ABSTRACT

TANADKITHIRUN, RAYWAT. Partition-Based Proposal Distributions for Importance Sampling . (Under the direction of Min Kang.)

Importance Sampling (IS) is a useful Monte Carlo based technique to estimate an expectation of a target function with respect to a distribution of interest using random samples drawn from another distribution, called a proposal distribution. The optimal proposal distribution minimizing the variance of the estimator is known, but it cannot be used in reality. This work summarizes basic theory for IS method including the convergence of the estimators as well as the optimal proposal distributions for both kinds of IS: basic and self-normalized IS. Moreover, the insufficiency of the widely used rule of thumb in choosing a proposal density is clarified. A partition-based method that utilizes the information of the known optimal proposal distribution is proposed in this work. The idea can also extend to the case of multidimensional spaces. IS was only done for known distributions in the past, but the partition-based method is not limited to known distributions. Furthermore, the optimal distribution for a simultaneous simulation using IS is identified and proved. An alternative scheme of sequential IS, which allows us to draw each component of a sample sequentially in time, using basic IS instead of self-normalized IS as the base step is also provided.

Partition-Based Proposal Distributions for Importance Sampling

by
Raywat Tanadkithirun

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2016

APPROVED BY:

Jeffrey Scroggs	Tao Pang
Paul Fackler	Min Kang Chair of Advisory Committee

DEDICATION

This dissertation is dedicated to my beloved family for their cheer, love and encouragement during my study, especially to my father, my mother and aunt Tuakow. I also dedicate this dissertation to the memory of my grandmother Arma who is one of my motivations to pursue a PhD.

BIOGRAPHY

Raywat Tanadkithirun was born August 19, 1985 in Bangkok, Thailand. In 2004, he finished high school from Triam Udom Suksa school, Bangkok, Thailand and received the National Excellent Academic in Science Scholarship to pursue a Bachelor of Science degree in Mathematics at Chulalongkorn University, Thailand. In 2008, he received a Bachelor of Science degree with first-class honors. He continued his study there in the Master of Science program in Mathematics with financial support from the Chulalongkorn University Graduate Scholarship to Commemorate the 72^{nd} Anniversary of His Majesty King Bhumibol Adulyadej to gain more knowledge in both pure and applied aspects of Mathematics and received a Master's degree in 2010. He also received the Royal Thai Government Scholarship to pursue Master's and Doctoral degrees in the US. In 2012, he received a Master of Financial Mathematics degree at North Carolina State University. From 2012 - 2016, he has been pursuing a PhD in Applied Mathematics at North Carolina State University.

ACKNOWLEDGEMENTS

My deepest and sincere gratitude is to my adviser, Dr. Min Kang, for her kind supervision and consistent encouragement. She gave me the freedom to explore what I like and, at the same time, the guidance to recover when my steps faltered. Without her constructive suggestion and knowledgeable guidance, this dissertation would never have been successfully completed. Also, I would like to express my sincere thank and deep appreciation to my committee members, Dr. Jeffrey Scroggs, Dr. Tao Pang and Dr. Paul Fackler, for their meticulous comments and helpful suggestions to improve my dissertation. In addition, I would like to thank Dr. Sirod Sirisup for his persistent help about computer problems and Denise Seabrooks for her assistance on paperwork. I am also grateful to the Royal Thai Government Scholarship for granting me financial support throughout my study at North Carolina State University. Last but not least, I would like to thank my beloved family and dear friends for their cheer and support during my study at North Carolina State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Importance Sampling (IS) is a Monte Carlo-based simulation technique to estimate an expectation with respect to a distribution of interest using random samples drawn from another distribution. Normally, this technique is used when we have problems arising from our distribution of interest, which will be called the *target distribution*, and wish that we could use another distribution, which will be called a *proposal distribution* or *importance distribution*, instead. We would like to use an ideal proposal distribution to generate samples to use in the estimation.

This technique is first used in rare event applications [17, 18, 14, 32, 2] where we want to estimate the expectation of a function that concentrates on the extremely-unlikely-to-be-visited region of the target distribution. In ordinary Monte Carlo method, we need to generate a huge number of samples directly from the target distribution to get just a sample falling inside that region. In many rare event applications, even when we set a pretty big number of samples, we could fail to have even one sample in that important region. Using IS, we can manage to estimate the expectation by sampling from another proposal distribution that has high probability around that important region of interest, hence the name *importance sampling*.

IS is very useful when we have difficulty in sampling from the target distribution especially in Bayesian Analysis [21, 33, 11, 36, 24]. Sometimes, it is hard or even impossible to sample from the target distribution. However, IS allows us to get samples from another easier-to-sample-from

distribution together with associated weights that can be used as an empirical estimate of that target distribution. We can also use this empirical estimate to approximate the expectation of any given function with respect to the target distribution.

Apart from rare event applications and Bayesian inference which are two primary applications of IS, there are more uses IS can offer. In some applications where we have more than one target distribution, we can sample a single set of samples that can be used for all target distributions [1, 35]. This greatly reduces the amount of work from the direct simulation where we have to sample a separate set of samples for each target distribution. IS can also be a tool for estimating derivatives of expectations with respect to parameters of the underlying distribution [26, 27, 13, 3, 8]. This usage allows IS to attack some problems to which simple Monte Carlo method can hardly be directly applied. Moreover, IS is a key ingredient to develop Sequential Monte Carlo simulation in which IS is performed sequentially in time with samples and associated weights from the past time steps [9, 31, 7, 10, 5].

IS is considered as a variance reduction method to the ordinary Monte Carlo method. This depends on how well we choose a proposal distribution for sampling. If we pick a bad proposal distribution, it could blow the variance and has worse performance compared to a plain Monte Carlo method. A good proposal distribution actually depends on the target distribution and the function of interest.

To see the derivation of IS, we need to talk about a classical Monte Carlo method first. Let X be a random variable with a probability distribution π, and f a $\pi-$integrable function. Define the true expectation

$$\mu = \mathbb{E}[f(X)] = \mathbb{E}_\pi(f).$$

Since we assume that f is a $\pi-$integrable function, μ is finite. This is a convention to apply the Monte Carlo method. Also, because we are not interested in the obvious case, we assume that f is not a constant function. A classical Monte Carlo estimator for μ is

$$\hat{\mu}_\pi = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \quad X_i \overset{\text{iid}}{\sim} \pi \tag{1.1}$$

where $X_i \overset{\text{iid}}{\sim} \pi$ denote that X_i's are independent and identically distributed with the distribution π. Also, we will denote by $X \sim \pi$ when X has distribution π. Please keep in mind that there are N number of samples in the formula of $\hat{\mu}_\pi$, but we will suppress N for a short notation and focus on the distribution from which the samples are drawn.

The concept of IS is to change the dominating measure from π to another probability distribution. There are two kinds of IS: basic IS and self-normalized IS. These names are from [25], and will be

used in this work. Most of the related work in the literature talk about just one kind of IS relying on their application, and most of foundation theory are available for only basic IS.

Throughout this work, we will express everything in the case of continuous space. For a discrete space, it can be derived in a similar manner and will be much easier to implement. To be able to apply IS, we assume that every probability distribution in this study is absolutely continuous with respect to the Lebesgue measure and has a density function. We will use the same notation for both probability measure and its probability density. When we write the integral without specifying the domain, we mean that the integral is taken on the whole domain which is allowed to be multidimensional.

1.1 Basic Importance Sampling

Consider a given pair of a target function and a target density (f, π). For a valid density q, we can have that

$$\begin{aligned}\mathbb{E}_\pi(f) &= \int f(x)\pi(x)\,dx \\ &= \int \frac{f(x)\pi(x)}{q(x)} q(x)\,dx \\ &= \mathbb{E}_q\left(\frac{f\pi}{q}\right).\end{aligned}$$

Define the weight function

$$w(\cdot) = \frac{\pi(\cdot)}{q(\cdot)}. \tag{1.2}$$

Then, the basic IS estimator for μ is

$$\hat{\mu}_q = \frac{1}{N}\sum_{i=1}^{N} w(X_i) f(X_i), \quad X_i \overset{\text{iid}}{\sim} q. \tag{1.3}$$

Now, let's discuss more on the validity of q. One may set an assumption on q to be $\pi \ll q$ (π is absolutely continuous with respect to q), and most current research use that $q(x) = 0 \implies \pi(x) = 0$ as the assumption for a proposal density function q, which is slightly stronger than $\pi \ll q$. However, we can relax $\pi \ll q$ to a weaker assumption. Since μ is finite, we can consider $\phi(dx) = f(x)\pi(x)\,dx$ as a finite measure. So, we can set an assumption for a legitimate q in the derivation step as $\phi \ll q$. We will denote this by

$$f\pi \ll q$$

and this is a weaker assumption than $\pi \ll q$.

Note that the weight function compensates the fact that we change the dominating measure from π to q. X_i's in (1.1) are drawn according to π, while X_i's in (1.3) are drawn according to q. The regular Monte Carlo method has equal weight $\frac{1}{N}$ to all samples, but the basic IS method has weight $\frac{1}{N} w(X_i)$ for sample X_i. The sample that properly represents the importance region will have high weight and become an important sample.

Remark 1.1. If we choose a proposal distribution q to be π itself, then it is equivalent to performing the classical Monte Carlo method.

1.2 Self-normalized Importance Sampling

In some situations especially in Bayesian inference applications where the normalizing constant for a posterior distribution is very hard to calculate, we can only deal with an unnormalized version of π. Consider a given pair of a target function and an unnormalized target density (f, p) where

$$\pi(x) = \frac{p(x)}{Z}$$

and $p(x)$ is known pointwise but the normalizing constant $Z = \int p(x)\, dx$ is not known. For a valid density function q, we can have that

$$\begin{aligned}
\mathbb{E}_\pi(f) &= \int f(x)\pi(x)\, dx \\
&= \frac{1}{Z} \int \frac{f(x)p(x)}{q(x)} q(x)\, dx \\
&= \frac{1}{Z} \mathbb{E}_q\left(\frac{fp}{q}\right) \\
&\approx \frac{1}{Z}\left(\frac{1}{N}\sum_{i=1}^{N} \tilde{w}(X_i) f(X_i)\right) \qquad (X_i \overset{\text{iid}}{\sim} q)
\end{aligned}$$

where

$$\tilde{w}(\cdot) = \frac{p(\cdot)}{q(\cdot)}.$$

Here, the self-normalized weight function $\tilde{w}$ is not defined the same as the basic weight function from (1.2) in Section 1.1 due to the unknown normalizing constant Z. Here, we need that $f\pi \ll q$. Assume further that $\pi \ll q$ or equivalently

$$p \ll q.$$

Then,

$$\begin{aligned} Z &= \int p(x)\, dx \\ &= \int \frac{p(x)}{q(x)} q(x)\, dx \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \tilde{w}(X_i). \qquad (X_i \overset{\text{iid}}{\sim} q) \end{aligned}$$

Thus, the self-normalized IS estimator for μ is

$$\tilde{\mu}_q = \frac{\sum_{i=1}^{N} \tilde{w}(X_i) f(X_i)}{\sum_{j=1}^{N} \tilde{w}(X_j)}, \quad X_i \overset{\text{iid}}{\sim} q. \tag{1.4}$$

Remark 1.2. We can define the self-normalized weight function to be $\tilde{w}(\cdot) = c\,\frac{p(\cdot)}{q(\cdot)}$ for any constant multiplier c. All of these weight functions can induce the same self-normalized IS estimator $\tilde{\mu}_q$ due to the cancellation in the ratio $\frac{\tilde{w}(X_i)}{\sum_{j=1}^{N} \tilde{w}(X_j)}$.

Observe that the self-normalized weights are $\frac{\tilde{w}(X_i)}{\sum_{j=1}^{N} \tilde{w}(X_j)}$ summing up to one. Recall that for the basic IS, the weights are $\frac{1}{N} w(X_i)$ and may not be summed up to one. The self-normalized IS can be used to simulate the distribution of interest π by getting some samples $\{X_i\}_{i=1}^{N}$ from a proposal distribution q with associated weights

$$\left\{ W_i = \frac{\tilde{w}(X_i)}{\sum_{j=1}^{N} \tilde{w}(X_j)} \right\}_{i=1}^{N}$$

and approximating the distribution π by the empirical distribution

$$\sum_{i=1}^{N} \frac{\tilde{w}(X_i)}{\sum_{j=1}^{N} \tilde{w}(X_j)} \delta_{X_i}(dx)$$

thanks to the self-normalized weights summing up to one. Self-normalized IS is used widely because the target distribution can be approximated by this empirical distribution, unlike basic IS. We can apply self-normalized IS even when we know the normalizing constant of the target distribution, so self-normalized IS can be applied to any applications that basic IS can be applied. In Statistics,

especially in Bayesian inference, the proposal distribution q is selected as close as possible to the target distribution π without much consideration to the target function f. However, the samples with associated weights can represent the distribution π, and the corresponding empirical distribution can be used to approximate

$$\begin{aligned}\mathbb{E}_\pi(f) &\approx \int f(x)\sum_{i=1}^{N}\frac{\tilde{w}(X_i)}{\sum_{j=1}^{N}\tilde{w}(X_j)}\delta_{X_i}(dx) \qquad (X_i \overset{\text{iid}}{\sim} q)\\ &= \sum_{i=1}^{N}\frac{\tilde{w}(X_i)}{\sum_{j=1}^{N}\tilde{w}(X_j)}f(X_i)\end{aligned}$$

which is (1.4).

1.3 Mathematical Theory

The derivation of IS is pretty simple, but this simulation technique has a huge benefit in many areas of applications described before. For whatever reasons or applications, we prefer to have the lowest possible variance for our IS. This section will provide mathematical statements about the convergence theorem in the form of Central Limit Theorem for each estimator: Monte Carlo, basic IS, and self-normalized IS. This obviously includes results on the expectation and the variance of each estimator. Moreover, the optimal proposal density for each kind of IS is identified here. Current work in this research area state these results without proper assumptions and precise details. In particular, the mathematical statements with proper assumptions and proofs of Theorem 1.5 and Theorem 1.10 have never been rigorously established in the past. The essential proofs, which are for Theorem 1.5, 1.7 and 1.10, will be provided in Chapter 2, and only the mathematical statements and some straightforward proofs are given in this Chapter.

1.3.1 Convergence Theorems

Let's consider the Monte Carlo estimator $\hat{\mu}_\pi$ and the basic IS estimator $\hat{\mu}_q$ defined by (1.1) and (1.3), respectively. We can easily acquire the following theorem.

Theorem 1.3. $\mathbb{E}(\hat{\mu}_\pi) = \mu$ *and* $\mathbb{E}(\hat{\mu}_q) = \mu$. *Also,*

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_\pi) &= \frac{1}{N}\int (f(x)-\mu)^2\pi(x)\,dx\\ &= \frac{1}{N}\left(\int f(x)^2\pi(x)\,dx - \mu^2\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_q) &= \frac{1}{N}\int \frac{(f(x)\pi(x)-\mu q(x))^2}{q(x)}\,dx \\ &= \frac{1}{N}\left(\int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx-\mu^2\right).\end{aligned}$$

Since $\mathbb{E}(\hat{\mu}_\pi)=\mathbb{E}(\hat{\mu}_q)=\mu$, both $\hat{\mu}_\pi$ and $\hat{\mu}_q$ are unbiased estimators. Also, by the Central Limit Theorem, we have the following corollary.

Corollary 1.4. *If* $\int f(x)^2\pi(x)\,dx<\infty$,

$$\sqrt{N}\left(\hat{\mu}_\pi-\mu\right)\xrightarrow[N\to\infty]{D}\mathcal{N}\left(0,\int f(x)^2\pi(x)\,dx-\mu^2\right).$$

If $\int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx<\infty$,

$$\sqrt{N}\left(\hat{\mu}_q-\mu\right)\xrightarrow[N\to\infty]{D}\mathcal{N}\left(0,\int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx-\mu^2\right).$$

Now, consider the self-normalized IS estimator $\tilde{\mu}_q$ defined by (1.4). Because of the ratio formula of $\tilde{\mu}_q$, $\mathbb{E}(\tilde{\mu}_q)$ and $\mathrm{Var}(\tilde{\mu}_q)$ cannot be directly derived. Generally, $\mathbb{E}(\tilde{\mu}_q)\neq\mu$, so the self-normalized IS estimator is biased. However, by the Strong Law of Large Number together with the continuous mapping theorem, $\tilde{\mu}_q$ converges to μ almost surely as $N\to\infty$. In the same way as the basic IS, we can also have a convergence theorem for self-normalized IS. The following theorem tells us about the asymptotic variance of the estimator in the form of Central Limit Theorem.

Theorem 1.5. *If* $\int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx$ *and* $\int \frac{\pi(x)^2}{q(x)}\,dx$ *are finite, then*

$$\sqrt{N}\left(\tilde{\mu}_q-\mu\right)\xrightarrow[N\to\infty]{D}\mathcal{N}\left(0,\int \frac{(f(x)-\mu)^2\pi(x)^2}{q(x)}\,dx\right).$$

$\mathrm{AVar}(\tilde{\mu}_q)=\frac{1}{N}\int \frac{(f(x)-\mu)^2\pi(x)^2}{q(x)}\,dx$ will be called the asymptotic variance of the self-normalized IS estimator $\tilde{\mu}_q$. From Corollary 1.4 and Theorem 1.5, the rate of convergence for Monte Carlo, basic IS, and self-normalized IS is $O\left(\frac{1}{\sqrt{N}}\right)$. Since $\pi(x)=\frac{p(x)}{Z}$, Theorem 1.5 can also be restated in terms of p which is what is actually given. However, it is common to write with the π version when we deal with theory. The p version of this theorem is as follow.

Corollary 1.6. *If* $\int \frac{f(x)^2 p(x)^2}{q(x)}\,dx$ *and* $\int \frac{p(x)^2}{q(x)}\,dx$ *are finite, then*

$$\sqrt{N}\left(\tilde{\mu}_q - \mu\right) \xrightarrow[N\to\infty]{D} \mathcal{N}\left(0, \int \frac{(f(x)-\mu)^2 p(x)^2}{Z^2 q(x)}\,dx\right).$$

1.3.2 Optimal Proposal Densities

Unlike the classical Monte Carlo estimator which has a fixed variance, the variance of the basic IS estimator and the asymptotic variance of the self-normalized IS estimator can vary and can depend on the proposal distribution. If we can choose a good proposal density q that makes $\mathrm{Var}(\hat{\mu}_q)$ or $\mathrm{AVar}(\tilde{\mu}_q)$ lower than $\mathrm{Var}(\hat{\mu}_\pi)$, this will increase the precision of the IS estimate over the simple Monte Carlo method. However, a wrong choice of proposal distribution can make the estimation worse than that of the simple Monte Carlo method. We wish to find a proposal density that minimizes the variance of the basic IS estimator, and another for the asymptotic variance of the self-normalized IS estimator.

Theorem 1.7. *The basic IS proposal density* q *that satisfies the validity condition* $f\pi \ll q$ *and minimizes* $\mathrm{Var}(\hat{\mu}_q)$ *is*

$$q_b^*(x) = \frac{|f(x)|\pi(x)}{\int |f(x)|\pi(x)\,dx}. \tag{1.5}$$

Remark 1.8. q_b^* always satisfies the validity for being a basic IS proposal density

$$f\pi \ll q_b^*$$

as well as the necessary condition for the convergence theorem for the basic IS estimators

$$\int \frac{f(x)^2\pi(x)^2}{q_b^*(x)}\,dx < \infty.$$

Remark 1.9. If f is non-negative, then $\mathrm{Var}(\hat{\mu}_{q_b^*}) = 0$. This fact can be easily verified by calculating $\mathrm{Var}(\hat{\mu}_{q_b^*})$ with $q_b^*(x) = \frac{f(x)\pi(x)}{\mu}$. Similarly, if f is non-positive, then $\mathrm{Var}(\hat{\mu}_{q_b^*}) = 0$.

Theorem 1.10. *The self-normalized IS proposal density* q *that minimizes* $AVar(\tilde{\mu}_q)$ *is*

$$q_{sn}^*(x) = \frac{|f(x)-\mu|p(x)}{\int |f(x)-\mu|p(x)\,dx} \tag{1.6}$$

provided that $p \ll q_{sn}^*$.

Remark 1.11. q^*_{sn} may not satisfy the self-normalized IS validity condition $p \ll q^*_{sn}$ nor the assumption of the convergence theorem for self-normalized IS estimators $\int \frac{f(x)^2\pi(x)^2}{q^*_{sn}(x)}dx, \int \frac{\pi(x)^2}{q^*_{sn}(x)}dx < \infty$. However, the validity condition $p \ll q^*_{sn}$ is satisfied for all target functions f that is not constant at μ on a set of finite measure. Hence, q^*_{sn} satisfies the self-normalized IS validity condition $p \ll q^*_{sn}$ in most of real-life problems.

q^*_b in Theorem 1.7 and q^*_{sn} in Theorem 1.10 will be called the optimal proposal densities for basic IS and self-normalized IS, respectively. Although we know what is the optimal proposal density for basic IS from Theorem 1.7, we cannot use it in reality because we do not know $\int |f(x)|\pi(x)\,dx$. Even in an easy case when f is always non-negative, this term equals to μ which is what we want to estimate at first, so we do not know $\int |f(x)|\pi(x)\,dx$ in advance for the problem that we want to apply the basic IS method. The same argument also apply to the self-normalized IS method. We do not know in advance both μ and the normalizing constant $\int |f(x)-\mu|p(x)\,dx$ needed for the optimal proposal density for self-normalized IS acquired from Theorem 1.10.

1.4 Current Approaches

As discussed in the previous section, we cannot use the theoretically optimal proposal density in IS method. So, what do people do?

This research area of IS are growing fast in Statistics. What statisticians currently do for basic IS is to select a class of known distributions and try to find a distribution within that class which minimizes $\text{Var}(\hat{\mu}_q)$. One of the best and recent methods is discussed in Subsection 1.4.1. To select a class of distributions for the proposal distribution q, statisticians focus on balancing weights of the samples, $w(X_i)$'s where $X_i \overset{\text{iid}}{\sim} q$, and try to minimize variance of these associated weights because dramatic fluctuation in weights usually results in a bad estimation. Hence, they want $\text{Var}_q(w)$ to be bounded. Consequently,

$$\int \frac{\pi(x)^2}{q(x)}\,dx < \infty \tag{1.7}$$

is used as a rule of thumb [23, 34] in selecting a class of distributions for q. This is also the case in the self-normalized IS, since the weight function is the scalar multiplication version of that of basic IS. The equivalent rule of thumb for self-normalized IS is

$$\int \frac{p(x)^2}{q(x)}\,dx < \infty,$$

and statisticians just rely on only this rule of thumb to select a good proposal density.

1.4.1 Efficient Importance Sampling

There is a widely cited method called *efficient IS* [28, 29] which may be praised as the best method for IS currently. It is an iterative method for basic IS that still has some limitations. It works only for a positive target function, and the proposal distribution has to be chosen in an exponential family of distributions. By the way, the rule of thumb (1.7) is broadly used to select such exponential family of distributions.

Suppose a class of known distributions indexed by a vector of auxiliary parameters

$$Q = \{q_a \mid a \in A\}$$

is carefully selected, where A is a space of auxiliary parameters. Known distributions are used because they provide a quick access in sampling by various available software. The expression of $\mathrm{Var}(\hat{\mu}_{q_a})$ from Theorem 1.3 can be re-expressed as

$$\sigma^2(a) = \mathrm{Var}(\hat{\mu}_{q_a}) = \int h\left[g_a^2(x)\right] f(x)\pi(x)\, dx$$

where

$$g_a(x) = \log\left[\frac{f(x)\pi(x)}{\mu\, q_a(x)}\right]$$

and

$$h(x) = e^{\sqrt{x}} + e^{-\sqrt{x}} - 2.$$

Note that h is monotone, convex on $\mathbb{R}_+$, and $h(x) \geq x$. Minimizing $\sigma^2(a)$ with respect to a calls for nonlinear optimization. Consider the simpler function

$$\begin{aligned} v(a) &= \int \left[g_a^2(x)\right] f(x)\pi(x)\, dx \\ &= \int \left\{\log[f(x)\pi(x)] - \log(\mu) - \log[q_a(x)]\right\}^2 f(x)\pi(x)\, dx. \end{aligned}$$

Let a^* and $\hat{a}$ be the optimal parameters minimizing $\sigma^2(a)$ and $v(a)$, respectively. Then, we can have that

$$\sigma^2(\hat{a}) \geq \sigma^2(a^*) \geq h[v(a^*)] \geq h[v(\hat{a})]$$

which gives an upper bound and a lower bound for $\sigma^2(a^*)$, the smallest variance over the selected class Q.

To seek for $\hat{a}$, an iterative method is proposed. First, one selects an initial distribution, say q_{a_0},

and draws a number of samples from this distribution: $x_i^0 \overset{\text{iid}}{\sim} q_{a_0}$ for all $i = 1, \ldots, R$. Then, minimizing

$$\hat{v}_R(a) = \frac{1}{R} \sum_{i=1}^{R} \left\{ \log\left[f(x_i^0) \pi(x_i^0) \right] - c - \log\left[q_a(x_i^0) \right] \right\}^2 f(x_i^0) \frac{\pi(x_i^0)}{q_{a_0}(x_i^0)}$$

with respect to a and c, an intercept in place of the unknown $\log(\mu)$, takes the form of a simple weighted linear least square problem, since Q is chosen to be an exponential family of distributions. Starting from a_0, a_1 is obtained as a result of solving the above generalized least square problem. The method can be iterated by using as initial sampler in any given round the optimized sampler from the previous round. The process continues until a stable solution for a is reached, and that final a is used as $\hat{a}$. It is said that by experience, no more than 3 to 4 iterations are required to produce the final $\hat{a}$. Eventually, $q_{\hat{a}}$ is used as a proposal distribution for basic IS.

There is also an extension of efficient IS using a mixture of distributions for a proposal distribution [20]. The use of a mixture of distributions can improve the performance of IS method in the case of heavy tailed or multi-modal target densities. Still, choosing a class of distributions at the beginning of the process is a serious issue. If an improper class of distributions is selected, the final result can be quite bad.

1.4.2 Truncated Importance Sampling

There is another readily applicable and theoretically justifiable approach for basic IS called *truncated IS* [16]. Since unfavorable IS estimation results usually come from harsh fluctuation in samples' weights, this method suggests to truncate extreme weights by some constant depending on N, the number of samples. The truncated IS estimator for μ is

$$\hat{\mu}'_q = \frac{1}{N} \sum_{i=1}^{N} w'(X_i) f(X_i), \quad X_i \overset{\text{iid}}{\sim} q. \tag{1.8}$$

where

$$w'(X_i) = w(X_i) \wedge \tau_N,$$

the minimum of $w(X_i)$ and τ_N.

This estimator is biased, but the bias goes to zero as $N \to \infty$, provided that $\lim_{N\to\infty} \tau_N = \infty$. Also, the variance of the estimator goes to zero as $N \to \infty$, if $\mathrm{Var}_\pi(f) < \infty$ and $\lim_{N\to\infty} \frac{\tau_N}{N} = 0$. The recommended truncation rate in the literature is $\tau_N = \sqrt{N}$. Although this method can reduce the sensitivity of importance sampling on the choice of the proposal distribution, it can slightly reduce the rate of convergence for the IS method and it is a biased method for basic IS.

Although this method can be easily implemented and work well in many applications, it may

overlook something that can cause extreme fluctuation of each term in the summation formula of the IS estimator. For example, consider basic IS with $f(x) = \frac{e^{\frac{x}{2}}}{\sqrt[4]{x}}$, $\pi(x) = 2e^{-2x}$, and $q(x) = \frac{5}{2}e^{-\frac{5}{2}x}$. Note that this q satisfies the rule of thumb (1.7). The basic IS weight function is $w(x) = \frac{4}{5}e^{\frac{x}{2}}$. So, if we truncate the weight function with some threshold, we will truncate weights only for very large samples. But, each term in the summation formula of the truncated IS estimator from (1.8), $w'(x)f(x) = \frac{4}{5}\frac{e^x}{\sqrt[4]{x}} \wedge \tau_N \frac{e^{\frac{x}{2}}}{\sqrt[4]{x}}$, can blow up for both too large and too small samples which can cause high variance in estimation. Thus, the truncation cannot detect the problem with very small samples, and the estimation can be poor. Furthermore, consider another example with $f(x) = x^3$, $\pi(x) = 2e^{-2x}$, and $q(x) = 4xe^{-2x}$ which satisfies the rule of thumb (1.7). This setup yields $w(x) = \frac{1}{2x}$ and $w'(X_i)f(X_i) = \frac{x^2}{2} \wedge \tau_N x^3$. The truncation will occur only for small samples, even though the real problem is from large samples. This method may fail because it too focuses on the fluctuation of samples' weights without taking into account the target function f. For truncated IS, we should truncate $w(x)f(x)$ instead of $w(x)$.

1.5 Examples

To understand better what is going on when IS is performed, the following example for IS is provided. This example also shows the insufficiency of the widely used rule of thumb (1.7) in choosing a proposal density.

Example 1.12. Let the target distribution π be an exponential distribution with parameter mean $\frac{1}{2}$, and a target function

$$f(x) = x^3.$$

The distribution π has density

$$\pi(x) = 2e^{-2x}.$$

Consider applying IS method with the following proposal distributions:

1. a gamma distribution with shape parameter 4 and scale parameter $\frac{1}{2}$
2. a no-name probability distribution whose density is directly proportional to a function $\left|x^3 - \frac{3}{4}\right| e^{-2x}$
3. a Weibull distribution with scale parameter $\frac{1}{2}$ and shape parameter 2
4. an equal mixture distribution between a gamma distribution with shape parameter 3 and scale parameter $\frac{1}{4}$ and an exponential distribution with parameter mean $\frac{1}{4}$

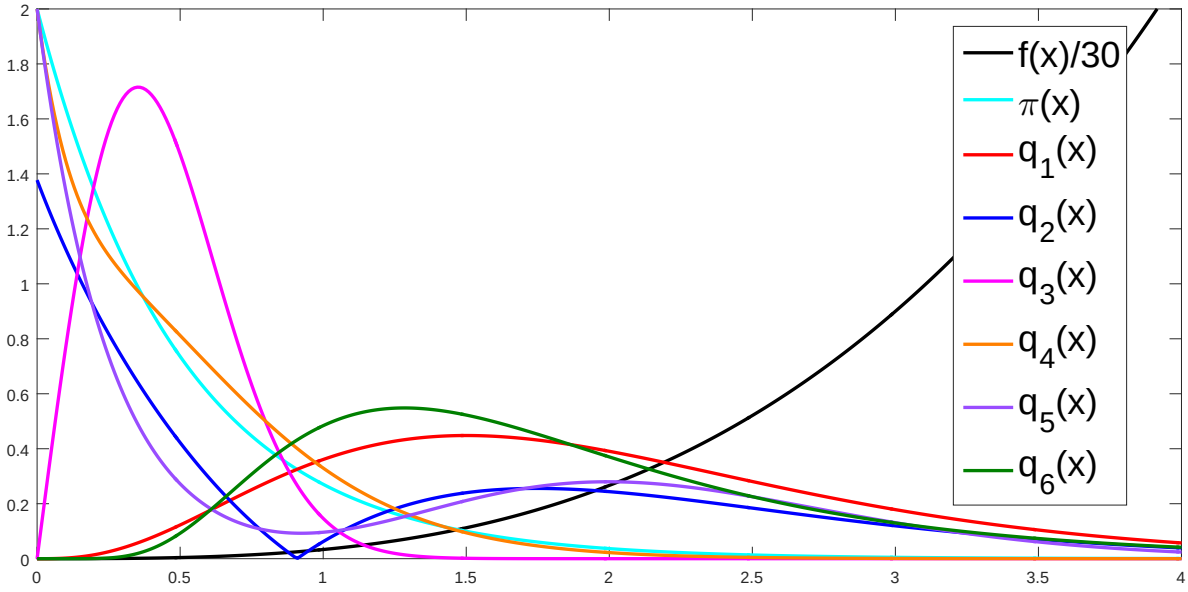

Figure 1.1 All PDFs in Example 1.12 with the scaled target function.

5. an equal mixture distribution between a gamma distribution with shape parameter 9 and scale parameter $\frac{1}{4}$ and an exponential distribution with parameter mean $\frac{1}{4}$
6. a log-normal distribution with parameter $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{4}$

which have respectively the following densities:

$$
\begin{aligned}
q_1(x) &= \tfrac{8}{3} x^3 e^{-2x} && \text{Gamma}(4,2) \\
q_2(x) &= \tfrac{1}{Z_2} \left| x^3 - \tfrac{3}{4} \right| e^{-2x} && \text{where } Z_2 = \tfrac{3}{2}\left(\tfrac{1}{2} + \left(\tfrac{3}{4}\right)^{\frac{1}{3}} + \left(\tfrac{3}{4}\right)^{\frac{2}{3}} \right) e^{-2(\frac{3}{4})^{\frac{1}{3}}} \\
q_3(x) &= 8x e^{-4x^2} && \text{Weibull}\left(\tfrac{1}{2}, 2\right) \\
q_4(x) &= 2(8x^2+1) e^{-4x} && \tfrac{1}{2}\,\text{Gamma}(3,4) + \tfrac{1}{2}\,\text{Exp}(4) \\
q_5(x) &= \left(\tfrac{1024}{315} x^8 + 2\right) e^{-4x} && \tfrac{1}{2}\,\text{Gamma}(9,4) + \tfrac{1}{2}\,\text{Exp}(4) \\
q_6(x) &= \tfrac{\sqrt{2}}{\sqrt{\pi} x} e^{-2(\log x - \frac{1}{2})^2} && \text{LogNormal}\left(\tfrac{1}{2}, \tfrac{1}{4}\right).
\end{aligned}
$$

Fig. 1.1 shows all the densities with the scaled target function f for ease of comparison. We observe that π has high probability around zero, but f is close to zero around that area. In this

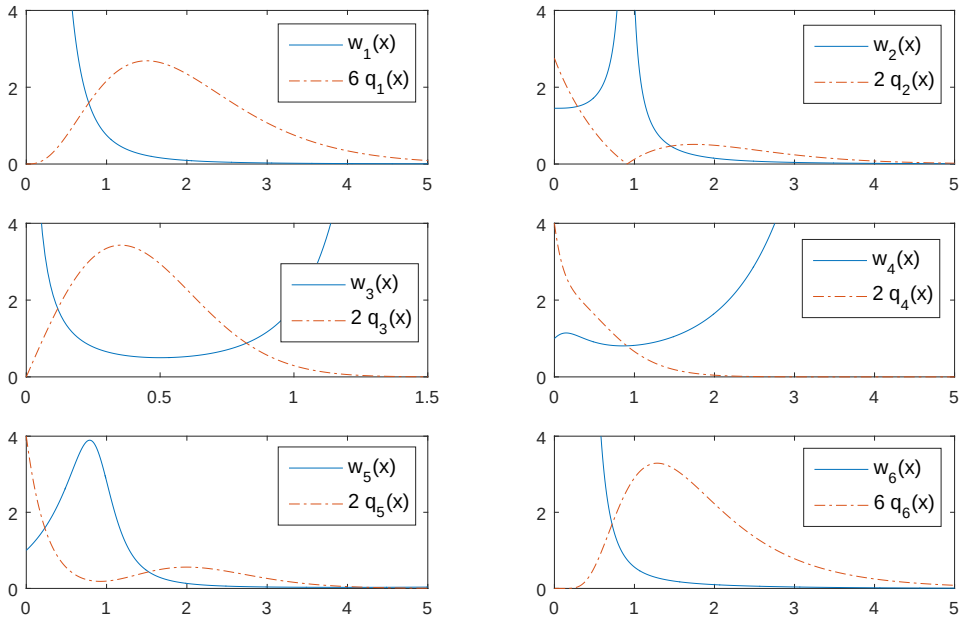

Figure 1.2 Basic IS weight functions

example, q_1 is the optimal basic IS proposal density acquired from (1.5) in Theorm 1.7, q_2 is the optimal self-normalized IS proposal density acquired from (1.6) in Theorm 1.10, and the true expectation is $\mu = \int_0^\infty f(x)\pi(x)\,dx = \frac{3}{4}$.

Basic IS method is considered first. All basic IS weight functions using Eq. 1.2 are presented in Fig. 1.2. We perform a simple Monte Carlo simulation and the basic IS with the proposals $q_1, \ldots, q_6$, and compare the results with the theoretical expected value.

In this simulation, we sample 10,000 samples from each distribution: $\pi, q_1, \ldots, q_6$. Fig. 1.3 shows histograms of these random samples. They tend to fit the corresponding densities very well. Since q_2 is a no-name distribution, we do not have a direct command in computer programming for sampling from this distribution. To sample from q_2, we use the acceptance-rejection method from Appendix A. Here, we use the exponential distribution with parameter mean 1.3, where its density is $r(x) = \frac{1}{1.3} e^{-\frac{1}{1.3}x}$, as a instrumental distribution to q_2 in the acceptance-rejection method with the bounding constant $c = 1.3 \times \frac{3}{4} \times \frac{1}{Z_2}$. We can easily verify that $\frac{q_2(x)}{c\,r(x)} \le 1$ for all $x \ge 0$. The graph of $q_2(x)$ with a scaled version of $r(x)$, $c\,r(x)$, is also shown in Fig. 1.3. As for sampling from q_4 and q_5, we first draw a Bernoulli random number with parameter $\frac{1}{2}$ to consider which distribution in the mixture is

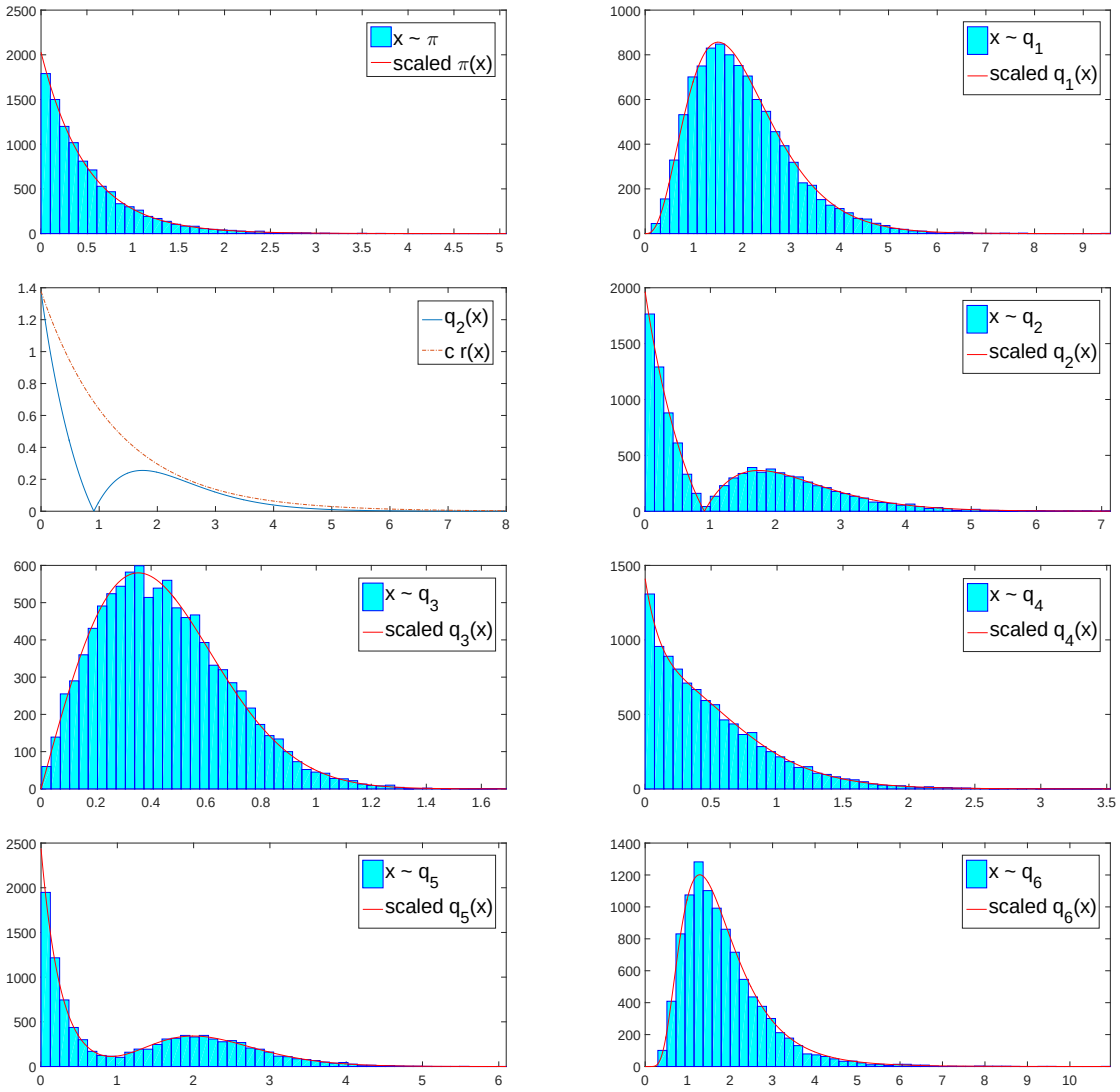

Figure 1.3 Histograms of samples from $\pi, q_1, \ldots, q_6$ with the graph of $c\,r(x)$

used, and then simply draw a sample according to that selected distribution in the mixture.

Fig. 1.4 shows basic IS performance of each proposal distribution. Each sub-graph has a difference scale for easier comparison. We sample 10,000 samples and approximate μ using (1.3). The x-axis is the number of samples, N, and the y-axis is the approximated expectation using the first N samples. We can see that the proposal q_1 has the best performance, and q_3 has the worst performance. However, this is just one simulation. To see more about the accuracy of each proposal

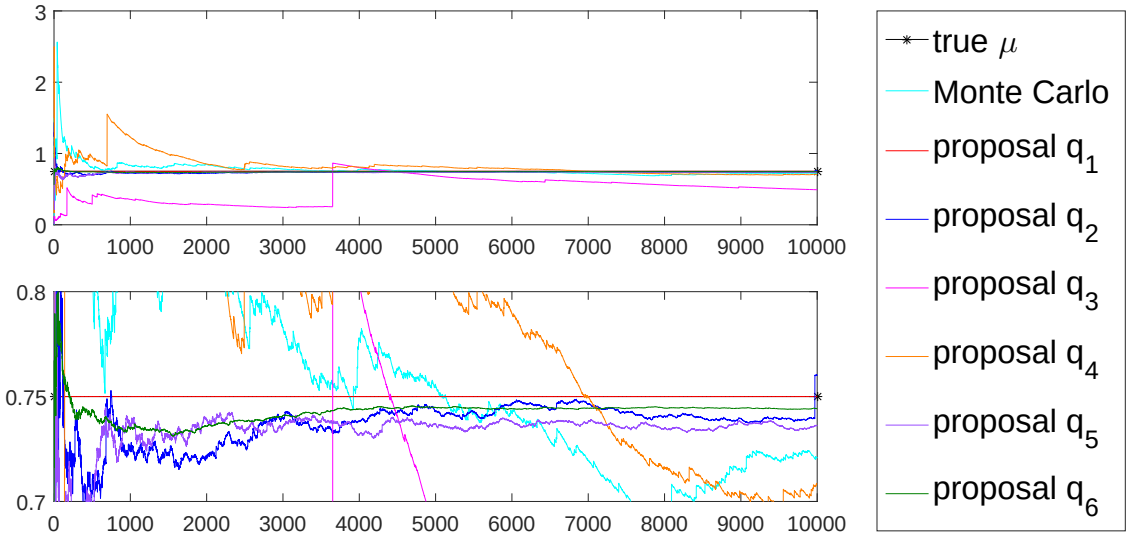

Figure 1.4 One simulation of basic IS

distribution, we perform many replications of this scenario. However, before we go into that simulation, the jump behavior of the proposal distributions q_3 and q_4 is interesting. Later, we will see more about this jump behavior, so we discuss this issue here using this example of the proposal q_3.

From Fig. 1.2, we can see that the weight function for the proposal q_3 will blow up when x is too small or too large and this area has low probability to be sampled from. From the basic IS formula (1.3), we need the product of the weight function and the target function in the summation. In this case it is $\frac{\pi(x)}{q_3(x)}f(x) = \frac{x^2 e^{4x^2-2x}}{4}$ and this amount will blow up for only large x. So, if we have a large sample which rarely happen here, that term will increase the total summation drastically and result in a jump-up behavior. However, this large sample will have less effect on the total summation if we use a large total number of samples, N. A very low sample does not make the factor $\frac{x^2 e^{4x^2-2x}}{4}$ high in the calculation, although the corresponding weight is pretty high. Note that in this case, a sample which is greater than 1.6 may be considered a large sample. From Fig. 1.3, we can see that it is very hard to get a sample with really high value by sampling from the proposal q_3. We will get small samples most of the time so that each term of $\frac{x^2 e^{4x^2-2x}}{4}$ in the summation will keep low, and when a big sample occurs, we get a jump. After a jump occurs, it will come back to keep getting lower and lower again until the next high sample appears and brings another jump. All information of this discussion is provided in Fig. 1.5.

Now, we arrive at the most interesting part of this example where we perform several scenarios of the basic IS method. Here, we generate 100 scenarios of the previous simulation and plot all

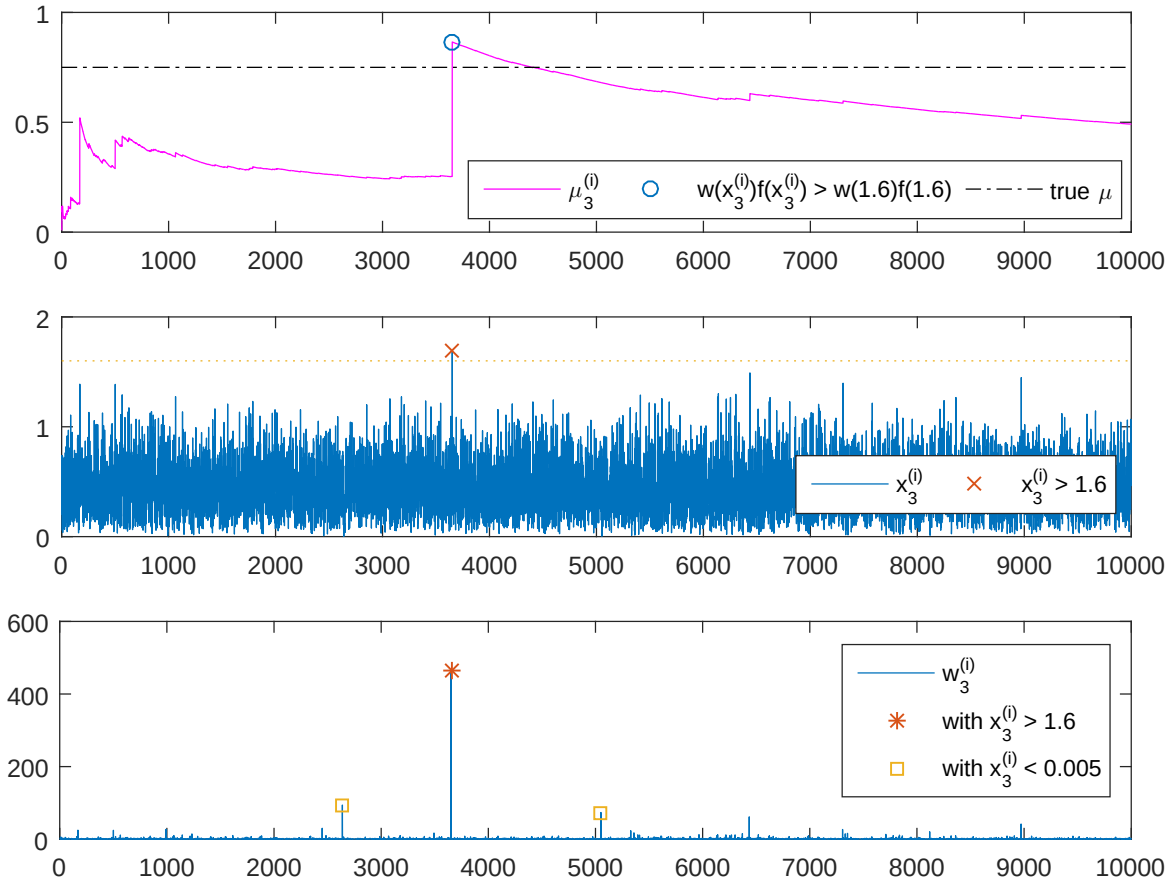

Figure 1.5 Jump analysis for q_3

results in the same graph for each proposal distribution to envisage variance of each estimator. The theoretical variance of the simple Monte Carlo estimator is $\text{Var}(\hat{\mu}_\pi) = \frac{1}{N}\left(\int 2x^6 e^{-2x}\, dx - \left(\frac{3}{4}\right)^2\right) = \frac{1}{N}\left(\frac{45}{4} - \frac{9}{16}\right) = \frac{171}{16N} = \frac{10.6875}{N}$. Note that the true variance is not known in real-world problems, but this example composing of easy functions need to know the true expectation and variance for comparison purpose. Fig. 1.6 shows the performance of Monte Carlo method together with the plot of $\mu \pm \sqrt{\text{Var}(\hat{\mu}_\pi)} = \frac{3}{4} \pm \frac{3\sqrt{19}}{4}\frac{1}{\sqrt{N}}$ indicating the true rate of convergence which is of the order $O(\frac{1}{\sqrt{N}})$. Fig. 1.7 shows the result of all proposal distributions. We can now visualize about the accuracy each distribution can offer.

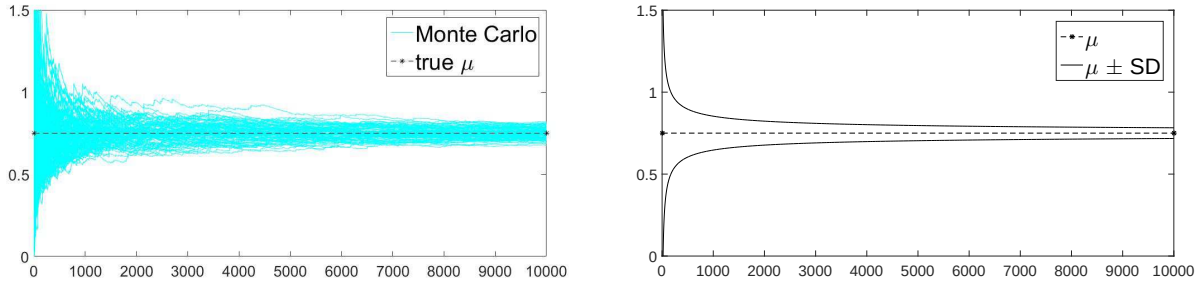

Figure 1.6 The performance of simple Monte Carlo method

From Fig. 1.7, the proposal q_1, which is the optimal basic IS proposal density q_b^* from Theorm 1.7, seems to give the ideal result which is zero variance. This coincides with the theory. From Theorem 1.3,

$$\mathrm{Var}(\hat{\mu}_q) = \frac{1}{N}\left(\int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx - \mu^2\right).$$

We can calculate that $\mathrm{Var}(\hat{\mu}_{q_1}) = \frac{1}{N}\left(\int \frac{3}{2}x^3 e^{-2x}\,dx - \left(\frac{3}{4}\right)^2\right) = \frac{1}{N}\left(\frac{3}{2}\frac{3}{8} - \frac{9}{16}\right) = 0$ for all N, the number of samples. This is the ideal proposal density for the basic IS with this specific f and π. How could it be possible to get a zero variance in a Monte Carlo approximation using random samples? This is because in calculation, f multiplied with π just perfectly cancels out q_1 and leaves just a constant: $\frac{f(x)\pi(x)}{q_1(x)} = \frac{3}{4}$ for any sample drawn from q_1. This explains why we obtain the true expectation $\frac{3}{4}$ for all N since the beginning $N = 1$, and this really coincides with the theory about zero variance for the optimal proposal density in the case of non-negative f.

Now, a very interesting observation is that according to Fig. 1.4, Fig. 1.6 and Fig. 1.7, the proposal q_6 tends to have much smaller variance comparing to the simple Monte Carlo method. However, we can check that $\int \frac{f(x)^2\pi(x)^2}{q_6(x)}\,dx = \infty$ which means $\mathrm{Var}(\hat{\mu}_{q_6}) = \infty$ for all N. This seems like a contradiction. Also, we can check that $\mathrm{Var}(\hat{\mu}_{q_2}) = \mathrm{Var}(\hat{\mu}_{q_3}) = \mathrm{Var}(\hat{\mu}_{q_4}) = \infty$. We can clearly notice the jump behavior in both q_3 and q_4 cases, and there are a small number of jumps in q_2, q_5 and q_6 cases. Normally, if we have a jump behavior in basic IS with a proposal distribution, then that proposal distribution may yield an infinite variance of the corresponding basic IS estimator. We will continue to discuss this issue together with the conflict about the simulation performance of q_6 versus its theoretical aspect in Chapter 3.

Now, the self-normalized IS is performed using the same setup of $f, p = \pi, q_1, \ldots, q_6$. From remark 1.2, we can also see that we can use p as π or any constant multiplication of π. The resulting calculation for using π multiplied by an arbitrary positive constant will turn out to be the same. Here, we choose to use p as π for simplicity. Note again that, the proposal distribution q_2 is the

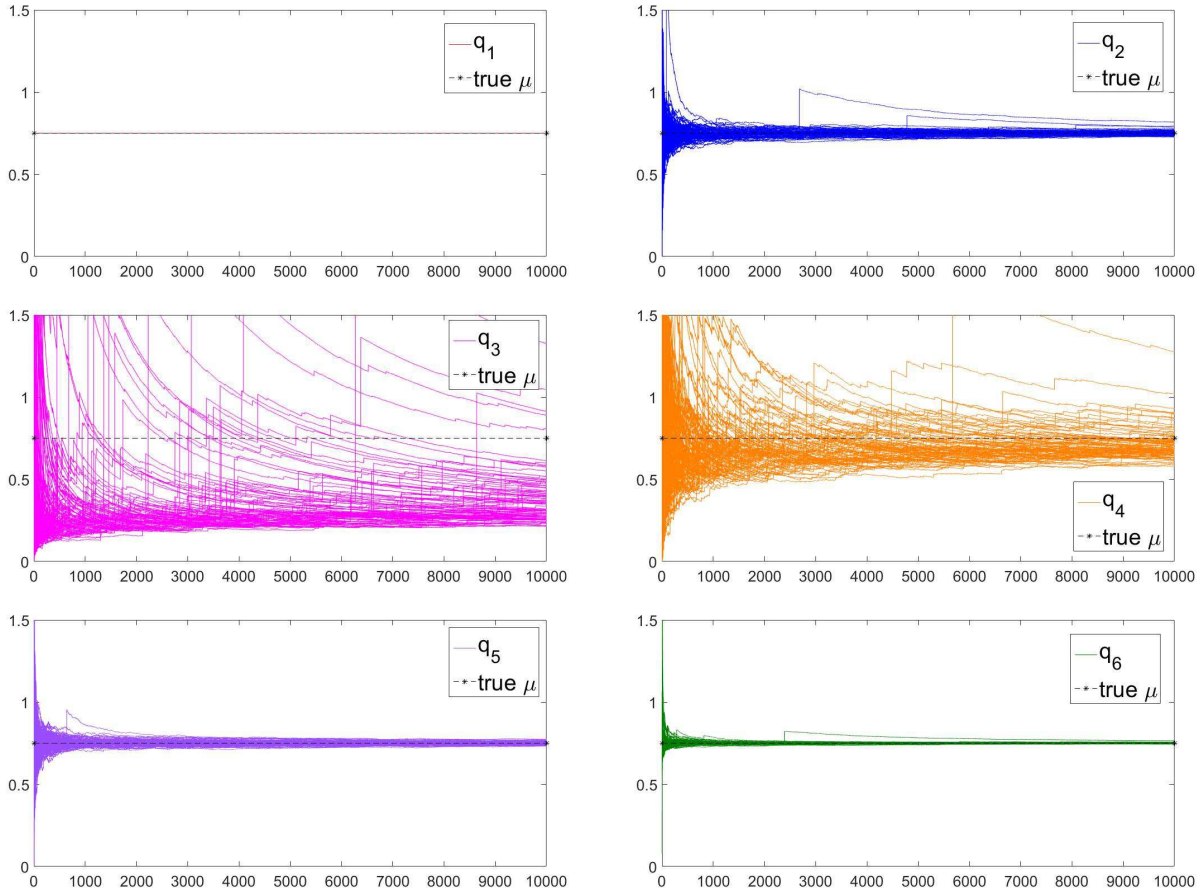

Figure 1.7 Basic importance sampling performance

optimal density q^*_{sn} acquired from Theorem 1.10.

Fig. 1.8 shows the self-normalized IS result of all proposal distributions using the same sampled data when we perform the basic IS method. From Theorem 1.5, the asymptotic variance for the proposal q is given by

$$\mathrm{AVar}(\tilde{\mu}_q) = \frac{1}{N}\int \frac{(f(x)-\mu)^2\pi(x)^2}{q(x)}\,dx.$$

We can check that the asymptotic variance for the proposals q_1, q_3, q_4 and q_6 are all infinite. Indeed, these proposal distributions or even q_2 which is q^*_{sn} do not satisfy the assumption of Theorem 1.5, the convergence theorem for self-normalized IS. Thus, they are not even pre-qualified to be considered to have an asymptotic variance. We will discuss this issue more in Chapter 3, and call their asymptotic variance by the above formula as $\mathrm{AVar}(\tilde{\mu}_q)$ for now. The asymptotic variance for the

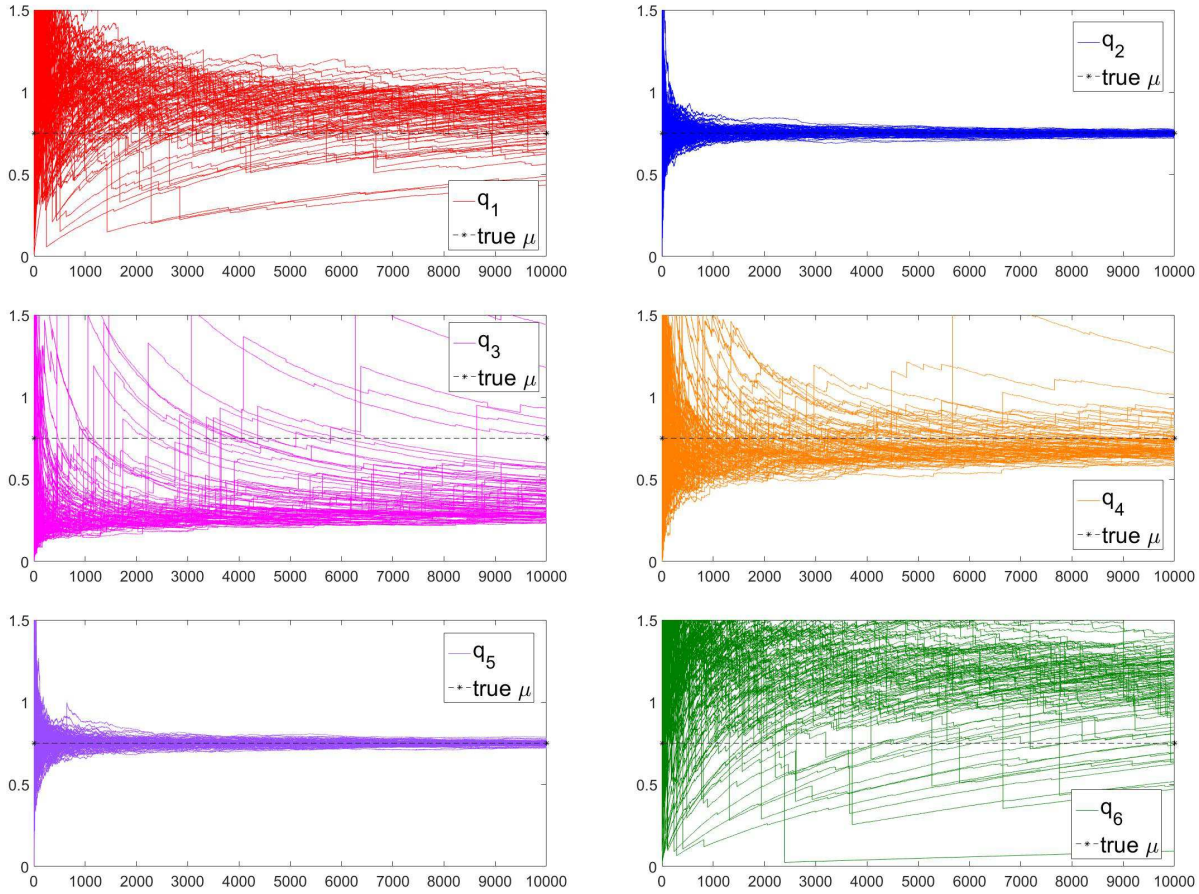

Figure 1.8 Self-normalized importance sampling performance

proposal q_2 is $\frac{9}{8N}\left(14+7(6^{\frac{1}{3}})+4(6^{\frac{2}{3}})\right)e^{-2(6^{\frac{1}{3}})} \approx \frac{1.18603}{N}$. We can see that q_1, q_3, q_4 and q_6 yield the jump behavior indicating very poor performance due to infinite asymptotic variance. We can see that there are both jump-up and jump-down behaviors unlike the basic IS method. The jump behavior is interesting and we will explain more here using this example.

For the case of non-negative f, in basic IS method, the jump behavior usually occurs when weights in the summation of the estimator formula are very high, and only the jump-up behavior can occur. We should observe that from Fig. 1.2, q_1 and q_6 have similar graphs of densities and weight functions, but q_1 which is the optimal proposal density does not have the jump behavior while q_6 does. This is because f also has an effect on the summation formula. From (1.4), if the new adding term, from N to $N+1$ iteration, of weights in the formula is comparatively high, the new summation can be jump-down because of the new much higher denominator. However, the target

function f may make the adding term in numerator so huge that the whole numerator dominates the increase in denominator which has no effect from f, hence the jump-up behavior. We can see that not only π and q but f has an effect on the performance of IS method. The jump-up or jump-down behavior really depends on all p, q, and f, but neither behaviors are good. The jump behavior usually suggests that the estimator has infinite variance. However, there are some cases where the variance or asymptotic variance can be bounded as shown in the case of our q_5 here. So, we cannot judge by just checking the appearance of the jump behavior.

It is quite hard to choose a good proposal density for self-normalized IS. The optimal one is usually an unnatural density as we may realize even in this one-dimensional problem. In the same way as basic IS, we cannot use the optimal density in reality due to the unknown μ in the first place.

An important thing to notice from this example is that q_4 satisfies the rule of thumb (1.7) in choosing a proposal density, but it obviously has a bad performance for both basic and self-normalized IS methods. Also, if the proposal distribution is not well chosen such as q_3 and q_4 , the IS method may be worse than the regular Monte Carlo method.

The purpose of this example is to illustrate both basic and self-normalized IS methods, so we need to know the theoretical answer to compare the results. However, more suitable problems where IS method should apply will be more complicated, and we cannot use the optimal proposal in reality. Moreover, this example shows that the rule of thumb (1.7) is not a perfect criterion for selecting a proposal density for IS. In addition, this example can spark an idea to carry on this work. We will discuss more using this example in Chapter 3.

CHAPTER

2

MATHEMATICAL PROOFS

The proofs for Theorem 1.5, 1.7 and 1.10 are separated from their statements in Section 1.3 and presented in this chapter.

2.1 Proof of Theorem 1.5

This section provides the proof of Theorem 1.5. The proof relies on the so-called delta method which is, provided by the aid of Taylor expansions, a generalization of Central Limit Theorem. In the proof, we will really see why we need some additional assumptions which require $\int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx$ and $\int \frac{\pi(x)^2}{q(x)}\,dx$ to be finite. Note that there is a proof given in [11], but that proof does not cite the reference properly.

Proof of Theorem 1.5. We will apply the delta method from Appendix B to $\tilde{\mu}_q$. Let

$$A_i = \tilde{w}(X_i)f(X_i) \quad \text{and} \quad B_i = \tilde{w}(X_i)$$

where $X_i \overset{iid}{\sim} q$ so that the random vectors (A_i, B_i) are independent and identically distributed. Then, $\mathbb{E}(A_1) = \mu Z$ and $\mathbb{E}(B_1) = Z$. Denote the variance of A_1, the variance of B_1, and the covariance between A_1 and B_1 by σ_A^2, σ_B^2, and σ_{AB}, respectively.

$$\begin{aligned}\sigma_A^2 &= \int \left(\frac{p(x)}{q(x)} f(x)\right)^2 q(x)\,dx - (\mu Z)^2 \\ &= Z^2\left(\int \frac{\pi(x)^2 f(x)^2}{q(x)}\,dx - \mu^2\right).\end{aligned}$$

Also,

$$\begin{aligned}\sigma_B^2 &= \int \left(\frac{p(x)}{q(x)}\right)^2 q(x)\,dx - Z^2 \\ &= Z^2\left(\int \frac{\pi(x)^2}{q(x)}\,dx - 1\right)\end{aligned}$$

and

$$\begin{aligned}\sigma_{AB} &= \mathbb{E}\left[\left(\tilde{w}(X_1) f(X_1) - \mu Z\right)\left(\tilde{w}(X_1) - Z\right)\right] \\ &= \mathbb{E}\left[\tilde{w}(X_1)^2 f(X_1) - Z\,\tilde{w}(X_1) f(X_1) - \mu Z\,\tilde{w}(X_1) + \mu Z^2\right] \\ &= Z^2\left(\int \frac{\pi(x)^2 f(x)}{q(x)}\,dx - \int \pi(x) f(x)\,dx - \mu \int \pi(x)\,dx + \mu\right) \\ &= Z^2\left(\int \frac{\pi(x)^2 f(x)}{q(x)}\,dx - \mu\right).\end{aligned}$$

Note that by Holder's inequality,

$$\begin{aligned}\int \frac{\pi(x)^2 f(x)}{q(x)}\,dx &= \int \left(\frac{\pi(x) f(x)}{\sqrt{q(x)}}\right)\left(\frac{\pi(x)}{\sqrt{q(x)}}\right) dx \\ &\leq \left(\int \frac{\pi(x)^2 f(x)^2}{q(x)}\,dx\right)^{\frac{1}{2}} \left(\int \frac{\pi(x)^2}{q(x)}\,dx\right)^{\frac{1}{2}}.\end{aligned}$$

Thus, by the assumption that $\int \frac{f(x)^2 \pi(x)^2}{q(x)}\,dx$ and $\int \frac{\pi(x)^2}{q(x)}\,dx$ are finite, all σ_A^2, σ_B^2 and σ_{AB} are finite. By Central Limit Theorem, we have that

$$\sqrt{N}\left(\left(\frac{1}{N}\sum_{i=1}^N \tilde{w}(X_i) f(X_i), \frac{1}{N}\sum_{i=1}^N \tilde{w}(X_i)\right) - (\mu Z, Z)\right) \xrightarrow[N\to\infty]{D} \mathcal{N}_2(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{bmatrix}.$$

Now, let $g(a,b)=\frac{a}{b}$. We have that

$$g\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{w}(X_i)f(X_i),\frac{1}{N}\sum_{i=1}^{N}\tilde{w}(X_i)\right)=\tilde{\mu}_q.$$

We simply calculate $g(\mu Z,Z)=\mu$ and $\nabla_{\vec{\mu}}^T:=\nabla g(\mu Z,Z)^T=(\frac{1}{Z},-\frac{\mu}{Z})$. Then,

$$\begin{aligned}\nabla_{\vec{\mu}}^T\,\Sigma\,\nabla_{\vec{\mu}}&=\frac{1}{Z^2}\sigma_A^2+\frac{\mu^2}{Z^2}\sigma_B^2-2\frac{\mu}{Z^2}\sigma_{AB}\\&=\left(\int\frac{\pi(x)^2f(x)^2}{q(x)}\,dx-\mu^2\right)+\mu^2\left(\int\frac{\pi(x)^2}{q(x)}\,dx-1\right)-2\mu\left(\int\frac{\pi(x)^2f(x)}{q(x)}\,dx-\mu\right)\\&=\int\frac{\pi(x)^2(f(x)-\mu)^2}{q(x)}\,dx.\end{aligned}$$

Applying the delta method, we obtain $\sqrt{N}\left(\tilde{\mu}_q-\mu\right)\xrightarrow[N\to\infty]{D}\mathcal{N}\left(0,\int\frac{(f(x)-\mu)^2\pi(x)^2}{q(x)}\,dx\right)$. □

2.2 Proof of Theorem 1.10

The proof of Theorem 1.10 is given in this section. Moreover, the proof of Theorem 1.7, which can be found in [19, 30, 25], is provided in details here. We will see the way to come up with the proof for Theorem 1.7 which is to seek for a nominee for the optimal proposal density and then show that such nominee really is the optimal one. Then, we can use this line of proof to prove Theorem 1.10. We now begin with the proof of Theorem 1.7.

Proof of Theorem 1.7. To minimize $\int\frac{f(x)^2\pi(x)^2}{q(x)}\,dx$ subject to a constraint $\int q(x)\,dx=1$, we apply the method of Lagrange multipliers for calculus of variations from Appendix C at least to find the candidate for the minimizer. Note that we also have constraints π and q being non-negative. Let

$$L(x,q,\lambda)=\frac{f(x)^2\pi(x)^2}{q(x)}+\lambda q(x).$$

Setting

$$\frac{\partial L}{\partial q}=-\frac{f(x)^2\pi(x)^2}{q(x)^2}+\lambda=0,$$

we have that

$$q(x)=\sqrt{\frac{f(x)^2\pi(x)^2}{\lambda}}=\frac{|f(x)|\pi(x)}{\sqrt{\lambda}}.$$

Since $\int q(x)\,dx=1$, we have that $q(x)=\frac{|f(x)|\pi(x)}{\int|f(x)|\pi(x)\,dx}$. Note that this method of Lagrange multipliers for calculus of variations does not give us a complete proof as Kahn and Marshall [19] claimed. Now,

we will show that this

$$q_b^*(x) = \frac{|f(x)|\pi(x)}{\int |f(x)|\pi(x)\,dx}$$

yields the minimum variance among all valid basic IS proposal densities. Obviously, $f\pi \ll q_b^*$. Let q be any density satisfying $f\pi \ll q$. Then,

$$\begin{aligned}
\int \frac{f(x)^2\pi(x)^2}{q_b^*(x)}\,dx &= \int \frac{f(x)^2\pi(x)^2}{\frac{|f(x)|\pi(x)}{\int |f(x)|\pi(x)\,dx}}\,dx = \left(\int |f(x)|\pi(x)\,dx\right)^2 \\
&= \left(\int \frac{|f(x)|\pi(x)}{q(x)} q(x)\,dx\right)^2 = \left(\mathbb{E}\left[\frac{|f(X)|\pi(X)}{q(X)}\right]\right)^2, \qquad X \sim q \\
&\le \mathbb{E}\left[\left(\frac{|f(X)|\pi(X)}{q(X)}\right)^2\right] = \int \frac{f(x)^2\pi(x)^2}{q(x)^2} q(x)\,dx \\
&= \int \frac{f(x)^2\pi(x)^2}{q(x)}\,dx
\end{aligned}$$

by Jensen's inequality. Hence,

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{q_b^*}) &= \frac{1}{N}\left(\int \frac{(f(x)^2\pi(x)^2}{q_b^*(x)}\,dx - \mu^2\right) \\
&\le \frac{1}{N}\left(\int \frac{(f(x)^2\pi(x)^2}{q(x)}\,dx - \mu^2\right) \\
&= \mathrm{Var}(\hat{\mu}_q)
\end{aligned}$$

which completes the proof. □

Now, we will follow analogous line of proof for Theorem 1.10. Note that there is a statement with proof given in [11], but that statement does not talk about the validity of the proposal density.

Proof of Theorem 1.10. Let

$$L(x,q,\lambda) = \frac{\pi(x)^2(f(x)-\mu)^2}{q(x)} + \lambda q(x).$$

Setting

$$\frac{\partial L}{\partial q} = -\frac{(f(x)-\mu)^2\pi(x)^2}{q(x)^2} + \lambda = 0,$$

we have that

$$q(x) = \sqrt{\frac{(f(x)-\mu)^2\pi(x)^2}{\lambda}} = \frac{|f(x)-\mu|\pi(x)}{\sqrt{\lambda}}.$$

Since $\int q(x)\,dx = 1$, we have that $q(x) = \frac{|f(x)-\mu|\pi(x)}{\int |f(x)-\mu|\pi(x)\,dx} = \frac{|f(x)-\mu|p(x)}{\int |f(x)-\mu|p(x)\,dx}$. Now, we will show that this candidate

$$q^*_{sn}(x) = \frac{|f(x)-\mu|p(x)}{\int |f(x)-\mu|p(x)\,dx}$$

yields the minimum asymptotic variance. Let q be any density function such that $p \ll q$.

$$\begin{aligned}
\int \frac{(f(x)-\mu)^2\pi(x)^2}{q^*_{sn}(x)}\,dx &= \int \frac{(f(x)-\mu)^2\pi(x)^2}{\frac{|f(x)-\mu|\pi(x)}{\int |f(x)-\mu|\pi(x)\,dx}}\,dx = \left(\int |f(x)-\mu|\pi(x)\,dx\right)^2 \\
&= \left(\int \frac{|f(x)-\mu|\pi(x)}{q(x)} q(x)\,dx\right)^2 = \left(\mathbb{E}\left[\frac{|f(X)-\mu|\pi(X)}{q(X)}\right]\right)^2, \qquad X \sim q \\
&\le \mathbb{E}\left[\left(\left(\frac{|f(X)-\mu|\pi(X)}{q(X)}\right)^2\right)\right] = \int \frac{(f(x)-\mu)^2\pi(x)^2}{q(x)^2} q(x)\,dx \\
&= \int \frac{(f(x)-\mu)^2\pi(x)^2}{q(x)}\,dx
\end{aligned}$$

by Jensen's inequality. Therefore, we get the desired result. □

CHAPTER

3

PARTITION-BASED METHOD

Before the partition-based method is presented, we continue the discussion of Example 1.12 here. Table 3.1 shows the related integrals for proposal distributions in Example 1.12. Surprisingly, the optimal densities for both basic ($q_1 = q_b^*$) and self-normalized ($q_2 = q_{sn}^*$) IS have integral $\int \frac{\pi(x)^2}{q_i(x)}\, dx = \infty$. This integral is the current rule of thumb (1.7) widely used by statisticians as a criterion to choose a good proposal density for both basic and self-normalized IS. More surprisingly, q_2 which is the

Table 3.1 Related integrals for proposal distributions in Example 1.12

q_i	$\int \frac{\pi(x)^2}{q_i(x)} dx$	$\int \frac{f(x)^2\pi(x)^2}{q_i(x)} dx$	$\int \frac{(f(x)-\mu)^2\pi(x)^2}{q_i(x)} dx$
$q_1 = q_b^*$	∞	0.5625	∞
$q_2 = q_{sn}^*$	∞	∞	≈ 1.18603
q_3	∞	∞	∞
q_4	≈ 1.11072	∞	∞
q_5	≈ 1.93144	≈ 1.34172	≈ 1.50409
q_6	∞	∞	∞

optimal density for self-normalized IS does not satisfy the convergence assumption of Theorem 1.5. This does not violate any theory. From Remark 1.11, q_{sn}^* is just the probability density function that minimizes $\int \frac{(f(x)-\mu)^2\pi(x)^2}{q(x)}\,dx$ and may not satisfy the self-normalized IS validity condition nor the assumption of the convergence theorem for self-normalized IS estimators. Here, q_2 attains the minimum of $\int \frac{(f(x)-\mu)^2\pi(x)^2}{q_2(x)}dx$ at $\frac{9}{8}(14+7\sqrt[3]{6}+4\sqrt[3]{36})e^{-2\sqrt[3]{6}} \approx 1.18603$. It satisfies $p \ll q_2$, but $\int \frac{f(x)^2\pi(x)^2}{q_2(x)}dx, \int \frac{\pi(x)^2}{q_2(x)}dx = \infty$. Another interesting point is that $\int \frac{\pi(x)^2}{q_4(x)}\,dx = \frac{\pi}{2\sqrt{2}} \approx 1.11072 < \infty$, so q_4 satisfies the rule of thumb (1.7). However, it is a bad proposal density for both basic IS and self-normalized IS as we can see from Fig. 1.7 and Fig. 1.8. Thus, the condition (1.7) is not a satisfying criterion for choosing a good proposal density for IS method. If we really pay attention to the convergence theorem of IS estimator, we can get closer to the answer. The criterion should rather be

$$\int \frac{f(x)^2\pi(x)^2}{q(x)}dx < \infty$$

for basic IS, and

$$\int \frac{f(x)^2p(x)^2}{q(x)}dx, \int \frac{p(x)^2}{q(x)}dx < \infty$$

for self-normalized IS. A good example for this is q_5 which has $\int \frac{f(x)^2p(x)^2}{q_5(x)}dx = \frac{3^{7/4}\,35^{7/8}}{2^9}\left(\frac{17}{2}+6\sqrt{2}\right)^{\frac{1}{8}}\pi \approx 1.34172 < \infty$ and $\int \frac{p(x)^2}{q_5(x)}dx = \frac{3^{1/4}\,35^{1/8}}{2^{25/8}}\left(4+2\sqrt{2}\right)^{\frac{1}{2}}\pi \approx 1.93144 < \infty$. Therefore, we have a finite variance, $\mathrm{Var}(\hat{\mu}_{q_5}) = \frac{1}{N}\left(\int \frac{f(x)^2p(x)^2}{q_5(x)}dx - \mu^2\right) \approx \frac{1.34172-0.5625}{N} = \frac{0.77922}{N}$ and a finite asymptotic variance, $\mathrm{AVar}(\tilde{\mu}_{q_5}) = \frac{1}{N}\int \frac{(f(x)-\mu)^2\pi(x)^2}{q_5(x)}dx = \frac{1}{N}\left(\frac{3^{1/4}\,35^{1/8}}{2^{1/8}}\left(4+2\sqrt{2}\right)^{\frac{1}{2}} + \frac{3^{3/4}35^{7/8}}{12}\left(\frac{17}{2}+6\sqrt{2}\right)^{\frac{1}{8}} - \left(\frac{35}{2}\right)^{\frac{1}{2}}\right)\frac{9\pi}{2^7} \approx \frac{1.50409}{N}$. Both basic IS and self-normalized IS performances of q_5 are really good.

Although we cannot always use the theoretically optimal proposal density in IS method in reality, we may be able to find a nice proposal density that gives a really pleasant outcome and at the same time satisfies all theoretical assumptions we need. To get an idea of how this work can come about, we consider the case of basic IS first. From Example 1.12, we can see from Fig. 1.7 that the proposal density q_6 has a good performance, and from Fig. 1.1 that the probability density function of q_6 is very closed to the optimal proposal density q_1. This gives us an idea that a good proposal density should be closed to the optimal one, q_b^* for basic IS and q_{sn}^* for self-normalized IS. A question that should come to one's mind is why the proposal q_6 brings about a nice result despite the fact that $\mathrm{Var}(\hat{\mu}_{q_6}) = \infty$. From Theorem 1.3, we have that $\mathrm{Var}(\hat{\mu}_{q_6}) = \frac{1}{N}\left(\int_0^\infty 2\sqrt{2\pi}x^7e^{2\left(\log x-\frac{1}{2}\right)^2-4x}\,dx - (\frac{3}{4})^2\right)$ and one can check that the part that make the variance infinite is the integral around the neighborhood of zero, say $\int_0^\epsilon 2\sqrt{2\pi}x^7e^{2\left(\log x-\frac{1}{2}\right)^2-4x}\,dx = \infty$, $\epsilon > 0$. From the basic IS formula (1.3),each term in the summation with the proposal density q_6 is $\frac{f(x)\pi(x)}{q_6(x)} = \sqrt{2\pi}x^4e^{2\left(\log x-\frac{1}{2}\right)^2-2x}$ and this amount will go

Table 3.2 All assumptions for proposal distributions/densities

	Basic IS	Self-normalized IS
Given	f, π	f, p
Validity	$f\pi \ll q$	$p \ll q$
Convergence	$\int \frac{f(x)^2\pi(x)^2}{q(x)} dx < \infty$	$\int \frac{f(x)^2 p(x)^2}{q(x)} dx$, $\int \frac{p(x)^2}{q(x)} dx < \infty$

to infinity as x goes to 0. So, the jump behavior will arise when we get a sample too closed to 0. According to the probability density function q_6 from Fig. 1.1, most of the time we will get random samples not too close to 0. Thus, the jump problem rarely happens in the simulation, and that makes q_6 look good enough to be used as a proposal density. However, we cannot deny the fact that it brings about infinite variance. Therefore, in this chapter, we propose a valid proposal density that gives a really pleasant outcome, guarantee finite variance (for basic IS or asymptotic variance for self-normalized IS) of the estimator, and satisfies the assumption for the associated convergence theorem.

3.1 Sufficient Conditions

The derivation of both kinds of IS from Section 1.1 and 1.2 needs the assumption for a legitimate proposal distribution. Also, Corollary 1.4 and Theorem 1.5 tell us all the assumptions for a proposal density to have the associated convergence theorem. Table 3.2 summarizes all of these assumptions. Note that the assumption $\int \frac{f(x)^2\pi(x)^2}{q(x)} dx < \infty$ also implies the boundedness of $\mathrm{Var}(\hat{\mu}_q)$, and the assumption $\int \frac{f(x)^2 p(x)^2}{q(x)} dx$, $\int \frac{p(x)^2}{q(x)} dx < \infty$ also implies the boundedness of $\mathrm{AVar}(\tilde{\mu}_q)$. Therefore, all we need for a good proposal distribution are

$$f\pi \ll q \tag{3.1}$$

and

$$\int \frac{f(x)^2\pi(x)^2}{q(x)} dx < \infty \tag{3.2}$$

for basic IS, and also

$$p \ll q \tag{3.3}$$

and

$$\int \frac{f(x)^2 p(x)^2}{q(x)} dx \, , \int \frac{p(x)^2}{q(x)} dx < \infty \tag{3.4}$$

for self-normalized IS.

The following proposition gives some sufficient conditions to satisfy the assumption for the convergence theorem for the IS estimators.

Proposition 3.1. *If $\frac{\pi}{q}$ is bounded almost everywhere, then*

$$\int \frac{\pi(x)^2}{q(x)} dx < \infty.$$

Also, if either

1. $\mathrm{Var}_\pi(f) < \infty$ *and $\frac{\pi}{q}$ is bounded almost everywhere, or*
2. $\frac{f\pi}{q}$ *is bounded almost everywhere*

then

$$\int \frac{f(x)^2 \pi(x)^2}{q(x)} dx < \infty.$$

Proof. The proof of this proposition is quite obvious once we note that $\int \pi(x) dx = 1$, $\mathrm{Var}_\pi(f) < \infty$ implies $\int f(x)^2 \pi(x) dx < \infty$, and f is π—integrable which means $\int |f(x)| \pi(x) dx < \infty$. Therefore, if $\frac{\pi}{q}$ is bounded almost everywhere by a constant M, then

$$\int \frac{\pi(x)^2}{q(x)} dx \leq M \int \pi(x) dx = M < \infty.$$

If $\mathrm{Var}_\pi(f) < \infty$ and $\frac{\pi}{q}$ is bounded almost everywhere by a constant M, then

$$\int \frac{f(x)^2 \pi(x)^2}{q(x)} dx \leq M \int f(x)^2 \pi(x) dx < \infty.$$

If $\frac{|f|\pi}{q}$ is bounded almost everywhere by a constant K, then

$$\int \frac{f(x)^2 \pi(x)^2}{q(x)} dx \leq K \int |f(x)| \pi(x) dx < \infty$$

which completes the proof. □

Table 3.3 Sufficient conditions for the convergence theorem for IS estimators.

Basic IS	Self-normalized IS
$\mathrm{Var}_\pi(f) < \infty$ and $\frac{\pi}{q}$ is bounded or $\frac{f\pi}{q}$ is bounded	$\mathrm{Var}_\pi(f) < \infty$ and $\frac{p}{q}$ is bounded or $\frac{fp}{q}$ and $\frac{p}{q}$ are bounded

In practice, we can replace "bounded almost everywhere" in this proposition by just "bounded". Thus, to practically meet the assumption for the convergence theorem for IS estimators, we want the proposal density q that satisfies the conditions given in Table 3.3. For a valid proposal distribution, we also need to be aware of the absolute continuity issue.

Some people may say that knowing the optimal density is useless because it cannot be used in reality. However, finding a proposal that is close to the optimal one is possible, and it should provide a very good result. For example, in Example 1.12, we would like to acquire a valid proposal density that satisfies the assumption for the convergence theorem like q_5 and is close to the optimal density like q_6 for basic IS and q_5 for self-normalized IS.

3.2 Classes of Functions

The classes of functions that can be handled by the partition-based method are discussed before we go into the partition-based method. The primary issue is about the oscillation of a function. For a function that has oscillatory behavior, we need to justify what kind of an oscillation can still be applied the partition-based method.

Definition 3.2. Let f be a real-valued function on a domain $D \subset \mathbb{R}$ that has countable discontinuities and countable local extrema. Then, the collection of all the discontinuities and all the local extrema can partition D into enumerated intervals $\{I_i^f\}_{i\in J}$ for some index set J. We say that f is *well oscillated*, if

$$\inf_{I_i^f \subset [-T,T]} |I_i^f| > 0, \quad \forall T > 0.$$

For a well oscillated function f, we call

$$\inf_{I_i^f \subset W} |I_i^f|$$

the *pseudo period* of f on W for any bounded subset $W \subset D$.

Definition 3.3. For a function f on a domain $D \subset \mathbb{R}^d$ for some $d > 1$, we say that f is *well oscillated*, if

$$f_k(x_k) = f(x_1, \ldots, x_d)|_{x_i=\alpha_i, \forall i \in \{1,\ldots,d\}\setminus\{k\}}$$

is well oscillated for all $k \in \{1, \ldots, d\}$ and for all $(\alpha_1, \ldots, \alpha_d) \in D$.

Denote the collection of all probability density functions on D by $\mathscr{P}(D)$. Define classes for the pairs of target functions and target densities:

$$\begin{aligned}
\mathscr{A}_b(D) &= \{(f,\pi) \in \mathbb{R}^D \times \mathscr{P}(D) \mid |f|\pi \text{ is well oscillated}\} \\
\mathscr{A}_{sn}(D) &= \{(f,\pi) \in \mathbb{R}^D \times \mathscr{P}(D) \mid |f-\mu|\pi \text{ is well oscillated}\} \\
\mathscr{B}(D) &= \{(f,\pi) \in \mathbb{R}^D \times \mathscr{P}(D) \mid \mathrm{Var}_\pi(f) < \infty\} \\
\mathscr{C}_1(D) &= \{(f,\pi) \in \mathbb{R}^D \times \mathscr{P}(D) \mid \pi \text{ is bounded}\} \\
\mathscr{C}_2(D) &= \{(f,\pi) \in \mathbb{R}^D \times \mathscr{P}(D) \mid f\pi \text{ is bounded}\}.
\end{aligned}$$

We will categorize the classes for the pairs of target functions and target densities that can be handled by the partition-based method into the following four groups.

1. $\mathscr{G}_b^b(D) = \mathscr{A}_b(D) \cap \big([\mathscr{B}(D) \cap \mathscr{C}_1(D)] \cup \mathscr{C}_2(D) \big)$
2. $\mathscr{G}_b^u(D) = \mathscr{A}_b(D) \cap \big([\mathscr{B}(D) \cap \mathscr{C}_1(D)^c] \cup \mathscr{C}_2(D)^c \big)$
3. $\mathscr{G}_{sn}^b(D) = \mathscr{A}_{sn}(D) \cap \big([\mathscr{B}(D) \cap \mathscr{C}_1(D)] \cup [\mathscr{C}_1(D) \cap \mathscr{C}_2(D)] \big)$
4. $\mathscr{G}_{sn}^u(D) = \mathscr{A}_{sn}(D) \cap \big([\mathscr{B}(D) \cap \mathscr{C}_1(D)^c] \cup [\mathscr{C}_1(D)^c \cap \mathscr{C}_2(D)] \cup [\mathscr{C}_1(D) \cap \mathscr{C}_2(D)^c] \cup [\mathscr{C}_1(D)^c \cap \mathscr{C}_2(D)^c] \big)$

The subscript in the notation refers to which kind of IS the partition-based method can be applied: b for basic IS, and sn for self-normalized IS. The superscript in the notation determines the boundedness of π and/or $f\pi$: b for bounded, and u for unbounded. Each class will be differently managed before applying the partition-based method. Note that most functions from real world applications are usually well-oscillated. We just want to clearly state the class of functions to which the partition-based method can be applied.

3.3 One-Dimensional Spaces

IS is capable of having a zero variance estimator for the case of basic IS with positive target function by using the optimal density, so we should find a way to obtain a proposal density that is close to

the optimal one by utilizing known theory. We will start with one-dimensional problem to get an idea of the proposed method, and generalize that to the multidimensional case later.

In most applications, there are three types of one-dimensional domain: bounded interval domain $[a,b]$, semi-infinite domain $[0,\infty)$, and infinite domain $(-\infty,\infty)$. We will illustrate examples mainly in semi-infinite domain because there are two issues that normally do not appear at the same time in the other two cases. The first issue is about the tail in which the bounded interval domain does not have. The second issue is the unboundedness of the target distribution π. A distribution with infinite domain usually has bounded density. Even if it is unbounded, we can adjust it to use the method proposed in the semi-infinite case.

Now, a method to get a proposal density for IS that has a good performance and satisfies all the required theoretical assumptions is proposed. The idea is very simple and can be easily simulated in 1-dimensional problem. The concept is to get a proposal density that close to the optimal density (1.5) from Theorem 1.7 for basic IS or (1.6) from Theorem 1.10 for self-normalized IS, and to ensure that it satisfies (3.1) and (3.2) for basic IS or (3.3) and (3.4) for self-normalized IS.

3.3.1 Basic Importance Sampling

In this subsection, we will focus on the case of basic IS. The classes of the pairs of the target functions and the target densities that can be handled by the partition-based method are $\mathscr{G}_b^b(D)$ and $\mathscr{G}_b^u(D)$. We will explain the process of the method for each class separately, and primarily focus on the class $\mathscr{G}_b^b(D)$.

3.3.1.1 Class $\mathscr{G}_b^b(D)$

Assume that $(f,\pi)\in\mathscr{G}_b^b(D)$. Consider the case when $D=[0,\infty)$. The first step in the method is to partition the domain. Let's fix a proper constant $M\in(0,\infty)$ somewhere in the domain. We will discuss how to choose the parameter M later. We will call $[0,M]$ the importance region, and $[M,\infty)$ the tail region. Let

$$0=x_0<x_1<\ldots<x_L=M<\infty=x_{L+1}$$

be a partition for $[0,\infty)$. We will discuss how to choose the parameter L later. For the sake of simplicity in programming, let

$$\Delta=\frac{M}{L}\quad\text{and}\quad x_i=i\Delta \text{ for } i=0,1,\ldots,L.$$

The first L subintervals of the partition have equal length Δ and their union is the importance region, and the last subinterval is $[M,\infty)$ which is the tail region. Note that subintervals in the importance

region do not need to have the same length, and we can use varying-length subintervals at the expense of more complexity and memory requirement in programming.

Next, we will approximate the optimal basic-IS proposal density from Theorem 1.7. For the importance region, we will use the idea of Riemann integral where the function value is constant on each subinterval. We choose representative points from each subinterval

$$x_i^* \in [x_{i-1}, x_i] \text{ for all } i = 1, \ldots, L$$

and evaluate with $|f|\pi$, the optimal function without the normalizing constant, to get the unnormalized proposal density $\tilde{q}$ in each subinterval, let

$$\tilde{h}_i = |f(x_i^*)|\pi(x_i^*) \text{ for all } i = 1, \ldots, L$$

$$\tilde{q}(x) = \tilde{h}_i \text{ for } x \in [x_{i-1}, x_i).$$

Recall that q needs to satisfy the absolute continuity (3.1). It is possible that the point x_i^* can cause $|f(x_i^*)|\pi(x_i^*) = 0$, but $f\pi$ is not zero on the entire subinterval $[x_{i-1}, x_i)$, and we will call this the zero problem. The support of the final proposal density should cover the support of $f\pi$. So, we just need nonzero $\tilde{h}_i$ to approximately represent q on $[x_{i-1}, x_i)$ and we can avoid this zero problem by reselecting x_i^*, or using several intermediate points and averaging the function values with these intermediate points. For instance, we select $x_{i1}^*, \ldots, x_{ik}^* \in [x_{i-1}, x_i]$ for some k and set $\tilde{h}_i = \frac{1}{k}\sum_{j=1}^{k} |f(x_{ij}^*)|\pi(x_{ij}^*)$ or even use a weighted average. Anyway, we may predetermine the domain where $f\pi$ is nonzero, and choosing x_i^* to be the mid-point of each interval $[x_{i-1}, x_i]$ seems to be an easy choice and works well. If $f\pi$ is zero on the entire subinterval $[x_{i-1}, x_i)$, then any choice of x_i^* will make $|f(x_i^*)|\pi(x_i^*) = 0$ which does not violate (3.1). Suppose for now that $x_i^* = x_i - \frac{\Delta}{2}$, the mid-point of the interval $[x_{i-1}, x_i]$ does not bring about the zero problem. Sometimes, we can make a transformation by adding or subtracting f with some constant to avoid zero problem.

According to the assumption of Proposition 3.1, if we assume that $\mathrm{Var}_\pi(f) < \infty$, we have to choose q that makes $\frac{\pi}{q}$ bounded (more practical and stronger than bounded almost everywhere), otherwise we need that $\frac{f\pi}{q}$ is bounded. For $(f, \pi) \in \mathscr{G}_b^b(D)$, we choose q to be a piecewise-constant function in the importance region, so the bounded condition in the importance region is satisfied for these cases.

For the tail region, a proper function $\tilde{q}_t$ will be chosen to be the tail of our proposal density. $\tilde{q}_t$ has to satisfy the condition that $\frac{\pi}{\tilde{q}_t}$ is bounded on $[M, \infty)$ for the case of $(f, \pi) \in \mathscr{A}_b(D) \cap \mathscr{B}(D) \cap \mathscr{C}_1(D)$; or $\frac{f\pi}{\tilde{q}_t}$ is bounded on $[M, \infty)$ for the case of $(f, \pi) \in \mathscr{A}_b(D) \cap \mathscr{C}_2(D)$. We will choose a continuous function $\tilde{q}_t$ such that $\lim_{x\to\infty} \frac{\pi(x)}{\tilde{q}_t} < \infty$ for the case of $(f, \pi) \in \mathscr{A}_b(D) \cap \mathscr{B}(D) \cap \mathscr{C}_1(D)$; or $\lim_{x\to\infty} \frac{f(x)\pi(x)}{\tilde{q}_t} < \infty$ for the case of $(f, \pi) \in \mathscr{A}_b(D) \cap \mathscr{C}_2(D)$, and $\int_M^\infty \tilde{q}_t \, dx$ can be computed. We suggest that $\tilde{q}_t$ has purely

exponential decay or purely polynomial decay of degree more than one. It is possible that purely polynomial decay is not enough. For example, we may need to choose functions like $\frac{1}{x(\log x)^{1+\alpha}}$ for $\alpha > 0$. Note that if exponentially decaying function is appliable, so is polynomially decaying function. We also suggest to match $\tilde{q}_t(M)$ with $\tilde{h}_L$ to get a sense of smoothness. Now, we have the unnormalized proposal density

$$\tilde{q}(x) = \begin{cases} \tilde{h}_i & \text{, if } x \in [x_{i-1}, x_i) \\ \tilde{q}_t(x) & \text{, if } x \in [M, \infty). \end{cases}$$

Useful functions to be used as $\tilde{q}_t$ are $\tilde{h}_L e^{-\alpha(x-M)}$, $\frac{\tilde{h}_L}{(x-M+1)^{1+\alpha}}$, and $\frac{\tilde{h}_L M^{1+\alpha}}{x^{1+\alpha}}$ for $\alpha > 0$ which are clearly integrable on $[M, \infty)$. For these $\tilde{q}_t$'s, $\int_M^\infty \tilde{q}_t \, dx$ equals $\frac{\tilde{h}_L}{\alpha}$ for the first two densities, and $\frac{\tilde{h}_L M}{\alpha}$ for the third one. Then, the calculation of Z_q and h_i's follows. The probability of getting a sample in $[x_{i-1}, x_i)$ for $i = 1, \ldots, L$ is Δh_i, and the probability of getting a sample in $[M, \infty)$ is $\frac{h_L}{\alpha}$ for the first two densities or $\frac{\tilde{h}_L M}{\alpha}$ for the last one. The probability of getting a sample in the tail region can be controlled by choosing parameter α, but we have to ensure the existence of $\lim_{x\to\infty} \frac{\pi(x)}{\tilde{q}_t(x)}$ for the case of $(f, \pi) \in \mathscr{A}_b(D) \cap \mathscr{B}(D) \cap \mathscr{C}_1(D)$; or $\lim_{x\to\infty} \frac{f(x)\pi(x)}{\tilde{q}_t(x)}$ for the case of $(f, \pi) \in \mathscr{A}_b(D) \cap \mathscr{C}_2(D)$. Note that $\frac{\tilde{h}_L}{(x-M+1)^{1+\alpha}}$ and $\frac{\tilde{h}_L M^{1+\alpha}}{x^{1+\alpha}}$ are the polynomial tail with the same order. We suggest to use $\tilde{h}_L e^{-\alpha(x-M)}$ and $\frac{\tilde{h}_L}{(x-M+1)^{1+\alpha}}$ as the tail function because the probability of getting a sample in the tail region does not depend on M.

Next, we calculate the normalizing constant Z_q for our $\tilde{q}$. We still stick to the plan of using equal length of Δ for each subintervals in the importance region. Then,

$$Z_q = \Delta \cdot \sum_{i=1}^{L} \tilde{h}_i + \int_M^\infty \tilde{q}_t(x) \, dx.$$

Consequently, we obtain the proposal density

$$q(x) = \begin{cases} h_i = \frac{\tilde{h}_i}{Z_q} & \text{, if } x \in [x_{i-1}, x_i) \text{ for } i = 1, \ldots, L \\ q_t(x) = \frac{\tilde{q}_t(x)}{Z_q} & \text{, if } x \in [M, \infty). \end{cases}$$

We can draw random samples according to this density by using the inverse transform method described in Appendix D. Most of the time, we will get samples in the importance region and this is why we call it importance region. To draw random samples, we set $I_0 = 0$, $I_j = I_{j-1} + \Delta h_j = \Delta \cdot \sum_{i=1}^{j} h_i$ for $j = 1, \ldots, L$, and $I_{L+1} = 1$ as cumulative sum of the integral over the subintervals. To get a sample, we generate a random number u from the uniform distribution over the unit interval. Then, we find

the lowest k such that $I_k > u$. If $k = L+1$ which means the sample will fall in the tail region, then we have the relationship

$$\begin{aligned} u &= I_L + \int_M^x h_L e^{-\alpha(z-M)}\, dz \\ &= I_L + \frac{h_L}{\alpha}(1 - e^{-\alpha(x-M)}) \\ x &= -\frac{1}{\alpha}\log\left(1 - \frac{\alpha(u-I_L)}{h_L}\right) + M \end{aligned}$$

for $q_t(x) = h_L e^{-\alpha(x-M)}$ or

$$\begin{aligned} u &= I_L + \int_M^x \frac{h_L}{(z-M+1)^{1+\alpha}}\, dz \\ &= I_L + \frac{h_L}{\alpha}\left(1 - \frac{1}{(x-M+1)^{\alpha}}\right) \\ x &= \left(1 - \frac{\alpha(u-I_L)}{h_L}\right)^{-1/\alpha} + M - 1 \end{aligned}$$

for $q_t(x) = \frac{h_L}{(x-M+1)^{1+\alpha}}$, otherwise the sample will fall in the importance region and we have that

$$\begin{aligned} u &= I_{k-1} + \int_{(k-1)\Delta}^x h_k\, dz \\ &= I_k - \Delta h_k + (x - (k-1)\Delta)h_k \\ x &= \frac{u - I_k}{h_k} + k\Delta \end{aligned}$$

with $q(x) = h_k$. This computed x is a sample drawn from the proposal distribution q, and the regular IS method can be directly applied.

Example 3.4. We will illustrate this method by applying it to the setup from Example 1.12. Here, we have $f(x) = x^3$ and $\pi(x) = 2e^{-2x}$. We will apply our method with parameters $M = 4, 10$ and $L = 10, 25, 80$. In each combination of M and L, we set $\Delta = \frac{M}{L}$ and choose $x_i^* = (i - \frac{1}{2})\Delta$, the mid-point of the i^{th} subinterval, for $i = 1, \ldots, L$. The height of the unnormalized proposal density in each subinterval can be straightforwardly calculated by $\tilde{h}_i = |f(x_i^*)|\pi(x_i^*)$ for $i = 1, \ldots, L$. Note that we must check if any of $\tilde{h}_i$ is zero, and if so, we have to reselect x_i^* or use the average of more than one point for each subinterval. However, for this setup of f and π, the subinterval's mid-points do not create the zero problem.

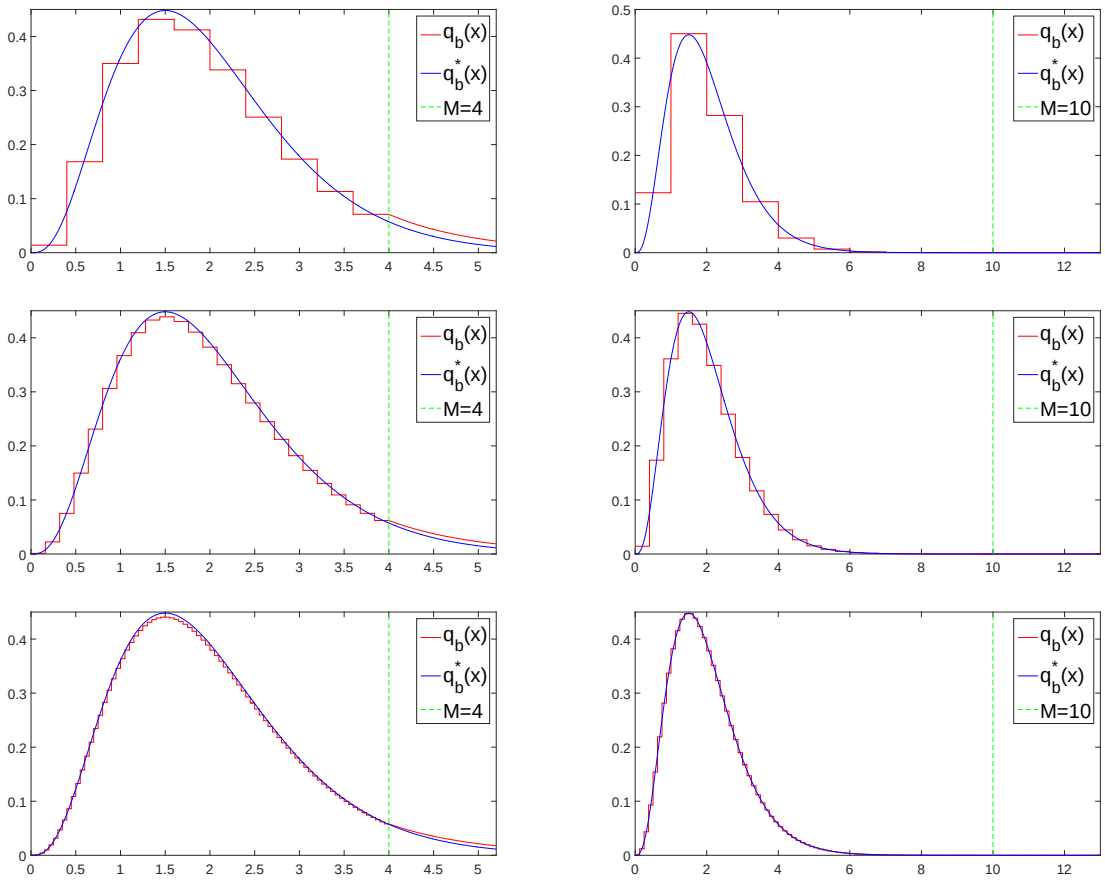

Figure 3.1 Proposal densities for parameters $M = 4, 10$ and $L = 10, 25, 80$ for Example 3.4

As for the tail region, we have two easy choices of the tail function. One is an exponentially decaying function $e^{-\alpha x}$ and the other is a polynomially decaying function $\frac{1}{(x-M+1)^{1+\alpha}}$ for some $\alpha > 0$, and both functions are multiplied by a constant to match up the height of the last subinterval in importance region. Here, we choose

$$\tilde{q}_t(x) = \tilde{h}_L e^{-(x-M)}.$$

Then, $\lim_{x\to\infty} \frac{\pi(x)}{\tilde{q}_t(x)} = 0$ and $\lim_{x\to\infty} \frac{f(x)\pi(x)}{\tilde{q}_t(x)} = 0$, so the bounded condition is satisfied. The integral of $\tilde{q}_t$ on the tail region is $\int_M^\infty \tilde{h}_L e^{-(x-M)}\, dx = \tilde{h}_L$. Thus, the normalizing constant for the proposal density is

$$Z_q = \Delta \cdot \sum_{i=1}^{L} \tilde{h}_i + \tilde{h}_L.$$

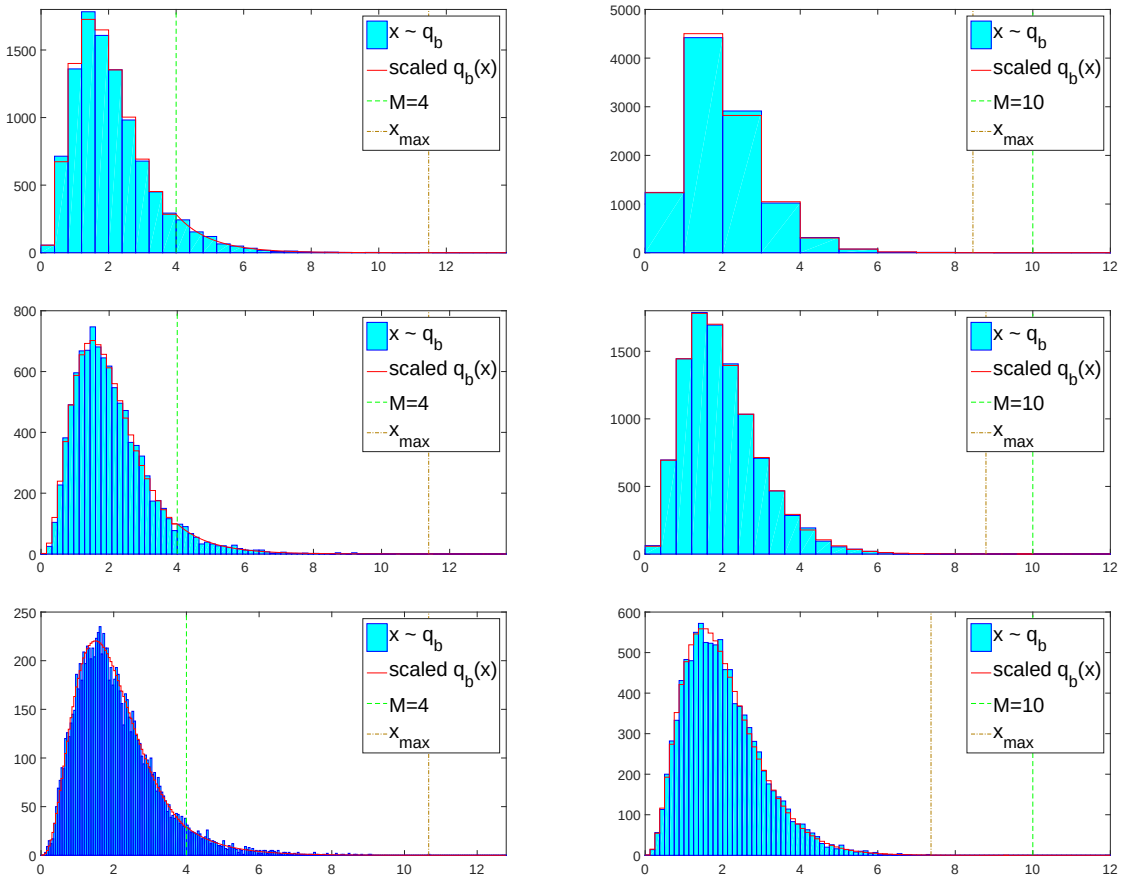

Figure 3.2 Histograms of the corresponding densities from Fig. 3.1

We set $h_i = \frac{\tilde{h}_i}{Z_q}$ for $i = 1, \ldots, L$. Then, the proposal density is

$$q_b(x) = \begin{cases} h_i & \text{, if } x \in [x_{i-1}, x_i) \text{ for } i = 1, \ldots, L \\ h_L e^{-(x-M)} & \text{, if } x \in [M, \infty). \end{cases}$$

Fig. 3.1 shows the proposal densities for all the combinations of parameters $M = 4, 10$ and $L = 10, 25, 80$. The optimal basic-IS density q_b^* is also plotted to show how close each proposal density is to the optimal one.

We generate 10,000 samples from each density using the inverse transform method. Then, the basic IS computation using (1.3) proceeds. Fig. 3.2 shows histograms of samples obtained from the corresponding densities in Fig. 3.1.

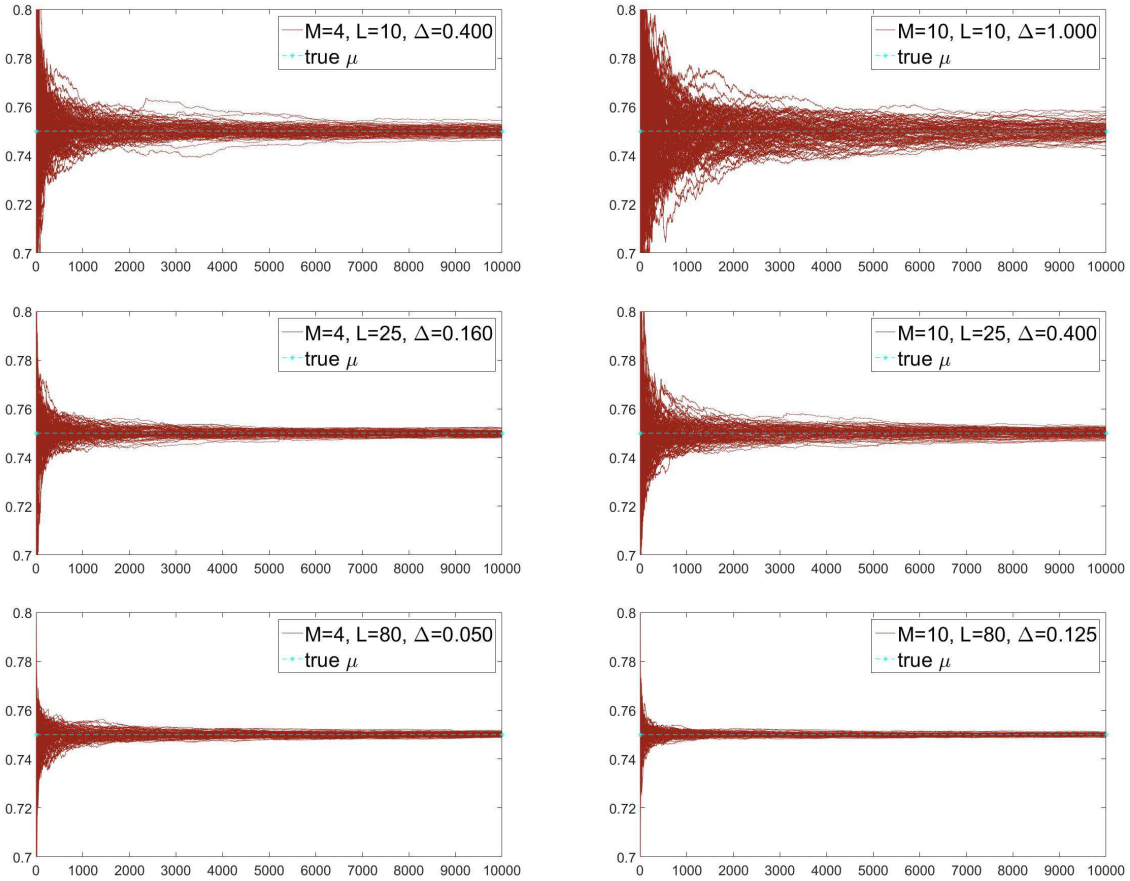

Figure 3.3 Basic IS performance of the corresponding densities from Fig. 3.1

Fig. 3.3 shows the basic-IS performance of the corresponding densities from Fig. 3.1 with $n = 100$ scenarios. We can see that the smaller Δ is, the closer our q_b tends to be to the optimal density; hence, we have a better approximation. As a comparison, the performance of proposal densities q_5 and q_6 from Example 1.12 is presented in Fig. 3.4 with the same scale as in Fig. 3.3. The optimal proposal density q_1 gives absolutely zero variance, so there is no need to reshow a rescaled graph of its performance. Note that q_6 yields infinite variance of estimation, while all of our q_b with various parameters M and L guarantee to have finite variance according to Proposition 3.1.

There is another way of sampling method using two random numbers for one output sample. In our method for 1-dimensional problem, the probability of getting a sample in each subinterval of the importance region and in the tail region can be easily calculated. We can express our proposal distribution as a linear combination of known distributions. For instance, the proposal density in

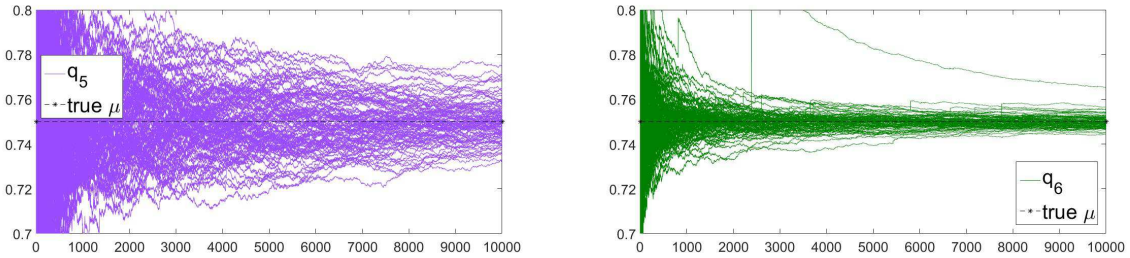

Figure 3.4 Rescaled basic IS performance of q_5 and q_6 from Example 1.12

Example 3.4 can be expressed as

$$\begin{aligned} q_b(x) &= \sum_{i=1}^{L} h_i \mathbb{1}_{[x_{i-1},x_i)}(x) \; + \; h_L e^{-(x-M)} \mathbb{1}_{[M,\infty)}(x) \\ &= \sum_{i=1}^{L} (h_i \Delta) \left(\frac{1}{\Delta} \mathbb{1}_{[x_{i-1},x_i)}(x) \right) \; + \; h_L \left(e^{-(x-M)} \mathbb{1}_{[M,\infty)}(x) \right) \end{aligned}$$

so that our proposal distribution can be expressed as

$$\sum_{i=1}^{L} (h_i \Delta) \text{Unif}(x_{i-1}, x_i) \; + \; h_L \text{GPD}(0,1,M)$$

where $\text{GPD}(\xi, \sigma, \theta)$ is the generalized Pareto distribution with shape parameter ξ, scale parameter σ, and threshold parameter θ described in Appendix E. The reason why we use the generalized Pareto distribution instead of the exponential distribution is because the generalized Pareto distribution is more general and includes polynomial-decaying distributions, which can be selected to be the tail part of the proposal density. It is often used to model the tails of diverse distributions. Recall that $\sum_{i=1}^{L} h_i \Delta + h_L = 1$. Thus, the proposal distribution can be expressed as a mixture distribution and be sampled by the composition method.

To perform the composition method, recall that we have $I_0 = 0$, $I_j = I_{j-1} + \Delta h_j$ for $j = 1, \ldots, L$, and $I_{L+1} = I_L + h_L = 1$. So, we generate u from $\text{Unif}(0,1)$ and find the first j that gives $I_j \geq u$. Then, we simply generate a desired sample x from the j^{th} corresponding distribution. That is, if $j \leq L$, we sample x from $\text{Unif}(x_{j-1}, x_j)$, and if $j = L+1$, we sample x from $\text{GPD}(0,1,M)$. This mixture distribution is a combination of distributions with different supports that form a partition for the resulting mixture distribution. This is different from the current glowing research where the mixture distribution is a combination of distributions with the same support.

As we can see from Example 3.4, the size of Δ and the threshold M can affect the performance of the method. Now, one way to determine the parameters in the partition-based method is discussed. By fixing the tail function to be a generalized-Pareto-density type, we can express our unnormalized tail function as

$$\tilde{q}(x) = \sum_{i=1}^{L} \Delta \tilde{h}_i \text{Unif}_{(x_{i-1},x_i)}(x) + \sigma \tilde{h}_L \text{GPD}_{(\xi,\sigma,M)}(x)$$

with parameter $\xi \geq 0$ and $\sigma > 0$. Here, $\text{Unif}_{(x_{i-1},x_i)}$ and $\text{GPD}_{(\xi,\sigma,M)}$ refer to their corresponding probability density functions:

$$\text{Unif}_{(x_{i-1},x_i)}(x) = \frac{1}{\Delta} \mathbb{1}_{[x_{i-1},x_i)}(x) \tag{3.5}$$

$$\text{GPD}_{(\xi,\sigma,M)}(x) = \begin{cases} \frac{1}{\sigma\left(1+\xi\frac{(x-M)}{\sigma}\right)^{1+\frac{1}{\xi}}} \mathbb{1}_{[M,\infty)}(x) & \text{, if } \xi > 0 \\ \frac{1}{\sigma} e^{-\frac{(x-M)}{\sigma}} \mathbb{1}_{[M,\infty)}(x) & \text{, if } \xi = 0. \end{cases} \tag{3.6}$$

Note that $\tilde{q}(M) = \tilde{h}_L$. Then, $\int \tilde{q}(x)\, dx = \sum_{i=1}^{L} \Delta \tilde{h}_i + \sigma \tilde{h}_L$ so that

$$q(x) = \frac{\tilde{q}(x)}{\sum_{i=1}^{L} \Delta \tilde{h}_i + \sigma \tilde{h}_L}.$$

Now, the first step is to choose parameters ξ and σ in order to satisfy all the required assumptions. Specifically, the tail function $\tilde{q}_t$ chosen from the generalized Pareto densities have to make $\frac{\pi}{\tilde{q}_t}$ bounded on $[M,\infty)$ for the case of $(f,\pi) \in \mathscr{A}_b(D) \cap \mathscr{B}(D) \cap \mathscr{C}_1(D)$; or $\frac{f\pi}{\tilde{q}_t}$ bounded on $[M,\infty)$ for the case of $(f,\pi) \in \mathscr{A}_b(D) \cap \mathscr{C}_2(D)$.

The second step is to choose Δ. The smaller Δ is, the better the approximation will be. However, if it is too small, the computation cost will be high. For $(f,\pi) \in \mathscr{A}_b(D)$, if $|f|\pi$ has no oscillation, Δ can be any small number, but if $|f|\pi$ has oscillation behavior, Δ should be chosen to be smaller than the pseudo period of $|f|\pi$ on $[0,M]$. In general real-world problems, the pseudo period of $|f|\pi$ on the entire domain is strictly positive, so we can just choose Δ to be smaller than this global pseudo period. Sometimes, Δ bigger than the pseudo period of $|f|\pi$ on $[0,M]$ can still work well.

Now, to find a cutting point M, we will seek L, the number of Δ-size intervals in the importance region. Let's introduce two more parameters ϵ_Δ and ϵ_0. Here, ϵ_Δ is an upper bound for the probability of getting a sample in the tail region and ϵ_0 is the machine precision bound. The probability of getting a sample in the tail region is

$$\frac{\int \sigma \tilde{h}_L \text{GPD}_{(\xi,\sigma,M)}(x)\, dx}{\sum_{i=1}^{L} \Delta \tilde{h}_i + \sigma \tilde{h}_L} = \frac{\sigma \tilde{h}_L}{\sum_{i=1}^{L} \Delta \tilde{h}_i + \sigma \tilde{h}_L}.$$

Thus, we want that

$$\epsilon_0 < \frac{\sigma \tilde{h}_L}{\sum_{i=1}^{L} \Delta \tilde{h}_i + \sigma \tilde{h}_L} < \epsilon_\Delta.$$

Simple calculation yields

$$\frac{\Delta}{\sigma} \frac{\epsilon_0}{1-\epsilon_0} < \frac{\tilde{h}_L}{\sum_{i=1}^{L} \tilde{h}_i} < \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta}.$$

Note that $\frac{x}{1-x}$ is an increasing function on $(0,1)$. We always have $\epsilon_0 > 0$ and choose $\epsilon_\Delta < 1$. Hence, if $\epsilon_0 < \epsilon_\Delta$, then $\frac{\Delta}{\sigma} \frac{\epsilon_0}{1-\epsilon_0} < \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta}$. The third step is to increase the number of intervals L and find the first L that makes $\frac{\tilde{h}_L}{\sum_{i=1}^{L} \tilde{h}_i}$ lower than the threshold $\frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta}$. Note that ϵ_0 is so small that we should get

$$\frac{\Delta}{\sigma} \frac{\epsilon_0}{1-\epsilon_0} < \frac{\tilde{h}_L}{\sum_{i=1}^{L} \tilde{h}_i}.$$

We should check for this condition, but we can ignore it in practice since ϵ_0 tends to be very small.

Consider a sequence $\frac{\tilde{h}_L}{\sum_{i=1}^{L} \tilde{h}_i}$ in L. The denominator $\sum_{i=1}^{L} \tilde{h}_i$ is increasing in L because $\tilde{h}_i = |f(x_i^*)|\pi(x_i^*) \geq 0$ for $i = 1, \ldots, L$. The numerator $\tilde{h}_L$ will converge to zero, if $\lim_{x\to\infty} f(x)\pi(x) = 0$. Recall that $\mu = \int f(x)\pi(x)\, dx < \infty$, so $\lim_{x\to\infty} f(x)\pi(x) = 0$ holds. One may construct a counterexample such as a function that takes value n on $(n, n + \frac{1}{n^3})$ for all $n \in \mathbb{N}$ and zero elsewhere. $\tilde{h}_L$ is not guaranteed to converge to zero because it may do not converge at all.

Claim. *There exists $L_0 \in \mathbb{N}$ such that*

$$\frac{\tilde{h}_{L_0}}{\sum_{i=1}^{L_0} \tilde{h}_i} < \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta}.$$

Proof. We must have $\tilde{h}_k > 0$ for some $k \in \mathbb{N}$. Thus, $\tilde{h}_k \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta} > 0$. Since $\int |f(x)|\pi(x)\, dx < \infty$, there exists $L_0 > k$ such that

$$\begin{aligned} \tilde{h}_{L_0} &= |f(x_{L_0}^*)|\pi(x_{L_0}^*) \\ &< \tilde{h}_k \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta} \\ &\leq \left(\sum_{i=1}^{L_0} \tilde{h}_i \right) \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta}. \end{aligned}$$

Hence, $\frac{\tilde{h}_{L_0}}{\sum_{i=1}^{L_0} \tilde{h}_i} < \frac{\Delta}{\sigma} \frac{\epsilon_\Delta}{1-\epsilon_\Delta}$. □

Note that this claim relies on the choice of x_i^*'s and we may have to choose x_i^*'s other than the interval-mid-points. Eventually, $\frac{\tilde{h}_L}{\sum_{i=1}^{L}\tilde{h}_i}$ will be lower than the threshold $\frac{\Delta}{\sigma}\frac{\epsilon_\Delta}{1-\epsilon_\Delta}$ for some L, and we will use this L as our parameter. Thus, we finally have $M = \Delta L$ and the proposed method can be carried on.

This method is self-determining where to cut for the tail and can also be applied with the scheme where we use one uniform random number for one output sample. We will call that scheme the *auto-sampling-once scheme*. The above scheme where we use one uniform random number and one generalized Pareto random number for one output sample will be called the *auto-sampling-twice scheme*. The summary for both schemes are shown in Table 3.4. Sometimes we need to take care of the choice of x_i^*'s, but the interval-mid-points usually work well in general. The algorithm can also be adjusted to be applied to the other kinds of domain in $\mathbb{R}$.

In Table 3.4, Step 1 - 4 compute the proposal density, and step 5 - 7 deliver one random sample x from the proposal density with the corresponding function value $q(x) = \text{qx}$ for the regular IS approximation using (1.3). One may force the cutting point M to be further than a certain point M_0 by computing the initial step before entering the while-loop until $\Delta L > M_0$.

Note that the parameters in the auto-sampling-once and auto-sampling-twice schemes can be equivalently changed to the other scheme. Table 3.5 shows the relationship of the parameters change between these two schemes. Both schemes should have the same performance, but the auto-sampling-twice scheme uses more computing time because it draws one more random number than the auto-sampling-once scheme for one sample. Thus, we recommend the auto-sampling-once scheme over the auto-sampling-twice scheme. We introduce the idea of the auto-sampling-twice scheme because it can be extended easily to the problems with multidimensional space. We will talk about this idea in Section 3.4.

Example 3.5. We redo Example 3.4 again with the auto-sampling-once and auto-sampling-twice schemes with $\Delta = 0.1, 0.2, 0.4, 0.8$ and $\epsilon_\Delta = 5\%$. For the auto-sampling-once scheme, we use the exponential tail with parameter $\alpha = 1$ which is equivalent to using parameters $\xi = 0$ and $\sigma = 1$ for the auto-sampling-twice scheme. The proposal densities q_b with different parameter Δ using the self-determined-where-to-cut method is shown in Fig. 3.5. Note that the auto-sampling-once scheme and the auto-sampling-twice scheme are different mainly on the sampling step, and have the same process of computing proposal densities. Fig. 3.6 and Fig. 3.7 show the performance of the auto-sampling-once scheme and the auto-sampling-twice scheme, respectively, with different Δ. Note that each case has $n = 100$ scenarios and each scenario uses $N = 10,000$ samples. The time to compute each proposal density (step 1 - 4 in Table 3.4), the time in sampling step (step 5 - 7 in Table 3.4 with $n \times N$ samples), and the time to calculate IS approximation (step 8 in Table 3.4)

Table 3.4 Algorithm of the partition-based method for basic IS on $[0,\infty)$

Step 1	For the auto-sampling-once scheme, choose a type of tail (polynomial or exponential) and parameter α for the algorithm. For the auto-sampling-twice scheme, choose parameters ξ and σ for the algorithm. Choose parameters Δ and ϵ_Δ for the algorithm and set Threshold $= \Delta\alpha\frac{\epsilon_\Delta}{1-\epsilon_\Delta}$ or $\frac{\Delta}{\sigma}\frac{\epsilon_\Delta}{1-\epsilon_\Delta}$ depending on the chosen scheme.
Step 2	Initially set $L=1$, $x^*=\frac{\Delta}{2}$, $h_1=\big(\lvert f\rvert\pi\big)(x^*)$ and SUMh $= h_1$.
Step 3	While $\frac{h_L}{\text{SUMh}} \geq$ Threshold Increase L by 1, set $x^*=x^*+\Delta$ and compute NEWh $=\big(\lvert f\rvert\pi\big)(x^*)$. Append NEWh to h so that $h_L =$ NEWh. Set SUMh $=$ SUMh $+$ NEWh. end while loop
Step 4	Compute $M=\Delta L$, and check for zero problem. Normalize all h_i by $\Delta\cdot$SUMh $+\frac{h_L}{\alpha}$ or $\Delta\cdot$SUMh $+\sigma h_L$ depending on the chosen scheme.
Step 5	Set $I_j=\Delta\cdot\sum_{i=1}^{j}h_i$ for $j=1,\ldots,L$ and $I_{L+1}=1$.
Step 6	Draw $u\sim$ Unif$(0,1)$ and find the first index k of I that $I_k>u$.

Step 7	For the auto-sampling-once scheme:
	If $k=L+1$
	If choose exponential tail,
	set $x=-\frac{1}{\alpha}\log\Big(1-\frac{\alpha(u-I_L)}{h_L}\Big)+M$ and qx $=h_L e^{-\alpha(x-M)}$.
	If choose polynomial tail,
	set $x=\Big(1-\frac{\alpha(u-I_L)}{h_L}\Big)^{-1/\alpha}+M-1$ and qx $=\frac{h_L}{(x-M+1)^{1+\alpha}}$.
	else
	Set $x=\frac{u-I_k}{h_k}+k\Delta$ and qx $=h_k$.
	end if
	For the auto-sampling-twice scheme:
	If $k=L+1$
	Set $x=$ GPD(ξ,σ,M) and qx $=\sigma h_L$GPD$_{(\xi,\sigma,M)}(x)$.
	else
	Set $x=$ Unif$((k-1)\Delta,k\Delta)$ and qx $=h_k$.
	end if

Step 8	Calculate IS approximation using (1.3) with a number of samples from step 5 - 7.

Table 3.5 Parameters change between auto-sampling-once and auto-sampling-twice schemes

auto-sampling-once	auto-sampling-twice
exponential tail	$\xi = 0$
polynomial tail	$\xi = \frac{1}{\alpha}$
$\alpha = \frac{1}{\sigma}$	$\sigma = \frac{1}{\alpha}$

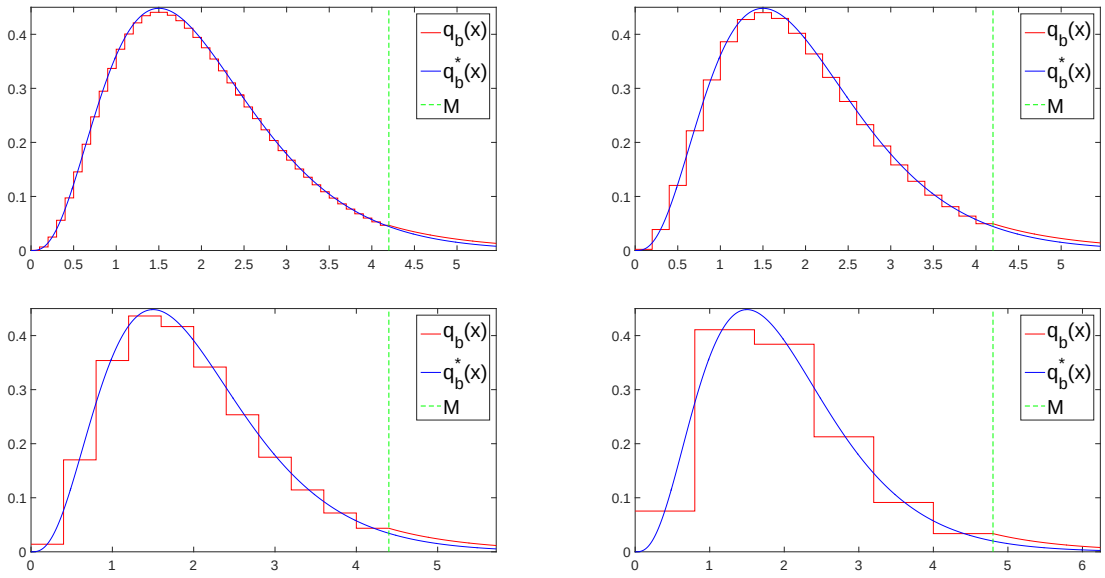

Figure 3.5 Proposal densities with $\Delta = 0.1, 0.2, 0.4, 0.8$ for Example 3.5

are presented in Table 3.6. All simulations in this work are on the same machine which is an Intel Core i5-3320M 2.60 GHz processor. The sampling time for auto-sampling-twice scheme is obviously much more than the sampling time for auto-sampling-once scheme, although the performance of both schemes are the same.

For comparison, the performance of the simple Monte Carlo method is shown in Fig. 3.8 with two different scales. With $N = 10,000$ and $n = 100$, the sampling time and the time to calculate Monte Carlo approximation using (1.1) are 0.0464 and 0.0892 seconds, respectively. For fair comparison, we should consider to rerun the Monte Carlo method with larger sample size or the partition-based method with lower sample size. We rerun the Monte Carlo method with sample size $N = 1,000,000$ and other parameters fixed. The performance of this simulation is presented in Fig. 3.9 which has

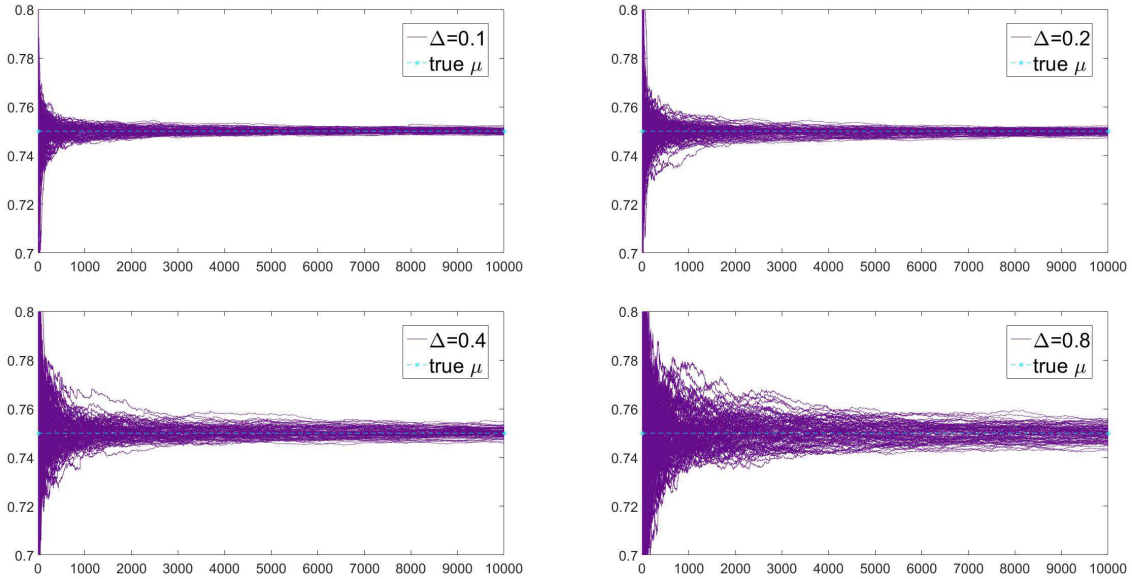

Figure 3.6 Performance of the corresponding densities from Fig. 3.5 with the auto-sampling-once scheme.

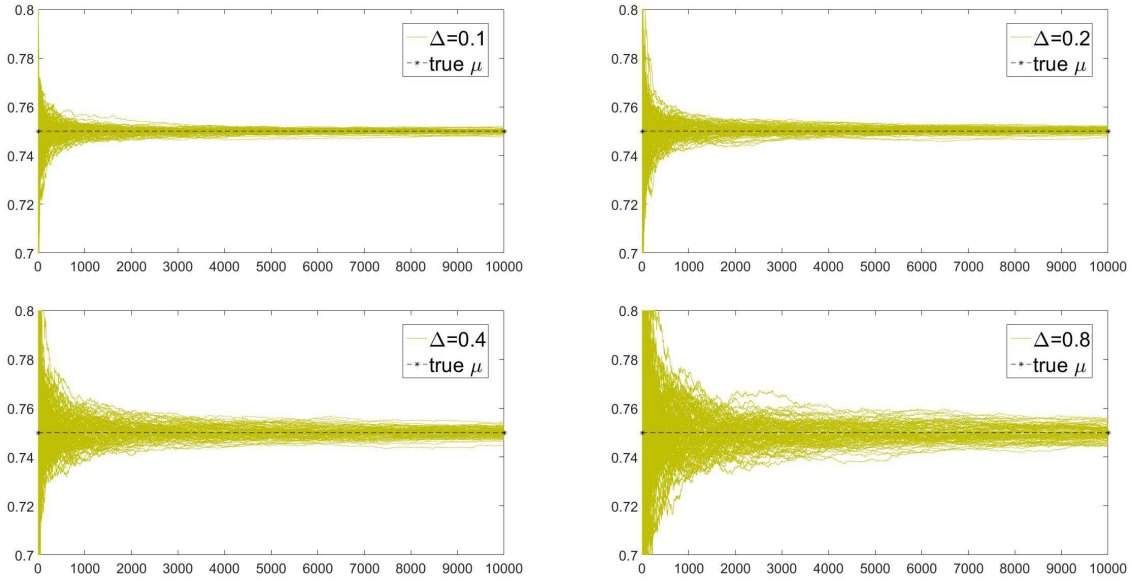

Figure 3.7 Performance of the corresponding densities from Fig. 3.5 with the auto-sampling-twice scheme.

the same y-axis scale as Fig. 3.6 and Fig. 3.7. With $N = 1,000,000$ and $n = 100$, the sampling time and the time to calculate Monte Carlo approximation are 3.5906 and 18.8331 seconds, respectively.

Table 3.6 Computing time in seconds for Example 3.5 with $N = 10,000$

	$\Delta = 0.1$	$\Delta = 0.2$	$\Delta = 0.4$	$\Delta = 0.8$
Computing the proposal density	0.0053	0.0051	0.0055	0.0012
Sampling for (I) auto-sampling-once scheme	1.8942	1.8211	1.6877	1.8462
Sampling for (II) auto-sampling-twice scheme	13.5899	13.4918	13.4355	13.3309
Calculating IS approximation for (I)	0.1158	0.1283	0.1102	0.1290
Calculating IS approximation for (II)	0.1212	0.1090	0.1069	0.1044

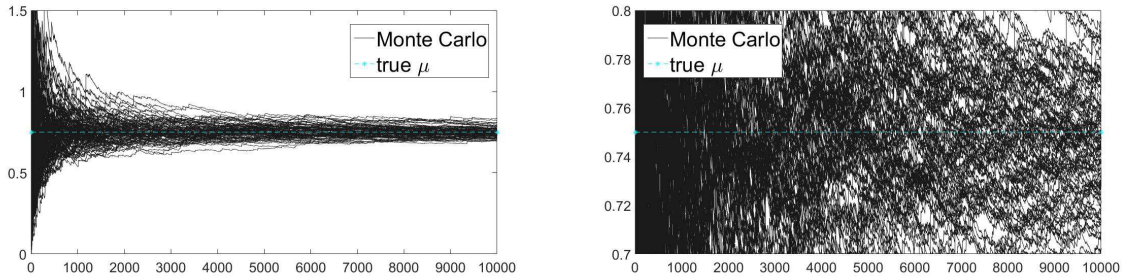

Figure 3.8 Performance of the simple Monte Carlo method for Example 3.5 with $N = 10,000$.

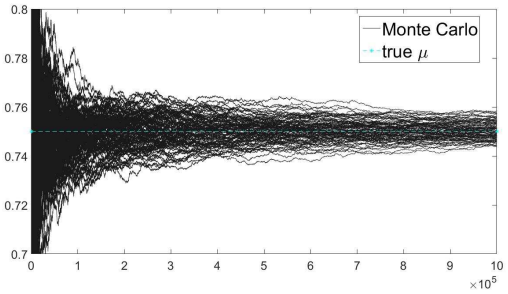

Figure 3.9 Performance of the simple Monte Carlo method for Example 3.5 with $N = 1,000,000$.

We can see that the partition-based method outperforms the Monte Carlo method in this example.

3.3.1.2 Class $\mathscr{G}_b^u(D)$

Assume that $(f,\pi) \in \mathscr{G}_b^u(D)$. We want that $\frac{\pi}{q}$ is bounded for $(f,\pi) \in \mathscr{A}_b(D) \cap \mathscr{B}(D) \cap \mathscr{C}_1(D)^c$, or $\frac{f\pi}{q}$ is bounded for $(f,\pi) \in \mathscr{A}_b(D) \cap \mathscr{C}_2(D)^c$. Note that when we say that a function is well-oscillated, the number of points where the function assumes infinity is finite. When π or $f\pi$ is unbounded, we cannot take q as constant around the area that π or $f\pi$ is unbounded. We have to choose a

function q_h that goes unbounded on that area but still make $\frac{\pi}{q_h}$ or $\frac{f\pi}{q_h}$ bounded there. It is similar to the concept of choosing the tail for the proposal density. The proposal density is tailored to almost fit the optimal density. We will illustrate this method with the adjusted auto-sampling-once scheme through the following example.

Example 3.6. Consider $f(x) = \sqrt[4]{x}$ and $\pi(x) = \frac{e^{-x}}{\sqrt{\pi x}}$ on $(0,\infty)$. We can see that π and $f\pi$ are unbounded around 0. We choose $x_i^* = (i-\frac{1}{2})\Delta$, the mid-point of the i^{th} subinterval, for $i = 2,\dots,L$. The height of the unnormalized proposal density in each subinterval can be straightforwardly calculated by $\tilde{h}_i = |f(x_i^*)|\pi(x_i^*)$ for $i = 2,\dots,L$.

Regarding the first subinterval, we choose the unnormalized head function to be

$$\tilde{q}_h(x) = \frac{\tilde{h}_2 \Delta^{1-\alpha_h}}{x^{1-\alpha_h}}$$

with $\alpha_h = \frac{1}{2}$ so that the bounded condition is satisfied. Note that it is not necessary to use the first subinterval with the same size as the other intervals. The integral of q_h on the first subinterval is $\int_0^\Delta \frac{\tilde{h}_2\Delta^{1-\alpha_h}}{x^{1-\alpha_h}}\, dx = \frac{\tilde{h}_2\Delta}{\alpha_h}$. As for the tail, we choose

$$\tilde{q}_t(x) = \tilde{h}_L e^{-\alpha_t(x-M)}$$

with $\alpha_t = 1$, so that the bounded condition is satisfied. The integral of $\tilde{q}_t$ on the tail region is $\int_M^\infty \tilde{h}_L e^{-\alpha_t(x-M)}\, dx = \frac{\tilde{h}_L}{\alpha_t}$. Thus, the normalizing constant for the proposal density is

$$Z_q = \frac{\tilde{h}_2\Delta}{\alpha_h} + \Delta \cdot \sum_{i=2}^{L} \tilde{h}_i + \frac{\tilde{h}_L}{\alpha_t}.$$

Setting $h_i = \frac{\tilde{h}_i}{Z_q}$ for $i = 2,\dots,L$, we have the proposal density

$$q_b(x) = \begin{cases} \frac{h_2\Delta^{1-\alpha_h}}{x^{1-\alpha_h}} & \text{, if } x \in [0,\Delta) \\ h_i & \text{, if } x \in [(i-1)\Delta, i\Delta) \text{ for } i = 2,\dots,L \\ h_L e^{-\alpha_t(x-M)} & \text{, if } x \in [M,\infty). \end{cases}$$

Now, to determine where to cut, M, we set a parameter ϵ_Δ to be the upper bound for the probability of getting a sample in the tail region. The probability of having a sample in the tail region is

$$\frac{\frac{\tilde{h}_L}{\alpha_t}}{\frac{\tilde{h}_2\Delta}{\alpha_h} + \Delta \cdot \sum_{i=2}^{L} \tilde{h}_i + \frac{\tilde{h}_L}{\alpha_t}}.$$

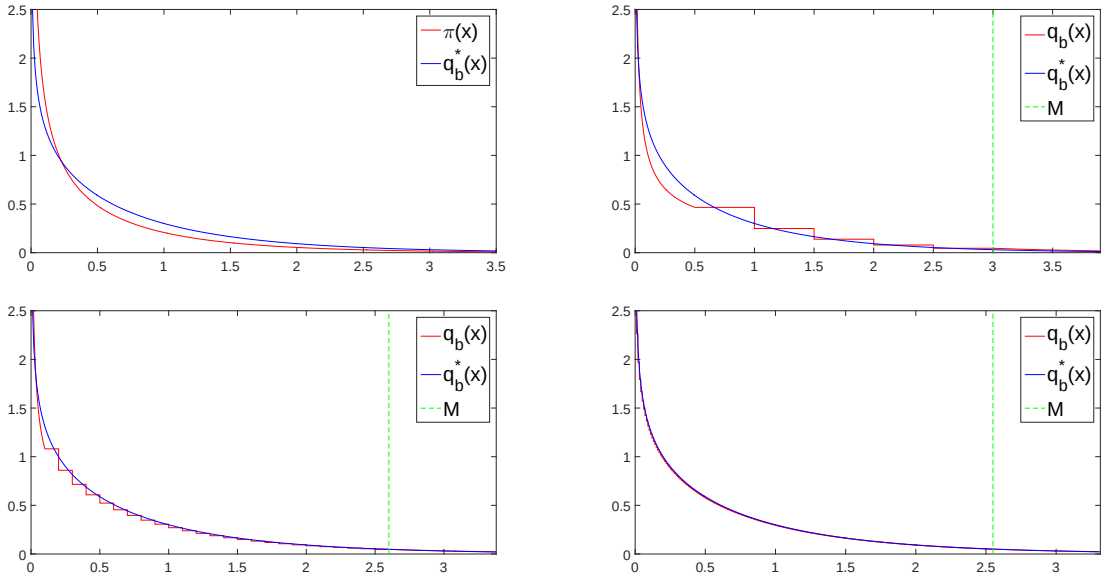

Figure 3.10 Target density and proposal densities with $\Delta = 0.5, 0.1, 0.01$ for Example 3.6

Then, the condition to determine L is that

$$\frac{\tilde{h}_L}{\frac{\tilde{h}_2}{\alpha_h} + \sum_{i=2}^{L} \tilde{h}_i} < \Delta\alpha_t \frac{\epsilon_\Delta}{1-\epsilon_\Delta}.$$

The algorithm starts with $L = 2$ and increases it until the condition is satisfied. We can define $\tilde{h}_1 = \frac{\tilde{h}_2}{\alpha_h}$ to get a full array h when we implement the method.

We apply the partition-based method with parameters $\Delta = 0.5, 0.1, 0.01$ and $\epsilon_\Delta = 5\%$. Fig. 3.10 shows graphs of the target density π and the proposal densities with these parameters. The optimal basic-IS density

$$q_b^*(x) = \frac{e^{-x}}{\Gamma(\frac{3}{4})\sqrt[4]{x}},$$

where Γ is the gamma function $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx$, is also plotted to show how close each proposal density to the optimal one.

In the sampling step, we set $I_0 = 0$, $I_1 = \frac{h_2\Delta}{\alpha_h}$, $I_j = I_{j-1} + \Delta h_j$ for $j = 2, \ldots, L$, and $I_{L+1} = 1$ as cumulative sum of the integral over the subintervals. To get a sample using the inverse transform method, we generate a random number u from the uniform distribution over the unit interval, $u \sim \text{Unif}(0,1)$, and find the lowest k such that $I_k > u$. If $k = L+1$, then we have

$$\begin{aligned} u &= I_L + \int_M^x h_L e^{-\alpha_t(z-M)}\, dz \\ &= I_L + \frac{h_L}{\alpha_t}(1-e^{-\alpha_t(x-M)}) \\ x &= -\frac{1}{\alpha_t}\log\left(1-\frac{\alpha_t}{h_L}(u-I_L)\right)+M. \end{aligned}$$

If $k=1$, we obtain

$$\begin{aligned} u &= \int_0^x \frac{h_2\Delta^{1-\alpha_h}}{z^{1-\alpha_h}}\, dz \\ &= \frac{h_2\Delta^{1-\alpha_h}}{\alpha_h}x^{\alpha_h} \\ x &= \Delta\left(\frac{u}{I_1}\right)^{\frac{1}{\alpha_h}}, \end{aligned}$$

otherwise we have

$$\begin{aligned} u &= I_{k-1} + \int_{(k-1)\Delta}^x h_k\, dz \\ &= I_k - \Delta h_k + (x-(k-1)\Delta)h_k \\ x &= \frac{u-I_k}{h_k} + k\Delta. \end{aligned}$$

This computed x is a sample drawn from the proposal distribution q_b. Then, the basic IS computation using (1.3) can proceed.

Fig. 3.11 shows the basic-IS performance of the corresponding densities from Fig. 3.10 with 100 scenarios of 10,000 samples. The ordinary Monte Carlo method performance is also shown in Fig. 3.11. From Fig. 3.10, we can see that π is close to q_b^*. Thus, the ordinary Monte Carlo method is already a good approximation for this problem. With $N = 10,000$ and $n = 100$, the original Monte Carlo method uses 0.0949 seconds in sampling step, and 0.0893 seconds to calculate Monte Carlo approximation using (1.1). The time used in the simulation for each proposal density of the partition-based method with $\Delta = 0.5, 0.1, 0.01$ is shown in Table 3.7.

Consider the partition-based method with $\Delta = 0.001$ and $N = 10,000$ versus the Monte Carlo method with $N = 1,000,000$. Their performances are presented in Fig. 3.12. The time used in the simulation of each method with $n = 100$ scenarios is presented in Table 3.8. With fair comparison, the partition-based method works better than the simple Monte Carlo method.

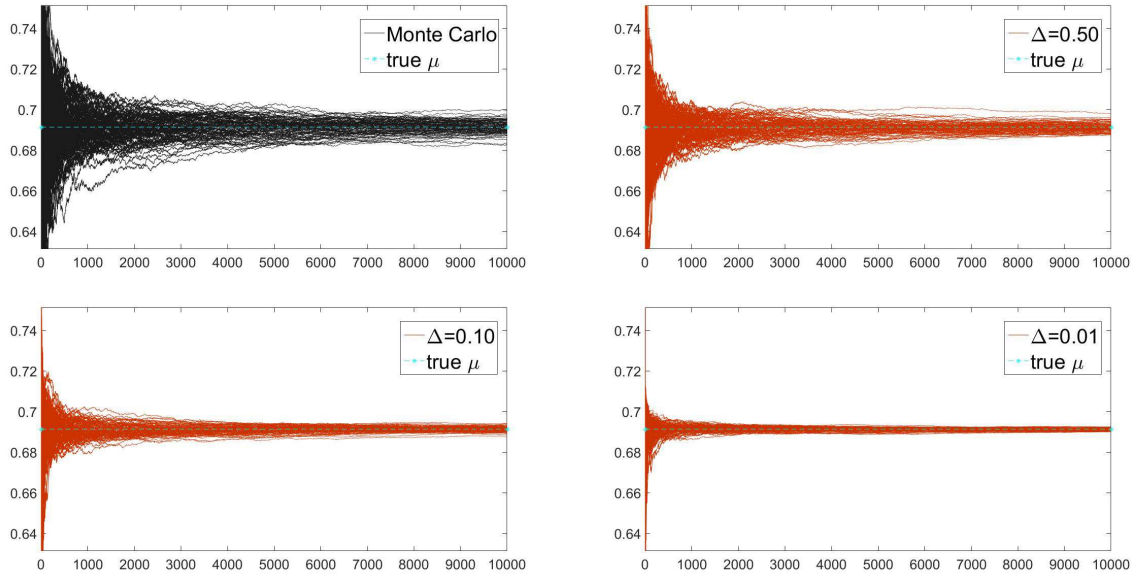

Figure 3.11 Performance of simple Monte Carlo method and basic IS with the corresponding densities from Fig. 3.10

3.3.2 Self-normalized Importance Sampling

Recall that $\pi(x) = \frac{p(x)}{Z}$, and $p(x)$ is known point-wise for self-normalized IS. For the basic IS scheme, we just approximate the optimal basic-IS density $q_b^*(x) = \frac{|f(x)|\pi(x)}{\int |f(x)|\pi(x)\, dx}$ and use the approximated density as the proposal density. The IS weights will theoretically correct the Monte Carlo approximation, and the proposed method guarantees finite variance of the approximation with an attempt to get closer to the optimal density. Now, we would like to do the same thing for the self-normalized IS. However, the problem is that the optimal self-normalized-IS density $q_{sn}^*(x) = \frac{|f(x)-\mu|p(x)}{\int |f(x)-\mu|p(x)\, dx}$ really requires the knowledge of μ, which is what we want to approximate. Thus, it is not straightforward

Table 3.7 Computing time in seconds of the partition-based method with $N = 10,000$, $n = 100$ and $\Delta = 0.5, 0.1, 0.01$ for Example 3.6

	$\Delta = 0.5$	$\Delta = 0.1$	$\Delta = 0.01$
Computing the proposal density	0.0077	0.0147	0.0746
Sampling	1.9806	1.9342	2.2803
Calculating IS approximation	0.6946	0.7424	0.7439

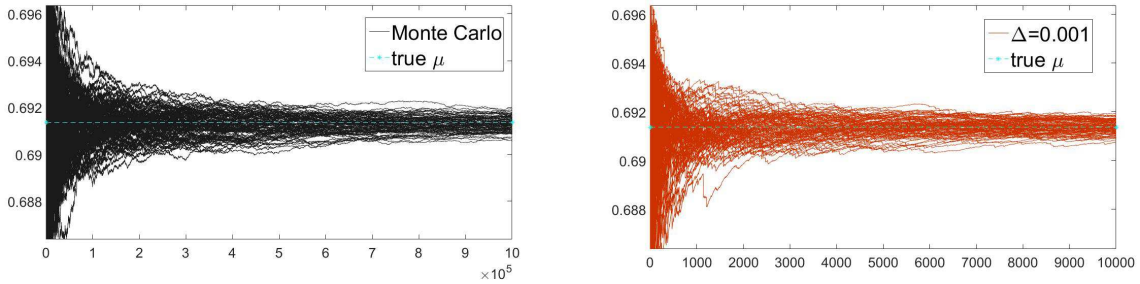

Figure 3.12 Performance of the simple Monte Carlo method with $N = 1,000,000$ and the partition-based method with $\Delta = 0.001$ and $N = 10,000$ for Example 3.6.

Table 3.8 Computing time in seconds of the simple Monte Carlo method with $N = 1,000,000$ and the partition-based method with $\Delta = 0.001$ and $N = 10,000$ for Example 3.6

	Monte Carlo with $N = 1,000,000$	$\Delta = 0.001$ with $N = 10,000$
Computing the proposal density	-	0.7459
Sampling	8.7935	7.1354
Calculating the approximation	10.3522	0.7744

to imitate the basic-IS partition-based method.

A possible way is to roughly approximate μ as the first step and use it in the optimal self-normalized-IS formula. To do this, we need to fix M and L at first to approximate μ. Since

$$\mu = \frac{\int f(x)\pi(x)\,dx}{\int \pi(x)\,dx} = \frac{\int f(x)p(x)\,dx}{\int p(x)\,dx},$$

we can approximate μ by the Riemann sum on the importance region

$$\mu \approx \frac{\sum_{i=1}^{L} f(x_i^*)p(x_i^*)}{\sum_{i=1}^{L} p(x_i^*)}.$$

By doing so, the resulting proposal density will be changed just a bit depending on the error in approximating μ in the first step. Even though we have error from approximating μ in the first step, the resulting density can do the job as a proposal density because any q such that $\pi \ll q$ can actually be used as a proposal density in self-normalized IS. The IS weights will automatically correct the Monte Carlo approximation. Note that in our regular procedure, all subintervals have the same size

Δ. If the subintervals are chosen to have different sizes, the approximated μ can still be achieved by the above formula with the adjustment of multiplying the interval sizes in both summations.

Example 3.7. We illustrate this idea by applying it to Example 1.12 which is a bounded case $(f,\pi) \in \mathscr{G}_{sn}^{b}(D)$. For an unbounded case $(f,\pi) \in \mathscr{G}_{sn}^{u}(D)$, we can still use this procedure with some adjustment in the area where the unboundedness occurs as demonstrated in Example 3.6. Here, we have $f(x) = x^3$ and $p(x) = 2e^{-2x}$. We will apply our method with the parameters $(M,L) = (3,5),(3,20),(3,80),(6,5),(6,20),(6,80),(8,5)$ and $(8,2000)$. In each combination of M and L, we set $\Delta = \frac{M}{L}$ and we choose $x_i^* = (i-\frac{1}{2})\Delta$, the mid-point of the i^{th} subinterval, for $i = 1,\ldots,L$. Then, we roughly estimate μ using these x_i^*'s and call the approximated result ν.

$$\nu = \frac{\sum_{i=1}^{L} f(x_i^*)p(x_i^*)}{\sum_{i=1}^{L} p(x_i^*)}.$$

Then, we calculate the unnormalized heights by

$$\tilde{h}_i = |f(x_i^*) - \nu| p(x_i^*) \text{ for all } i = 1,\ldots,L$$

and check if any $\tilde{h}_i$ causes the zero problem. Here, q is chosen to be nonzero finite-piecewise constant in the importance region, so the bounded condition in the importance region according to Proposition 3.1 is satisfied. Next, we choose $\tilde{q}_t(x) = \tilde{h}_L e^{-(x-M)}$ for the tail region, which corresponds to choosing $\xi = 0, \sigma = 1$ for the generalized Pareto distribution, so that the bounded condition in Proposition 3.1 is satisfied. The normalizing constant for the proposal density is $Z_q = \Delta \cdot \sum_{i=1}^{L} \tilde{h}_i + \tilde{h}_L$. We set $h_i = \frac{\tilde{h}_i}{Z_q}$ for $i = 1,\ldots,L$. Then, the proposal density is

$$q_{sn}(x) = \begin{cases} h_i & \text{, if } x \in [x_{i-1}, x_i) \text{ for } i = 1,\ldots,L \\ h_L e^{-(x-M)} & \text{, if } x \in [M,\infty). \end{cases}$$

Fig. 3.13 shows the graphs of the proposal densities with the optimal self-normalized-IS density q_{sn}^*. We generate 10,000 samples from each density using the inverse transform method in the same way as we did in Example 3.4. Then, the self-normalized IS computation using (1.4) proceeds. Fig. 3.14 shows the self-normalized IS performance of the corresponding densities from Fig. 3.13. All the proposal densities with various parameters M and L guarantee to give finite variance according to Proposition 3.1.

We can clearly see from Fig. 3.13 that the graph of the proposal density with parameters $(M,L) = (3,80)$ is shifted from q_{sn}^* or the pivot point x such that $f(x) = \mu$, due to the error in approximating μ in

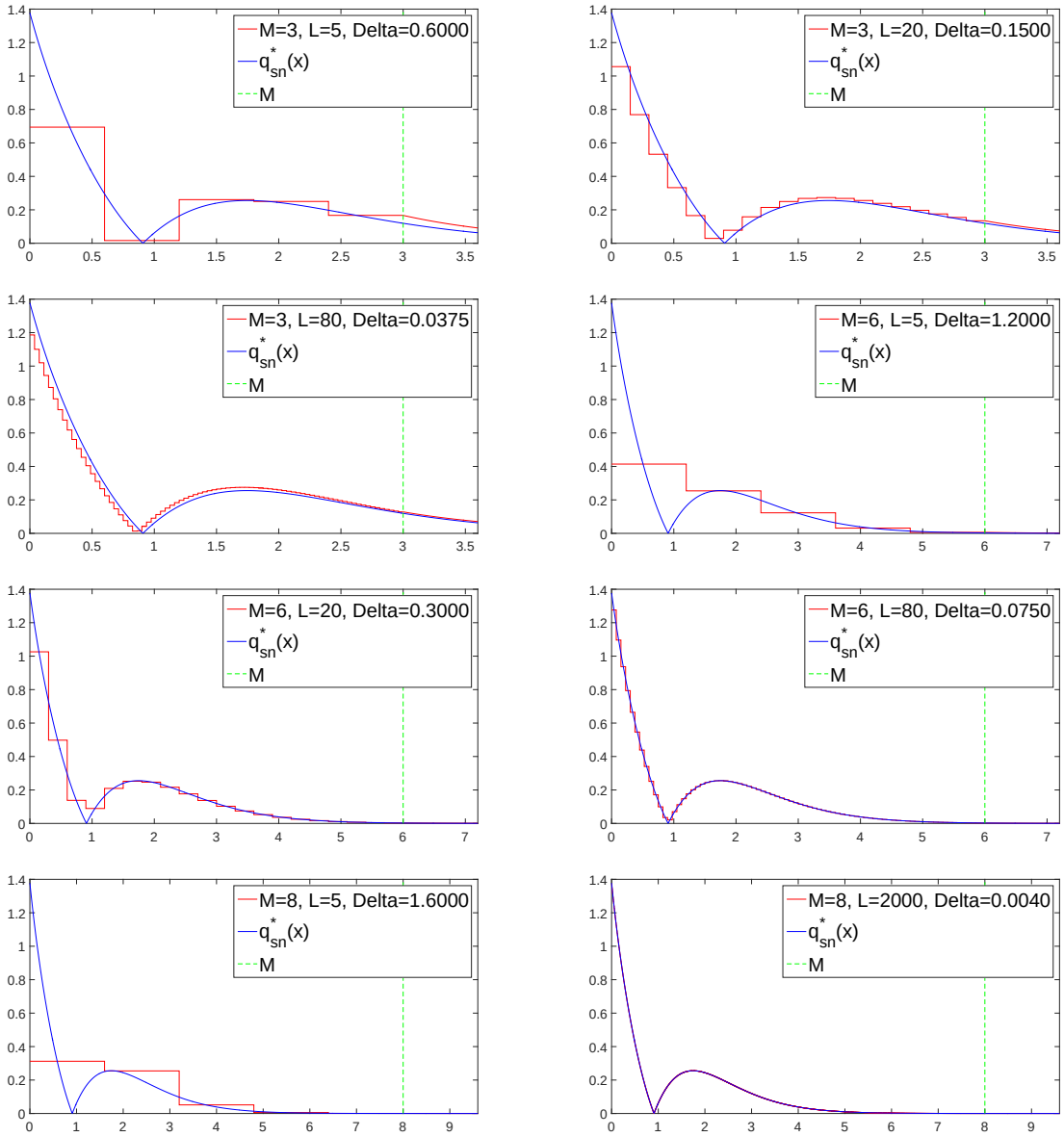

Figure 3.13 Proposal densities with various parameters for Example 3.7

the beginning step. Also, the graphs of the proposal densities with parameters $(M, L) = (3, 20), (6, 20)$ are slightly shifted from the pivot point, and it is hard to notice that the graph of the proposal density with parameters $(M, L) = (3, 5)$ is actually slightly shifted from the pivot point because of the bigger subinterval size, Δ. When Δ is bigger than the pseudo period of q^*_{sn} on $[0, M]$, the proposal density

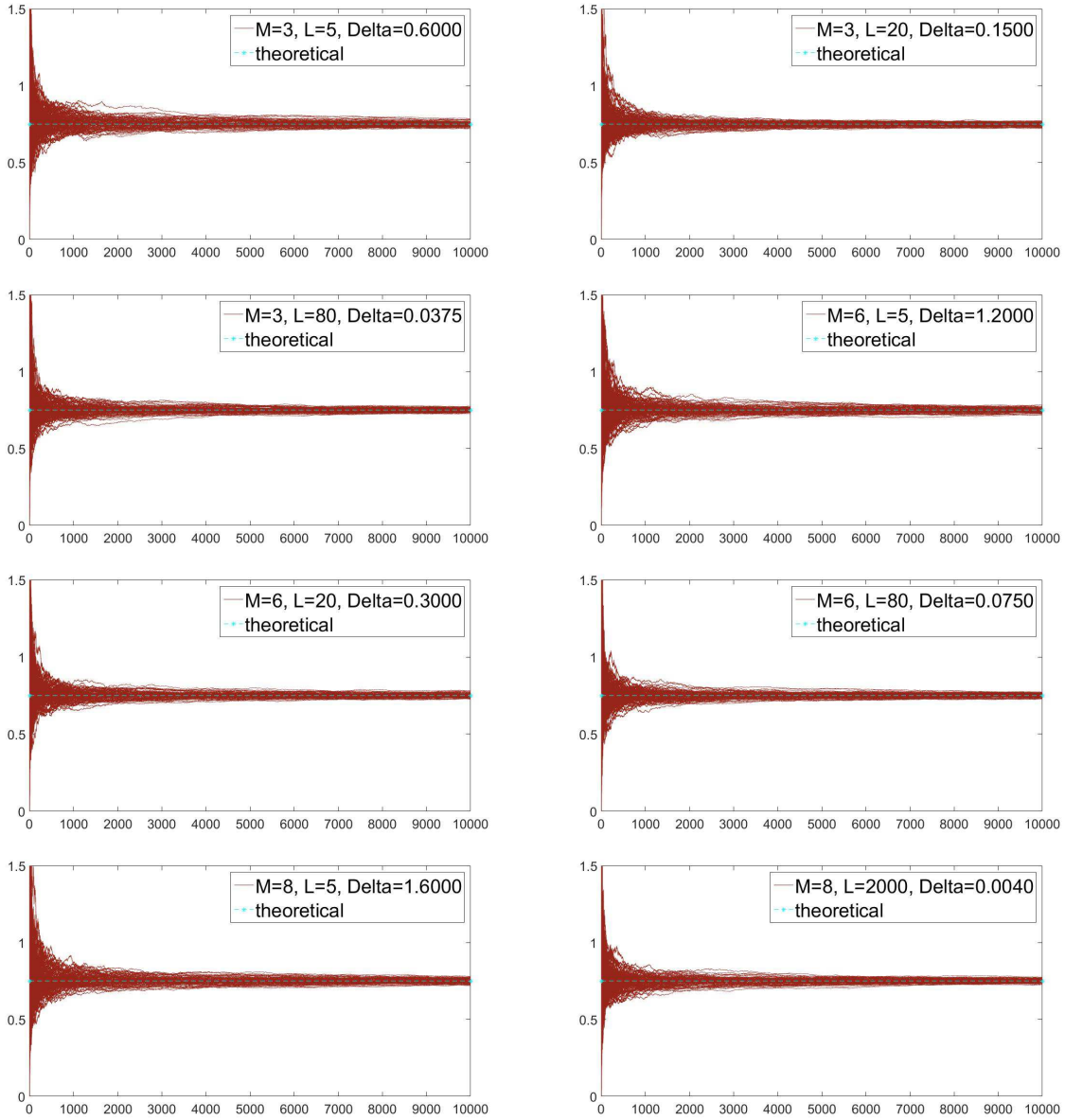

Figure 3.14 Self-normalized IS performance of the corresponding densities in Fig. 3.13

may not approximate q^*_{sn} well. We can see that the proposal densities with parameters $(M,L) = (6,5),(8,5)$ do not have the reflection behavior of q^*_{sn} because the first subinterval is so big that it eats the pivot point. As we expect, the proposal densities with parameters $(M,L) = (6,80),(8,2000)$ can approximate q^*_{sn} very well thanks to the large enough parameters M and L. However, from

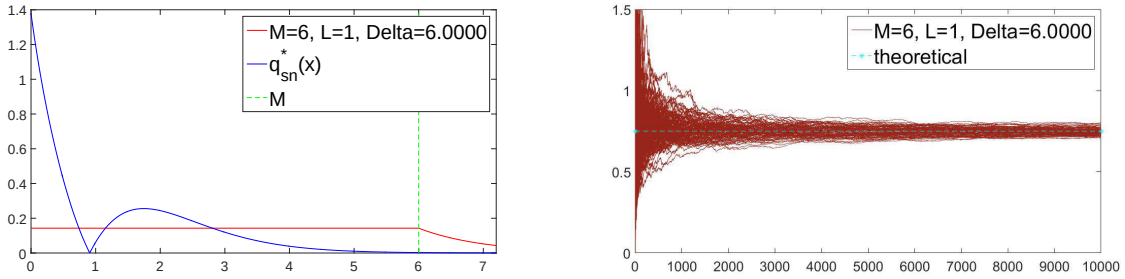

Figure 3.15 Self-normalized IS performance with $(M, L) = (6, 1)$

Fig. 3.14, all the proposal densities surprisingly have insignificantly different performance including even the one with $(M, L) = (8, 2000)$. If we look back at Fig. 1.8, we will see that even the optimal proposal density $q_2 = q^*_{sn}$ has the same performance as these proposal densities obtained from the partition-based method.

Let's take a look at the result with another pair of fixed parameters $(M, L) = (6, 1)$ with the same tail function. We may say that the variance of the estimator with $(M, L) = (6, 1)$ seems to be slightly bigger than the others, but the result still surprisingly have insignificantly different performance. It seems like every density that satisfies all the conditions for self-normalized IS from Table 3.2 should work. Note that some proposal densities used in this example do not bring Δ less than the pseudo period of $|f - \mu|p$ on $[0, M]$, but they still work well.

From Example 3.7, we should question ourselves whether it is worth putting lots of effort to approximate q^*_{sn}. These efforts include creating the self-determined-where-to-cut scheme for self-normalized IS. The big difference between basic IS and self-normalized IS is that for a positive (or negative) target function f, q^*_b yields the zero-variance estimator for basic IS, but q^*_{sn} never gives the zero-variance estimator for self-normalized IS. Thus, we suggest to choose parameters M and L large enough and proceed with the partition-based method as we did in this example for the self-normalized IS. We just have to make sure that the tail function will make the proposal density satisfy all the required conditions for self-normalized IS from Table 3.2. Specifically, we want our proposal density to satisfy Table 3.3 and the absolute continuity condition. Note that the current development in choosing proposal densities for IS method focuses only on basic IS. Currently, there is only the rule of thumb (1.7) in choosing proposal densities for self-normalized IS, and that rule of thumb may fail as we can see from Example 1.12. Perhaps, statisticians just rely on the rule of thumb for self-normalized IS because the performances of various proposal densities are not significantly different.

3.4 Multidimensional Spaces

In this section, an algorithm for choosing a proposal density for IS method in multi-dimensional space that has a good performance with finite variance is proposed. The concept is still to obtain a proposal density that is close to the optimal density, and to ensure that it yields a finite variance estimate. In Section 3.3 which talks about only 1-dimensional space, the domain can be easily separated into two regions: the importance region and the tail region. The importance region in which most of the samples lie is partitioned into many subintervals, and a proposal density is approximated to be piecewise constant on each subinterval based on the optimal density formula from Theorem 1.7 or 1.10. A proper function, which is easy to handle and guarantees finite variance estimate, is selected to be the tail region for the proposal density. With this proposal density, the IS method can be carried out properly. Now, for a multi-dimensional space, we would like to stay on this concept. However, there are more messy problems we have to cope with. The primary problems are how to separate the importance region and the tail region, and how to choose a workable tail function that allows the capability to be sampled from the final proposal density, not just any function that is integrable over the tail region and satisfies the bounded condition.

We will explain a partition-based method in a multi-dimensional space D for $(f,\pi)\in\mathscr{G}_b^b(D)$. The other three classes $\mathscr{G}_b^u(D), \mathscr{G}_{sn}^b(D), \mathscr{G}_{sn}^u(D)$ can also be adjusted to adopt the procedure as we did in Section 3.3. Recall the auto-random-twice scheme which is an alternative sampling method for 1-dimensional problem that uses two random numbers for one sample. We will generalize that idea to a multi-dimensional problem. We will explain for a d-dimensional space $D=[0,\infty)^d$. For other kinds of multi-dimensional space such as $\mathbb{R}^d$ or $[a,b]^d$, we can still adjust this proposed method.

Let $D_i=[0,\infty)$ be the domain for the i^{th} coordinate so that the domain $D=\prod_{i=1}^d D_i=[0,\infty)^d$. Similar to what we did in the 1-dimensional problem, we fix a proper constant $M_i\in(0,\infty)$ somewhere in D_i to get importance regions $[0,M_i]$ and tails $[M_i,\infty)$. Then, choose L_i the number of partitions in the importance region in each dimension. Let

$$\Delta_i=\frac{M_i}{L_i} \quad \text{and} \quad x_i^j=j\Delta_i \text{ for } j=0,1,\ldots,L_i$$

and $x_i^{L_i+1}=\infty$. Here, i is the index indicating which dimension to focus, and j in x_i^j is the index indicating the partition

$$0=x_i^0<x_i^1<\ldots<x_i^{L_i}=M_i<\infty=x_i^{L_i+1}$$

for $D_i=[0,\infty)$. Let's name each subinterval

$$J_i^j=[x_i^{j-1},x_i^j)=[(j-1)\Delta_i,j\Delta_i) \quad \text{for } j=1,\ldots,L_i$$

and

$$J_i^{L_i+1} = [M_i, \infty).$$

Then, we have a partition

$$\left\{ J_1^{j_1} \times \ldots \times J_d^{j_d} \right\}_{j_i \in \{1,\ldots,L_i+1\} \forall i=1,\ldots,d}$$

for the domain D. We will call $\prod_{i=1}^{d}[0, M_i)$ the importance region, and its compliment the tail region of our proposal distribution.

Now, we can imitate the idea of mixture distribution in Section 3.3. In each subcell of the partition, we will use 1-dimensional independent distribution for each coordinate. We will use uniform distributions in all coordinates for a subcell in the importance region, and a mixture distribution of uniform distributions and generalized Pareto distributions for a subcell in the tail region. To do this, we choose representative points from each subcell in the importance region

$$x^*_{j_1,\ldots,j_d} \in J_1^{j_1} \times \ldots \times J_d^{j_d}$$

for $j_i \in \{1,\ldots,L_i\}$ for all $i = 1,\ldots,d$ and evaluate with $|f|\pi$, the optimal function without the normalizing constant,

$$\tilde{h}_{j_1,\ldots,j_d} = \left| f\left(x^*_{j_1,\ldots,j_d}\right) \right| \pi\left(x^*_{j_1,\ldots,j_d}\right)$$

for $j_i \in \{1,\ldots,L_i\}$ for all $i = 1,\ldots,d$ to get the unnormalized proposal density $\tilde{q}$ on each subcell in the importance region

$$\tilde{q}(\vec{x}) = \tilde{h}_{j_1,\ldots,j_d} \quad \forall \vec{x} \in J_1^{j_1} \times \ldots \times J_d^{j_d} \tag{3.7}$$

for $j_i \in \{1,\ldots,L_i\}$ for all $i = 1,\ldots,d$. Note that this is well defined because we have the partition for the importance region

$$\bigcup_{\substack{j_i \in \{1,\ldots,L_i\} \\ \forall i=1,\ldots,d}}^{\cdot} J_1^{j_1} \times \ldots \times J_d^{j_d}.$$

Thus, for each $\vec{x}$ in the importance region, there exists uniquely a cell $J_1^{j_1} \times \ldots \times J_d^{j_d}$ in which $\vec{x}$ lives, and we will have $\tilde{h}_{j_1,\ldots,j_d}$ for the corresponding indices $j_1,\ldots,j_d$. We need to check whether $\tilde{h}_{j_1,\ldots,j_d}$'s cause the zero problem, and may need to justify a new choice of $x^*_{j_1,\ldots,j_d}$. According to the assumption of Proposition 3.1, if we assume that $\mathrm{Var}_\pi(f) < \infty$, we have to choose q that makes $\frac{\pi}{q}$ bounded, otherwise we need that $\frac{f\pi}{q}$ is bounded. For $(f,\pi) \in \mathscr{G}_b^b(D)$, we choose q to be a finite piecewise-constant function in the importance region, so the bounded condition in importance region is satisfied for these cases.

For the tail region, we choose the tail function

$$\tilde{q}_t(\vec{x}) = \tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,\, j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \prod_{\substack{i=1 \\ j_i=L_i+1}}^{d} \sigma_i \mathrm{GPD}_{(\xi_i,\sigma_i,M_i)}(x_i) \tag{3.8}$$

for $\vec{x} = (x_1,\ldots,x_d) \in J_1^{j_1} \times \ldots \times J_d^{j_d}$, for appropriate parameters ξ_i's and σ_i's. Let's contemplate this complicated-looking function. For $\vec{x} = (x_1,\ldots,x_d)$ in the tail region, there must be at least one coordinate, say x_i, falling in the tail part of that coordinate, $[M_i,\infty)$. For such coordinate, the corresponding index j_i will be L_i+1. The generalized Pareto distributions are used for those coordinates which fall into the tail region, so we have the product for tail-regioned coordinates

$$\prod_{\substack{i=1 \\ j_i=L_i+1}}^{d} \sigma_i \mathrm{GPD}_{(\xi_i,\sigma_i,M_i)}(x_i).$$

The constants σ_i's are put into the product because we want to control the height at the connected part between the importance region and the tail region of the final proposal density. Since $\mathrm{GPD}_{(\xi_i,\sigma_i,M_i)}(M_i) = \frac{1}{\sigma_i}$ for each i, the above product at the connected part of each dimension will be 1. For the other coordinates falling in the importance region, we simply use the uniform distribution over the corresponding subinterval for each coordinate. Thus, it is just constant for the importance-regioned coordinates, say 1.

Now, an appropriate height is picked and put into the product formula. For a tail-regioned coordinate s, we want the height $\tilde{h}$ with the corresponding index L_s indicating the last interval in importance region of that coordinate. Since $j_s = L_s+1$ for the tail-regioned coordinate, j_s-1 will be our wanted index. For an importance-regioned coordinate r, the stay-still index j_r is simply used. For such coordinate, $j_r < L_r+1$. Thus, we have that

$$j_i - \mathbb{1}_{\{L_i+1\}}(j_i) = \begin{cases} j_i - 1 = L_i & \text{, if } x_i \in [M_i,\infty) \\ j_i & \text{, if } x_i \in [0,M_i). \end{cases}$$

Hence, $\tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,\, j_d-\mathbb{1}_{\{L_d+1\}}(j_d)}$ is our wanted height for controlling smoothness of the result function $\tilde{q}$.

Recalling the uniform and generalized Pareto densities from (3.5) and (3.6), we have, in this situation,

$$\mathrm{Unif}_{J_i^{j_i}}(x_i) = \mathrm{Unif}_{[(j_i-1)\Delta_i,\, j_i\Delta_i)}(x_i) = \frac{1}{\Delta_i}\mathbb{1}_{[(j_i-1)\Delta_i,\, j_i\Delta_i)}(x_i)$$

and

$$\mathrm{GPD}_{(\xi_i,\sigma_i,M_i)}(x_i)=\begin{cases}\frac{1}{\sigma_i\left(1+\xi_i\frac{(x_i-M_i)}{\sigma_i}\right)^{1+\frac{1}{\xi_i}}}\mathbb{1}_{[M_i,\infty)}(x_i) & \text{, if } \xi_i>0\\ \frac{1}{\sigma_i}e^{-\frac{(x_i-M_i)}{\sigma_i}}\mathbb{1}_{[M_i,\infty)}(x_i) & \text{, if } \xi_i=0.\end{cases}$$

Combining (3.7) and (3.8), we have the unnormalized proposal density

$$\begin{aligned}\tilde{q}(\vec{x}) \;=\; & \tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,\, j_d-\mathbb{1}_{\{L_d+1\}}(j_d)}\\ & \cdot\prod_{\substack{r=1\\ j_r<L_r+1}}^{d}\Delta_r\mathrm{Unif}_{J_r^{j_r}}(x_r)\;\cdot\prod_{\substack{s=1\\ j_s=L_s+1}}^{d}\sigma_s\mathrm{GPD}_{(\xi_s,\sigma_s,M_s)}(x_s)\end{aligned}\tag{3.9}$$

for $\vec{x}=(x_1,\ldots,x_d)\in J_1^{j_1}\times\ldots\times J_d^{j_d}$. Note that the product with the dummy variable r refers to the importance-regioned coordinates, and the product with the dummy variable s refers to the tail-regioned coordinates. Also, note that the empty product is by convention equal to 1. Note again that this is well defined because we have the partition for D

$$D=\bigcup_{\substack{j_i\in\{1,\ldots,L_i+1\}\\ \forall i=1,\ldots,d}}^{\cdot} J_1^{j_1}\times\ldots\times J_d^{j_d}.$$

Thus, for each $\vec{x}=(x_1,\ldots,x_d)\in D$, there exists uniquely determined cell $J_1^{j_1}\times\ldots\times J_d^{j_d}$ in which $\vec{x}$ lives; hence, we obtain the corresponding indices $j_1,\ldots,j_d$. Since this unnormalized density can be defined through this partition, the matching unnormalized distribution can be expressed as the mixture distribution

$$\tilde{q}\;=\;\sum_{j_1=1}^{L_1+1}\cdots\sum_{j_d=1}^{L_d+1}\tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,\, j_d-\mathbb{1}_{\{L_d+1\}}(j_d)}\;\cdot\prod_{\substack{r=1\\ j_r<L_r+1}}^{d}\Delta_r\mathrm{Unif}(J_r^{j_r})\;\cdot\prod_{\substack{s=1\\ j_s=L_s+1}}^{d}\sigma_s\mathrm{GPD}(\xi_s,\sigma_s,M_s).$$

Finally, $\tilde{q}$ can be normalized by the normalizing constant Z_q obtained from integrating (3.9) over the whole domain

$$\begin{aligned}Z_q \;&=\; \int\tilde{q}(\vec{x})d\vec{x}\\ &=\;\sum_{j_1=1}^{L_1+1}\cdots\sum_{j_d=1}^{L_d+1}\tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,\, j_d-\mathbb{1}_{\{L_d+1\}}(j_d)}\;\cdot\prod_{\substack{r=1\\ j_r<L_r+1}}^{d}\Delta_r\;\cdot\prod_{\substack{s=1\\ j_s=L_s+1}}^{d}\sigma_s\\ &=\;\Delta\cdot\sum_{j_1=1}^{L_1+1}\cdots\sum_{j_d=1}^{L_d+1}\tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,\, j_d-\mathbb{1}_{\{L_d+1\}}(j_d)}\;\cdot\prod_{\substack{s=1\\ j_s=L_s+1}}^{d}\frac{\sigma_s}{\Delta_s}\end{aligned}$$

where $\Delta = \prod_{i=1}^{d} \Delta_i$. Setting

$$h_{j_1,\ldots,j_d} = \frac{1}{Z_q} \tilde{h}_{j_1,\ldots,j_d}$$

for $j_i \in \{1,\ldots,L_i\}$ for all $i = 1,\ldots,d$, we acquire the proposal density

$$q(\vec{x}) = h_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \cdot \prod_{\substack{r=1 \\ j_r<L_r+1}}^{d} \Delta_r \mathrm{Unif}_{J_r^{j_r}}(x_r) \cdot \prod_{\substack{s=1 \\ j_s=L_s+1}}^{d} \sigma_s \mathrm{GPD}_{(\xi_s,\sigma_s,M_s)}(x_s)$$

for $\vec{x} = (x_1,\ldots,x_d) \in J_1^{j_1} \times \ldots \times J_d^{j_d}$, with the corresponding mixture distribution

$$\begin{aligned} q = & \sum_{j_1=1}^{L_1+1} \cdots \sum_{j_d=1}^{L_d+1} h_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \\ & \cdot \prod_{\substack{r=1 \\ j_r<L_r+1}}^{d} \Delta_r \mathrm{Unif}(J_r^{j_r}) \cdot \prod_{\substack{s=1 \\ j_s=L_s+1}}^{d} \sigma_s \mathrm{GPD}(\xi_s,\sigma_s,M_s). \end{aligned}$$

Now, let's discuss how to choose parameters M_i and L_i for $i = 1,\ldots,d$. This also relates to writing computer programming for the method. The idea is the same as the one-dimensional case, but the tail region is also partitioned into many pieces now.

The first step is to choose parameters ξ_i and σ_i for $i = 1,\ldots,d$ in order to satisfy all the desired assumptions. Specifically, the generalized Pareto densities for the tail function $\tilde{q}_t$ have to make $\frac{\pi}{\tilde{q}_t}$ bounded on the tail region for $(f,\pi) \in \mathscr{A}_b(D) \cap \mathscr{B}(D) \cap \mathscr{C}_1(D)$; or $\frac{f\pi}{\tilde{q}_t}$ bounded on $[M,\infty)$ for $(f,\pi) \in \mathscr{A}_b(D) \cap \mathscr{C}_2(D)$. The second step is to choose Δ_i for $i = 1,\ldots,d$. The smaller Δ_i is, the better the approximation will be. The next crucial step is to find the cutting point M_i. We will seek L_i, the number of subintervals in the i^{th} coordinate. Let ϵ_Δ be an upper bound for the probability of getting a sample in the tail region and ϵ_0 the machine precision bound. The probability of getting a sample in the tail region is

$$\frac{S_{\text{tail}}}{S_{\text{imp}} + S_{\text{tail}}}$$

where

$$\begin{aligned} S_{\text{imp}} &= \Delta \cdot \sum_{j_1=1}^{L_1} \cdots \sum_{j_d=1}^{L_d} \tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \cdot \prod_{\substack{s=1 \\ j_s=L_s+1}}^{d} \frac{\sigma_s}{\Delta_s} \\ &= \Delta \cdot \sum_{j_1=1}^{L_1} \cdots \sum_{j_d=1}^{L_d} \tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \end{aligned}$$

and

$$S_{\text{tail}} = \left(\Delta \cdot \sum_{j_1=1}^{L_1+1} \cdots \sum_{j_d=1}^{L_d+1} \tilde{h}_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \cdot \prod_{\substack{s=1 \\ j_s=L_s+1}}^{d} \frac{\sigma_s}{\Delta_s} \right) - S_{\text{imp}}.$$

Thus, we want to have that

$$\epsilon_0 < \frac{S_{\text{tail}}}{S_{\text{imp}} + S_{\text{tail}}} < \epsilon_\Delta.$$

Simple calculation yields

$$\frac{\epsilon_0}{1-\epsilon_0} < \frac{S_{\text{tail}}}{S_{\text{imp}}} < \frac{\epsilon_\Delta}{1-\epsilon_\Delta}.$$

The machine precision bound ϵ_0 is so small that having $\frac{\epsilon_0}{1-\epsilon_0} < \frac{S_{\text{tail}}}{S_{\text{imp}}}$ should not be a problem. With the same reason as how we deal with the 1-dimensional case, we will increase L_i, for all $i = 1,\ldots,d$, from 1 until we get

$$\frac{S_{\text{tail}}}{S_{\text{imp}}} < \frac{\epsilon_\Delta}{1-\epsilon_\Delta}.$$

Once we determine L_i, we have $M_i = \Delta_i L_i$ and the proposed method can be carried on.

Since the probability to get a sample in each cell $J_1^{j_1} \times \ldots \times J_d^{j_d}$ for $j_i \in \{1,\ldots,L_i+1\}$ for all $i = 1,\ldots,d$ is known, we can easily obtain a sample by first drawing a uniform random number to decide which cell the sample falls into, and then drawing d random numbers corresponding to each dimensional distribution. For the dimension that the sample falls into the importance region, the corresponding uniform distribution is used, and for the dimension that the sample falls into the tail region, the corresponding generalized Pareto distribution is used.

We summarize the algorithm in Table 3.9. Note that the variable Pr, which will finally be a multidimensional matrix, collects the probability of getting a sample in each partitioned cell divided by Δ. Also, PrIR and PrTail are the probability of getting a sample in the importance region and in the tail region, respectively, divided by Δ. Dividing by Δ helps moderately reduce the running time for programming. Step 1 - 5 talk about getting the proposal density, and step 6 - 8 produce one random sample from this density. The function cumsum is the cumulative sum. As for other types of the domain such as $\mathbb{R}^d$, the idea of this algorithm can still be imitated by extending each layer in all directions.

Example 3.8. Consider a 4-dimensional problem with

$$f(\vec{x}) = x_1^2 x_2 x_3 \left(x_4 - \frac{1}{\sqrt{2}} \right)$$

and

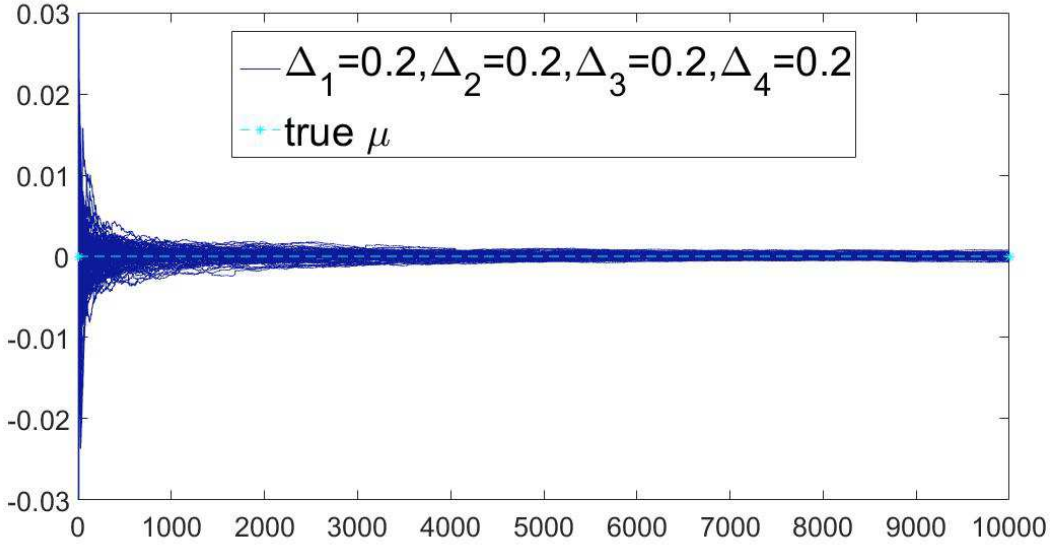

Figure 3.16 The performance of a 4-dimensional partition-based proposal density in Example 3.8

$$\pi(\vec{x}) = \frac{384\sqrt{2}}{\left(4+3\sqrt{\pi}\right)\pi} \frac{(x_1+x_2)e^{-(x_1+2x_3)^2}}{(x_2+1)^5(x_4^4+1)}.$$

The true expectation is 0, and $\mathrm{Var}_\pi(f) = \frac{5}{768} + \frac{401}{6720(4+3\sqrt{\pi})} \approx 0.0129149$.

The algorithm from Table 3.9 is applied to this 4-dimensional problem. Choosing $\xi_1 = 0, \xi_2 = \frac{1}{2}, \xi_3 = 0, \xi_4 = \frac{1}{2}$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$ will guarantee the bounded condition. Here, we choose $\epsilon_\Delta = 5\%$ and $\Delta_1 = \Delta_2 = \Delta_3 = \Delta_4 = 0.2$. The performance of the proposed method is shown in Fig. 3.16 with 100 scenarios of 10,000 samples. The time for computing the proposal density is 8.1254 seconds, and the sampling time for a million samples here is 479.4017 seconds.

Observe that this setting problem may be difficult to apply the classical Monte Carlo method because of the difficulty in sampling directly from π. Also, in order to use the efficient IS, a family of distributions must be appropriately selected.

Table 3.9 Algorithm of the partition-based method for basic IS in the multidimensional spaces $[0,\infty)^d$

Step 1	Choose parameters $\xi_i, \sigma_i, \Delta_i, \epsilon_\Delta$ for the algorithm. Set $\Delta = \prod_{i=1}^{d} \Delta_i$ and Threshold $= \frac{\epsilon_\Delta}{1-\epsilon_\Delta}$.
Step 2	Initially set $L_i = 1$ for all i and assign $\Pr_{1,\ldots,1} = \big(\lvert f\rvert\pi\big)\big(\frac{\Delta_1}{2},\ldots,\frac{\Delta_d}{2}\big)$. Also set PrIR $= \Pr_{1,\ldots,1}$ and PrTail $= 0$.
Step 3	Assign only additional layer (L_i+1) to Pr according to $\Pr_{j_1,\ldots,j_d} = \Pr_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \cdot \prod_{\substack{s=1 \\ j_s=L_s+1}}^{d} \frac{\sigma_s}{\Delta_s}$ and add this $\Pr_{j_1,\ldots,j_d}$ to PrTail for each additional cell.
Step 4	While $\frac{\text{PrTail}}{\text{PrIR}} \geq$ Threshold Increase all L_i's by 1, and reset PrTail $= 0$. Reassign only the previous layer (L_i now) to Pr according to $\Pr_{j_1,\ldots,j_d} = \big(\lvert f\rvert\pi\big)\big((j_1-\frac{1}{2})\Delta_1,\ldots,(j_d-\frac{1}{2})\Delta_d\big)$ and add this $\Pr_{j_1,\ldots,j_d}$ to PrIR. Assign only additional layer (L_i+1) to Pr according to $\Pr_{j_1,\ldots,j_d} = \Pr_{j_1-\mathbb{1}_{\{L_1+1\}}(j_1),\ldots,j_d-\mathbb{1}_{\{L_d+1\}}(j_d)} \cdot \prod_{\substack{s=1 \\ j_s=L_s+1}}^{d} \frac{\sigma_s}{\Delta_s}$ and add this $\Pr_{j_1,\ldots,j_d}$ to PrTail for each additional cell. end while loop
Step 5	Compute $M_i = L_i \cdot \Delta_i$, and check for zero problem. Normalize all $\Pr_{j_1,\ldots,j_d}$ by $\Delta \cdot \sum_{j_1=1}^{L_1+1} \cdots \sum_{j_d=1}^{L_d+1} \Pr_{j_1,\ldots,j_d}$.
Step 6	Set $I = \Delta \cdot \text{cumsum}(\text{Pr})$.
Step 7	Draw a uniform random number Unif$(0,1)$ and find indices $i_1,\ldots,i_d$ of I that the random number falls into.

Step 8	Set qx $= \Pr_{i_1,\ldots,i_d}$.
	For $t = 1$ to d
	If $i_t \leq L_t$
	$x_t = \text{Unif}((i_t-1)\Delta_t, i_t\Delta_t)$
	else
	$x_t = \text{GPD}(\xi_t, \sigma_t, M_t)$
	qx $=$ qx $\cdot\, \Delta_t \cdot \text{GPD}_{(\xi_t,\sigma_t,M_t)}(x_t)$
	end if
	end for loop

Step 9	Calculate IS approximation using (1.3) with a number of samples from step 6 - 8.

CHAPTER

4

PRACTICAL EXAMPLES

4.1 Option Greeks

In finance, option prices can change due to directional price shifts in the underlying asset, prices changes in the implied volatility, time decay, and even changes in the interest rates. The option Greeks [15] are the quantities representing the sensitivity of the price of the options to change in these parameters. They have also been called the risk sensitivities, risk measures or hedge parameters. They play an important role in understanding the sensitivity of prices to relevant parameters, constructing a hedging portfolio, and approximating the loss distribution for risk management. The most commonly used Greeks are delta (Δ), vega (ν), theta (Θ), rho (ρ) and gamma (Γ). We will write their full names instead of using Greek letters to avoid redundancy in using the Greek letter Δ.

Consider the Black-Scholes model [15] for the European call option with initial stock price S_0, strike price K, expiration T, risk-free interest rate r, and volatility constant σ. The underlying stock price follows the geometric Brownian motion

$$dS_t = rS_t\,dt + \sigma S_t\,dB_t$$

which has the solution

$$S_t = S_0 e^{\left(r-\frac{\sigma^2}{2}\right)t+\sigma B_t} \tag{4.1}$$

where B_t is a Brownian motion. The probability density function for S_T is

$$\begin{aligned} f_{S_T}(x) = f_{S_T}(x; S_0, r, \sigma, T) &= \frac{1}{\sigma\sqrt{T}\,x}\ \phi\left(\zeta(x)\right) \\ &= \frac{1}{\sigma\sqrt{2\pi T}\,x}\ e^{-\frac{\left(\log\left(\frac{x}{S_0}\right)-\left(r-\frac{\sigma^2}{2}\right)T\right)^2}{2\sigma^2 T}} \end{aligned}$$

where ϕ is the standard normal density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{4.2}$$

and

$$\zeta(x) = \zeta(x; S_0, r, \sigma, T) = \frac{\log\left(\frac{x}{S_0}\right)-\left(r-\frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}.$$

The price of the European call option is

$$\begin{aligned} C = C(S_0, r, \sigma, T, K) &= \mathbb{E}\left[e^{-rT}\,(S_T-K)_+\right], \qquad S_T \sim f_{S_T} \\ &= \int_0^\infty e^{-rT}\,(x-K)_+\, f_{S_T}(x)\,dx \end{aligned}$$

where $a_+ = \max(a, 0)$.

Example 4.1 (Black-Scholes Delta)**.** Delta of an option measures the rate of change of the theoretical option value with respect to changes in the underlying asset's price. It is the first derivative of the value of the option with respect to the underlying asset's price, i.e. $\frac{\partial C}{\partial S_0}$. We can manage to interchange the order of differentiation and integration

$$\begin{aligned} \frac{\partial C}{\partial S_0} = \frac{\partial}{\partial S_0} C(S_0, r, \sigma, T, K) &= \frac{\partial}{\partial S_0} \int_0^\infty e^{-rT}(x-K)_+\, f_{S_T}(x)\,dx \\ &= \int_0^\infty e^{-rT}\,(x-K)_+\, \frac{\partial}{\partial S_0} f_{S_T}(x)\,dx. \end{aligned}$$

We can consider $e^{-rT}\,(x-K)_+\,\frac{\partial}{\partial S_0} f_{S_T}(x)\,dx$ as a finite measure. Then, importance sampling can be used to estimate delta of the option by the change of measure

$$\begin{aligned}\frac{\partial C}{\partial S_0} &= \int_0^\infty \left(\frac{e^{-rT}(x-K)_+ \frac{\partial}{\partial S_0} f_{S_T}(x)}{f_{S_T}(x)} \right) f_{S_T}(x)\, dx \\ &= \int_0^\infty e^{-rT}(x-K)_+ \left(-\zeta(x) \frac{\partial \zeta(x)}{\partial S_0} \right) f_{S_T}(x)\, dx \\ &= \int_0^\infty e^{-rT}(x-K)_+ \left(\frac{\zeta(x)}{S_0 \sigma \sqrt{T}} \right) f_{S_T}(x)\, dx \\ &= \mathbb{E}\left[e^{-rT}(S_T-K)_+ \left(\frac{\zeta(S_T)}{S_0 \sigma \sqrt{T}} \right) \right].\end{aligned}$$

This is called a likelihood ratio method [12], and $\frac{\frac{\partial}{\partial S_0} f_{S_T}(x)}{f_{S_T}(x)}$ often written as $\frac{\partial \log f_{S_T}(x)}{\partial S_0}$ is called a score function. If S_T is generated with fixed parameters S_0, r, σ, T from (4.1) using a standard normal random variable Z with

$$S_T = S_0 e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z}, \quad Z \sim \mathcal{N}(0,1),$$

then $\zeta(S_T) = Z$ and the estimator simplifies to

$$e^{-rT}(S_T - K)_+ \frac{Z}{S_0 \sigma \sqrt{T}}.$$

Therefore,

$$\frac{\partial C}{\partial S_0} = e^{-rT}\, \mathbb{E}\left[\left(S_0 e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z} - K \right)_+ \frac{Z}{S_0 \sigma \sqrt{T}} \right], \quad Z \sim \mathcal{N}(0,1)$$

and the Monte Carlo estimator for the option delta is

$$e^{-rT}\, \frac{1}{N} \sum_{i=1}^{N} \left[\left(S_0 e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z_i} - K \right)_+ \frac{Z_i}{S_0 \sigma \sqrt{T}} \right], \quad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

However, instead of changing the measure to the original f_{S_T}, we can use the partition-based method describe in Chapter 3. To avoid unnecessary zero h_i in the partition-based method, we can change the variable of the integral

$$\begin{aligned}\frac{\partial C}{\partial S_0} &= \int_K^\infty e^{-rT}(x-K) \left(\frac{\zeta(x)}{S_0 \sigma \sqrt{T}} \right) f_{S_T}(x)\, dx \\ &= \int_0^\infty e^{-rT}\, x \left(\frac{\zeta(x+K)}{S_0 \sigma \sqrt{T}} \right) f_{S_T}(x+K)\, dx\end{aligned}$$

and use the proposal density q_b acquired from the partition-based method instead of using f_{S_T}

$$\begin{aligned}\frac{\partial C}{\partial S_0} &= \int_0^\infty e^{-rT}\, x \left(\frac{\zeta(x+K)}{S_0\sigma\sqrt{T}}\right)\left(\frac{f_{S_T}(x+K)}{q_b(x)}\right) q_b(x)\, dx \\ &= e^{-rT}\, \mathbb{E}\left[X\left(\frac{\zeta(X+K)}{S_0\sigma\sqrt{T}}\right)\frac{f_{S_T}(X+K)}{q_b(X)}\right], \quad X \sim q_b\end{aligned}$$

with

$$q_b(x) \approx q_b^*(x) \propto \left| x\left(\frac{\zeta(x+K)}{S_0\sigma\sqrt{T}}\right)\right| f_{S_T}(x+K).$$

Here, the target function is $f(x) = x\left(\frac{\zeta(x+K)}{S_0\sigma\sqrt{T}}\right)$ and the unnormalized target density is $\pi(x) = f_{S_T}(x+K)$ on the domain $(0,\infty)$. The partition-based method still works for unnormalized target density because the approximated proposal density will be eventually normalized. Then, the partition-based Monte Carlo estimator for the option delta is

$$e^{-rT}\, \frac{1}{N}\sum_{i=1}^{N}\left[X_i\left(\frac{\zeta(X_i+K)}{S_0\sigma\sqrt{T}}\right)\frac{f_{S_T}(X_i+K)}{q_b(X_i)}\right], \quad X_i \overset{\text{iid}}{\sim} q_b.$$

The partition-based method is taken with the auto-sampling-once scheme using the polynomial tail with parameters $\alpha = 0.01, \Delta = 1.5, \epsilon_\Delta = 0.01$. Note that this auto-sampling-once scheme is equivalent to the auto-sampling-twice scheme parameters $\xi = 100, \sigma = 100$. Both methods are simulated with $n = 100$ scenarios and the number of samples in each scenario is $N = 10,000$. The performance of both the likelihood ratio method and the partition-based method is given in Fig. 4.1 using the Black-Scholes model with $S_0 = 100, K = 110, r = 0.05, \sigma = 0.2, T = 1$. The theoretical Black-Scholes delta is known to be $\Phi(d_1)$ where Φ is the standard normal cumulative distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy \tag{4.3}$$

and

$$d_1 = \frac{\log\left(\frac{S_0}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}. \tag{4.4}$$

For the likelihood ratio method, the time used in sampling step is 0.0671 seconds, and the time used in calculating Monte Carlo approximation is 0.0591 seconds. For the partition-based method, the time used in computing the proposal density is 0.0070 seconds, the time used in sampling step is 2.1997 seconds, and the time used in calculating IS approximation is 0.1653 seconds. Although the total time for the partition-based method is more than the total time for the likelihood ratio method, the accuracy of the partition-based method is far more satisfying than that of the likelihood ratio method. From Fig. 4.1, it seems that 100 samples from the partition-based method is better than

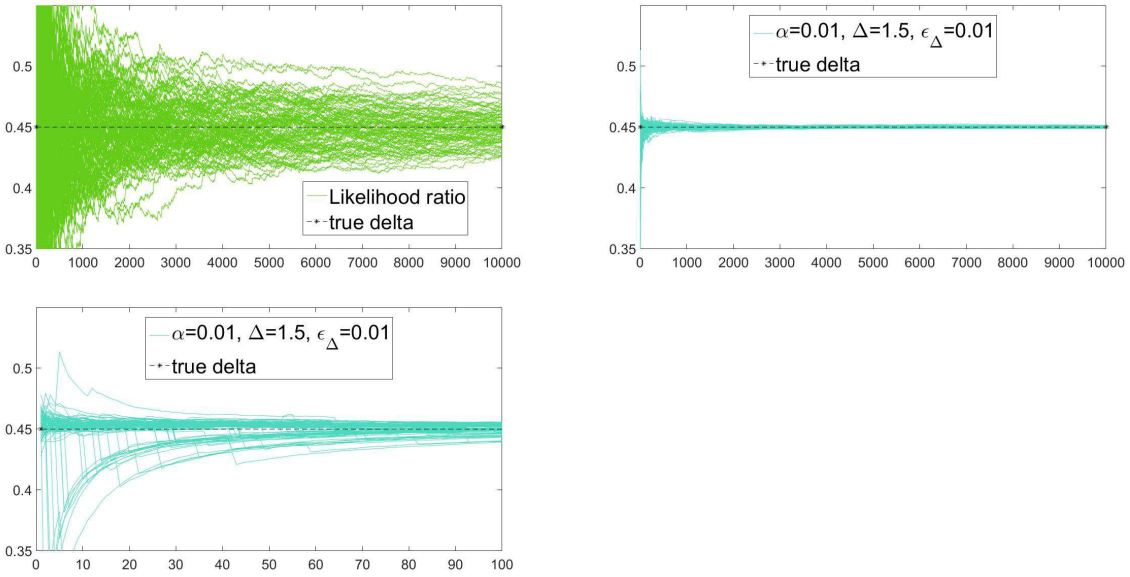

Figure 4.1 Estimating delta of European call option

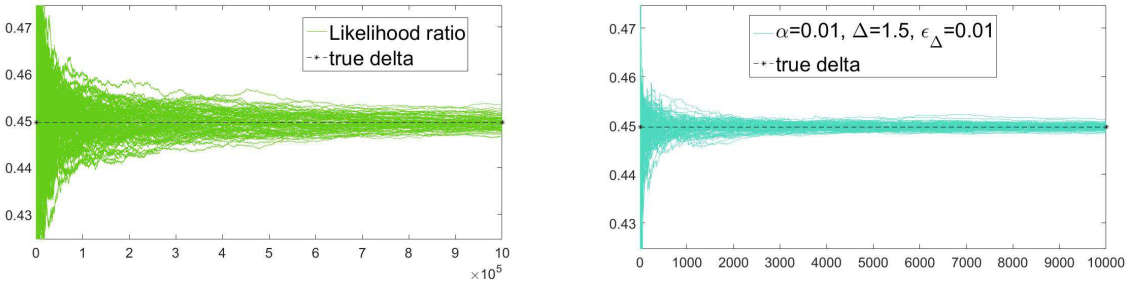

Figure 4.2 Comparison between the likelihood ratio method and the partition-based method for estimating delta of European call option

10,000 samples from the likelihood ration method. For this example, if we roughly calculate the time used in both methods with comparable accuracy, it appears that the partition-based method is more favorable.

For fairer comparison, we rerun the likelihood ratio method with sample size $N = 1,000,000$ and other parameters fixed. The performance of this simulation is presented in Fig. 4.2 against the previous result from the partition-based method with the same y-axis scale. The time used in the simulation of each method with $n = 100$ scenarios is presented in Table 4.1.

Example 4.2 (Black-Scholes Vega)**.** Vega of an option measures the rate of change of the theoretical

Table 4.1 Computing time in seconds of the likelihood ratio method with $N = 1,000,000$ and the partition-based method with $\Delta = 1.5, \epsilon_\Delta = 0.01, \xi = 100, \sigma = 100$ and $N = 10,000$ corresponding to Fig. 4.2

	Likelihood ratio with $N = 1,000,000$	$\Delta = 1.5$ with $N = 10,000$
Computing the proposal density	-	0.0070
Sampling	3.9659	2.1997
Calculating the approximation	21.5785	0.1653

option value with respect to changes in the volatility of the underlying asset, i.e. it is

$$\begin{aligned}\frac{\partial C}{\partial \sigma} = \frac{\partial}{\partial \sigma} C(S_0, r, \sigma, T, K) &= \frac{\partial}{\partial \sigma} \int_0^\infty e^{-rT}(x-K)_+ \, f_{S_T}(x)\, dx \\ &= \int_0^\infty e^{-rT}\,(x-K)_+ \, \frac{\partial f_{S_T}(x)}{\partial \sigma}\, dx.\end{aligned}$$

By simple calculation, we have that

$$\frac{\partial f_{S_T}(x)}{\partial \sigma} = \left(-\frac{1}{\sigma} - \zeta(x)\,\frac{\partial \zeta(x)}{\partial \sigma}\right) f_{S_T}(x).$$

with

$$\begin{aligned}\frac{\partial \zeta(x)}{\partial \sigma} &= \frac{\log\left(\frac{S_0}{x}\right) + \left(r + \frac{\sigma^2}{2}\right)T}{\sigma^2\sqrt{T}} \\ &= -\frac{\zeta(x)}{\sigma} + \sqrt{T}.\end{aligned}$$

Therefore, we can derive the Monte Carlo estimator for the option vega:

$$\begin{aligned}\frac{\partial C}{\partial \sigma} &= e^{-rT}\,\mathbb{E}\left[(S_T - K)_+ \left(-\frac{1}{\sigma} - \zeta(S_T)\left(-\frac{\zeta(S_T)}{\sigma} + \sqrt{T}\right)\right)\right] \\ &= e^{-rT}\,\mathbb{E}\left[\left(S_0 e^{\left(r-\frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z} - K\right)_+ \left(\frac{Z^2-1}{\sigma} - \sqrt{T}Z\right)\right], \quad Z \sim \mathcal{N}(0,1) \\ &\approx e^{-rT}\,\frac{1}{N}\sum_{i=1}^{N}\left[\left(S_0 e^{\left(r-\frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z_i} - K\right)_+ \left(\frac{Z_i^2-1}{\sigma} - \sqrt{T}Z_i\right)\right], \quad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).\end{aligned}$$

Also, the partition-based estimator for the option vega is

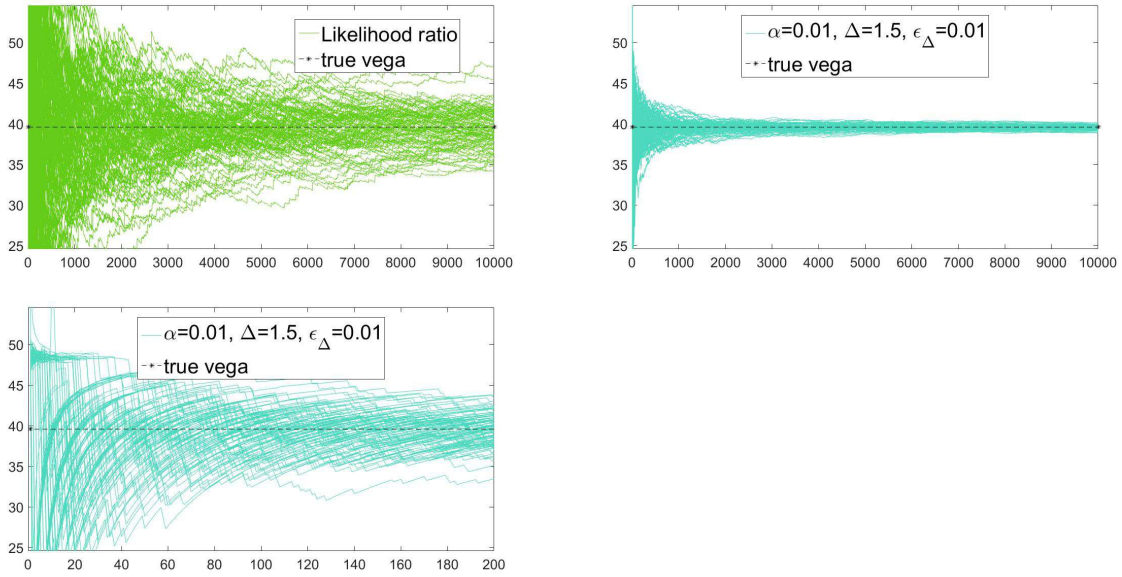

Figure 4.3 Estimating vega of European call option

$$\frac{\partial C}{\partial \sigma} = e^{-rT}\,\mathbb{E}\left[X\left(\frac{\zeta^2(X+K)-1}{\sigma}-\sqrt{T}\zeta(X+K)\right)\frac{f_{S_T}(X+K)}{q_b(X)}\right], \quad X\sim q_b$$
$$\approx e^{-rT}\,\frac{1}{N}\sum_{i=1}^{N}\left[X_i\left(\frac{\zeta^2(X_i+K)-1}{\sigma}-\sqrt{T}\zeta(X_i+K)\right)\frac{f_{S_T}(X_i+K)}{q_b(X_i)}\right], \quad X_i \overset{\text{iid}}{\sim} q_b.$$

The partition-based method is taken with the auto-sampling-once scheme using the polynomial tail with parameters $\alpha = 0.01, \Delta = 1.5, \epsilon_\Delta = 0.01$. Both methods are simulated with $n = 100$ scenarios and the number of samples in each scenario is $N = 10,000$. The performance of both the likelihood ratio method and the partition-based method is given in Fig. 4.3 using the Black-Scholes model with $S_0 = 100, K = 110, r = 0.05, \sigma = 0.2, T = 1$. The theoretical Black-Scholes vega is known to be $S_0\sqrt{T}\,\phi(d_1)$ where d_1 and ϕ are from (4.4) and (4.2). For the likelihood ratio method, the time used in sampling step is 0.0295 seconds, and the time used in calculating Monte Carlo approximation is 0.0269 seconds. For the partition-based method, the time used in computing the proposal density is 0.0089 seconds, the time used in sampling step is 1.7750 seconds, and the time used in calculating IS approximation is 0.1264 seconds.

For fairer comparison, we rerun the likelihood ratio method with sample size $N = 500,000$ and other parameters fixed. The performance of this simulation is presented in Fig. 4.4 against the previous result from the partition-based method with the same y-axis scale. The time used in the

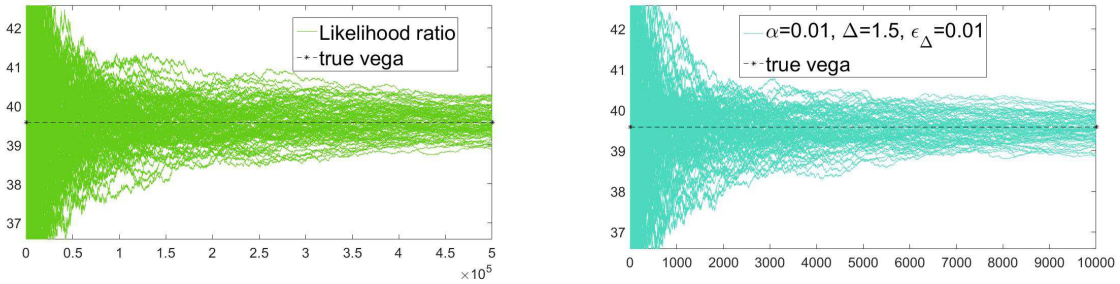

Figure 4.4 Comparison between the likelihood ratio method and the partition-based method for estimating vega of European call option

Table 4.2 Computing time in seconds of the likelihood ratio method with $N = 500,000$ and the partition-based method with $\Delta = 1.5, \epsilon_\Delta = 0.01, \xi = 100, \sigma = 100$ and $N = 10,000$ corresponding to Fig. 4.4

	Likelihood ratio with $N = 500,000$	$\Delta = 1.5$ with $N = 10,000$
Computing the proposal density	-	0.0089
Sampling	1.5942	1.7750
Calculating the approximation	1.2863	0.1264

simulation of each method with $n = 100$ scenarios is presented in Table 4.2.

Example 4.3 (Black-Scholes Theta)**.** Theta of an option measures the sensitivity of the option price to the time decay, i.e. it is

$$\begin{aligned} -\frac{\partial C}{\partial T} = -\frac{\partial}{\partial T} C(S_0, r, \sigma, T, K) &= -\frac{\partial}{\partial T} \int_0^\infty e^{-rT} (x-K)_+ \, f_{S_T}(x) \, dx \\ &= -\int_0^\infty (x-K)_+ \, \frac{\partial}{\partial T} \left(e^{-rT} \, f_{S_T}(x) \right) \, dx. \end{aligned}$$

By simple calculation, we have that

$$\frac{\partial}{\partial T} \left(e^{-rT} \, f_{S_T}(x) \right) = e^{-rT} \left(-r - \frac{1}{2T} - \zeta(x) \, \frac{\partial \zeta(x)}{\partial T} \right) f_{S_T}(x).$$

with

$$\frac{\partial \zeta(x)}{\partial T} = \frac{\log\left(\frac{S_0}{x}\right) - \left(r - \frac{\sigma^2}{2}\right) T}{2\sigma T^{\frac{3}{2}}} = -\frac{\zeta(x)}{2T} - \frac{r - \frac{\sigma^2}{2}}{\sigma \sqrt{T}}.$$

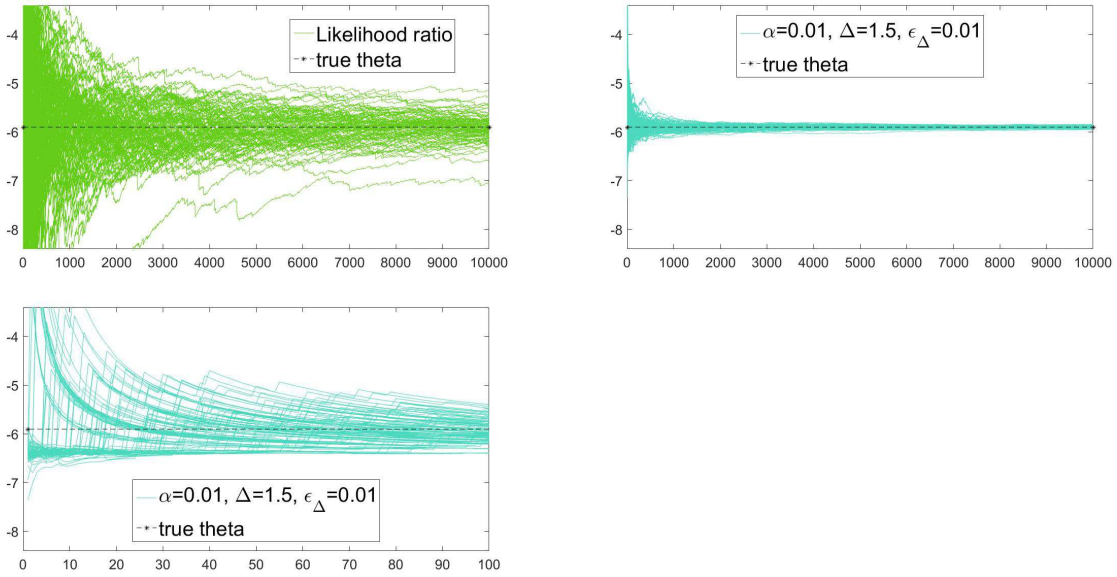

Figure 4.5 Estimating theta of European call option

Therefore, we can derive the Monte Carlo estimator for the option theta:

$$
\begin{aligned}
-\frac{\partial C}{\partial T} &= -e^{-rT}\,\mathbb{E}\left[(S_T-K)_+\left(-r-\frac{1}{2T}-\zeta(S_T)\left(-\frac{\zeta(S_T)}{2T}-\frac{r-\frac{\sigma^2}{2}}{\sigma\sqrt{T}}\right)\right)\right] \\
&= -e^{-rT}\,\mathbb{E}\left[\left(S_0 e^{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}Z}-K\right)_+\left(\frac{Z^2-1}{2T}+\frac{r-\frac{\sigma^2}{2}}{\sigma\sqrt{T}}Z-r\right)\right], \quad Z\sim\mathcal{N}(0,1) \\
&\approx -e^{-rT}\,\frac{1}{N}\sum_{i=1}^{N}\left[\left(S_0 e^{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}Z_i}-K\right)_+\left(\frac{Z_i^2-1}{2T}+\frac{r-\frac{\sigma^2}{2}}{\sigma\sqrt{T}}Z_i-r\right)\right], \quad Z_i \overset{\text{iid}}{\sim}\mathcal{N}(0,1).
\end{aligned}
$$

Also, the partition-based estimator for the option theta is

$$
\begin{aligned}
-\frac{\partial C}{\partial T} &= -e^{-rT}\,\mathbb{E}\left[X\left(\frac{\zeta^2(X+K)-1}{2T}+\frac{r-\frac{\sigma^2}{2}}{\sigma\sqrt{T}}\zeta(X+K)-r\right)\frac{f_{S_T}(X+K)}{q_b(X)}\right], \quad X\sim q_b \\
&\approx -e^{-rT}\,\frac{1}{N}\sum_{i=1}^{N}\left[X_i\left(\frac{\zeta^2(X_i+K)-1}{2T}+\frac{r-\frac{\sigma^2}{2}}{\sigma\sqrt{T}}\zeta(X_i+K)-r\right)\frac{f_{S_T}(X_i+K)}{q_b(X_i)}\right], \quad X_i \overset{\text{iid}}{\sim} q_b.
\end{aligned}
$$

The partition-based method is taken with the auto-sampling-once scheme using the polynomial tail with parameters $\alpha=0.01, \Delta=1.5, \epsilon_\Delta=0.01$. Both methods are simulated with $n=100$ scenarios and the number of samples in each scenario is $N=10,000$. The performance of both the likelihood

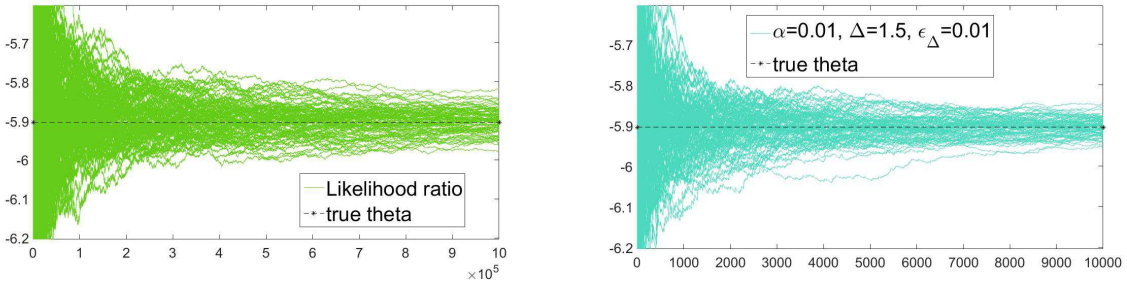

Figure 4.6 Comparison between the likelihood ratio method and the partition-based method for estimating theta of European call option

Table 4.3 Computing time in seconds of the likelihood ratio method with $N = 1,000,000$ and the partition-based method with $\Delta = 1.5, \epsilon_\Delta = 0.01, \xi = 100, \sigma = 100$ and $N = 10,000$ corresponding to Fig. 4.6

	Likelihood ratio with $N = 1,000,000$	$\Delta = 1.5$ with $N = 10,000$
Computing the proposal density	-	0.0042
Sampling	3.1676	1.7728
Calculating the approximation	18.3153	0.1336

ratio method and the partition-based method is given in Fig. 4.5 using the Black-Scholes model with $S_0 = 100, K = 110, r = 0.05, \sigma = 0.2, T = 1$. The theoretical Black-Scholes theta is known to be $-\frac{S_0\sigma\phi(d_1)}{2\sqrt{T}} - rKe^{-rT}\Phi(d_2)$ where ϕ, Φ and d_1 are defined in (4.2), (4.3) and (4.4), respectively, and

$$d_2 = d_1 - \sigma\sqrt{T} = \frac{\log\left(\frac{S_0}{K}\right) + \left(r - \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}. \tag{4.5}$$

For the likelihood ratio method, the time used in sampling step is 0.0310 seconds, and the time used in calculating Monte Carlo approximation is 0.0250 seconds. For the partition-based method, the time used in computing the proposal density is 0.0042 seconds, the time used in sampling step is 1.7728 seconds, and the time used in calculating IS approximation is 0.1336 seconds.

For fairer comparison, we rerun the likelihood ratio method with sample size $N = 1,000,000$ and other parameters fixed. The performance of this simulation is presented in Fig. 4.6 against the previous result from the partition-based method with the same y-axis scale. The time used in the simulation of each method with $n = 100$ scenarios is presented in Table 4.3.

Example 4.4 (Black-Scholes Rho)**.** Rho of an option measures the sensitivity of the option price to

the interest rate, i.e. it is

$$\begin{aligned}\frac{\partial C}{\partial r} = \frac{\partial}{\partial r} C(S_0, r, \sigma, T, K) &= \frac{\partial}{\partial r} \int_0^\infty e^{-rT} (x-K)_+ \, f_{S_T}(x) \, dx \\ &= \int_0^\infty (x-K)_+ \, \frac{\partial}{\partial r} \left(e^{-rT} \, f_{S_T}(x) \right) \, dx.\end{aligned}$$

By simple calculation, we have that

$$\frac{\partial}{\partial r} \left(e^{-rT} \, f_{S_T}(x) \right) = e^{-rT} \left(-T - \zeta(x) \, \frac{\partial \zeta(x)}{\partial r} \right) f_{S_T}(x).$$

with

$$\frac{\partial \zeta(x)}{\partial r} = -\frac{\sqrt{T}}{\sigma}.$$

Therefore, we can derive the Monte Carlo estimator for the option rho:

$$\begin{aligned}\frac{\partial C}{\partial r} &= e^{-rT} \, \mathbb{E}\left[(S_T - K)_+ \left(\frac{\sqrt{T}}{\sigma} \, \zeta(S_T) - T \right) \right] \\ &= e^{-rT} \, \mathbb{E}\left[\left(S_0 e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z} - K \right)_+ \left(\frac{\sqrt{T}}{\sigma} \, Z - T \right) \right], \quad Z \sim \mathcal{N}(0,1) \\ &\approx e^{-rT} \, \frac{1}{N} \sum_{i=1}^N \left[\left(S_0 e^{\left(r - \frac{\sigma^2}{2}\right)T + \sigma\sqrt{T}Z_i} - K \right)_+ \left(\frac{\sqrt{T}}{\sigma} \, Z_i - T \right) \right], \quad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).\end{aligned}$$

Also, the partition-based estimator for the option rho is

$$\begin{aligned}\frac{\partial C}{\partial r} &= e^{-rT} \, \mathbb{E}\left[X \left(\frac{\sqrt{T}}{\sigma} \, \zeta(X+K) - T \right) \frac{f_{S_T}(X+K)}{q_b(X)} \right], \quad X \sim q_b \\ &\approx e^{-rT} \, \frac{1}{N} \sum_{i=1}^N \left[X_i \left(\frac{\sqrt{T}}{\sigma} \, \zeta(X_i+K) - T \right) \frac{f_{S_T}(X_i+K)}{q_b(X_i)} \right], \quad X_i \overset{\text{iid}}{\sim} q_b.\end{aligned}$$

The partition-based method is taken with the auto-sampling-once scheme using the polynomial tail with parameters $\alpha = 0.01, \Delta = 1.5, \epsilon_\Delta = 0.01$. Both methods are simulated with $n = 100$ scenarios and the number of samples in each scenario is $N = 10,000$. The performance of both the likelihood ratio method and the partition-based method is given in Fig. 4.7 using the Black-Scholes model with $S_0 = 100, K = 110, r = 0.05, \sigma = 0.2, T = 1$. The theoretical Black-Scholes rho is known to be $KTe^{-rT}\Phi(d_2)$ where Φ and d_2 are defined in (4.3) and (4.5), respectively. For the likelihood ratio method, the time used in sampling step is 0.0301 seconds, and the time used in calculating Monte Carlo approximation is 0.0243 seconds. For the partition-based method, the time used in computing

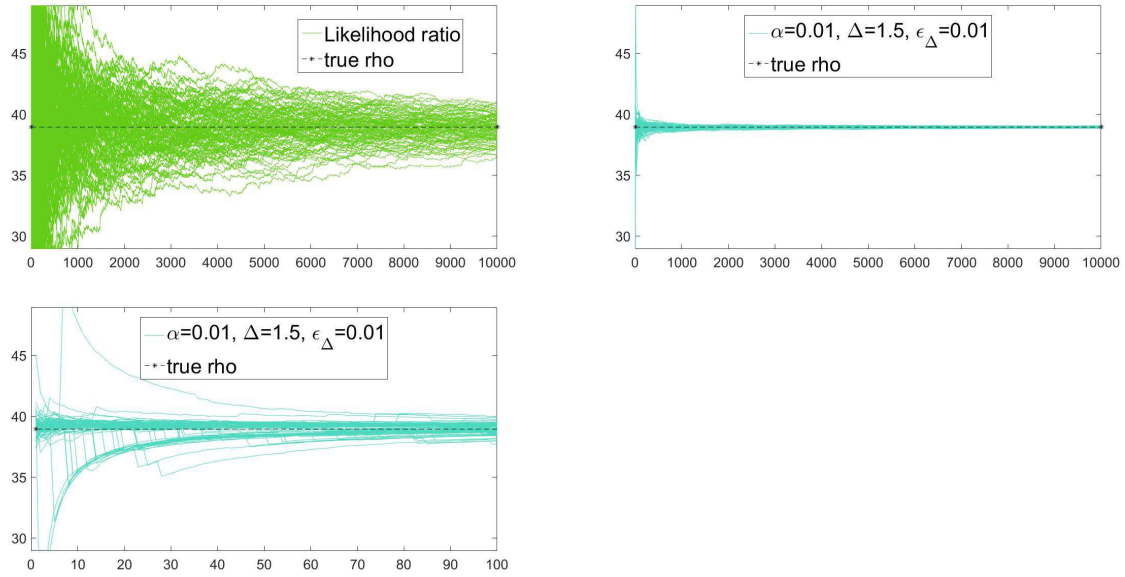

Figure 4.7 Estimating rho of European call option

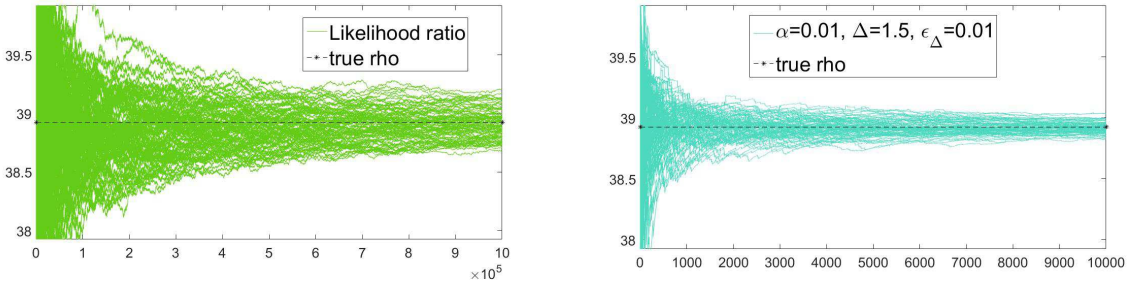

Figure 4.8 Comparison between the likelihood ratio method and the partition-based method for estimating rho of European call option

the proposal density is 0.0081 seconds, the time used in sampling step is 1.7712 seconds, and the time used in calculating IS approximation is 0.1024 seconds.

For fairer comparison, we rerun the likelihood ratio method with sample size $N = 1,000,000$ and other parameters fixed. The performance of this simulation is presented in Fig. 4.8 against the previous result from the partition-based method with the same y-axis scale. The time used in the simulation of each method with $n = 100$ scenarios is presented in Table 4.4.

Example 4.5 (Black-Scholes Gamma)**.** Gamma measures the rate of change in the delta with respect to changes in the underlying asset's price, i.e. it is

Table 4.4 Computing time in seconds of the likelihood ratio method with $N = 1,000,000$ and the partition-based method with $\Delta = 1.5, \epsilon_\Delta = 0.01, \xi = 100, \sigma = 100$ and $N = 10,000$ corresponding to Fig. 4.8

	Likelihood ratio with $N = 1,000,000$	$\Delta = 1.5$ with $N = 10,000$
Computing the proposal density	-	0.0081
Sampling	3.1781	1.7712
Calculating the approximation	17.1659	0.1024

$$\begin{aligned}\frac{\partial^2 C}{\partial S_0^2} = \frac{\partial^2}{\partial S_0^2} C(S_0, r, \sigma, T, K) &= \frac{\partial}{\partial S_0} \int_0^\infty e^{-rT} (x-K)_+ \left(\frac{\zeta(x)}{S_0\sigma\sqrt{T}}\right) f_{S_T}(x)\, dx \\ &= \int_0^\infty e^{-rT} (x-K)_+ \frac{\partial}{\partial S_0} \left(\left(\frac{\zeta(x)}{S_0\sigma\sqrt{T}}\right) f_{S_T}(x)\right) dx.\end{aligned}$$

By simple calculation, we have that

$$\frac{\partial}{\partial S_0}\left(\left(\frac{\zeta(x)}{S_0\sigma\sqrt{T}}\right) f_{S_T}(x)\right) = \left(\frac{\zeta^2(x)-1}{S_0^2\sigma^2 T} - \frac{\zeta(x)}{S_0^2\sigma\sqrt{T}}\right) f_{S_T}(x).$$

Therefore, we can derive the Monte Carlo estimator for the option gamma:

$$\begin{aligned}\frac{\partial^2 C}{\partial S_0^2} &= e^{-rT}\, \mathbb{E}\left[(S_T - K)_+ \left(\frac{\zeta^2(S_T)-1}{S_0^2\sigma^2 T} - \frac{\zeta(S_T)}{S_0^2\sigma\sqrt{T}}\right)\right] \\ &= e^{-rT}\, \mathbb{E}\left[\left(S_0 e^{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}Z} - K\right)_+ \left(\frac{Z^2-1}{S_0^2\sigma^2 T} - \frac{Z}{S_0^2\sigma\sqrt{T}}\right)\right], \quad Z \sim \mathcal{N}(0,1) \\ &\approx e^{-rT}\, \frac{1}{N}\sum_{i=1}^N \left[\left(S_0 e^{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}Z_i} - K\right)_+ \left(\frac{Z_i^2-1}{S_0^2\sigma^2 T} - \frac{Z_i}{S_0^2\sigma\sqrt{T}}\right)\right], \quad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).\end{aligned}$$

Also, the partition-based estimator for the option gamma is

$$\begin{aligned}\frac{\partial^2 C}{\partial S_0^2} &= e^{-rT}\, \mathbb{E}\left[X \left(\frac{\zeta^2(X+K)-1}{S_0^2\sigma^2 T} - \frac{\zeta(X+K)}{S_0^2\sigma\sqrt{T}}\right) \frac{f_{S_T}(X+K)}{q_b(X)}\right], \quad X \sim q_b \\ &\approx e^{-rT}\, \frac{1}{N}\sum_{i=1}^N \left[X_i \left(\frac{\zeta^2(X_i+K)-1}{S_0^2\sigma^2 T} - \frac{\zeta(X_i+K)}{S_0^2\sigma\sqrt{T}}\right) \frac{f_{S_T}(X_i+K)}{q_b(X_i)}\right], \quad X_i \overset{\text{iid}}{\sim} q_b.\end{aligned}$$

The partition-based method is taken with the auto-sampling-once scheme using the polynomial tail with parameters $\alpha = 0.01, \Delta = 1.5, \epsilon_\Delta = 0.01$. Both methods are simulated with $n = 100$ scenarios

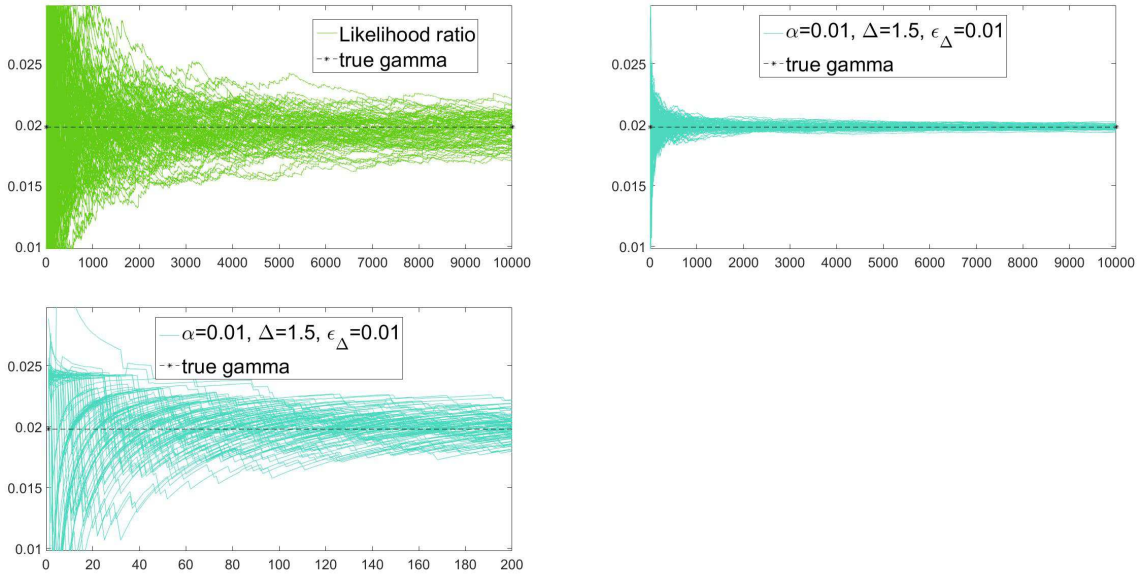

Figure 4.9 Estimating gamma of European call option

and the number of samples in each scenario is $N = 10,000$. The performance of both the likelihood ratio method and the partition-based method is given in Fig. 4.9 using the Black-Scholes model with $S_0 = 100, K = 110, r = 0.05, \sigma = 0.2, T = 1$. The theoretical Black-Scholes gamma is known to be $\frac{\phi(d_1)}{S_0\sigma\sqrt{T}}$ where ϕ and d_2 are defined in (4.2) and (4.5), respectively. For the likelihood ratio method, the time used in sampling step is 0.0368 seconds, and the time used in calculating Monte Carlo approximation is 0.0312 seconds. For the partition-based method, the time used in computing the proposal density is 0.0090 seconds, the time used in sampling step is 1.7872 seconds, and the time used in calculating IS approximation is 0.1619 seconds.

For fairer comparison, we rerun the likelihood ratio method with sample size $N = 500,000$ and other parameters fixed. The performance of this simulation is presented in Fig. 4.10 against the previous result from the partition-based method with the same y-axis scale. The time used in the simulation of each method with $n = 100$ scenarios is presented in Table 4.5.

Note that apart from delta, vega, theta, rho and gamma, which are the most common Greeks, other useful Greeks are vanna $= \frac{\partial^2 C}{\partial S_0 \partial \sigma}$, vomma $= \frac{\partial^2 C}{\partial \sigma^2}$, charm $= -\frac{\partial^2 C}{\partial S_0 \partial T}$, veta $= \frac{\partial^2 C}{\partial \sigma \partial T}$, vera $= \frac{\partial^2 C}{\partial \sigma \partial r}$, speed $= \frac{\partial^3 C}{\partial S_0^3}$, zomma $= \frac{\partial^3 C}{\partial S_0^2 \partial \sigma}$, color $= \frac{\partial^3 C}{\partial S_0^2 \partial T}$ and ultima $= \frac{\partial^3 C}{\partial \sigma^3}$.

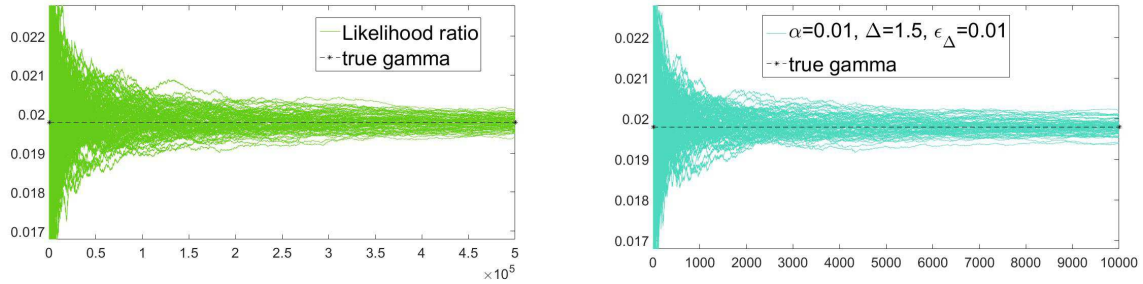

Figure 4.10 Comparison between the likelihood ratio method and the partition-based method for estimating gamma of European call option

Table 4.5 Computing time in seconds of the likelihood ratio method with $N = 500,000$ and the partition-based method with $\Delta = 1.5, \epsilon_\Delta = 0.01, \xi = 100, \sigma = 100$ and $N = 10,000$ corresponding to Fig. 4.10

	Likelihood ratio with $N = 500,000$	$\Delta = 1.5$ with $N = 10,000$
Computing the proposal density	-	0.0090
Sampling	1.4613	1.7872
Calculating the approximation	1.3958	0.1619

4.2 Simultaneous Simulation

For given target functions and distributions, $f_1, f_2, \ldots, f_R$ and $\pi_1, \pi_2, \ldots, \pi_R$, the expectations $\mathbb{E}_{\pi_1}(f_1)$, $\mathbb{E}_{\pi_2}(f_2), \ldots, \mathbb{E}_{\pi_R}(f_R)$ can be approximated using IS method with the same proposal distribution q. Although we can approximate each expectation separately with individual optimal proposal density, we may want to consider using just a single set of samples to approximate all the expectations at once when R is large. For $j = 1, 2, \ldots, R$,

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} w_j(X_i) f_j(X_i), \quad X_i \overset{\text{iid}}{\sim} q. \tag{4.6}$$

where

$$w_j(\cdot) = \frac{\pi_j(\cdot)}{q(\cdot)}. \tag{4.7}$$

The following theorem provides the optimal common proposal density which can be used to approximate a good proposal density by the partition-based method.

Theorem 4.6. *The proposal density q that satisfies $f_j\pi_j \ll q$ for all $j = 1,2,\ldots,R$ and minimizes*

$$\sum_{j=1}^{R} \alpha_j \text{Var}(\hat{\mu}_j)$$

for given constants α_j's such that $\alpha_j > 0$ for all $j = 1,2,\ldots,R$ is

$$q_R^*(x) = \frac{\sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}}{\int \sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}\, dx}.$$

Proof. Minimizing $\sum_{j=1}^{R} \alpha_j \text{Var}(\hat{\mu}_j)$ is equivalent to minimizing

$$\sum_{j=1}^{R} \alpha_j \int \frac{f_j(x)^2 \pi_j(x)^2}{q(x)}\, dx$$

and the minimizing constraint is $\int q(x)\, dx = 1$. We also have constraints q and π_j being non-negative for all $j = 1,2,\ldots,R$. Applying the method of Lagrange multipliers for calculus of variations from Appendix C, we set

$$L(x,q,\lambda) = \sum_{j=1}^{R} \frac{\alpha_j f_j(x)^2 \pi_j(x)^2}{q(x)} + \lambda q(x).$$

Setting

$$\frac{\partial L}{\partial q} = -\sum_{j=1}^{R} \frac{\alpha_j f_j(x)^2 \pi_j(x)^2}{q(x)^2} + \lambda = 0,$$

we have that

$$q(x) = \frac{\sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}}{\sqrt{\lambda}}.$$

Since $\int q(x)\, dx = 1$, we have that $q(x) = \frac{\sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}}{\int \sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}\, dx}$. Now, we will show that this candidate

$$q_R^*(x) = \frac{\sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}}{\int \sqrt{\sum_{j=1}^{R} \alpha_j f_j(x)^2 \pi_j(x)^2}\, dx}$$

indeed yields the minimum variance among all valid basic IS proposal densities. Clearly, $f_j\pi_j \ll q_R^*$ for all $j = 1,2,\ldots,R$. Let q be any density satisfying $f_j\pi_j \ll q$ for all $j = 1,2,\ldots,R$. Then,

$$\begin{aligned}
\sum_{j=1}^{R}\alpha_j\int\frac{f_j(x)^2\pi_j(x)^2}{q_R^*(x)}\,dx &= \left(\int\sqrt{\sum_{j=1}^{R}\alpha_j f_j(x)^2\pi_j(x)^2}\,dx\right)\sum_{j=1}^{R}\alpha_j\int\frac{f_j(x)^2\pi_j(x)^2}{\sqrt{\sum_{i=1}^{R}\alpha_i f_i(x)^2\pi_i(x)^2}}\,dx \\
&= \left(\int\sqrt{\sum_{j=1}^{R}\alpha_j f_j(x)^2\pi_j(x)^2}\,dx\right)^2 \\
&= \left(\int\frac{\sqrt{\sum_{j=1}^{R}\alpha_j f_j(x)^2\pi_j(x)^2}}{q(x)}q(x)\,dx\right)^2 \\
&= \left(\mathbb{E}\left[\frac{\sqrt{\sum_{j=1}^{R}\alpha_j f_j(X)^2\pi_j(X)^2}}{q(X)}\right]\right)^2, \qquad X\sim q.
\end{aligned}$$

By Jensen's inequality, we have that

$$\begin{aligned}
\sum_{j=1}^{R}\alpha_j\int\frac{f_j(x)^2\pi_j(x)^2}{q_R^*(x)}\,dx &\le \mathbb{E}\left[\left(\frac{\sqrt{\sum_{j=1}^{R}\alpha_j f_j(X)^2\pi_j(X)^2}}{q(X)}\right)^2\right] \\
&= \int\frac{\sum_{j=1}^{R}\alpha_j f_j(x)^2\pi_j(x)^2}{q(x)^2}q(x)\,dx \\
&= \sum_{j=1}^{R}\alpha_j\int\frac{f_j(x)^2\pi_j(x)^2}{q(x)}\,dx
\end{aligned}$$

which completes the proof. □

Example 4.7 (Simultaneous Greeks)**.** Consider applying the proposal density using Theorem 4.6 to approximate delta, vega, theta, rho and gamma from Example 4.1, 4.2, 4.3, 4.4, 4.5 with the same parameters $S_0 = 100, K = 110, r = 0.05, \sigma = 0.2, T = 1$. The partition-based method is taken with the auto-sampling-once scheme using the polynomial tail with parameters $\alpha = 0.01, \Delta = 1.5, \epsilon_\Delta = 0.01$. The simultaneous simulation uses $n = 100$, the number of scenarios, and $N = 10,000$, the number of samples in each scenario. The performance of the partition-based method using a common set of samples for each kind of option Greeks is given in Fig. 4.11 with the same scales of Fig. 4.1, 4.3, 4.5, 4.7 and 4.9 for comparison. The simultaneous simulation seems to have the very good performance. The partition-based proposal densities from Example 4.1, 4.2, 4.3, 4.4, 4.5 and this example as well as their corresponding optimal densities are presented in Fig. 4.12. The time used in computing the common proposal density is 0.0586 seconds, the time used in sampling the common samples is

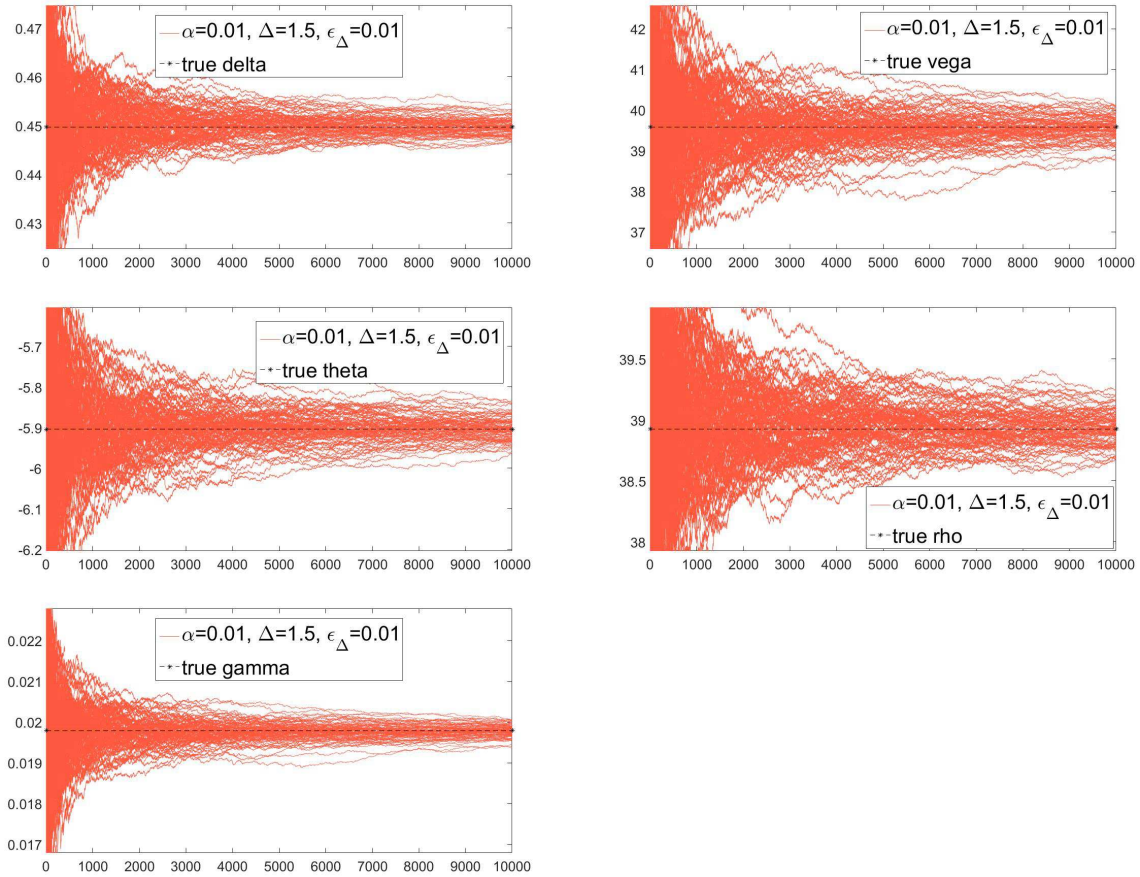

Figure 4.11 Estimating delta, vega, theta, rho and gamma of European call option simultaneously by partition-based method

1.9928 seconds, and the time used in calculating IS approximation for delta, vega, theta, rho and gamma is 0.1404, 0.1621, 0.1338, 0.1016 and 0.1663 seconds, respectively. Since the most consuming time in the approximation is used in the sampling step, simultaneous simulation can save time and memory by sampling just a single set of sample. We can also see that the partition-based method can work well even with the bimodal shape of the optimal proposal densities. It can be easily applied to any multimodal-shape optimal proposal density.

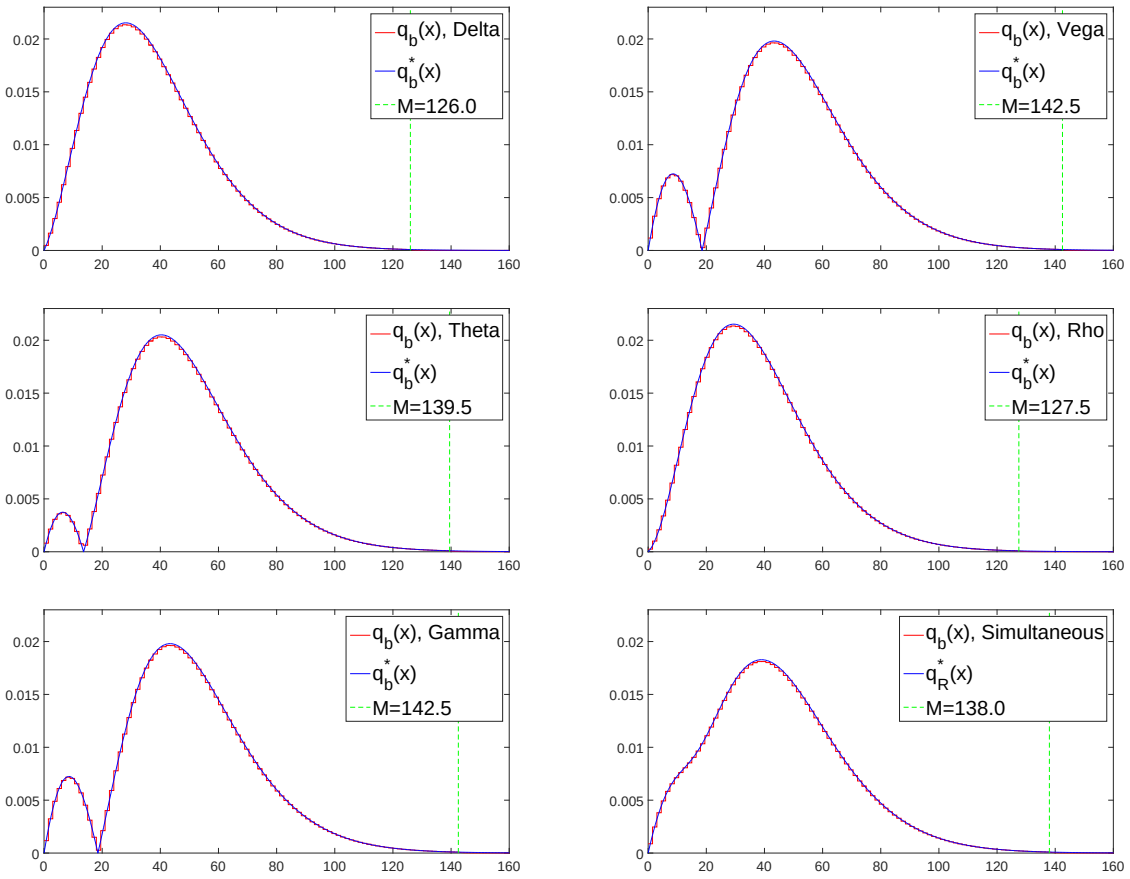

Figure 4.12 Partition-based proposal densities in Example 4.7 and their corresponding optimal densities

CHAPTER

5

SEQUENTIAL IMPORTANCE SAMPLING

This chapter briefly talks about the sequential importance sampling (SIS) method. First, we will introduce the SIS procedure [9] used in general application which uses the self-normalized IS as a base step in each time increment. Then, the SIS method using the basic IS as the base step is introduced. Finally, a possible extension of the partition-based method to the line of SIS is discussed.

5.1 Original Procedure

Let $(E, \mathscr{E})$ be a measurable space and $\{\pi_t\}_{t\in\mathbb{N}}$ a sequence of probability distributions defined on the product space $\{(E^t, \mathscr{E}^t)\}_{t\in\mathbb{N}}$. We denote by $x_{1:t}$ a vector $(x_1, \ldots, x_t) \in E^t$. Note that E can be a multidimensional space, and for that situation each x_i is also a vector and $x_{1:t}$ is a vector of vectors. Assume that π_t is known up to normalizing constants Z_t:

$$\pi_t(x_{1:t}) = \frac{p_t(x_{1:t})}{Z_t}.$$

We want to approximate expectations of π_t-integrable function $f_t : E^t \to \mathbb{R}$:

$$\mu_t = \mathbb{E}_{\pi_t}(f_t) = \int f_t(x_{1:t})\pi_t(x_{1:t})d\,x_{1:t}$$

as well as the normalizing constant Z_t.

At time $t = 1$, we apply self-normalized IS method with a chosen valid proposal density q_1.

$$\tilde{\mu}_1 = \sum_{i=1}^{N} W_1^{(i)} f_1(X_1^{(i)}) \quad \text{and} \quad \tilde{Z}_1 = \frac{1}{N} \sum_{i=1}^{N} w_1(X_1^{(i)}), \qquad X_1^{(i)} \sim q_1$$

where

$$W_1^{(i)} = \frac{w_1(X_1^{(i)})}{\sum_{i=1}^{N} w_1(X_1^{(i)})} \quad \text{and} \quad w_1(X_1^{(i)}) = \frac{p_1(X_1^{(i)})}{q_1(X_1^{(i)})}.$$

For time $t \geq 2$, we can just apply IS method on the space E^t with a valid proposal density $q_t(x_{1:t})$. Doing this, we will have, for each time t, new N t-dimensional-vectors $X_{1:t}^{(i)}$ drawn from q_t and their corresponding weights $W_t^{(i)}$. One may want to reuse the previous vectors $X_{1:t-1}^{(i)}$ and their weights to recursively compute $X_{1:t}^{(i)}$ and $W_t^{(i)}$. To do this, we need one more component to add on to the on-hand vectors $X_{1:t-1}^{(i)}$. So, instead of sampling $X_{1:t}^{(i)} \sim q_t(x_{1:t})$, we conditionally draw

$$X_t^{(i)} | X_{1:t-1}^{(i)} \sim q_t(x_t | x_{1:t-1})$$

where

$$\begin{aligned} q_t(x_{1:t}) &= q_{t-1}(x_{1:t-1})\, q_t(x_t | x_{1:t-1}) \\ &= q_1(x_1)\, q_2(x_2|x_1) \ldots\, q_t(x_t | x_{1:t-1}). \end{aligned}$$

Thus, we have to select $q_t(x_t | x_{1:t-1})$ instead of choosing importance sampling density $q_t(x_{1:t})$ in order to be able to reuse the past simulated trajectories. To calculate the corresponding weights, we simply do some algebra:

$$\begin{aligned} w_t(x_{1:t}) &= \frac{p_t(x_{1:t})}{q_t(x_{1:t})} \\ &= \frac{p_t(x_{1:t})}{q_{t-1}(x_{1:t-1})\, q_t(x_t | x_{1:t-1})} \\ &= \frac{p_{t-1}(x_{1:t-1})}{q_{t-1}(x_{1:t-1})} \frac{p_t(x_{1:t})}{p_{t-1}(x_{1:t-1})\, q_t(x_t | x_{1:t-1})} \\ &= w_{t-1}(x_{1:t-1}) \frac{p_t(x_{1:t})}{p_{t-1}(x_{1:t-1})\, q_t(x_t | x_{1:t-1})}. \end{aligned}$$

Define

$$\alpha_1(x_1) = w_1(x_1) = \frac{p_1(x_1)}{q_1(x_1)}$$

and for $t \geq 2$

$$\alpha_t(x_{1:t}) = \frac{p_t(x_{1:t})}{p_{t-1}(x_{1:t-1})\, q_t(x_t|x_{1:t-1})}.$$

Then, for $t \geq 2$

$$\tilde{\mu}_t = \sum_{i=1}^{N} W_t^{(i)} f_t(X_{1:t}^{(i)}) \quad \text{and} \quad \tilde{Z}_t = \frac{1}{N}\sum_{i=1}^{N} w_t(X_{1:t}^{(i)}), \qquad X_t^{(i)} \sim q_t(\,\cdot\,|X_{1:t-1}^{(i)})$$

where

$$W_t^{(i)} = \frac{w_t(X_{1:t}^{(i)})}{\sum_{i=1}^{N} w_t(X_{1:t}^{(i)})}$$

and

$$w_t(X_{1:t}^{(i)}) = w_{t-1}(X_{1:t-1}^{(i)})\, \alpha_t(X_{1:t}^{(i)}).$$

This SIS method has been developed primarily from the field of filtering estimation and Bayesian analysis [9]. It uses the self-normalized IS as a base step in each time increment. As a result, empirical distributions representing target distributions in each time step can be studied through McKean interpretations of Feynman-Kac models [5]. The nonuniqueness of McKean interpretations of Feynman-Kac models can lead correspondingly to interacting particle system (IPS) schemes, particularly a more specific scheme called SIS with resampling (SISR) scheme [9] which is practically and widely used. With the strong support theory of this IPS model, This original SIS method and SISR are typically used in many applications even when the target density is fully known, i.e. the normalizing constant Z_i is known. Instead of using self-normalized IS which brings biased estimators, we should consider using basic IS as a base step for SIS.

5.2 Alternative Procedure

We would like to introduce the SIS procedure using basic IS as the base step which should work better than the original SIS when the target density π_t is known for all t.

At time $t = 1$, we just apply basic IS method with a chosen valid proposal density q_1.

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} w_1(X_1^{(i)}) f_1(X_1^{(i)}), \qquad X_1^{(i)} \sim q_1$$

where

$$w_1(X_1^{(i)}) = \frac{\pi_1(X_1^{(i)})}{q_1(X_1^{(i)})}.$$

For time $t \geq 2$, we conditionally draw

$$X_t^{(i)} | X_{1:t-1}^{(i)} \sim q_t(x_t | x_{1:t-1})$$

where

$$\begin{aligned} q_t(x_{1:t}) &= q_{t-1}(x_{1:t-1})\, q_t(x_t | x_{1:t-1}) \\ &= q_1(x_1)\, q_2(x_2|x_1) \,\ldots\, q_t(x_t|x_{1:t-1}). \end{aligned}$$

The weight function

$$\begin{aligned} w_t(x_{1:t}) &= \frac{\pi_t(x_{1:t})}{q_t(x_{1:t})} \\ &= w_{t-1}(x_{1:t-1})\, \frac{\pi_t(x_{1:t})}{\pi_{t-1}(x_{1:t-1})\, q_t(x_t|x_{1:t-1})} \end{aligned}$$

can be used sequentially by defining

$$\beta_1(x_1) = w_1(x_1) = \frac{\pi_1(x_1)}{q_1(x_1)}$$

and for $t \geq 2$

$$\beta_t(x_{1:t}) = \frac{\pi_t(x_{1:t})}{\pi_{t-1}(x_{1:t-1})\, q_t(x_t|x_{1:t-1})}.$$

Then, for $t \geq 2$

$$\hat{\mu}_t = \frac{1}{N} \sum_{i=1}^{N} w_t(X_{1:t}^{(i)}) f_t(X_{1:t}^{(i)}), \qquad X_t^{(i)} \sim q_t(\,\cdot\,|X_{1:t-1}^{(i)})$$

where

$$w_t(X_{1:t}^{(i)}) = w_{t-1}(X_{1:t-1}^{(i)})\, \beta_t(X_{1:t}^{(i)}).$$

Unlike the original SIS method, this proposed SIS procedure does not have empirical distributions representing the target distributions in each time step, so it may be hard to develop the associated IPS theory. This is definitely a good future work to explore. However, this SIS can be developed to SISR scheme where the resampling is performed for every time step using weighted empirical measure defined by the resampling weights to establish the associated IPS theory [4, 6]. Actually, that developed scheme is equivalent to the original SIS with resampling in every time step though.

5.3 Partition-Based Method

Recall from Section 5.1 and 5.2 that we must choose the proposal densities satisfying

$$q_t(x_{1:t}) = q_{t-1}(x_{1:t-1})\, q_t(x_t|x_{1:t-1}). \tag{5.1}$$

For the SIS method using basic IS as described in Section 5.2, we can use the partition-based method to obtain the proposal density q_1 approximating the optimal density $(q_1)_b^*$

$$q_1(x_1) \approx (q_1)_b^*(x_1) = \frac{|f_1(x_1)|\, \pi_1(x_1)}{\int |f_1(x_1)|\, \pi_1(x_1)\, d\, x_1}.$$

For $t \geq 2$, we would like to acquire

$$q_t(x_{1:t}) \approx (q_t)_b^*(x_{1:t}) = \frac{|f_t(x_{1:t})|\, \pi_t(x_{1:t})}{\int |f_t(x_{1:t})|\, \pi_t(x_{1:t})\, d\, x_{1:t}}. \tag{5.2}$$

Suppose we use the partition-based method to obtain $q_1(x_1)$ and $q_2(x_{1:2})$. There is no guarantee that q_1 and q_2 can have the relation (5.1). In other words, $\int q_2(x_{1:2})\, d\, x_2$ may not equal $q_1(x_1)$. However, in a special case where we are interested only at the final time step, we can create the sequence of q_t satisfying (5.1). Unfortunately, it may require more work than directly applying a partition-based method on the terminal time space.

Suppose we consider approximating $\mathbb{E}_{\pi_m}(f_m) = \int f_m(x_{1:m})\pi_m(x_{1:m})\, d\, x_{1:m}$. We can use the partition-based method in Section 3.4 to get $q_m(x_{1:m}) \approx (q_m)_b^*(x_{1:m})$. The proposal density q_m is a function on E^m which is partitioned into many subcells. We can have the probabilities of getting a sample in each subcell which can be stored in a multidimensional matrix. Thus, we can easily sum it up in proper dimensions to get a way to conditionally draw each component of a sample from time 1 to m sequentially. The sampling step will finally use m uniform random numbers for one sample in E^m. Each uniform random number will create each component x_i of the sample $x_{1:m}$. Recall that applying partition-based IS directly at time m will use $m+1$ uniform random numbers for one sample in E^m, so the sequential partition-based scheme use one less uniform random number. However, it require more work in the sequential procedure. It may depend on the problem which method is better, but we suggest to use the partition-based IS directly at time m.

CHAPTER

6

CONCLUSION

IS is a useful Monte Carlo technique that can attack some problems that the classical Monte Carlo method or other techniques cannot be applied to. These problems include rare-event applications and situations when sampling directly from the target function is difficult especially in Bayesian Analysis. Both basic IS and self-normalized IS allow us to use samples that better locate in the rare-event region, or are easier to draw. Moreover, IS can be beneficial in estimating results under multiple target distributions simultaneously. Furthermore, it can be a useful tool to estimate derivatives of expectations. Sequential Monte Carlo simulation also uses IS as a key step to draw each component of a sample sequentially. In addition, self-normalized IS can be used to simulate the target distribution by getting a set of samples together with their associated weights summing up to one.

IS is a well-known variance reduction method. In spite of its high complexity, it is possibly the best variance reduction method among the known approaches. For a single-signed target function, basic IS can have the zero-variance estimator. To get a great variance reduction, it depends on how well we choose a proposal distribution for sampling. Both kinds of IS have the optimal proposal densities. Unfortunately, the optimal densities cannot be put to use in general problems due to the need of knowing the answer at first. Most people stick to using known distributions for the proposal distributions. There is a rule of thumb (1.7) in choosing proposal densities that is widely used. However, that criterion neglects information from the target functions, and may cause failure in

approximation for some target functions. In addition, the approximation can be at best only among the chosen family of distributions. In this work, we propose a way to get a proposal distribution that is not a known distribution. There are five primary contributions in this work.

The first contribution is to summarize basic mathematical theory for IS method including the convergence of the estimators as well as the optimal proposal distributions for both basic and self-normalized IS. The proofs of the convergence theorem of a self-normalized IS estimator and the optimal proposal distribution for self-normalized IS are provided.

The second contribution is to point out a blind spot of the current rule of thumb (1.7) for choosing a proposal density in IS method. A counter example that really shows the failure of that rule of thumb is given. From the summarized theory, proper criteria for proposal distributions are summed up.

The third contribution is to provide a partition-based method that utilizes the information of optimal proposal densities. A class of functions that can be covered by this method is discussed. An outcome proposal density will be close to the optimal proposal density. Thus, it yields the finite-variance IS estimator, and satisfies the validity to be a proposal density as well as the assumption for the convergence theorem leading to CLT. The concept of the partition-based method is to break the domain into pieces and try to approximate the optimal density on each piece of domain. The method still depends on choices of parameters in the method. Criteria to partition the domain are discussed. For basic IS, the process to determine where to cut for the tail region is provided. The partition-based idea also extends to the case of multidimensional spaces. Some practical examples using the partition-based method are provided, and the efficiency of the partition-based method seems very good. Computing time may be high, but we can reduce the number of samples much lower than using other methods and still secure the same level of accuracy.

The fourth contribution is to provide an optimal density for the simultaneous simulation. Here, the optimality is in the sense of a linear combination of estimators' variances. The statement and proof are given.

The fifth contribution is to provide the SIS method using the basic IS as the base step when all target densities are known. It does not have the supported IPS theory like the original SIS method which uses the self-normalized IS as the base step. A possible extension of the partition-based method to the SIS procedure is discussed. However, the partition-based method is actually designed specifically to a pair of target function and target density. It is not suitable for the SIS method but should rather be applied directly for a pair of target function and target density.

The partition-based method does not work for all the Monte Carlo problems. The obvious drawback of the partition-based method is that the explicit forms of the target function and the target density are needed. We may deal with this problem in some situations with some cost. For

example, when the target function or the target density are defined through an integral, we can increase the dimension of the problem by adding the integrator variables into the problem. However, this issue is at the same time an advantage of this method. The partition-based method can be applied to multimodal distributions or weird functions that the other methods in the line of IS may have
high dimensional problem because of the cost in computing the proposal density and sampling. The partition-based method in this work can be studied further. The sensitivity of parameters in the method is interesting. We may improve the criteria on how to partition the domain. Changing the ordinary rectangular partition to polar-coordinate partition or other system may improve the performance of the estimation.

IS is capable of getting the zero-variance estimation and the optimal proposal density is known, so we should utilize this fact. This work tries to best approximate the expectation by first finding the optimal proposal density and then approximating it. Considering the proof of an optimal proposal density for basic IS, self-normalized IS, and simultaneous simulation of IS, we can have a pattern to find an optimal proposal density for a given IS scheme which may come out in the future. Approximating the optimal proposal density, the partition-based method proposed in this work should be able to adapt to any reasonable pair of target function and target density, and it allows IS to be fruitfully applied in more general setting of applications.

REFERENCES

[1] BECKMAN, R. J., AND MCKAY, M. D. Monte Carlo Estimation under Different Distributions Using the Same Simulation. *Technometrics 29*, 2 (1987), 153–160.

[2] BLANCHET, J., AND LAM, H. State-Dependent Importance Sampling for Rare-Event Simulation: An Overview and Recent Advances. *Surveys in Operations Research and Management Science 17* (2012), 38–59.

[3] BROADIE, M., AND GLASSERMAN, P. Estimating Security Price Derivatives Using Simulation. *Management Science 42*, 2 (1996), 269–285.

[4] CHAN, H. P., AND LAI, T. L. A Sequential Monte Carlo Approach to Computing Tail Probabilities in Stochastic Models. *The Annals of Applied Probability 21*, 6 (2011), 2315–2342.

[5] DEL MORAL, P. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag New York, 2004.

[6] DEL MORAL, P., AND GARNIER, J. Genealogical Particle Analysis of Rare Events. *The Annals of Applied Probability 15*, 4 (2005), 2496–2534.

[7] DEL MORAL, P., JASRA, A., AND DOUCET, A. Sequential Monte Carlo Samplers. *Journal Of The Royal Statistical Society. Series B (Statistical Methodology) 68*, 3 (2006), 411–436.

[8] DETEMPLE, J., AND RINDISBACHER, M. Monte Carlo Methods for Derivatives of Options with Discontinuous Payoffs. *Computational Statistics & Data Analysis 51* (2007), 3393–3417.

[9] DOUCET, A., DE FREITAS, N., AND GORDON, N. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag New York, 2001.

[10] DOUCET, A., GODSILL, S., AND ANDRIEU, C. On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and Computing 10*, 3 (2000), 197–208.

[11] GEWEKE, J. Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica 57*, 6 (1989), 1317–1339.

[12] GLASSERMAN, P. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag New York, 2003.

[13] GLYNN, P. W. Stochastic Approximation for Monte Carlo Optimization. *Proceedings of the 1986 Winter Simulation Conference* (1986), 356–364.

[14] HEIDELBERGER, P. Fast Simulation of Rare Events in Queueing and Reliability Models. *Acm Transactions on Modeling and Computer Simulation 5*, 1 (1995), 43–85.

[15] HULL, J. C. *Options, Futures, and Other Derivatives*, 7th ed. Prentice-Hall New Jersey, 2008.

[16] IONIDES, E. L. Truncated Importance Sampling. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 295–311.

[17] KAHN, H. Random Sampling (Monte Carlo) Techniques in Neutron Attenuation Problems - I. *Nuclenonics 6*, 5 (1950), 27–37.

[18] KAHN, H. Random Sampling (Monte Carlo) Techniques in Neutron Attenuation Problems - II. *Nuclenonics 6*, 6 (1950), 60–65.

[19] KAHN, H., AND MARSHALL, A. W. Methods of Reducing Sample Size in Monte Carlo Computations. *Journal of the Operations Research Society of America 1*, 5 (1953), 263–278.

[20] KLEPPE, T. S., AND LIESENFELD, R. Efficient Importance Sampling in Mixture Frameworks. *Computational Statistics and Data Analysis 76* (2014), 449–463.

[21] KLOEK, T., AND VAN DIJK, H. K. Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo. *Econometrica 46*, 1 (1978), 1–19.

[22] LEHMANN, E. L., AND CASELLA, G. *Theory of Point Estimation*, 2nd ed. Springer-Verlag New York, 1998.

[23] LIESENFELD, R., AND RICHARD, J.-F. Monte Carlo Methods and Bayesian Computation: Importance Sampling. *International Encyclopedia of the Social & Behavioral Sciences* (2001), 10000âĂŞ10004.

[24] O'NEILL, B. Importance Sampling for Bayesian Sensitivity Analysis. *International Journal of Approximate Reasoning 50*, 2 (2009), 270–278.

[25] OWEN, A. B. Monte Carlo Theory, Methods and Examples. unpublished book, 2013.

[26] REIMAN, M. I., AND WEISS, A. Sensitvity Analysis Via Likelihood Ratios. *Proceedings of the 1986 Winter Simulation Conference* (1986), 285–289.

[27] REIMAN, M. I., AND WEISS, A. Sensitivity Analysis for Simulations via Likelihood Ratios. *Operations Research 37*, 5 (1989), 830–844.

[28] RICHARD, J.-F., AND ZHANG, W. Efficient High-Dimensional Monte Carlo Importance Sampling. *Mimeo, University of Pittsburgh, PA* (1998).

[29] RICHARD, J.-F., AND ZHANG, W. Efficient High-Dimensional Importance Sampling. *Journal of Econometrics 141* (2007), 1385–1411.

[30] RUBINSTEIN, R. Y. *Simulation and the Monte Carlo Method*. John Wiley & Sons, 1981.

[31] RUBINSTEIN, R. Y., AND KROESE, D. P. *Simulation and the Monte Carlo Method*, 2nd ed. John Wiley & Sons, 2008.

[32] SRINIVASAN, R. Importance Sampling - the Simulation Theory of Rare Events and its Applications. *Defence Science Journal 49*, 1 (1999), 9–17.

[33] STEWART, L. Multiparameter Univariate Bayesian Analysis. *Journal of American Statistical Association 74*, 367 (1979), 684–693.

[34] TOKDAR, S. T., AND KASS, R. E. Importance Sampling: A Review. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 1 (2010), 54–60.

[35] TUKEY, J. W. Configural Polysampling. *SIAM Review 29*, 1 (1987), 1–20.

[36] YUAN, C., AND DRUZDZEL, M. J. Theoretical Analysis and Practical Insights on Importance Sampling in Bayesian Networks. *International Journal of Approximate Reasoning 46*, 2 (2007), 320–333.

APPENDICES

APPENDIX

A

ACCEPTANCE-REJECTION METHOD

Let f and g be probability densities such that there exists a constant $c > 0$ with $f(x) \leq c\,g(x)$ for all x such that $f(x) > 0$. We can generate a random number from the density f as follows:

1. Generate $X \sim g$.
2. Generate $U \sim \text{Unif}(0,1)$, the uniform distribution over the unit interval.
3. Check whether or not $U \leq \frac{f(X)}{c\,g(X)}$.
    - If this holds, accept and return X.
    - If not, reject X and go back to step 1.

APPENDIX

B

DELTA METHOD

Theorem B.1 (The Multivariate Delta Method [22])**.** *Let* $\vec{X}_n = (X_n^{(1)}, \ldots, X_n^{(k)})$ *be a sequence of random vectors such that*

$$\sqrt{n}\left(\vec{X}_n - \vec{\mu}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}_k\left(\vec{0}, \Sigma\right).$$

Let $g : D \subseteq \mathbb{R}^k \to \mathbb{R}$ *be defined and has continuous first partial derivatives in a neighborhood of* $\vec{\mu}$*. If the elements of* $\nabla_{\vec{\mu}} = \left(\frac{\partial g}{\partial x_1}(\vec{x}), \ldots, \frac{\partial g}{\partial x_k}(\vec{x})\right)\Big|_{\vec{x}=\vec{\mu}}$ *are nonzero, then*

$$\sqrt{n}\left(g(\vec{X}_n) - g(\vec{\mu})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \nabla_{\vec{\mu}}\, \Sigma\, \nabla_{\vec{\mu}}^T\right).$$

APPENDIX

C

CALCULUS OF VARIATIONS

We can extend the method of Lagrange multipliers to calculus of variations. To solve an isoperimetric problem

$$\text{Optimize}: I(y(x)) = \int_{x_0}^{x_1} F(x, y, y')dx$$
$$\text{Subject to}: \int_{x_0}^{x_1} G(x, y, y')dx = J$$

where J is a known constant, we set the Lagrangian function

$$L(x, y, y', \lambda) = F(x, y, y') + \lambda G(x, y, y')$$

and the following unconstrained Euler equation is solved along with the constraint equation

$$\frac{d}{dx}\left(\frac{\partial L}{\partial y'}\right) - \frac{\partial L}{\partial y} = 0.$$

APPENDIX

D

INVERSE TRANSFORM METHOD

Let F be a continuous cumulative distribution function and F^{-1} its inverse function defined by $F^{-1}(u) = \inf\{x|F(x) \geq u\}$. We can generate a random number distributed according to the distribution described by F as follows:

1. Generate $U \sim \text{Unif}(0,1)$, the uniform distribution over the unit interval.
2. Compute and return $X = F^{-1}(U)$.

APPENDIX

E

GENERALIZED PARETO DISTRIBUTION

The probability density function for the generalized Pareto distribution GPD(ξ, σ, θ) with shape parameter $\xi \neq 0$, scale parameter $\sigma > 0$, and location parameter θ is

$$f(x|\xi,\sigma,\theta) = \frac{1}{\sigma\left(1+\xi\frac{(x-\theta)}{\sigma}\right)^{1+\frac{1}{\xi}}}$$

for $x > \theta$ when $\xi > 0$, or for $\theta < x < \theta - \frac{\sigma}{\xi}$ when $\xi < 0$.

For $\xi = 0$, the density is

$$f(x|\xi,\sigma,\theta) = \frac{1}{\sigma}e^{-\frac{(x-\theta)}{\sigma}}$$

for $x > \theta$.

If $\xi = 0$ and $\theta = 0$, the generalized Pareto distribution is equivalent to the exponential distribution. If $\xi > 0$ and $\theta = \frac{\sigma}{\xi}$, the generalized Pareto distribution is equivalent to the Pareto distribution.