \frac{1}{p_\theta} \right) \left(c \sum_{i=1}^{X_\theta} T_i \right) - \frac{1}{p_\theta}\left(\theta\, f(\theta) - p'_\theta\, Q_{X_\theta} \right).$$

⁂

7.2.2 Products of Measures

In the following we will discuss applying the score function method to vectors of random variables, or, equivalently, to products of measures. Let f_θ be a μ-density of probability measure ν_θ and let h_θ be a μ-density of probability measure η_θ. Provided that f_θ and h_θ are μ-almost surely differentiable it holds μ-almost surely that

$$\begin{aligned} \frac{\partial}{\partial\theta}(f_\theta(x)h_\theta(x)) &= \left(\frac{\partial}{\partial\theta} f_\theta(x) \right) h_\theta(x) + f_\theta(x) \frac{\partial}{\partial\theta} h_\theta(x) \\ &= \left(\frac{\partial}{\partial\theta}\log(f_\theta(x)) + \frac{\partial}{\partial\theta}\log(h_\theta(x)) \right) f_\theta(x)h_\theta(x). \end{aligned}$$

Denoting the score function of $f_\theta(x)$ by $S_f(\theta, x)$ and the score function of $h_\theta(x)$ by $S_h(\theta, x)$, we arrive at following computational rule for the score function of the product of f_θ and h_θ:

$$\frac{\partial}{\partial\theta}\log(f_\theta(x)h_\theta(x)) = S_f(\theta, x) + S_h(\theta, x). \tag{7.16}$$

In the light of (7.16) the extension of Theorem 7.3 to higher dimensional problems can be stated as follows, which provides the SF formula for the static gradient estimation problem.

Theorem 7.4 *Let μ be a measure. For $\theta \in \Theta$, let $\nu_i(\theta)$, for $1 \leq i \leq n$, be probability measures with μ-density $f_i(\theta)$. If*

(i) $f_{i,\theta}$ are μ-almost everywhere differentiable with respect to θ and have common support as mapping of θ, for $1 \leq i \leq N$,

(ii) for a measurable mapping h it holds that

$$|h(x_1, \ldots, x_N)| \left| \prod_{i=1}^{N} f_{i,\theta+\Delta}(x_i) - \prod_{i=1}^{N} f_{i,\theta}(x) \right| \leq |\Delta| k(x_1, \ldots, x_N),$$

for all Δ such that $|\Delta| < \epsilon$, where ϵ is some small number independent of x, and

$$\int k(x_1, \ldots, x_N) \prod_{i=1}^{N} \mu(dx_i) < \infty,$$

then

$$\frac{d}{d\theta} \int h(x_1, \ldots, x_N) \prod_{i=1}^{N} \nu_i(\theta, dx_i) = \int h(x_1, \ldots, x_N) \sum_{i=1}^{N} S_i(\theta, x_i) \prod_{i=1}^{N} \mu(dx_i),$$

where $S_i(\theta, x) = \frac{\partial}{\partial\theta} \log(f_{i,\theta}(x))$, for $1 \leq i \leq n$.

EXAMPLE 7.13. We revisit Example 7.5 and use the notation defined therein. The $X_n(\theta)$'s have Lebesgue density $f_\theta(x) = \exp(-x/\theta)/\theta$. The derivative of f_θ is given in (6.13) and it is straightforward to verify that that f_θ is Lipschitz continuous. By simple algebra it now follows that the n-fold product of the densities is Lipschitz continuous as well. By Theorem 7.4, it holds that

$$\begin{aligned}
\frac{d}{d\theta}\mathbb{E}[T_\theta(n)] &= \frac{d}{d\theta} \int \left(\sum_{i=1}^{n} x_i \right) \prod_{i=1}^{n} f_\theta(x_i)\, dx_1, \ldots, dx_n \\
&= \int \left(\sum_{i=1}^{n} x_i \right) \sum_{j=1}^{n} \left(\frac{\partial}{\partial\theta} f_\theta(x_j) \right) \prod_{i=1, i \neq j}^{n} f_\theta(x_i)\, dx_1, \ldots, dx_n \\
&= \int \left(\sum_{i=1}^{n} x_i \right) \sum_{j=1}^{n} S(\theta, x_j) \prod_{i=1}^{n} f_\theta(x_i)\, dx_1, \ldots, dx_n \\
&= \mathbb{E}\left[T_\theta(n) \sum_{j=1}^{n} S(\theta, X_j(\theta)) \right].
\end{aligned}$$

Like the IPA estimator, the SF estimator is thus unbiased. However, as we will show in the next chapter, SF yields an unbiased estimator for the Poisson counting process $N_\theta(T)$, for which we have already shown that IPA is biased; see Example 7.5.

⁂

EXAMPLE 7.14. Our final example considers again the sojourn time problem of Example 6.10. The function $L_N(X_1, \ldots, X_n)$ can be expressed as a function h depending only on the random variables

$A_1, \ldots, A_{N+1}; S_1, \ldots, S_N$. It may be a complicated function containing recursive computations, but clearly is it a function of those variables only. Thus the expectation is of the form

$$\mathbb{E}[L_N(X_1, \ldots, X_N)] = \int h(a_1, \ldots, a_{N+1}; x_1, \ldots, x_N) \prod_{j=1}^{N+1} g(a_i) \prod_{i=1}^{n} f_\theta(x_i)\, dx_1, \ldots, dx_n\,; da_1, \ldots da_n.$$

Suppressing the non-θ-denpendent arrival times in notation and letting $S_i = X_i$, for $1, \leq i \leq N$, the Score Function estimator is given by

$$\hat{G}^{\mathrm{SF}} = L_N(X_1, \ldots, X_n) \sum_{i=1}^{N} S(\theta, X_i) = \frac{1}{\theta} L_N(X_1, \ldots, X_n) \sum_{i=1}^{N} \left(\frac{X_i}{\theta} - 1 \right)$$

Comparing with Example 7.6 it should be apparent that developing the stochastic derivative in IPA is highly dependent on the function L, but the same expression applies to any distribution for which θ is a scale parameter. In contrast, the formula for the SF gradient estimator is applicable for any bounded function L (it may even have discontinuities).

It remains to be shown that $\hat{G}^{\mathrm{SF}}$ is unbiased. For this we use Theorem 7.4 and as differentiability of the product density is trivial, we turn to the Lipschitz continuity condition. As the density of the interarrival times does not depend on θ, it suffices to show Lipschitz continuity of the product of the service times densities. Using the Mean Value Theorem, we need to verify that there exists a $k(x_1, \ldots, x_n)$ such that

$$\sup_{\theta \in \Theta_0} \left| \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f_\theta(x_i) \right| \leq k(x_1, \ldots, x_n)$$

such that $|h| k_2$ is integrable. By computation,

$$\frac{\partial}{\partial \theta} \prod_{i=1}^{n} f_\theta(x_i) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} f_\theta(x_i) \right) \prod_{j=1, i \neq j}^{n} f_\theta(x_j)$$

Letting $[\theta_0, \theta_1]$ be a neighbourhood of θ and using the bound provided in Example 7.10 yields

$$\left| \frac{\partial}{\partial \theta} f_\theta(x) \right| \leq \left(\frac{x}{\theta_0^2} - 1 \right) \frac{1}{\theta_0^2} \exp(-x/\theta_1) \tag{7.17}$$

and

$$f_\theta(x) = \frac{1}{\theta} e^{-\theta x} \leq \frac{1}{\theta_0} e^{-\theta_1 x} = \frac{\theta_1}{\theta_0} f_{\theta_1}(x),\ x \geq 0.$$

Following the train of thought in Example 7.10 condition (ii) in Theorem 7.4 is satisfied provided that

$$\mathbb{E}\left[L_N(X_1(\theta), \ldots, X_n(\theta)) \sup_{\theta \in \Theta_0} \left| \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f_\theta(x_i) \right| \right] \leq \frac{\theta_1^n}{\theta_0^n} \mathbb{E}\left[L_N(X_1, \ldots, X_n) \sum_{i=1}^{n} \left(\frac{X_i}{\theta_0^2} - 1 \right) \right],$$

where X_i is exponentially distrbuted with mean θ_1. Noting that L_N can be bounded by the summing the interrival and service times

$$L_N(X_1, \ldots, X_N) \leq \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{i} (A_i + X_i),$$

and using the independence assumption, a sufficient condition for condition (ii) is that

$$\mathbb{E}\left[X_1\left(\frac{X_1}{\theta_0^2}-1\right)\right]<\infty,$$

for X_1 exponentially distributed with mean θ_1.

⁂

7.3 Measure-Valued Differentiation (MVD)

For $\theta \in \Theta \subset \mathbb{R}$, let μ_θ denote a probability measure, and denote the set of absolutely μ_θ integrable mappings for any $\theta \in \Theta$ by $L^1(\mu_\theta, \Theta)$, i.e.,

$$h \in L^1(\mu_\theta,\Theta) \quad \Leftrightarrow \quad \forall \theta \in \Theta : \quad \int |h(x)|\mu_\theta(dx) < \infty.$$

A measure is called a *signed measure* if it assigns negative mass to certain measurable sets.

Definition 7.4 *Let $\mathcal{D} \subset L^1(\mu_\theta, \Theta)$. The probability measure μ_θ is called $\mathcal{D}$-*differentiable *if a signed measure μ'_θ exists such that for all $h \in \mathcal{D}$*

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\left(\int h(s)\mu_{\theta+\Delta}(ds)-\int h(s)\mu_\theta(ds)\right)=\int h(s)\mu'_\theta(ds).$$

Let c_θ be a constant and μ_θ^+ and μ_θ^- two probability measures such that

$$\int h(s)\mu'_\theta(ds)=c_\theta\left(\int h(s)\mu_\theta^+(ds)-\int h(s)\mu_\theta^-(ds)\right),$$

then $(c_\theta,\mu_\theta^+,\mu_\theta^-)$ is called a $\mathcal{D}$-derivative of μ_θ.

The fact that μ'_θ is a signed measure, allows to write it as the difference between two positive measures. This fact is known as *Hahn-Jordan decomposition*. In the case that μ_θ has a differentiable density, the Hahn-Jordan decomposition can be constructed explicitly. To see this, let f_θ be the Lebesgue density of μ_θ such that f_θ is differentiable with respect to θ, and let $\mathcal{D} \subset L^1(\mu_\theta,\Theta)$ be such that for all $h \in \mathcal{D}$ it holds that

$$\frac{d}{d\theta}\int h(x)f_\theta(x)dx=\int h(x)\frac{\partial}{\partial\theta}f_\theta(x)\,dx, \tag{7.18}$$

see Theorem 7.3 for sufficient conditions. Then, with c_θ, f_θ^+ and f_θ^- as constructed in (6.9), (6.10) and (6.11) in Section 6.2, (7.18) implies

$$\forall h \in \mathcal{D}: \qquad \frac{d}{d\theta}\int h(x)f_\theta(x)dx=c_\theta\left(\int h(x)f_\theta^+(x)\,dx-\int h(x)f_\theta^-(x)\,dx\right). \tag{7.19}$$

In words, μ_θ is $\mathcal{D}$-differentiable, with $\mathcal{D}$ derivative $(c_\theta,\mu_\theta^+,\mu_\theta^-)$, where $\mu_\theta^\pm$ has density $f_\theta^\pm$. In the above construction the support of the measures μ_θ^+ and μ_θ^- is disjunct (revisit Figure 6.3 for illustration). More specifically, there exists a measurable set A such that for any measurable set B either (i) $\mu_\theta^+(B\setminus A)>0$ and $\mu_\theta^-(B\setminus A)=0$, or (ii) $\mu_\theta^+(B\cap A)=0$ and $\mu_\theta^-(B\cap A)>0$. This implies that $c_\theta\mu_\theta^+$ together with $c_\theta \mu_\theta^-$ is the Hahn-Jordan decomposition of μ'_θ. Note that *any* decomposition of $df_\theta/d\theta$ into the difference of two densities will yield a weak derivative representation and, having simulation in mind, one seeks a representation that leads to ease of coding. Fortunately, for most distributions used in applications, convenient representations can be found as we will show later on in the text.

The above construction presupposes differentiability of f_θ and is thus in essence equivalent to SF with the only difference that in SF $df_\theta/d\theta$ is scaled by $1/f_\theta$ yielding the score function and in MVD it is written as the difference between two densities. To understand the power of the concept of weak differentiation it is important to realize that differentiability of f_θ is not essential for this approach. This is explained in the following example.

EXAMPLE 7.15. Let U_θ be uniformly distributed on $[0,\theta]$. Letting $U_\theta = \theta U$, with U uniformly distributed on $[0,1]$, the sample path derivative becomes

$$\frac{d}{d\theta} U_\theta = U = \frac{1}{\theta} U_\theta .$$

For the distributional approach, note that U_θ has density

$$f_\theta(x) = \frac{1}{\theta}\, \mathbf{1}_{\{x \in [0,\theta]\}},$$

which fails to be differentiable with respect to θ, and consequently SF doesn't apply. For MVD, note that for any continuous mapping h basic analysis yields

$$\begin{aligned}
\frac{d}{d\theta}\int h(x) f_\theta(x)dx &= \frac{d}{d\theta}\left(\frac{1}{\theta}\int_0^\theta h(x)dx\right) = -\frac{1}{\theta^2}\int_0^\theta h(x)dx + \frac{1}{\theta}h(\theta) \\
&= \frac{1}{\theta}\left(h(\theta) - \frac{1}{\theta}\int_0^\theta h(x)dx\right) \\
&= \frac{1}{\theta}\left(h(\theta) - \int_0^\theta h(x) f_\theta(x)dx\right).
\end{aligned}$$

Hence, we obtain for the uniform distribution the estimators

$$\begin{aligned}
\psi^{\mathrm{IPA}}(h, U(\theta), \theta) = & \quad \frac{U_\theta}{\theta}\, h'(U_\theta), \\
\psi^{\mathrm{MVD}}(h, U(\theta), \theta) = & \quad \frac{1}{\theta}\,(h(\theta) - h(U_\theta)).
\end{aligned}$$

In summary, for this example

$$\text{IPA} \qquad \frac{1}{\theta}\mathbb{E}\big[U_\theta h'(U_\theta)\big] = \frac{d}{d\theta}\mathbb{E}\big[h(U_\theta)\big] = \begin{cases} \text{not applicable} & \text{SF} \\ & \\ \frac{1}{\theta}\mathbb{E}\big[h(\theta) - h(U_\theta)\big] & \text{MVD.} \end{cases}$$

Note that for IPA to be unbiased h has to differentiable, whereas for the weak differentiation approach continuity of h is sufficient.

***

The set $L^1(\mu_\theta, \Theta)$ is the maximal set of mappings h having the property that the integral $\int h(x)\mu_\theta(dx)$ is differentiable. In practice, the set of mappings that are feasible for differentiation is smaller than

$L^1(\mu, \theta)$. Recall, for example, that the uniform distribution is differentiable with respect to continuous functions only; see Example 7.15.

In applications, one usually has a certain class of mappings in mind. For example, one might be interested in continuous mappings only, denoted by $\mathcal{C}$, or in the much wider class of measurable mappings, denoted by $\mathcal{B}$. Let $\mathcal{H} \subset L^1(\mu_\theta, \Theta)$ denote the range of mappings one is interested in from the onset. To characterize analytical properties of $\mathcal{H}$ it is helpful to choose a mapping $v \in L^1(\mu_\theta, \Theta)$, and consider the set

$$\mathcal{D}_v \overset{\text{def}}{=} \mathcal{D}_v(\mathcal{H}) = \{h \in \mathcal{H} : |h(x)| \leq cv(x) \text{ for all } x \text{ and some finite constant } c\}.$$

The set $\mathcal{H}$ is called the *base set* (usually $\mathcal{C}$ or $\mathcal{B}$). Notice that the subset $\mathcal{D}_v$ corresponds to all functions $h \in \mathcal{H}$ with growth limited by v, that is, functions that are $\mathcal{O}(v(x))$. A typical choice for v in applications, is

$$v_p(x) = 1 + |x|^p,$$

for $p \in \mathbb{N}$. Generally speaking, for any $v(x)$ such that $v(x) \geq 1$ for all $x \in \mathbb{R}$ (respectively, $x \in [0, \infty)$) the *weighted supremum norm*, or v-norm for short, of a real-valued mapping h is defined by

$$\sup_x \frac{|h(x)|}{v(x)}.$$

Note that $\mathcal{D}_v$ is the set of all $h \in \mathcal{H}$ with finite v_p-norm. Furthermore, note that $\mathcal{C}_{v_0} \overset{\text{def}}{=} \mathcal{D}(\mathcal{C})_{v_0}$ is the set of continuous and bounded mappings. For example, the exponential distribution is weakly differentiable with respect to the set $\mathcal{B}_{v_p} \overset{\text{def}}{=} \mathcal{D}(\mathcal{B})_{v_p}$ for any p, which stems from the fact that (6.14) holds for any polynomially bounded h. In the next chapter we will encounter another choice for v that is useful when differentiating stationary characteristics, and this is $v_\alpha(x) = \alpha^x$.

In the light of the above discussion, application of MVD to a probability measure μ_θ leads to the following question: Given a family of distributions $\{\mu_\theta\}$, what is the largest set of cost functions $\mathcal{D}$ for which the derivative $(c_\theta, \mu_\theta^+, \mu_\theta^-)$ can be used for simulation? Fortunately, this question is easy to answer for distributions common in applications and an overview is provided in the following.

Table 7.1 summarizes representations of weak derivatives that are straightforward to use in applications. We also state the range of mappings the weak derivative applies to. The term ds-Maxwell (m,s^2) denotes the double-sided Maxwell distribution with mean μ and shape parameter s having density

$$\frac{1}{s^3\sqrt{2\pi}}(x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{s}\right)^2}, \quad x \in \mathbb{R}.$$

Furthermore, by [Gamma(2, θ)]$^{1/\alpha}$ we denote the distribution of the α-root of a Gamma (α, θ) random variable, and (Poisson (θ)+1) denotes the shifted Poisson distribution, i.e., a sample from this distribution is obtained by adding 1 to a sample of the Poisson(θ) distribution. By Pareto' (θ, β) we denote the Pareto type I distribution with cdf $1 - (\theta/x)^\beta$, $x \geq \theta$. And finally, $\theta\pm$ Weibull(2,1/2 σ^2) denotes the distribution of a Weibull(2,1/2 σ^2) random variable W that is transformed by $\theta \pm W$.

Expect for the Pareto I distribution, all distributions list in Table 7.1 are $\mathcal{D}$-differentiable with respect to

$$\mathcal{D} = \{h \in \mathcal{H} : ||h||_v < \infty\},$$

where v may be v_p, for any $p \in \mathbb{N}$, as well as $v = v_\alpha$, and $\mathcal{H}$ as indicated in the utmost right column in Table 7.1. The Pareto I distribution is only differentiable for v_p with $p \leq \beta$.

Table 7.1: Differentiability of Common Distributions

μ_θ	c_θ	μ_θ^+	μ_θ^-	$\mathcal{H}$
Bernoulli(θ) on $\{0,1\}$	1	Dirac(0)	Dirac(1)	$\mathcal{B}$
Poisson(θ)	1	Poisson(θ)+1	Poisson(θ)	$\mathcal{B}$
Uniform $(0,\theta)$	$1/\theta$	Dirac(θ)	Uniform $(0,\theta)$	$\mathcal{C}$
Normal(θ,σ^2)	$1/\sigma\sqrt{(2\pi)}$	θ + Weibull(2,1/2 σ^2)	θ - Weibull(2,1/2σ^2)	$\mathcal{B}$
Normal(m,θ^2)	$1/\theta$	ds-Maxwell(m,θ^2)	Normal(m,θ^2)	$\mathcal{B}$
Exponential (θ)	$1/\theta$	Gamma($2,\theta$)	Exponential(θ)	$\mathcal{B}$
Gamma (α,θ)	α/θ	Gamma($\alpha+1,\theta$)	Gamma(α,θ)	$\mathcal{B}$
Pareto I (θ,β)	β/θ	Pareto I (θ,β)	Dirac(θ)	$\mathcal{C}$
Weibull(α,θ)	$1/\theta$	Weibull(α,θ)	$[\text{Gamma}(2,\theta)]^{1/\alpha}$	$\mathcal{B}$

REMARK. Recall that by Lemma 7.1 under appropriate smoothness, theIPA derivative of $X(\theta)$, denoted by $X'(\theta)$, can be obtained from

$$-\frac{\frac{\partial}{\partial\theta}F_\theta(X(\theta))}{f_\theta(X(\theta))},$$

which can be written as

$$\mathbb{E}[X'(\theta)|X(\theta)] = -\frac{\frac{\partial}{\partial\theta}F_\theta(X(\theta))}{f_\theta(X(\theta))}.$$

Using the MVD expression

$$\frac{\partial}{\partial\theta}F_\theta(x) = c_\theta(F_\theta^+(x) - F_\theta^-(x))$$

we obtain

$$\mathbb{E}[X'(\theta)|X(\theta)] = -\frac{c_\theta}{f_\theta(X(\theta))}(F_\theta^+(X(\theta)) - F_\theta^-(X(\theta))).$$

Alternatively, from $F_\theta(x) = \mathbb{E}[\mathbf{1}_{\{X(\theta)\le x\}}]$ we obtain the SF representation

$$\mathbb{E}[X'(\theta)|X(\theta)] = -\frac{c_\theta}{f_\theta(X(\theta))}\mathbb{E}\left[\mathbf{1}_{\{\tilde{X}(\theta)\le X(\theta)\}}\mathrm{SF}_\theta(\tilde{X}(\theta))\right],$$

where $\tilde{X}(\theta)$ is an iid copy of $X(\theta)$.

EXAMPLE 7.16. We calculate here the MVD estimator for Example 6.9. In this example, we have five independent random variables $T_i \sim G_i, i \neq 3$ are independent of θ, and $T_3 \sim$ Exponential $(1/\theta)$. Let $T = (T_1, \ldots, T_5)$ and call $g_i(t_i)$ the density of G_i. Then the joint density is given by:

$$f(t) = f(t_1, \ldots t_5) = \frac{1}{\theta}\, e^{-t_3/\theta} \prod_{i\neq 3} g_i(t_i).$$

Therefore, for any function $L\colon \mathbb{R}^5 \to \mathbb{R}$ that is absolutely integrable: $\mathbb{E}[|L(T)|] < \infty$,

$$\mathbb{E}[L(\theta,T)] = \int_t \left(L(t_1,\ldots,t_5) \prod_{i\neq 3} g_i(t_i) \right) \frac{1}{\theta}\, e^{-t_3/\theta}\, dt. \tag{7.20}$$

Because the term in brackets is independent of θ and has a finite integral, we can interchange derivative w.r.t. θ and expectation. This follows from the fact that the exponential density has an absolutely bounded derivative, provided that $0 < \theta < \infty$, that is, assuming that the values of θ of interest are contained in a finite and positive interval. In terms of the bounding function $v_p(x)$, we use the fact that the exponential density has bounded moments, so for *any* p, if $|L(t)| \leq v_p(x)$, then the MVD estimator is unbiased. For this example, the life of the system is linear in T_3. By direct differentiation, we have:

$$\frac{d}{d\theta}\left(\frac{1}{\theta}\, e^{-t_3/\theta}\right) = \frac{1}{\theta}\left(\frac{t_3}{\theta^2} e^{-t_3/\theta} - \frac{1}{\theta}\, e^{-t_3/\theta}\right)$$

which corresponds to a scaled difference between the Gamma$(2, 1/\theta)$ and the Exponential $1/\theta)$ diistribution. The corresponding MVD estimator is:

$$\widehat{L}^{\text{MVD}} = \frac{1}{\theta}\Big(L(T_1, T_2, T_3(\theta) + X(\theta), T_4, T_5) - L(T_1, T_2, T_3(\theta), T_4, T_5)\Big), \tag{7.21}$$

with $X(\theta) \sim$ Exponential$(1/\theta)$.

This is not the only way to represent this derivative. For example, instead of using $T_3 + X$, any other random variable with Gamma distribution will yield the same expectation for L. As well, any other decomposition for the derivative of the density as a scaled difference of densities will work as well. Remark that if we want to estimate the derivative of the *variance* of L, for example, then MVD is also applicable, because it is polynomially bounded. Moreover, the estimator has the same form, by appropriately modifying the definition of L above.

Finally, comparing with expression (7.32), we note the following. The IPA formula established in (7.32) is unbiased when θ is a scale parameter of the distribution and does not require explicit knowledge of the distribution of T_3. In contrast, (7.21) is valid for $T_3 \sim \exp(1/\theta)$, and it requires modifying when T_3 has a different distribution. However, it also applies to other performance criteria. For example, if instead of the life of the system L we were interested in the risk measure $\mathbb{P}(L \leq \ell)$, for some given time ℓ, then (7.21) is applicable replacing L by $\mathbf{1}_{\{L \leq \ell\}}$. In contrast, the corresponding IPA formula needs to be re-evaluated for different performance functions (actually, IPA is biased for the risk measure, as indicated in Exercise 7.13).

⁂

Given that μ_θ and ν_θ are weakly differentiable, does it then hold that the product probability measure $\mu_\theta \times \nu_\theta$ is weakly differentiable as well? In case that μ_θ and ν_θ have differentiable densities, this question can be answered along the lines detailed in the proof of Theorem 7.4. In the general case the proof becomes more elaborate. However, if the functional space $(\mathcal{D}_v, ||\cdot||_v)$ is a Banach space, i.e., a complete normed space, then weak differentiability of product measures can be established in a mathematically comprehensive way. The supporting theory for this is provided in the next section.

An interesting feature of MVD is that $\mathcal{D}$-differentiability of products of probability measures can be deduced from that of the elements of the products without further assumptions, provided that the set $\mathcal{D}$ is well-chosen. For the analysis we require an extension of the v-norm defined in the previous section to measures.

Definition 7.5 *Let μ be a measure on $\mathbb{R}$, and let $v(x) \geq 1$ for all x in the support of μ. The* weighted supremum norm, *or v-*norm *for short, is defined as*

$$||\mu||_v = \int v(x)|\mu|(dx).$$

Note that $|\mu|(\cdot)$ denotes the absolute measure. In case that μ is a positive measure, such as a probability measure, it holds that $|\mu| = \mu$, in case that μ is a signed measure we consider the Hahn-Jordan decomposition of μ and let

$$\int v(x)|\mu|(dx) = \int v(x)\mu^+(dx) + \int v(x)\mu^-(dx).$$

The following result shows that under appropriate conditions, weak differentiability implies norm Lipschitz continuity, which is a property that helps simplifying the technical analysis as well as illustrating the weak differentiation behaves "almost" as strong (i.e., normwise) differentiability.

Lemma 7.4 *Let μ_θ be $\mathcal{D}$-differentiable, such that $\mathcal{D}$ equipped with norm $|| \cdot ||_v$ becomes a Banach space. Then a finite constant M exists such that for all Δ such that $\theta + \Delta \in \Theta$ it holds that*

$$||\mu_{\theta+\Delta} - \mu_\theta||_v \leq |\Delta| M.$$

In words, $\mathcal{D}$-differentiability implies Lipschitz continuity.

Proof: Using the shorthand notation $\langle h, \mu_\theta \rangle$ for the μ_θ integral of h, it holds, under the assumption in the lemma, that the sequence

$$\frac{1}{\Delta}(\langle h, \mu_{\theta+\Delta}\rangle - \langle h, \mu_\theta \rangle)$$

converges for any h in $\mathcal{D}$ as Δ tends to zero. Hence,

$$\sup_{\Delta \neq 0} \left| \frac{1}{\Delta}(\langle h, \mu_{\theta+\Delta}\rangle - \langle h, \mu_\theta \rangle) \right| < \infty$$

for any $h \in \mathcal{D}$. The Banach Steinhaus Theorem then implies that the above set is also bounded in norm-sense, i.e.,

$$\sup_{\Delta \neq 0} \left\| \frac{1}{\Delta}(\mu_{\theta+\Delta} - \mu_\theta) \right\|_v \stackrel{\text{def}}{=} M < \infty,$$

which proves the claim.

QED

7.3.1 Differentiability of Products of Measures

We now turn to the proof of the product rule for differentiation. It establishes that the "usual" chain rule for functions also applies to measure differentiation. Consider a set $\mathcal{T}$ of real-valued mappings on $\mathbb{R}^2$. Furthermore, let μ and ν be (probability) measures on $\mathbb{R}$, and denote the product measure on $\mathbb{R}^2$ by $\mu \otimes \nu$. Note that if μ has Lebesgue density f and if ν has Lebesgue density g, then $\mu \otimes \nu$ has Lebesgue density $f(x)g(y)$ for $(x, y)^\top \in \mathbb{R}^2$. The product rule will have to answer under what conditions $\mu \otimes \nu$ is differentiable provided μ and ν are. As differentiability is denfied in the weak sense and thus relative to the set of test functions, it may happen that μ is $\mathcal{D}^\mu$ differentiable and ν is $\mathcal{D}^\nu$-differentiable with $\mathcal{D}^\mu \neq \mathcal{D}^\nu$. The set of mappings the product measure can differentiate will thus be a subset of the mappings $h \in \mathcal{T}$ so that $h(\cdot, y) \in \mathcal{D}^\mu$ for all y and $h(x, \cdot) \in \mathcal{D}^\nu$ for all x. As it turns out the v-norm is a nice tool for handling this situation. While the product rule can be established in a more general setting, see [18], we will prove the statement in a simpler form that is sufficient for the current text. For this, we write $f \otimes g$ for the product of real-valued mappings f and g, i.e.,

$(f \otimes g)(x, y) = f(x)g(y)$, for $x, y \in \mathbb{R}$. The following product rule of weak differentiation can be obtained, where we use the fact that $(\mathcal{D}, || \cdot ||_v)$ is a Banach space, if $||h||_v < \infty$ for any $h \in \mathcal{D}$. For Banach spaces $(\mathcal{D}^\mu, || \cdot ||_v)$ and $(\mathcal{D}^\nu, || \cdot ||_w)$, we call $(\mathcal{D}, || \cdot ||_{v\otimes w})$ the corresponding product space if for all $h \in \mathcal{D}$ it holds that $h(\cdot, y) \in \mathcal{D}^\mu$ for all y and $h(x, \cdot) \in \mathcal{D}^\nu$ for all x.

Theorem 7.5 *Let $(\mathcal{D}^\mu, || \cdot ||_v)$ and $(\mathcal{D}^\nu, || \cdot ||_w)$ be Banach spaces on $\mathbb{R}$, with corresponding product space $(\mathcal{D}, || \cdot ||_{v\otimes w})$.*

If μ_θ is $\mathcal{D}^\mu$-differentiable and ν_θ is $\mathcal{D}^\nu$-differentiable, then the product measure $\mu_\theta \otimes \nu_\theta$ is $\mathcal{D}$-differentiable, that is, for any $h \in \mathcal{D}$ it holds that

$$\frac{d}{d\theta} \int h(x, y)\mu_\theta(dx)\nu_\theta(dy) = \int h(x, y)\mu'_\theta(dx)\nu_\theta(dy) + \int h(x, y)\mu_\theta(dx)\nu'_\theta(dy).$$

Moreover, if in addition, μ_θ has $\mathcal{D}^\mu$-derivative $(c^\mu_\theta, \mu^+_\theta, \mu^-_\theta)$ and that ν_θ has $\mathcal{D}^\nu$-derivative $(c^\nu_\theta, \nu^+_\theta, \nu^-_\theta)$. Then, it holds that for any $h \in \mathcal{D}$ that

$$\begin{aligned} &\frac{d}{d\theta} \int h(x, y)\mu_\theta(dx)\nu_\theta(dy) \\ &= (c^\mu_\theta + c^\nu_\theta)\Bigg(\frac{c^\mu_\theta}{c^\mu_\theta + c^\nu_\theta} \int h(x, y)\mu^+_\theta(dx)\nu_\theta(dy) + \frac{c^\nu_\theta}{c^\mu_\theta + c^\nu_\theta} \int h(x, y)\mu_\theta(dx)\nu^+_\theta(dy) \\ &\qquad - \frac{c^\mu_\theta}{c^\mu_\theta + c^\nu_\theta} \int h(x, y)\mu^-_\theta(dx)\nu_\theta(dy) + \frac{c^\nu_\theta}{c^\mu_\theta + c^\nu_\theta} \int h(x, y)\mu_\theta(dx)\nu^-_\theta(dy)\Bigg). \end{aligned}$$

Proof: For Δ such that $\theta + \Delta \in \Theta$, set

$$\bar{\mu}_\Delta = \frac{\mu_{\theta+\Delta} - \mu_\theta}{\Delta} - \mu'_\theta; \quad \bar{\nu}_\Delta = \frac{\nu_{\theta+\Delta} - \nu_\theta}{\Delta} - \nu'_\theta.$$

To simplify notation we write $\mu_n \overset{\mathcal{H}}{\Longrightarrow} \nu$ for

$$\lim_{n\to\infty} \langle h, \mu_n \rangle = \langle h, \nu \rangle, \qquad \text{for all } h \in \mathcal{H}.$$

By hypothesis, $\bar{\mu}_\Delta \overset{\mathcal{D}^\mu}{\Longrightarrow} \varnothing$ and $\bar{\nu}_\Delta \overset{\mathcal{D}^\nu}{\Longrightarrow} \varnothing$, for $\Delta \to 0$, where $\varnothing$ denotes the null measure. Simple algebra shows that the proof of the claim follows from

$$\Delta(\bar{\mu}_\Delta + \mu'_\theta) \times (\bar{\nu}_\Delta + \nu'_\theta) + \mu_\theta \times \bar{\nu}_\Delta + \bar{\mu}_\Delta \times \nu_\theta \overset{\mathcal{D}}{\Longrightarrow} \varnothing, \tag{7.22}$$

for $\Delta \to 0$. Hence, to conclude the proof, we show that each term on the left side of (7.22) converges weakly to null measure $\varnothing$.

Since $\bar{\mu}_\Delta + \mu'_\theta \overset{\mathcal{D}^\mu}{\Longrightarrow} \mu'_\theta$ an $\bar{\nu}_\Delta + \nu'_\theta \overset{\mathcal{D}^\nu}{\Longrightarrow} \nu'_\theta$, applying Lemma 7.4 yields

$$\sup_{\Delta\in V\setminus\{0\}} \|\bar{\mu}_\Delta + \mu'_\theta\|_v < \infty \quad \text{and} \quad \sup_{\Delta\in V\setminus\{0\}} \|\bar{\nu}_\Delta + \nu'_\theta\|_w < \infty,$$

for any compact neighborhood V of 0. By simple algebra,

$$\begin{aligned} &\left|\Delta \int h(s, t)\big((\bar{\mu}_\Delta + \mu'_\theta) \times (\bar{\nu}_\Delta + \nu'_\theta)\big)(ds, dt)\right| \\ &\qquad \le |\Delta| \left|\int \frac{h(s, t)}{v(s)w(t)}\big((v(s)|(\bar{\mu}_\Delta + \mu'_\theta|) \times (\ w(t)|\bar{\nu}_\Delta + \nu'_\theta|)\big)(ds, dt)\right| \\ &\qquad \le |\Delta| \int ||h||_{v\otimes w}\big((v(s)|(\bar{\mu}_\Delta + \mu'_\theta|) \times (\ w(t)|\bar{\nu}_\Delta + \nu'_\theta|)\big)(ds, dt) \\ &\qquad \le |\Delta| \cdot \|h\|_{v\otimes u} \cdot \|\bar{\mu}_\Delta + \mu'_\theta\|_v \cdot \|\bar{\nu}_\Delta + \nu'_\theta\|_w. \end{aligned}$$

Letting $\Delta \to 0$ in the above inequality it follows that the first term in (7.22) converges weakly to $\varnothing$.

The second and the third terms in (7.22) are symmetric so they can be treated similarly. For instance, for the second term in (7.22) note that

$$\int h(s,t)(\mu_\theta \times \bar{\nu}_\Delta)(ds,dt) = \int \int h(s,t)\mu_\theta(ds)\, \bar{\nu}_\Delta(dt) = \int H_\theta(h,t)\bar{\nu}_\Delta(dt),$$

where $H_\theta(h,t) = \int h(s,t)\mu_\theta(ds)$ for all t and for all h. Therefore,

$$\forall t \in T: \ \frac{|H_\theta(h,t)|}{w(t)} \le \frac{\|h(\cdot,t)\|_v}{w(t)}\|\mu_\theta\|_v \le \|h\|_{v\otimes w}\, \|\mu_\theta\|_v,$$

where the second inequality follows from

$$\forall s \in S, t \in T: \ |h(s,t)| \le \|h\|_{v\otimes w} v(s)w(t);$$

Consequently, $H_\theta(h,\cdot) \in \mathcal{D}^\nu$, for $h \in (\mathcal{D}, ||\cdot||_{v\otimes w})$. We have assumed that ν_θ is $\mathcal{D}^\nu$-differentiable, which yields $\bar{\nu}_\Delta \overset{\mathcal{D}^\nu}{\Longrightarrow} \varnothing$. Hence,

$$\lim_{\Delta\to 0} \int H_\theta(h,t)\bar{\nu}_\Delta(dt) \to 0,$$

which shows that the second term in (7.22) converges weakly to $\varnothing$. This concludes the proof.

QED

REMARK. One of the virtues of the v-norm is that one can show that if $\mathcal{D}$ is either the set of continuous or measurable mappings from $\mathbb{R}^2$ to $\mathbb{R}$ with finite $v \otimes v$ norm, then $(\mathcal{D}, ||\cdot||_{v\otimes v})$ is the product space corresponding to $(\hat{\mathcal{D}}, ||\cdot||_v)$, with $\hat{\mathcal{D}}$ being the set of continuous or measurable mappings from $\mathbb{R}$ to $\mathbb{R}$ with finite v norm, see [18] for a proof.

The derivative representation in Theorem 7.5 can be expressed in terms of random variables as follows. Let the conditions of Theorem 7.5 be satisfied and let $X^\pm(\theta)$ have distribution $\mu_\theta^\pm$ and let $Y^\pm(\theta)$ have distribution $\nu_\theta^\pm$. Then it holds that

$$\begin{aligned} \frac{d}{d\theta}\mathbb{E}[h(X(\theta),Y(\theta))] \quad &= \quad (c_\theta^\mu + c_\theta^\nu)\left(\mathbb{E}\left[\frac{c_\theta^\mu}{c_\theta^\mu + c_\theta^\nu}h(X(\theta)^+,Y(\theta)) + \frac{c_\theta^\nu}{c_\theta^\mu + c_\theta^\nu}h(X(\theta),Y(\theta)^+)\right]\right. \\ &\qquad \left. - \,\mathbb{E}\left[\frac{c_\theta^\mu}{c_\theta^\mu + c_\theta^\nu}h(X(\theta)^-,Y(\theta)) + \frac{c_\theta^\nu}{c_\theta^\mu + c_\theta^\nu}h(X(\theta),Y(\theta)^-)\right]\right). \end{aligned}$$

Alternatively, one can introduce a random variable Z_θ^+ such that

$$Z_\theta^+ = \begin{cases} (X(\theta)^+, Y(\theta)) & \text{with probability } \frac{c_\theta^\mu}{c_\theta^\mu + c_\theta^\nu} \\ (X(\theta), Y(\theta)^+) & \text{with probability } \frac{c_\theta^\nu}{c_\theta^\mu + c_\theta^\nu} \end{cases}$$

and

$$Z_\theta^- = \begin{cases} (X(\theta)^-, Y(\theta)) & \text{with probability } \frac{c_\theta^\mu}{c_\theta^\mu + c_\theta^\nu} \\ (X(\theta), Y(\theta)^-) & \text{with probability } \frac{c_\theta^\nu}{c_\theta^\mu + c_\theta^\nu}. \end{cases}$$

With this notation it holds that

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta),Y(\theta))] = (c_\theta^\mu + c_\theta^\nu)\,\mathbb{E}\left[h(Z_\theta^+) - h(Z_\theta^-)\right].$$

Alternatively, we may let $Z_\theta^{[\sigma]} = Z_\theta^+$, for $\sigma = 1$, and $Z_\theta^{[\sigma]} = Z_\theta^-$, for $\sigma = 2$. For $\mathbb{P}(\sigma = 1) = 1/2 = \mathbb{P}(\sigma = 2)$, we have

$$\frac{d}{d\theta}\mathbb{E}[h(X(\theta),Y(\theta))] = 2(c_\theta^\mu + c_\theta^\nu)\,\mathbb{E}\left[h\left(Z_\theta^{[\sigma]}\right)\right].$$

Rather than developing the theroy of differentiation of n-fold products of probability measures we turn in the next section to the differentiation theory of Markov chains. Since the product of independent probability measures can be seen as a Markov chain, these results are contained in the presentation provided in the next section.

7.3.2 Differentiability of Markov Chains

The analysis of the previous section carries over to Markov chains P_θ as detailed in this section. For ease of presentation we establish the product rule of differentiation for Markov chains living on the same state space. Furthermore, we introduce the concept of a transition kernel, which is a generalization of the concept a Markov chain. More specifically, we call R a *transition kernel* if $R(x,\cdot)$ is a finite (possibly signed) measure for all x, and $R(\cdot, B)$ is a measurable mapping for all measurable sets B. Hence, a transition kernel R is a Markov chain when $R(x,\cdot)$ is a probability measure for all x.

We will illustrate the concepts and results on this section by means of the following example.

Example 7.17. We revisit our basic sojourn times example with the notation as introduced in Example 6.10. Recall that the n-th interarrival time is denoted by A_n and the n-th service time by $S_n(\theta)$. We assume that the service system starts initially empty and that the typical independence assumptions apply. Moreover, we denote the density of A_n by f_A and the density of $S_n(\theta)$ by $f_S(\theta)$, where we assume that $f_S(\theta) = (1/\theta)e^{-x/\theta}$ is the pdf of the exponential distribution with mean θ. Consecutive sojourn times $X_n(\theta)$ then follow the recursive relation:

$$X_{n+1}(\theta) = \max(0, X_n(\theta) - A_n(\theta)) + S_{n+1}(\theta),\ n \geq 1, \tag{7.23}$$

and $X_1(\theta) = S_1(\theta)$. We denote the transition kernel of the sojourn time chain by P_θ, i.e.,

$$P_\theta(\mathcal{A}, x) = \mathbb{P}(X_{n+1}(\theta) \in \mathcal{A} | X_n(\theta) = x)$$

for $x \geq 0$ and $\mathcal{A} \subset [0,\infty)$ a (Borel) measurable set, or, more specifically

$$P_\theta(\mathcal{A}, x) = \int_0^\infty \left(\int_0^\infty 1_{\{\max(x-a,0)+s \in \mathcal{A}\}} f_S(\theta, s)\,ds \right) f_A(a)\,da.$$

We denote by $\mathcal{D}_\alpha$ the set of all measurable mappings $g : [0,\infty) \mapsto \mathbb{R}$ with finite v_α norm. For $g \in \mathcal{D}_\alpha$, it holds

$$|g| \leq ||g||_\alpha \alpha^x,$$

for $x \in [0, \infty)$. With this inequality we obtain,

$$\begin{aligned}\int g(y)P_\theta(x,dy) &= \int_0^\infty \left(\int_0^\infty g(\max(x-a,0)+s) f_S(\theta,s)ds\right) f_A(a)da \\ &\le \int_0^\infty \left(\int_0^\infty ||g||_\alpha \alpha^{x+a+s} f_S(\theta,s)ds\right) f_A(a)da \\ &\le \alpha^x ||g||_\alpha \int_0^\infty \left(\int_0^\infty \alpha^{a+s} f_S(\theta,s)ds\right) f_A(a)da \\ &= \alpha^x ||g||_\alpha \mathbb{E}\left[\alpha^A\right]\mathbb{E}\left[\alpha^{S(\theta)}\right] < \infty, \end{aligned} \tag{7.24}$$

provided the interarrival times have a finite moment generating function (which is true for the exponential distribution). To summarize,

$$||P_\theta||_\alpha \le ||g||_\alpha \mathbb{E}\left[\alpha^A\right]\mathbb{E}\left[\alpha^{S(\theta)}\right]$$

and

$$\int g(y)P_\theta(\cdot,dy) \in \mathcal{D}_\alpha.$$

⁂

As next, we present the extension of the concept of $\mathcal{D}$-differentiability of a measure to Markov chains.

Definition 7.6 *We call a Markov chain P_θ is $\mathcal{D}$-differentiable with respect to θ with $\mathcal{D}$-derivative P_θ if a transition kernel P'_θ exists such that for all x and $h \in \mathcal{D}$ it holds that*

$$\frac{d}{d\theta}\int h(y)P_\theta(x,dy) = \int h(y)P'_\theta(x,dy).$$

Furthermore, we call the triple $(c_\theta(\cdot), P^+_\theta, P^-_\theta)$ a $\mathcal{D}$-derivative of P_θ if for all x and all $h \in \mathcal{D}$

$$\int h(y)P_\theta(x,dy) = c_\theta(x)\left(\int h(y)P^+_\theta(x,dy) - \int h(y)P^-_\theta(x,dy)\right),$$

where $c_\theta(\cdot)$ is a measurable mapping and $P^\pm_\theta$ are Markov kernels.

We illustrate the concept of the measure-valued derivative of a Markov chain with our sojourn time example.

EXAMPLE 7.18. Represent $f_S(\theta)'$ by the triple $(c_\theta, f^+_S(\theta), f^-_S(\theta)$, see (6.13) for the case of the exponential density and (7.19) for the general principle. As detailed in Example 6.7

$$\frac{\partial}{\partial\theta} f_S(\theta,x) = \frac{1}{\theta}\Big(f^\gamma_S(\theta,x)\ - f_S(\theta,x)\Big),$$

where $f^\gamma(\theta,x) = \frac{x}{\theta^2}e^{-\frac{x}{\theta}}$ denotes the pdf of Gamma-(2,1/θ)-distribution.

We denote by $\mathcal{D}_\alpha$ the set of all measurable mappings $g : [0,\infty) \mapsto \mathbb{R}$ with finite v_α norm. Ror given x, the transition kernel $P_\theta(x,dy)$ can be written via integration with respect to the proudct measure of the exponential distribution with mean θ, denoted by μ_θ, and the distribution of the interarrival times ν. Note that μ_θ is $\mathcal{D}_\alpha$-differentiable and ν is $\mathcal{D}_\alpha$-differentiable as ν is independent of θ (take as ν' as the null measure and $(1,\nu,\nu)$ as weak derivative). For $g \in \mathcal{D}_\alpha$ we let

$$h_g(x,a,s) = g(\max(x-a,0)+s),$$

for $x,a,s \in [0,\infty)$. Since $g \in \mathcal{D}_\alpha$, we have that $h_g(x,\cdot,\cdot)$ has finite norm for $v_\alpha \otimes v_\alpha$, where $v_\alpha \otimes v_\alpha(x,y) = \alpha^{x+y}$. Then, by Theorem 7.5,

$$\begin{aligned}\frac{d}{d\theta}\int g(y)P_\theta(x,dy) &= \frac{d}{d\theta}\int\int h_g(x,a,s)\mu_\theta(ds)\otimes\nu(da)\\ &= \int\int h_g(x,a,s)\mu'_\theta(ds)\otimes\nu(da) =\\ &= \frac{1}{\theta}\left(\int_0^\infty\left(\int_0^\infty h_g(x,a,s)f_S^\gamma(\theta,s)ds\right)f_A(a)da - \int_0^\infty\left(\int_0^\infty h_g(x,a,s)f_S(\theta,s)ds\right)f_A(a)da\right),\end{aligned}$$

and we obtain as $\mathcal{D}$-derivative of P_θ the triple $(1/\theta, P_\theta^+, P_\theta^-)$, with $P_\theta^- = P_\theta$ and , or, more formally

$$P_\theta^+(A,x) = \int_0^\infty\left(\int_0^\infty 1_{\{\max(x-a,0)+s\in A\}}f_S^\gamma(\theta,s)ds\right)f_A(a)da.$$

for $x \geq 0$ and $A \subset [0,\infty)$ a (Borel) measurable set.

***

The v-norm is extended to transition kernels (and thereby to Markov chains) through the operator norm, that is, for R not necessarily sa tochastic transition kernel we let

$$||R||_v = \sup_{x\in\mathbb{R}}\frac{1}{v(x)}\int v(y)|R(dy,x)|,$$

for $v(x) \geq 1$. Note that with this definition we have

$$\begin{aligned}|(Rh)(x)| = \left|\int h(y)R(x,dy)\right| &= \left|\int \frac{h(y)}{v(y)}v(y)R(x,dy)\right|\\ &\leq \int\left(\sup_x\frac{|h(y)|}{v(y)}\right)v(y)|R(x,dy)|\\ &\leq ||h||_v\int v(y)|R(x,dy)|\\ &= v(x)||h||_v\frac{1}{v(x)}\int v(y)|R(x,dy)|\\ &\leq v(x)\,||h||_v\,||R||_v, \qquad (7.25)\end{aligned}$$

for all x. In the following we let $v(x) = v_\alpha(x) = \alpha^{|x|+1}$, for $x \in \mathbb{R}$, and for $x \in [0,\infty)$ it suffices $v(x) = \alpha^x$.

EXAMPLE 7.19. The v_α-norm of the sojourn time Markov chain can be bounded by straightforward computation

$$\begin{aligned}||P_\theta||_\alpha &= \sup_{x\geq 0}\alpha^{-x}\mathbb{E}[\alpha^{x-A+S(\theta)}1_{\{\max(0,x-A)>0\}} + \alpha^{S(\theta)}1_{\{\max(0,x-A)\leq 0\}}]\\ &\leq \sup_{x\geq 0}\left(\mathbb{E}[\alpha^{A+S(\theta)}1_{\{\max(0,x-A)>0\}}] + \mathbb{E}[\alpha^{S(\theta)-x}]\mathbb{P}(\max(0,x-A)\leq 0)\right)\\ &\leq \mathbb{E}[\alpha^{S(\theta)-A}] + \sup_{x\geq 0}\mathbb{E}[\alpha^{S(\theta)-x}]\\ &\leq \alpha^{\mathbb{E}[S(\theta)-A]} + \alpha^{\mathbb{E}[S(\theta)]} < \infty,\end{aligned}$$

where the first inequality follows from independence. Following the same line of argument, we obtain

$$\left\| \sup_{\hat{\theta} \in \Theta_0} P_{\hat{\theta}} \right\|_{\alpha} < \infty$$

for Θ_0 a neighborhood of θ.

⁂

Let P_θ and Q_θ be transition kernels on the same state space. Then we use $P_\theta Q_\theta$ as shorthand notation for the prodcut of the two kernels, that is,

$$\int P_\theta(x, dy) Q_\theta(y, B) \;=\; (P_\theta Q_\theta)(x, B),$$

for all x and measurable sets B.

Theorem 7.6 *Suppose that P_θ and Q_θ are Markov chains defined on a common states space and both $\mathcal{D}$-differentiable for some Banach space $(\mathcal{D}, || \cdot ||_v)$. If*

(i) $||P_\theta||_v$ and $||Q_\theta||_v$ are finite,

(ii) for $h \in \mathcal{D}$, it holds that $\int h(y) P_\theta(\cdot, dy) \in \mathcal{D}$,

(iii) P_θ and Q_θ are norm Lipschitz continuous:

$$||P_{\theta+\Delta} - P_\theta||_v \leq |\Delta| M_P \quad \text{and} \quad ||Q_{\theta+\Delta} - Q_\theta||_v \leq |\Delta| M_Q$$

for some finite constants M_P, M_Q.

Then

(1) $P_\theta Q_\theta$ is norm Lipschitz continuous:

$$||P_{\theta+\Delta} Q_{\theta+\Delta} - P_\theta Q_\theta||_v \leq |\Delta| M$$

for some finite constant M, and

(2) $P_\theta Q_\theta(x, dy)$ is $\mathcal{D}$-differentiable and it holds

$$(P_\theta Q_\theta)' = P_\theta Q'_\theta + P'_\theta Q_\theta.$$

Proof: By simple algebra

$$P_{\theta+\Delta} Q_{\theta+\Delta} - P_\theta Q_\theta = (P_{\theta+\Delta} - P_\theta) Q_\theta + P_\theta (Q_{\theta+\Delta} - Q_\theta) + (P_{\theta+\Delta} - P_\theta)(Q_{\theta+\Delta} - Q_\theta). \tag{7.26}$$

Applying norms and using *(i)* together with *(iii)* yields

$$||P_{\theta+\Delta} Q_{\theta+\Delta} - P_\theta Q_\theta||_v \leq |\Delta| M_P ||Q_\theta|| + |\Delta| M_Q ||P_\theta|| + \Delta^2 M_P M_Q.$$

For Δ small enough, the above right hand side can be bounded by $|\Delta| M$ for some finite M, which proofs the first part of the theorem.

We now turn to proof the second part of the theorem. For the first term of (7.26), we consider $h \in \mathcal{D}$. Then by condition *(ii)* we have that $\int h(y)P_\theta(x,dy)$ considered as a mapping of x is in $\mathcal{D}$ we obtain from $\mathcal{D}$-differentiability of P_θ:

$$\lim_{\Delta \to 0} \frac{1}{\Delta} \int \left(\int h(z) P_\theta(y,dz) \right) (Q_{\theta+\Delta} - Q_\theta)(x,dy) = \left(\int h(z) P_\theta(y,dz) \right) Q'_\theta(x,dy) = Q'_\theta P_\theta h.$$

For the second term of (7.26), we have by *(iii)* together with (7.25) that for Δ sufficiently small

$$\left| \int h(y)(P_{\theta+\Delta}(x,dy) - P_\theta(x,dy)) \right| \leq |\Delta| \, ||h||_v M_P v(x) \tag{7.27}$$

for some finite constant c. We have assumed in *(i)* that $\int v(y)Q_\theta(x,dy)$ is finite. By Dominated Convergence we thus obtain

$$\lim_{\Delta \to 0} \frac{1}{\Delta} \int \left(\int h(z)(P_{\theta+\Delta} - P_\theta)(y,dz) \right) Q_\theta(x,dy) = \left(\int h(z) P'_\theta(y,dz) \right) Q_\theta(x,dy) = P_\theta Q'_\theta h.$$

It remains to be shown that the third part of (7.26) vanishes. Applying (7.25) to $\int h(z)(P_{\theta+\Delta}(y,dz) - P_\theta y, dz)(P_{\theta+\Delta}(x,dy) - P_\theta(x,dy)$ yields

$$\frac{1}{|\Delta|} \left| \int h(z)(P_{\theta+\Delta}(y,dz) - P_\theta y, dz)(P_{\theta+\Delta}(x,dy) - P_\theta(x,dy) \right| \leq |\Delta| M_P M_Q v(x),$$

which shows that

$$\lim_{\Delta \to 0} \frac{1}{\Delta} \int h(z)(P_{\theta+\Delta}(y,dz) - P_\theta y, dz)(P_{\theta+\Delta}(x,dy) - P_\theta(x,dy) = 0.$$

Hence, we obtain

$$(P_\theta Q_\theta)' = P'_\theta Q_\theta + Q_\theta P'_\theta,$$

which concludes the proof.

QED

Note that a sufficient condition for *(iii)* in Theorem 7.6 (v-norm Lipschitz continuity) is

$$\left\| \sup_{\hat{\theta} \in \Theta_0} P_{\hat{\theta}} \right\|_v < \infty, \tag{7.28}$$

where Θ_0 is a neighborhood of θ. The proof is left as an excercise.

In the following we will establish the corresponding result for n-fold product of Markov chains. For sake of simplicity, we present the result for products of identical Markov chains. For this, we denote $P^n_\theta = P_\theta P^{n-1}_\theta$ for $n \geq 2$. Moreover, we denote by P^0_θ the identity operator.

Before we present the theorem we recall the randomization of a summation, which will be used in onne of statements of the theorem.

EXAMPLE 7.20. Let a_i, $1 \leq i \leq n$, be some finite numbers. Then

$$\sum_{i=1}^{n} (a_i - b_i) = 2n\mathbb{E}[r_\sigma],$$

where σ is uniformly ddistributed on $\{1, \ldots, 2n\}$ and

$$r_i = \begin{cases} a_i & i \leq n \\ -b_{i-n} & i > n. \end{cases}$$

Indeed, it holds that

$$\sum_{i=1}^{n}(a_i - b_i) = 2n\left(\sum_{i=1}^{n} a_i \frac{1}{2n} - \sum_{i=n+1}^{2n} b_{i-n}\frac{1}{2n}\right) = 2\left(n\sum_{i=1}^{n} a_i \mathbb{P}(\sigma = i) - \sum_{i=n+1}^{2n} b_{i-n}\mathbb{P}(\sigma = i)\right) = 2n\mathbb{E}[r_\sigma].$$

⁂

The overall theorem now reads as follows. The proof is left as an excerise.

Theorem 7.7 *Suppose that P_θ is $\mathcal{D}$-differentiable for some Banach space $(\mathcal{D}, ||\cdot||_v)$. If*

(i) $||P_\theta||_v$ *is finite,*

(ii) for $h \in \mathcal{D}$, it holds that $\int h(y)P_\theta(\cdot, dy) \in \mathcal{D}$,

(iii) P_θ is norm Lipschitz continuous:

$$||P_{\theta+\Delta} - P_\theta||_v \leq |\Delta| M_P$$

for some finite constant M_P.

Then, for any $h \in \mathcal{D}$ it holds that

$$(P_\theta^n)' = \sum_{j=0}^{n-1} P_\theta^{n-j-1} P_\theta' P_\theta^j.$$

Moreover, if in addition, P_θ has $\mathcal{D}$-derivative $(c_\theta(\cdot), P_\theta^+, P_\theta^-)$, then, it holds that for any $h \in \mathcal{D}$ that

$$\int h(y)P_\theta^n(x, dy) \quad = \quad \sum_{j=0}^{n-1}\int\left(\int h(z)P_\theta^{n-j-1}P_\theta^+(y, dz) - h(z)P_\theta^{n-j-1}P_\theta^-(y, dz)\right)c_\theta(y)P_\theta^j(x, dy).$$

or, equivalently,

$$\frac{d}{d\theta}\int h(y)P_\theta^n(x, dy) \quad = \quad 2nr(\sigma)\int\left(\int h(z)P_\theta^{n-\eta}P_\theta^{[\sigma]}P_\theta^{n-\eta-1}P_\theta^-(y, dz)\right)c_\theta(y)P_\theta^{\eta-1}(x, dy),$$

where σ is uniformly distributed on $\{1, \ldots, 2n\}$ independent of everything else, $\eta = \sigma \bmod n$,

$$P_\theta^{[\sigma]} = \begin{cases} P_\theta^+(y, dz) & i \leq n \\ P_\theta^-(y, dz) & i > n, \end{cases}$$

and $r(\sigma) = 1$ for $\sigma \leq n-1$ and -1 otherwise.

The third representation in Theorem 7.7 is called the *randomized MVD* and essentially states that an MVD estimator can be obtained as single-run estimator. This is a consequence of the basic randomization of sums as explained in Example 7.20. Note that randomization comes at the price of a higher variance of the estimator. Randomization is readily extended to the case of randomly picking

k, with $k \leq n$, points of differentation. Balancing memory requirement and computational burden with variance of the estimator, one can identify the optimal value of k. For details we refer to [19].

Example 7.21. We now apply Theorem 7.7 to our sojourn time example, where we assume that the interarrival time distribution has finite moment generating function. That condition *(i)* and *(ii)* are satisfied has been established in Example 7.17. For condition *(iii)* we use the Lipschitz modulus in (7.28) so that we conclude that *(iii)* holds from Example 7.19. leads to the following estimator for $d\mathbb{E}[g(X_n(\theta)]/d\theta$.

For $i \leq n$ and $\eta \in \{+,-\}$, let $X_n(\theta;i,\eta)$ be defined as follows. Up to time i, $X_n(\theta;i,\eta)$ behaves just like $X_n(\theta)$, i.e., following (7.23)

$$X_{n+1}(\theta;i,\eta) = \max(0, X_n(\theta;i,\eta) - A_n(\theta)) + S_{n+1}(\theta),$$

for $i \leq n$. At the transition from $X_{n-1}(\theta;i,\eta)$ to $X_n(\theta;i,\eta)$ service time $S_n(\theta)$ is perturbed for the "-" version of the sample path, and we let for $\eta =$ "+"

$$X_i(\theta;i,+) = \max(0, X_{i-1}(\theta;i,+) - A_{i-1}(\theta)) + S_i(\theta) + Y_i(\theta),$$

with $\{Y_i(\theta)\}$ iid exponential random variables with mean θ and independent of everything else, and for $\eta =$ "$-$"

$$X_i(\theta;i,-) = \max(0, X_{i-1}(\theta;i,-) - A_{i-1}(\theta)) + S_i(\theta).$$

For $n > i$, the transitions of $X_n(\theta;i,\eta)$ follow the standard update formula

$$X_{n+1}(\theta;i,\eta) = \max(0, X_n(\theta;i,\eta) - A_n(\theta)) + S_{n+1}(\theta),$$

Then, by Theorem 7.7

$$\begin{aligned} \frac{d}{d\theta}\mathbb{E}[g(X_n(\theta)] &= \frac{1}{\theta}\sum_{i=1}^{n} \mathbb{E}\left[g(X_n(\theta;i,+)) - g(X_n(\theta;i,-))\right] \qquad (7.29) \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{\theta} g(X_n(\theta;i,+))\right] - \frac{n}{\theta}\mathbb{E}[g(X_n(\theta))]. \end{aligned}$$

Using randomization, we randomly select σ out of $\{1,\ldots 2n\}$ and for $\sigma \leq n$, we generate the positive version of $X_n(\theta;i,\eta)$ perturbed at transition i and for $\sigma > n$ we generate the negative version of $X_n(\theta;i,\eta)$ perturbed at transition i, the leads to the following single run estimator

$$\frac{d}{d\theta}\mathbb{E}[g(X_n(\theta)] = \frac{2n}{\theta}\mathbb{E}\big[r(\sigma)g(X_n(\theta;\sigma \bmod n,[\sigma])\big],$$

where σ is uniformly distributed on $\{1,\ldots,2n\}$, $[\sigma] =$ "+ ", for $\sigma \leq n$, and $[\sigma] =$ "- ", for $\sigma > n$. It is worth noting that the above single run version has larger variance than the estimator in (7.29) but comes at the ease of having to only perturb once for a single sample path. Even though randomization leads to an increase in variance, the variance of the resulting estimator typically is still significantly lower that that of the corresponding SF estimator.

***

It is worth noting that P^n_θ in Theorem 7.7 allows for two interpretations. Firstly, as defined in the above way P^n_θ is the transition probabilty from $X_\theta(0)$ to $X_\theta(n)$ and thus is a conditional probability measure on the state space of X_θ, i.e,

$$\mathbb{E}[h(X_\theta(n))|X_\theta(0) = x] = \int h(y)P^n_\theta(x,dy)$$

for any measurable mapping such that the above expression are fininte. Secondly, P_θ^n can be interpreted as a conditional probability measure on the sample path $(X_\theta(1), \ldots, X_\theta(n))$ for given initial value $X_\theta(0) =$, i.e.,

$$\begin{aligned}\mathbb{E}[h(x, X_\theta(1), \ldots, X_\theta(n)) | X_\theta(0) = x] &= \int_{\mathbb{R}^n} h(y) P_\theta^n(x, dy) \\ &= \int \left(\int \cdots \left(\int h(x, x_1, \ldots, x_n) P_\theta(x_{n-1}, dx_n) \right) P_\theta(x_{n-2}, dx_{n-1}) \cdots \right) P_\theta(x, dx_1)\end{aligned}$$

for any measurable mapping such that the above expression are fininte. To simplify the notation we write $x_{i:j}$ for $(x_i, \ldots, x_j)$, for $i \leq j$, and

$$P(x, dx_i, \ldots, dx_j) = P(x, dx_i) P(x_i, dx_i) \cdots P(x_{j-1}, dx_j)$$

for $i \leq j$.

For the analysis that follows we don't have to distinguish between these two interpretation as all results developed hold for both interpretations under a mild regularity condition. Indeed, provided that P_θ is $\mathcal{D}$-differentiable with $\mathcal{D}$ a Banach space such that $\mathcal{D} = \{h : ||h||_v < \infty\}$, then the product extends to the sample path measure P_θ^n where P_θ^n is $\hat{\mathcal{D}}$-differentiable for all $h : S \to \mathbb{R}$ such that for all $(s_1, \ldots, s_n)$ the mapping $h(s_1, \ldots, s_{j-1}, \cdot, s_{j+1}, \ldots, s_n) \in \mathcal{D}$ and $||h||_w < \infty$, where $w(s_1, \ldots, s_n) = v(s_1) \cdots v(s_n)$.

Theorem 7.8 *Suppose that P_θ is $\mathcal{D}$-differentiable for some Banach space $(\mathcal{D}, || \cdot ||_v)$. If*

(i) $||P_\theta||_v$ is finite,

(ii) for $h \in \mathcal{D}$, it holds that $\int h(y) P_\theta(\cdot, dy) \in \mathcal{D}$,

(iii) P_θ is norm Lipschitz continuous:

$$||P_{\theta+\Delta} - P_\theta||_v \leq |\Delta| M_P$$

for some finite constant M_P.

Then, for any $h \in \mathcal{D}$ it holds that

$$(P_\theta^n)' = \sum_{j=0}^{n-1} P_\theta^{n-j-1} P'_\theta P_\theta^j.$$

Moreover, if in addition, P_θ has $\mathcal{D}$-derivative $(c_\theta(\cdot), P_\theta^+, P_\theta^-)$, then, it holds that for any $h \in \mathcal{D}$ that

$$\begin{aligned}\int_{\mathbb{R}^n} h(x, y) P_\theta^n(x, dy) &= \sum_{j=0}^{n-1} \int_{\mathbb{R}^j} \left(\int_{\mathbb{R}} \int_{\mathbb{R}^{n-j-1}} h(x, y, z, r) P_\theta^{n-j-1}(z, dr) P_\theta^+(y, dz) \right. \\ &\qquad \left. - h(x, y, z, r) P_\theta^{n-j-1}(z, dr) P_\theta^-(y, dz) \right) c_\theta(y) P_\theta^j(x, dy).\end{aligned}$$

or, equivalently,

$$\frac{d}{d\theta} \int h(y) P_\theta^n(x, dy) = 2nr(\sigma) \int \left(\int h(x, x_1, \ldots, x_n) P_\theta^{n-\eta} P_\theta^{[\sigma]} P_\theta^{n-\eta-1} P_\theta^-(y, dz) \right) c_\theta(y) P_\theta^{\eta-1}(x, dy),$$

where σ is uniformly distributed on $\{1, \ldots, 2n\}$ independent of everything else, $\eta = \sigma \mod n$,

$$P_\theta^{[\sigma]} = \begin{cases} P_\theta^+(y, dz) & i \leq n \\ P_\theta^-(y, dz) & i > n, \end{cases}$$

and $r(\sigma) = 1$ for $\sigma \leq n - 1$ and -1 otherwise.

Proof: We prove the theorem by induction. For $n = 2$, the proof resembels the one for Theorem 7.6. Specifically, we use the algebraic decompostion put forward in (7.26) and argue for the individual terms as in the proof of Theorem 7.6. In order use the same argument we need for the first part that $\int h(x, x_1, x_2) P_\theta(x_1, dx_2)$ as a mapping of x_1 belongs to $\mathcal{D}$. For the second term to converge, we use $|h(x, x_1, x_2)| \leq v(x)v(x_1)v(x_2)$ to argue that

$$\left| \int h(x, x_1, x_2) P_{\theta+\Delta}((x, dx_1) - P_\theta(x, dx_1)) \right| \leq |\Delta| \, ||P_{\theta+\Delta} - P_\theta||_v \, v(x)v(x_2).$$

7.3.3 $*$ The Weak Differentiation Approach

Let $X \in S$ have distribution μ on some measurable space $(S, \mathcal{S})$. The expected value of $h(X))$ reads

$$\mathbb{E}[h(X)] = \int_S h(x) \, \mu(dx)$$

and it can be interpreted as a bi-linear mapping $\langle \cdot, \cdot \rangle : (h, \mu) \mapsto \mathbb{E}[f(X)]$. Suppose that μ depends on some parameter θ and write μ_θ and X_θ, respectively. Properties of μ_θ such as continuity or, as we will see later on, differentiability can be introduced via families of test functions. For example, the sequence of measures $\{\mu_{\theta_n}\}$ is said to be weakly convergent towards a measure μ_θ if for any $\{\theta_n\}$ such that $\theta_n \to \theta$ as n tends to ∞ it holds that

$$\lim_{n\to\infty} \langle h, \mu_{\theta_n} \rangle = \langle h, \mu_\theta \rangle \quad \forall h \in \mathcal{C}^b(S), \tag{7.30}$$

where $\mathcal{C}^b(S)$ is the set of continuous and bounded mappings from S to $\mathbb{R}$. Note that weak convergence is also called weak continuity. A natural question is why not simply define continuity of measures via set-wise continuity $\lim_{n\to\infty} \mu_{\theta_n}(A) = \mu_\theta(A)$ for all $A \in \mathcal{S}$? The reason is that this defintion is too restrictive as it makes too few sequences convergent. Indeed, (7.30) may hold for a sequence $\{\mu_{\theta_n}\}$ that does not satisfy set-wise continuity. To see this, let μ_{θ_n} denote the uniform distribution on $[1 - \theta_n, 1]$ and consider convergence of μ_{θ_n} as $\theta_n \to 1$. Then

$$\lim_{n\to\infty} \langle h, \mu_{\theta_n} \rangle = \frac{1}{1 - \theta_n} \int_{1-\theta_n}^1 h(x) \, dx = f(1) = \int_S h(x) \, \delta_1(dx),$$

where δ is the Dirac measure with unit measure in point 1. Hence, μ_{θ_n} converges weakly towards δ_1. Set-wise convergence however does not hold. Indeed, let $A = \{1\}$, i.e., a singleton. Then

$$\lim_{n\to\infty} \mu_{\theta_n}(A) = 0 \neq 1 = \delta_1(1).$$

Apart from this, admittedly somewhat technical argument, the definition in (7.30) has the advantage to capture the core application of continuity of measures, namely that of continuity of expected values for a pre-defined class of functions.

REMARK. We will frequently work with *measures*, as opposed to cumulative distribution functions in this section. The difference being that a measure lives on a general measure space $(S, \mathcal{S})$ (and this is the same set S the cost function lives on) describing the phenomenon under study, e.g., $S = \mathbb{N}$ in a single server queue-length model or $S = [0, \infty)$ in the sojourn time model, where the cdf is a mapping from $\mathbb{R}$ to $[0, 1]$, whatever the underlying model is. Let F be cdf of X and let μ denote the measure of X, then

$$F(x) = \mathbb{P}(X \leq x) = \mu\Big(\{s \in S : X(s) \leq x\}\Big),$$

for all x. Denoting the probability density function of X by f_X, we have

$$\mathbb{E}[h(X)] = \int h(x) f_X(x)\, dx = \int h(x) \mu(dx),$$

where we assume that the expression on the above left hand side exists. A subtle but important point in the relation between measures and cdf's is that (pointwise) convergence of sequence of cdf's F_n towards a cdf F only characterizes F on continuity points of F. In words, if mass is shifted towards a single point when taking the limit, then the limit of F_n bears no information on F at that particular point. This is exactly what happened in the uniform example above. All F_n's were continuous distributions and taking the limit as n tends to ∞ the mass shifted towards the single point $\{1\}$, which results in a non-continuous limit F.

In this section we have developed a functional analytic approach to differentiation. Starting point was the view on expected values as displayed in (7.30). Our project was (i) to develop a concept of differentiation of measures that is rich enough so that all relevant measures "become differentiable" and (ii) to enrich the set of test-functions, as $\mathcal{C}^b(S)$ is rather limited for applications as one typically is not only interested in bounded functions. This project brought us to studying the interplay between norms of functions and measures and convergence of measure integrals. While this is an admittedly more demanding topic than that of IPA and SF (which rely on basic tools from analysis and applied probability), the reward of the theoretical foundations laid here is that the concepts easily transfer to the general setting of Markov chains and even operators, as we will explain in the next chapter.

7.4 Exercises

EXERCISE 7.1. Refer to the queuing model of Examples 6.10 and 7.5. Let $J(\theta) = \mathbb{E}[L(X_1(\theta, \ldots, X_N(\theta))$, where $L(X_1, \ldots, X_N)$ is such that $\frac{\partial L}{\partial x_i}(X_1, \ldots, X_N)$ is continuous and uniformly bounded in absolute value by a constant l for all i. Show that if θ is a scale parameter of the service distribution satisfying $\mathbb{E}[S(\theta)/\theta] < \infty$, then the IPA derivative is unbiased for $\nabla_\theta J(\theta)$.

EXERCISE 7.2. Provide a proof of Theorem 7.7.

EXERCISE 7.3. Refer to Examples 6.10 and 7.5. Assume that θ is a location parameter of the service times in a FCFS GI/G/1 queue. Specifically, assume that the sevice times have a representation of the form $S_n(\theta) \overset{\mathcal{L}}{=} \theta + F_0^{-1}(U_n)$, where $\{U_n\}$ is a sequence of iid uniform random variables $U(0, 1)$. For the queueing model, we consider only positive values: $\theta > 0$.

(a) Specify under which assumptions for $F_0(x)$ is $X_n(\theta)$ Lipschitz continuous w.p.1 for all $n \leq N$

(b) Calculate the derivative of the sample average

$$L(X_1(\theta), \dots, X_N(\theta)) = \frac{1}{N} \sum_{n=1}^{N} X_n(\theta).$$

(c) Under the assumptions in (a), prove that the IPA estimator of $\mathbb{E}[L]$ is unbiased.

EXERCISE 7.4. We revisit the sojourn time example as introduced in Examples 6.10 ad 7.5. with notations and basic assumptions as detailed in these examples. Provided that the service times $S_n(\theta)$ are almost surely Lipschitz continuous on some fixed interval Θ, show that

$$\frac{d}{d\theta} \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} X_n(\theta) \right] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \frac{dX_n(\theta)}{d\theta} \right]$$

for almost all $\theta \in \Theta$.

EXERCISE 7.5. Refer to Section 7.1.3 (optional). Show that a Lipschitz-continuous function is absolutely continuous.

EXERCISE 7.6. Let P_θ be $\mathcal{D}$-differentiabble for $\theta \in \Theta$, with $\mathcal{D} = \{h : ||h||_v < \infty\}$. Show that

$$||P_{\theta+\Delta} - P_\theta||_v \leq |\Delta| \sup_{\theta \in \Theta} ||P'_\theta||_v.$$

EXERCISE 7.7. Show that if μ_θ has a ν-density f_θ for all $\theta \in \Theta \subset \mathbb{R}$, then μ_θ has μ_{θ_0}-density f_θ / f_{θ_0} for any $\theta_0 \in \Theta$ provided that the support of f_θ is a subset of that of f_{θ_0} for all $\theta \in \Theta$.

EXERCISE 7.8. Let f denote a density of random variable X with support $(-\infty, \infty)$. Show that

$$f(x) = \mathbb{E}\left[\mathbf{1}_{\{X \leq x\}} \frac{f'(X)}{f(X)} \right].$$

This result shows that the value of the density at a point can be obtained via the score function.

EXERCISE 7.9. Revisit Example 7.2. Establish the corresponding SF and MVD quantile sensitivity estimators.

EXERCISE 7.10. Let $\{X_n\}$ be a finite Markov chain with continuously differentiable transition probabilities $P_{i,j}(\theta)$ and known initial state $X_0 = x_0$.

(a) Show that the Score Function for a horizon N is:

$$S_N(\theta; X_1, \dots, X_N) = \sum_{n=1}^{N} \frac{d}{d\theta} \log P_{X_{n-1},X_n}(\theta)$$

(b) Let $c\colon S \to \mathbb{R}$ be a bounded function and consider the cost function:

$$C(\theta) = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} c(X_n)\right). \tag{7.31}$$

Show that:

$$\frac{d}{d\theta}C(\theta) = \mathbb{E}\left[S_N(\theta; X_1, \dots, X_N)\frac{1}{N}\sum_{n=1}^{N} c(X_n)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} c(X_n) S_n(\theta; X_1, \dots, X_n)\right],$$

provided that the integrands above have finite expectation.

(c) Argue by inspection that the second estimator, namely

$$\frac{1}{N}\sum_{n=1}^{N} c(X_n) S_n(\theta; X_1, \dots, X_n)$$

must necessarily have smaller variance than the one using $S_N(\theta; X_1, \dots, X_N)$. This is a usual approach for variance reduction whenever applying SF to Markov Chains.

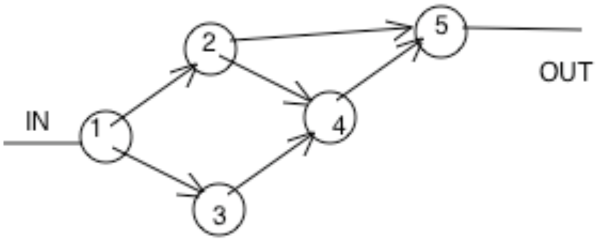


Figure 7.1: Reliability Network

EXERCISE 7.11. Consider the reliability model of Examples 6.9, 7.4 and 7.16, illustrated in Figure 7.1. The life of the system is

$$L(\theta) = \max\left(\min(T_1, T_2, T_5), \min(T_1, T_2, T_4, T_5), \min(T_1, T_3(\theta), T_4, T_5)\right),$$

where the components' lives are five independent random variables: $T_i \sim G_i, i \neq 3$ (independent of θ) and $T_3(\theta) \sim$ Exponential $(1/\theta)$.

(a) Calculate the SF estimator for the derivative of $\mathbb{E}[L(\theta)]$ and show that it is unbiased.

(b) Program a simulation to generate independent samples of $L(\theta)$ and of the three derivative estimators: the IPA (given in Example 7.4), the SF and the MVD (given in Example 7.16). For the same amount of samples, compute a table to compare variances and CPU times.

EXERCISE 7.12. Refer to the reliability model of Examples 6.9, 7.4 and 7.16, illustrated in Figure 7.1. Consider now the model where the lifetimes have a Weibull distribution with parameters (λ_i, k). In particular, for component 3 we have: $\mathbb{P}(T_3 \leq x) = G_3(x) = 1 - e^{-(x/\theta)^k}$, so that now $\mathbb{E}(T_3) = \theta\,\Gamma(1 + 1/k)$.

(a) Find the representation $T_3 = G_3^{-1}(U)$ and show that the IPA estimator:

$$\widehat{L}^{\text{IPA}}(\theta, \omega) = \begin{cases} \frac{T_3(\theta,\omega)}{\theta} & T_2(\omega) < T_3(\theta, \omega) < \bar{T}(\omega) = \min(T_1(\omega), T_4(\omega), T_5(\omega)) \\ 0 & \text{otherwise} \end{cases} \tag{7.32}$$

is valid for this case. For what family of distributions G_3 will the IPA estimator remain unchanged?

(b) Explain how the SF and MVD estimators of Exercise 7.11 should be modified for this problem.

EXERCISE 7.13. Consider the reliability model of Examples 6.9, 7.4 and 7.16, illustrated in Figure 7.1. Suppose now that we are interested in estimating the sensitivity:

$$\frac{d}{d\theta}\mathbb{P}(L > \bar{l}), \tag{7.33}$$

for some specific bound $\bar{l}$, and assume that the lifetime distributions are exponential: $G_i \sim$ Exponential $(1/\lambda_i)$, with $\lambda_3\theta = 1$.

(a) Use the same representation as in Example 6.14 and show that IPA is biased. Explain why.

(b) Explain how the SF and MVD estimators of Exercise 7.11 should be modified for this problem.

EXERCISE 7.14. We now present the reliability problem related to a network of logically interconnected components, illustrated in Figure 7.1. For such systems, maintenance rules must be specified in order to repair or replace failed components. Suppose that the system has been operating for some time, may be with some components having been replaced already, so not all of them have the same age. If one looks at such systems at any arbitrary time, there is a probability $p_i \in (0, 1)$ that one finds component i working. Let X_i be the indicator that component i is working. We assume that $\{X_i\}$ are independent, Bernoulli(p_i) random variables defined on a common probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. The *reliability* of the system is defined as the probability that the system works, that is, $\mathbb{P}(\phi(X_1, \ldots, X_5) = 1)$, where the reliability function $\phi(\cdot)$ is the indicator that the system works. Suppose that $\theta = p_3$ is the variable of interest, and that we wish to estimate the sensitivity:

$$\frac{d}{d\theta}\mathbb{P}[\phi(X_1, \ldots, X_5) = 1] = \frac{d}{d\theta}\mathbb{E}[\phi(X_1, \ldots, X_5)]. \tag{7.34}$$

(a) Show that the reliability ϕ of Figure 7.1 can be expressed as:

$$\phi(X_1, \ldots, X_n) = X_1 X_5 \max(X_2, X_3\, X_4).$$

(b) Consider the representation $X_3 = \mathbf{1}_{\{U \leq \theta\}}$ and show that IPA is biased for (7.34).

(c) Calculate the SF estimator for (7.34) and show that it is unbiased for (7.34).

(d) Calculate the MVD estimator for (7.34) and show that it is unbiased for (7.34).

(e) Use the properties of conditional expectation, namely $\mathbb{E}[\phi] = \mathbb{E}[\mathbb{E}(\phi \,|\, \mathcal{G})]$ for any $\mathcal{G} \subset \mathfrak{F}$ and show that $\mathbb{E}[\phi \,|\, X_1, X_2, X_4, X_5]$ is now absolutely continuous in θ. Find the IPA for this new representation and compare with the MVD estimator found in (d). Conditioning can in general "smooth out" discontinuities of functions, allowing for an interchange between derivative and expectation, and the method is known as Smoothed Perturbation Analysis (SPA).

EXERCISE 7.15. Revisit Example 7.5. Derive the MVD estimator for the Poisson counting process and establish unbiasedness of the estimator.