ESTIMATING THE DENSITY OF A CONDITIONAL EXPECTATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Samuel George Steckley

January 2006

ESTIMATING THE DENSITY OF A CONDITIONAL EXPECTATION

Samuel George Steckley, Ph.D.

Cornell University 2006

Given uncertainty in the input model and parameters of a stochastic simulation study, the goal of the study often becomes the estimation of a conditional expectation. The conditional expectation is expected performance conditioned on the selected model and parameters. The distribution of this conditional expectation describes precisely, and concisely, the impact of input uncertainty on performance prediction. In this thesis we estimate the density of a conditional expectation using ideas from the field of kernel density estimation. We show that our estimator converges under reasonable conditions and present results on optimal rates of convergence. We present two modifications of this estimator, a local estimator and a bias-corrected estimator. Convergence results are given for these estimators. We study the performance of our estimators on a number of test cases.

We also study the problem of computing performance for a service system in which there is input uncertainty. It is commonly assumed that the arrival process of customers to a service system is a nonhomogeneous Poisson process. We consider the case in which the rate function for the Poisson process is unknown. We also consider a related problem in which the rate function for the Poisson process varies from day to day. For each of these problems, we develop steady-state approximations for both the long-run fraction of calls answered quickly, and the distribution of the fraction of calls answered quickly within a short period. We also describe the corresponding simulation-based estimates. We perform a computational study

to evaluate the approximations and simulation-based estimates and improve our understanding of such systems.

BIOGRAPHICAL SKETCH

Samuel G. Steckley was born and raised in Galesburg, Illinois. In 1999 he received his Bachelor's degree in Economics from Oberlin College. He entered the doctoral program in Operations Research at Cornell University in 2000. He graduated with a Ph.D. in Operations Research in 2005, majoring in Applied Probability.

ACKNOWLEDGEMENTS

I am truly grateful to my advisor Shane Henderson for his excellent guidance, remarkable patience, and generosity. He has been and will continue to be a model for me in many ways.

I would like to thank David Ruppert for his invaluable contributions to this thesis and his kind advise. And I would like to thank Robert G. Bland, for his encouraging and helpful suggestions throughout my stay here at Cornell. I am also appreciative of the many contributions Vijay Mehrotra made to Chapter 4 of this thesis.

For the splendid education and support, I would like to thank the Operations Research department faculty, staff, and graduate students. And to my family and friends outside of the department, I am deeply grateful for your unwavering support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Problem Motivation

In this thesis, we study the estimation of the density of a conditional expectation and apply this work in part to the problem of computing service system performance. Let X be a real-valued random variable with $\mathrm{E}(|X|) < \infty$. Let Z be some other random object. The object Z could be a real-valued random variable, a random vector, or even a sample path of a stochastic process. The conditional expectation $\mathrm{E}(X|Z)$ is a random variable that represents the average value of X given only the value of the random variable Z. Then $\mathrm{E}(X|Z)$ can be thought of as one's best guess as to the value of X knowing only the value of Z. In this thesis we assume that the random variable $\mathrm{E}(X|Z)$ has a density with respect to Lebesgue measure and develop a method for estimating it. Our main assumptions are that

1. we can generate i.i.d. replicates of the random object Z, and

2. we can generate i.i.d. observations from the conditional distribution

$$P(X \in \cdot \mid Z = z)$$

   for any z in the range of Z.

Our primary motivation for studying this problem stems from the issue of input model uncertainty in stochastic simulation. This form of uncertainty arises when one is not completely certain what input distributions and associated parameters should be used in a stochastic simulation model. There are many methods for dealing with such uncertainty; see Henderson [2003] for a review. Many of these methods impose a probability distribution on the unknown input distributions and parameters. For example, this is the case in the papers Cheng [1994], Cheng and Holland [1997, 1998, 2003], Chick [2001], Zouaoui and Wilson [2003, 2004]. See Henderson [2003] for further discussion.

The input model uncertainty problem maps to the setting in this thesis as follows. The random object Z corresponds to a selection of input distributions and associated parameters for a simulation experiment. The random variable X corresponds to output from the stochastic simulation model such that performance is measured by $\mathrm{E}(X)$. The distribution of X is dependent on the choice Z of input distributions and parameters. The conditional expectation $\mathrm{E}(X|Z)$ represents the performance measure as a function of the input distributions and parameters. Notice that it is still a random variable due to the uncertainty in the input distributions and parameters. A density of $\mathrm{E}(X|Z)$ concisely captures the distribution of performance given the uncertainty in the inputs. It gives a sense of the uncertainty in the performance measure due to the uncertainty in the values of the input distributions and parameters.

For example, consider a service system such as a call center in which the arrival process of calls over the course of a day is thought to be a nonhomogeneous Poisson process (NHPP) with time-dependent rate function $\lambda = (\lambda(t) : t \geq 0)$. Let us focus on a single period (e.g., 10am-10:15am) in the day and assume that the rate function takes on the constant value λ^* in that period. Suppose we are uncertain of the arrival rate λ^* and we can model our uncertainty in terms of a random variable Λ. For reasons explained in Chapter 4, one particular performance measure of interest is $\mathrm{E}S/\mathrm{E}N$ where S is the number of calls in the period that wait less than some prescribed time before being answered and N is the total number of calls in the period. For a given arrival rate λ, $\mathrm{E}N = \lambda t$ where t is the length of the period. Then the performance measure for a given λ is

$$\frac{\mathrm{E}S}{\mathrm{E}N} = \mathrm{E}\left(\frac{S}{\lambda t}\right).$$

It follows that the performance measure given the uncertainty in the arrival rate modeled by Λ is

$$\mathrm{E}\left[\frac{S}{\Lambda t}\middle|\Lambda\right]. \tag{1.1}$$

The distribution of $S/\Lambda t$ is dependent on the arrival rate. If we generate observations of $S/(\Lambda t)$ for a given realization of the arrival rate Λ via simulation, then this is precisely an example of input model uncertainty in stochastic simulation described above.

The density of the random variable in (1.1) captures the distribution of the performance in the period given our uncertainty in the arrival rate. It gives general insight into how the uncertainty in the arrival rate translates into uncertainty about performance. Of particular interest to the call center manager, the density captures the effect of uncertainty on the risk of poor performance.

In this thesis, we study this problem as well as a related problem in which the arrival rate in the period is randomly varying from day to day. Though these problems are related in that they both have a kind of uncertainty in the arrival process, we stress they are different and require different approaches. We will discuss this problem further in Section 1.3.

Little work has been done on the estimation of the distribution of a conditional expectation. The most closely related work to ours involves the estimation of the *distribution function* of the conditional expectation $\mathrm{E}(X|Z)$. Lee and Glynn [1999] considered the case where Z is a discrete random variable. This work was an outgrowth of Chapter 2 of Lee [1998], where the case where Z is continuous is also considered. We prefer to directly estimate the density because we believe that the density is more easily interpreted (visually) than a distribution function. Steckley and Henderson [2003] estimated the density of $\mathrm{E}(X|Z)$ for the case in which Z is univariate and expected performance is monotone in Z. In this thesis, results are presented for a case that applies to a multivariate Z. The results presented here are important in the context of the input uncertainty problem since it is rarely the case that input uncertainty is restricted to only one parameter.

Andradóttir and Glynn [2003] discuss a certain estimation problem that, in our setting, is essentially the estimation of EX. Their problem is complicated by the fact that they explicitly allow for bias in the estimator. Such bias can arise in steady-state simulation experiments, for example.

1.2 Kernel Density Estimation

In this thesis we apply kernel smoothing methods to the estimation of the density of the conditional expectation. There has been a great deal of work done on kernel density estimation beginning with the seminal papers of Rosenblatt [1956] and Parzen [1962]. See Wand and Jones [1995] for an introduction and review of the subject. The standard setting for kernel density estimation is as follows: Suppose Y is a random variable with an unknown density g and $(Y_i : 1 \leq i \leq n)$ is a sequence of i.i.d. copies of the random variable Y. The standard kernel density estimator, sometimes called the "naive" estimator, is

$$\hat{g}(x;h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-Y_i}{h}\right), \tag{1.2}$$

where the kernel K is typically chosen to be a unimodal probability density function (p.d.f.) that is symmetric about zero, and the smoothing parameter h, often referred to as the bandwidth, is a positive number (e.g., Wand and Jones [1995]).

For all $x \in \mathbb{R}$, define the mean squared error (mse) of the estimator evaluated at x as

$$\text{mse}(\hat{g}(x;h)) = \text{E}\big(\hat{g}(x;h) - g(x)\big)^2.$$

Define the mean integrated squared error (mise) of the estimator as

$$\text{mise}(\hat{g}(\cdot,h)) = \text{E}\int \big(\hat{g}(x,h) - g(x)\big)^2 \, \text{d}x.$$

It is well known that if g is continuous, the naive estimator is consistent in quadratic mean. That is to say, $\text{mse}(\hat{g}(x;h))$ converges to zero for all $x \in \mathbb{R}$. It is also well known (e.g., Rao [1983]) that if g is twice continuously differentiable such that g'' is bounded and square integrable, mise converges to zero at an optimal rate of $n^{-4/5}$ where n here is the sample size. In Section 2.2, we will show that our density estimator is consistent in quadratic mean and mise converges to zero at an optimal rate of $c^{-4/7}$ where c is the computer budget and can be thought of as the sample size. This is the same rate that Steckley and Henderson [2003] computed for the case in which Z is univariate.

We also consider a local version of our estimator. A local kernel density estimator is a modification of the naive estimator that is often quite effective in practice. It allows the bandwidth to be a function of the point at which g is being estimated. Let $\hat{g}_L$ denote the local estimator in the standard density estimation setting. It has the form (e.g., Jones [1990])

$$\hat{g}_L(x;h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x)} K\left(\frac{x - Y_i}{h(x)}\right). \tag{1.3}$$

It has been shown that mise convergence for this estimator is better than that of the naive estimator. The optimal *rate* of convergence is actually the same but the constant multiplier of the rate is smaller (e.g., Jones [1990]).

We will also consider a bias-corrected version of our estimator. Jones and Signorini [1997] review bias-correction in kernel density estimation. We implement a method similar to jackknife bias-correction. For an introduction to the jackknife, see Efron and Tibshirani [1993]. In Section 2.4, we show that mse converges to zero at a rate of $c^{-8/11}$.

Kernel smoothing methods require the selection of the smoothing parameter which essentially controls the width of the kernel which in turn determines how much of the neighboring data is used in estimation at a particular location. The performance of kernel smoothing is quite dependent on this choice of the bandwidth. A good part of the kernel density estimation literature is devoted to the selection of the smoothing parameter. For a review of this literature see Wand and Jones [1995]. Schulman [1998] reviews some modern bandwidth selection methods in the context of local polynomial regression, a type of kernel regression. One such method is the empirical-bias bandwidth selection (EBBS) developed by Ruppert [1997]. In our setting, we must choose the bandwidth but equally important, we must choose the number of samples of Z and conditional on a given Z, the number of samples of X for a given computer budget. Applying the ideas from EBBS, we develop a data-driven method to select each of these parameters.

In the next section we introduce the problem of computing service system performance. As we mentioned before, given an unknown customer arrival process, this problem is an example of the kind of input model uncertainty problem to which our kernel density estimator can be applied.

1.3 Service System Performance Given Uncertainty in the Arrival Process

Finding the optimal staffing level in service systems involves computing performance and cost for a range of potential staffing levels. Computing the cost is straightforward. Computing performance is more difficult. We explore the question of how to compute performance for a given staffing level, focusing on the situation when the arrival rate of calls to the call center cannot be determined with certainty. More specifically we examine the questions of *what* to compute, *how* to compute it, and what are the probable implications of ignoring uncertainty associated with the arrival process.

We define performance to be the fraction of customers that wait less than a prescribed amount of time in the queue before being served. This is a commonly used metric and is sometimes referred to as the "service level" metric.

As we alluded to in Section 1.1, the term "uncertainty" has several possible interpretations. One interpretation is as follows. On any given day (the choice of days as a time scale is arbitrary but seems appropriate) it is quite reasonable to model the arrival process of calls as a nonhomogeneous Poisson process (NHPP) with time-dependent rate function $\Lambda = (\Lambda(t) : t \geq 0)$. This follows from the Palm-Khintchine theorem (e.g., Whitt [2002, p. 318]) that states that the superposition of arrivals from a large number of independent potential customers is well approximated by a Poisson process. The rate function depends on the propensity of customers to call, which in turn can depend on factors that cannot be planned for in advance, such as weather, marketing promotions, and competitor behavior. This situation can be modeled by viewing the rate function Λ as random. Once the rate function is realized for a given day (e.g., after the weather, marketing promotions, or behavior of competitors is revealed), the arrival rate function is then fixed, and call arrivals follow a NHPP with the realized rate function. We call this interpretation the *randomly varying arrival rate* (RVAR) case. This situation seems to be quite common (Avramidis et al. [2004], Brown et al. [2005], Steckley et al. [2005]).

A second interpretation relates to forecast error. In this setting we believe that there is a true deterministic arrival rate function $\lambda = (\lambda(t) : t \geq 0)$ but we do not know what it is. Unlike the RVAR case, the arrival rate function does not vary from day to day. The uncertainty here arises due to our lack of perfect knowledge of λ. For example, the response to a one-time marketing campaign fits this framework. If we model our uncertainty through a random function $\Lambda = (\Lambda(t) : t \geq 0)$ then we again have a random arrival rate, but the interpretation is quite different to the RVAR case. We call this case the *unknown arrival rate* (UAR) case. The example in Section 1.1 is a UAR case. It is this type of situation, one that is characterized by input uncertainty, that is the motivation for the work on estimating the density of a conditional expectation.

A hybrid situation where the *distribution* of the arrival rate in the RVAR case is unknown is also possible, if not typical, but while it may be the "correct" abstraction it also seems unwieldy. We do not consider that possibility further.

For each of these interpretations of uncertainty in the arrival process, we identify the appropriate measures for both long-run and short-run performance. Long-run performance measures the long-run fraction of satisfactory calls over a large number of days. Short-run performance measures performance for a single instance of a day. It answers the question: "What might happen tomorrow?"

We also discuss *how* to compute each of these performance measures. A common approach is to use closed-form expressions based on steady-state results for simple queueing models. When such approximations are inaccurate or infeasible, simulation provides an alternative way to compute performance. We discuss both steady-state approximations and simulation-based estimates.

Grassmann [1988] modeled forecast errors using a random arrival rate. Thompson [1999] and Jongbloed and Koole [2001] gave methods for staffing when the arrival rate is random. Whitt [1999] suggested a particular form of the random arrival rate for capturing forecast uncertainty. Chen and Henderson [2001] studied the potential impact of ignoring the issue on predictions. Ross [2001, Chapter 4] developed extensions to the "square-root staffing rule" to account for a random arrival rate. Avramidis et al. [2004] developed several different arrival process models and compared their fit to call center data. They also found that performance measures depend fairly strongly on the arrival process. Deslauriers et al. [2004] show that it is appropriate in their setting to weight performance by the arrival rate. Gans et al. [2003] discuss this issue as part of a survey of the area of call center design and management. Brown et al. [2005] developed an autoregressive model for the arrival rate that can capture correlation across different days. Harrison and Zeevi [2005] developed an economic model based on attaching costs to abandonment and agent levels. Mathematical support for their model is given in Bassamboo et al. [2004]. Whitt [2004] gives an economic analysis for a special case of the Harrison-Zeevi model, offering 2 computational approaches for estimating performance. Both the Harrison-Zeevi and Whitt papers address the RVAR case. We do not adopt an economic model here, instead working directly with performance measures associated with the waiting time distribution of a "typical" customer. In addition to a random arrival rate, Whitt [2004] deals explicitly with absenteeism, which he models through a random number of servers being available. We do not consider a random number of servers, although it is possible to capture that phenomenon in a straightforward manner in the RVAR case.

1.4 Thesis Organization

In the remaining section of this introductory chapter, Section 1.5, we present some mathematical preliminaries. We give some basic definitions and elementary results that serve as the foundation for this work.

In Chapter 2 we formulate estimators for the density of the conditional expectation and present convergence results. In Section 2.1, we formulate our "naive" estimator of the conditional expectation. This estimator is motivated by the naive estimator for standard kernel density estimation. In Section 2.2, we present convergence results and proofs for our naive estimator. We show the estimator is consistent in quadratic mean and, under stronger assumptions, derive expressions for the asymptotic mse and mise, respectively. Using the expression for the asymptotic mise, we compute the optimal rate of convergence of mise. In Sections 2.3 and 2.4 we present estimators that are modifications of the one formulated in Section 2.1. In Section 2.3 we consider a local version and in Section 2.4 we consider a bias-corrected version. We derive asymptotic expressions of the mse for both estimators and in the case of the bias-corrected estimate, we see an improvement in the optimal rate of convergence.

In Chapter 3, we address the implementation of the estimators for the density of the conditional expectation discussed in Chapter 2 and study their performance. Section 3.1 considers the implementation which requires specifying a number of inputs. For some of these inputs, we develop a data-based selection method based on the ideas of empirical-bias bandwidth selection (EBBS) used in local polynomial regression (Ruppert [1997]). We discuss the reasons for choosing this method and present the algorithm. In Section 3.2, we then compare the performance of the estimators for some simulated test cases. We consider representative plots and study the behavior of estimated mise for each of the estimators.

In Chapter 4 we turn to the computation of service system performance in the presence of an uncertain arrival rate. In Section 4.1 we consider the RVAR case and the performance measure giving the long-run fraction of customers that wait less than a prescribed amount of time in queue before receiving service. We give an expression for this quantity, and then consider approximations given by steady-state expectations. We also show that performance will typically be over-estimated if a randomly-varying arrival rate is ignored. We then turn to short-run performance, which is the distribution of the fraction of calls answered in the given time limit for a single instance of a period. We give a steady-state approximation based on a central limit theorem. The section concludes by discussing how one can use simulation to estimate both short-run and long-run performance measures efficiently. In Section 4.2 we consider the UAR case and again suggest appropriate performance measures for the short-run and long-run. We again consider approximations based on steady-state expectations. The section concludes with a discussion of simulation procedures to estimate the performance measures, including one which involves the density of a conditional expectation. In Section 4.3 we describe a set of experiments designed to examine performance for the RVAR case. Specifically, we wanted to determine which factors impact the performance measures discussed in Section 4.1, assess the quality of the approximations as compared to the simulation-based estimates of performance, and learn more about the behavior of systems with a random arrival rate.

1.5 Mathematical Preliminaries

In this section we give basic definitions and results for the conditional expectation (Section 1.5.1), the conditional distribution (Section 1.5.2), and the conditional density (Section 1.5.3).

1.5.1 Conditional Expectation

The following definition of a conditional expectation can be found in, e.g., Billingsley [1995].

Definition 1 *Let X be an integrable random variable on the probability space $(\Omega, \mathscr{F}, P)$ and let $\mathscr{G}$ be a σ-field in $\mathscr{F}$. There exists a random variable $E(X|\mathscr{G})$ called the conditional expected value of X given $\mathscr{G}$, having these two properties:*

(i) $E(X|\mathscr{G})$ is measurable $\mathscr{G}$ and integrable;

(ii) $E(X|\mathscr{G})$ satisfies the functional equation

$$\int_G E(X|\mathscr{G})\,\mathrm{d}P = \int_G X\,\mathrm{d}P, \qquad G \in \mathscr{G}.$$

In general, there will be many random variables that satisfy the properties of the definition, and each of these random variables is considered a version of the conditional expected value of X given $\mathscr{G}$. Any two versions are equal with probability one.

Let Z be a random variable defined on the probability space $(\Omega, \mathscr{F}, P)$. We write $\mathrm{E}(X|Z)$ for $\mathrm{E}(X|\sigma(Z))$.

1.5.2 Conditional Distribution Function

A conditional expectation can also be defined in terms of the conditional probability distribution. The conditional probability distribution itself is important to us. Breiman [1968] gave the following definition.

Definition 2 *Let $\mathscr{G}$ be a σ-field in $\mathscr{F}$. The conditional probability of $A \in \mathscr{F}$ given $\mathscr{G}$ is a random variable $P(A|\mathscr{G})$ on $(\Omega, \mathscr{G})$ satisfying*

$$\int_G P(A|\mathscr{G})\,\mathrm{d}P = P(A \cap G), \qquad G \in \mathscr{G}.$$

Once again there could be many random variables that satisfy the definition, all of which are referred to as versions of the conditional probability of $A \in \mathscr{F}$ given $\mathscr{G}$. There may not exist a version $\mathrm{P}^*(A|\mathscr{G})$ that gives a probability distribution on $\mathscr{F}$ for every $\omega \in \Omega$. When such a version does exist, it is called a regular conditional probability.

Let $\mathscr{B}_1$ denote the Borel σ-field in one-dimension. Breiman [1968] defined a Borel space $(\Omega, \mathscr{B})$ as follows.

Definition 3 *Call $(\Omega, \mathscr{B})$ a Borel space if there is an $E \in \mathscr{B}_1$ and a one-to-one mapping $\phi : \Omega \leftrightarrow E$ such that ϕ is $\mathscr{B}$-measurable and ϕ^{-1} is $\mathscr{B}_1$-measurable.*

In a certain sense, a space $(\Omega, \mathscr{B})$ is a Borel space if it looks like the measurable space $(\mathbb{R}, \mathscr{B}_1)$. Breiman [1968] showed that if X takes values in a Borel space then there is a regular conditional distribution for X given Z. Of course, if X is a (real-valued) random variable, then there exists a regular conditional distribution. The existence of a regular conditional distribution will be important to us beyond simply allowing us to define the conditional expectation. For example, in assuming that we can generate i.i.d. observations from the conditional distribution $P(X \in$

$\cdot|Z = z)$ for any z in the range of Z it is enough to assume that the regular conditional distribution exists.

1.5.3 Conditional Density

The existence of densities is also of obvious importance. Billingsley [1995] gave the following definition of the density of a measure.

Definition 4 *A measure ν on $(\Omega, \mathscr{F})$ is said to have density δ with respect to μ on $(\Omega, \mathscr{F})$ if δ is a nonnegative measurable function that satisfies*

$$\nu(A) = \int_A \delta \, \mathrm{d}\mu, \qquad A \in \mathscr{F}.$$

Again, there can be multiple versions of the density but any two versions differ on a set of μ-measure 0. The Radon-Nikodym theorem states that if ν and μ are σ-finite and ν is absolutely continuous with respect to μ, then the density δ exists.

We are concerned only with the case that ν is some probability measure P and μ is Lebesgue measure. In this case, P has density f if f is a nonnegative measurable function that satisfies

$$\mathrm{P}(A) = \int_A f(x) \, \mathrm{d}x, \qquad A \in \mathscr{F}.$$

Chapter 2

Estimating the Density of the Conditional Expectation and Convergence Results

In this chapter we formulate the estimators for the conditional expectation and present convergence results. In Section 2.1, we formulate a naive estimator that is motivated by the naive estimator for standard kernel density estimation. In Section 2.2, we present convergence results and proofs for the naive estimator. In Section 2.2.1, we show the estimator is consistent in quadratic mean. In Section 2.2.2 expressions for the asymptotic mse and mise are derived under stronger assumptions. In practice these assumptions will be difficult to verify but the results suggest the kind of asymptotic behavior that we might expect. Using the expression for the asymptotic mise, we compute the optimal rate of convergence of mise. The proofs of the results are given in Section 2.2.3. In Sections 2.3 and 2.4 we present estimators that are modifications of the one formulated in 2.1. In Section 2.3 we consider a local version and in Section 2.4 we consider a bias-corrected version. We derive asymptotic expressions of the mse for both estimators and in the case of the bias-corrected estimate, we establish an improvement in the optimal rate of convergence.

2.1 Estimation Methodology

Our problem is very similar in structure to that of Lee and Glynn [2003]. Accordingly, we adopt much of their problem structure and assumptions in what follows. We assume the ability to

1. draw samples from the distribution $P(Z \in \cdot)$, and
2. for any z in the range of Z, to draw samples from the conditional distribution $P(X \in \cdot | Z = z)$.

Let f denote the (target) density of $\mathrm{E}(X|Z)$, which we assume exists. Let $(Z_i : 1 \leq i \leq n)$ be a sequence of independent, identically distributed (i.i.d.) copies of the random variable Z. Conditional on $(Z_i : 1 \leq i \leq n)$, the sample $(X_j(Z_i) : 1 \leq i \leq n, \quad 1 \leq j \leq m)$ consists of independent random variables in which $X_j(Z_i)$ follows the distribution $P(X \in \cdot \ | Z = Z_i)$. For ease of notation define $\mu(\cdot) \equiv \mathrm{E}(X|Z = \cdot)$ and $\sigma^2(\cdot) \equiv \mathrm{var}(X|Z = \cdot)$.

In Section 1.2 we introduced the naive kernel estimator in the standard density estimation setting. In this setting we have a random variable Y with unknown density g, and a sequence $(Y_i : 1 \leq i \leq n)$ of i.i.d. copies of the random variable Y. In (1.2) we gave the formula of the naive kernel density estimator $\hat{g}$ evaluated at x. We repeat the formula here:

$$\hat{g}(x; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - Y_i}{h}\right).$$

The kernel K is typically chosen to be a unimodal probability density function (p.d.f.) that is symmetric about zero, and the smoothing parameter h, often referred to as the bandwidth, is a positive number (e.g., Wand and Jones [1995]). The estimator can be crudely described as the sum of equally weighted kernels centered at each realization Y_i. If the kernel is a p.d.f., the kernel spreads out the mass of $1/n$ symmetrically about the neighborhood of Y_i. In the case that K is the p.d.f. of a standard normal random variable, h is the standard deviation and thus gives the spread of the kernels.

This estimator immediately suggests that we can estimate $f(x)$, the density of $\mathrm{E}(X|Z)$ evaluated at x, by

$$\hat{f}(x; m, n, h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - \bar{X}_m(Z_i)}{h}\right), \tag{2.1}$$

where

$$\bar{X}_m(Z_i) = \frac{1}{m} \sum_{j=1}^{m} X_j(Z_i) \quad \text{for} \quad i = 1, \ldots, n.$$

Note that the values $\bar{X}_m(Z_i)$ at which the kernels are centered are not realizations of the random variable $\mathrm{E}(X|Z)$ as in the standard kernel density estimation setting described above, but rather estimates thereof. The observations used in the kernel density estimator are thus measured with error.

This measurement error will lead to additional smoothing in our estimator. For the kernel density estimator (1.2), the data is smoothed by the kernel K. The extent to which the data is smoothed is determined by the bandwidth h. The larger the bandwidth, the more the data is smoothed. For the kernel estimator of the conditional expectation (2.1), along with the smoothing from the kernel K, there is additional smoothing that results from the measurement error. A similar double smoothing was noted in Staudenmayer [2000]. He considered the problem of local polynomial regression in which the covariates are measured with error. The double smoothing increases the bias of our estimator given in (2.1) as compared with the estimator (1.2). Specifically, the additional smoothing results in an additional leading term in the bias expansion, and thus an additional leading term in the mse and mise expansions given in Theorems 2 and 3 of Section 2.2.2.

There has been work on a related problem known as the deconvolution problem. Similar to our problem, the goal is density estimation when the observations are measured with error. See Wand and Jones [1995] for a discussion of this work. But to the best of my knowledge, as the name suggests, all such work on the deconvolution problem makes a crucial assumption that we do not. In particular, let $(X_i : 1 \leq i \leq n)$ be a random sample with common density g which is to be estimated. Let $(Y_i : 1 \leq i \leq n)$ be the actual observed data. In the deconvolution problem it is assumed that

$$Y_i = X_i + U_i, \qquad i = 1, \ldots, n,$$

where, for each i, the random variable U_i is independent of X_i and has a common distribution (Carrol and Hall [1988]). In our setting,

$$\bar{X}_m(Z_i) = \mathrm{E}(X|Z_i) + U_i, \qquad i = 1, \ldots, n.$$

We assume that U_i has a known distribution but we do not assume that U_i is independent of $\mathrm{E}(X|Z_i)$.

2.2 Convergence Results

In this section we study the error in the estimator $\hat{f}(x; m(c), n(c), h)$ as the computer budget c goes to infinity. For any fixed c, the number of internal samples m, and the number of external samples n, must be chosen so that the total computational effort required to generate the estimator is approximately c. Note that $m(c)$ and $n(c)$ are thus functions of the computer budget c. Following Lee and Glynn [2003], the computational effort required to compute $\hat{f}(x; m(c), n(c), h)$ is taken as

$$\delta_1 n(c) + \delta_2 n(c) m(c),$$

where δ_1 and δ_2 are the average computational effort used to generate Z_i and $X_j(Z_i)$ conditional on Z_i, respectively.

We assume that $m(c) \to \infty$ as $c \to \infty$ so that $\bar{X}_{m(c)}(z_0) \to \mathrm{E}(X|z_0)$ almost surely. Assuming $m(c) \to \infty$, $\delta_1 n(c) + \delta_2 n(c)m(c) \approx \delta_2 n(c)m(c)$. One can assume $\delta_2 = 1$ without loss of generality. Then $m(c)$ and $n(c)$ must be chosen to satisfy the asymptotic relationship $m(c)n(c)/c \to 1$ as $c \to \infty$.

The bandwidth $h = h(c)$ is also a function of c. To keep the notation less cumbersome, the dependence of m, n, and h on c will be suppressed in the calculations.

We will present results concerning the convergence of the estimator as the computer budget c tends to ∞. We consider the following two error criterion. For all $x \in \mathbb{R}$, define the mean squared error (mse) of the estimator evaluated at x as

$$\mathrm{mse}(\hat{f}(x;m,n,h)) = \mathrm{E}\big(\hat{f}(x;m,n,h) - f(x)\big)^2.$$

Define the mean integrated squared error (mise) of the estimator as

$$\mathrm{mise}(\hat{f}(\cdot;m,n,h)) = \mathrm{E}\int \big(\hat{f}(x;m,n,h) - f(x)\big)^2 \,\mathrm{d}x.$$

These error criteria are not without drawbacks (see Devroye and Lúgosi [2001]) but the mathematical simplicity is appealing.

Before stating our results, we consider the distribution of the observations $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ and in doing so, we will collect some of the assumptions needed for the results. Let $N(\alpha_1, \alpha_2)$ denote a normally distributed random variable with mean α_1 and variance α_2. For two random objects X and Y, define the notation $X =_d Y$ to mean X and Y are equal in distribution. Recall $\mu(\cdot) \equiv \mathrm{E}(X|Z = \cdot)$ and $\sigma^2(\cdot) \equiv \mathrm{var}(X|Z = \cdot)$. Throughout this thesis we assume the following:

A1. Conditional on $(Z_i : 1 \leq i \leq n)$, $\bar{X}_m(Z_i) =_d N(\mu(Z_i), m^{-1}\sigma^2(Z_i))$ for $i = 1, \ldots, n$ and $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ are conditionally independent.

This essentially implies that the internal samples $X(Z)$ conditional on Z are unbiased and normally distributed. Of course, if the central limit theorem holds, then for large m this assumption is approximately true.

We now turn to the distribution of the unconditional observations $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ (unless otherwise specified, the random variables $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ are taken to be unconditional on $(Z_i : 1 \leq i \leq n)$). First note that $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ are i.i.d., Second, observe that under Assumption A1,

$$\bar{X}_m(Z_i) =_d Y_i + S_i \frac{1}{m} \sum_{j=1}^{m} U_{ij} \quad \text{for} \quad i = 1, \ldots, n,$$

where

(i) $((Y_1, S_1), \ldots, (Y_n, S_n))$ are i.i.d. with $(Y_i, S_i) =_d (\mu(Z), \sigma(Z))$;

(ii) $(U_{ij} : 1 \leq i \leq n, \quad 1 \leq j \leq m)$ are i.i.d. with $U_{ij} =_d N(0, 1)$.

Let $U_i^m = m^{-1/2} \sum_{j=1}^{m} U_{ij}$ so that for $i = 1, \ldots, n$,

$$\begin{aligned} \bar{X}_m(Z_i) &=_d Y_i + S_i \frac{1}{m} \sum_{j=1}^{m} U_{ij} \\ &= Y_i + \frac{S_i}{\sqrt{m}} U_i^m. \end{aligned}$$

Note that $U_i^m =_d N(0, 1)$ for $i = 1, \ldots, n$, and $(U_i^m : 1 \leq i \leq n)$ are i.i.d.,

Let F_m denote the distribution function of $\bar{X}_m(Z_i)$. Assuming $\mathrm{P}(S = 0) = 0$,

$$\begin{aligned} F_m(x) &= \mathrm{P}\left(Y_i + \frac{S_i}{\sqrt{m}} U_i^m \leq x\right) \\ &= \mathrm{P}\left(U_i^m \leq \frac{(x - Y_i)\sqrt{m}}{S_i}\right). \end{aligned}$$

The following is also assumed throughout:

A2. For each $y \in \mathbb{R}$ such that $f(y) > 0$, the conditional density with respect to Lebesgue measure of the conditional distribution $\mathrm{P}(\sigma(Z) \in \cdot \; | \mu(Z) = y)$ exists. Denote this density $g(\cdot|y)$.

Since $\sigma(Z)$ and $\mu(Z)$ are random variables we know that the regular conditional distribution $\mathrm{P}(\sigma(Z) \in \cdot \; |\mu(Z) = y)$ exists for all $y \in \mathbb{R}$. This assumption simply requires that for each $y \in \mathbb{R}$ such that $f(y) > 0$, $\mathrm{P}(\sigma(Z) \in \cdot \; |\mu(Z) = y)$ is absolutely continuous with respect to Lebesgue measure.

We believe that when Z is of dimension 2 or greater, there will be many cases in which A2 is satisfied. However, for univariate Z, A2 will rarely hold. By assuming A2 in this thesis, we focus on the case in which Z is of dimension 2 or greater. Steckley and Henderson [2003] treat the case in which Z is univariate and μ is monotone. Their results for mise are very similar to the ones presented in this thesis. The proofs are somewhat simpler but require different methods. For the sake of space, we omit these results and proofs and refer the reader to Steckley and Henderson [2003].

Assuming A2,

$$\begin{aligned} F_m(x) &= \mathrm{P}\left(U_i^m \leq \frac{(x - Y_i)\sqrt{m}}{S_i}\right) \\ &= \int\int \mathrm{P}\left(U_i^m \leq \frac{(x-y)\sqrt{m}}{s}\right) g(s|y) f(y)\, \mathrm{d}s\, \mathrm{d}y, \end{aligned}$$

where $g(\cdot|y)$ can be defined arbitrarily for $y \in \mathbb{R}$ such that $f(y) = 0$. Let Φ and ϕ denote the standard normal cumulative distribution function and density, respectively. In this notation,

$$\begin{aligned} F_m(x) &= \int\int \Phi\left(\frac{(x-y)\sqrt{m}}{s}\right) g(s|y) f(y)\, \mathrm{d}s\, \mathrm{d}y \\ &= \mathrm{E}\left(\Phi\left(\frac{(x-Y)\sqrt{m}}{S}\right)\right). \end{aligned}$$

Assuming we can differentiate the RHS, and interchange the derivative and expectation, we have that the density f_m of the distribution function F_m exists and is given by

$$f_m(x) = \int\int \frac{\sqrt{m}}{s} \quad \phi\left(\frac{(x-y)\sqrt{m}}{s}\right) g(s|y)f(y) \, \mathrm{d}s \, \mathrm{d}y. \tag{2.2}$$

A sufficient condition for the interchange is

A3. $\int\int (1/s)\, g(s|y)f(y) \, \mathrm{d}s \, \mathrm{d}y < \infty.$,

as can be seen from Lemma 1 below, which comes from a result given by L'Ecuyer [1990] and L'Ecuyer [1995] (see also Glasserman [1988]).

Lemma 1 Let G be a probability measure on a measurable space $(\Omega, \mathcal{F})$. Define

$$h(x) = \int h(x, \omega) \, \mathrm{d}G(\omega)$$

and

$$h'(x_0, \omega) = \frac{\mathrm{d}}{\mathrm{d}x} h(x, \omega)|_{x=x_0}.$$

Let $x_0 \in S_0$, where S_0 is an open interval. Let H be such that $G(H^c) = 0$. Assume that for all $\omega \in H$, there exists $D(\omega)$ where $D(\omega)$ is at most countable, such that

1. for all $\omega \in H$, $h(\cdot, \omega)$ is continuous everywhere in S_0;

2. for all $\omega \in H$, $h(\cdot, \omega)$ is differentiable everywhere in $S_0 \setminus D(\omega)$;

3. for all $\omega \in H$, $|h'(x, \omega)| \leq b(\omega)$ for all $x \in S_0 \setminus D(\omega)$ and $\int b(\omega) \, \mathrm{d}G(\omega) < \infty$;

4. $h(x, \omega)$ is almost surely differentiable at $x = x_0$.

Then h is differentiable at x_0 and

$$h'(x_0) = \int h'(x_0, \omega) \, \mathrm{d}G(\omega).$$

Returning to the density of the observations $\bar{X}_m(Z)$ given in (2.2), the change of variable $z = (x-y)\sqrt{m}$, gives

$$f_m(x) = \int\int \frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x - \frac{z}{\sqrt{m}}) f(x - \frac{z}{\sqrt{m}})\,\mathrm{d}s\,\mathrm{d}z.$$

Suppose $f(\cdot)$ is continuous. For y such that $f(y) = 0$, suppose that $g(\cdot|y)$ can be defined so that $g(s|\cdot)$ is continuous for all $s \in \mathbb{R}$. We assume the following:

A4. For almost all $y \in \mathbb{R}$, $g(\cdot|y)$ is nonnegative;

A5. For almost all $y \in \mathbb{R}$, $g(s|y) = 0$ for $s < 0$.

The Assumptions A4 and A5 are certainly true for y such that $f(y) > 0$ since in that case $g(\cdot|y)$ is a density for a nonnegative random variable. Under A4, the order of integration can be changed so that

$$f_m(x) = \int\int \frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x - \frac{z}{\sqrt{m}}) f(x - \frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s. \tag{2.3}$$

It will be useful to think in terms of the joint density of $\mu(Z)$ and $\sigma(Z)$. Let us denote this density by α. Of course

$$\alpha(x, s) = g(s|x) f(x). \tag{2.4}$$

Define for nonnegative integer k,

$$\alpha^{(k+1)}(x, s) = \frac{d}{dy}\alpha^{(k)}(y, s)\Big|_{y=x}, \tag{2.5}$$

where $\alpha^{(0)}(x, s) = \alpha(x, s)$. Also define for nonnegative integer k,

$$g^{(k+1)}(s|x) = \frac{d}{dy} g^{(k)}(s|y)\Big|_{y=x},$$

where $g^0(s|x) = g(s|x)$.

For ease of notation we define the following set of Assumptions parameterized by nonnegative integer k as A6(k).

1. $f(\cdot)$ is k times continuously differentiable;

2. for all $s \in \mathbb{R}$, $g(s|\cdot)$ is k times continuously differentiable;

3. $\exists B_f > 0$ such that $|f^{(j)}(\cdot)| \leq B_f$ for $j = 0, 1, \ldots, k$;

4. $\exists B_g > 0$ such that $|g^{(j)}(\cdot|\cdot)| \leq B_g$ for $j = 0, 1, \ldots, k$;

5. $\exists B_S > 0$ such that $\sigma^2(\cdot) \leq B_S$ everywhere.

Note that $f^{(0)}$ and $g^{(0)}$ are simply f and g, respectively, and when $k = 0$, Assumptions 1 and 2 imply that $f(\cdot)$ and $g(s|\cdot)$ are continuous.

2.2.1 Consistency in quadratic mean

The following theorem gives sufficient conditions for the consistency in quadratic mean for the estimator formulated in (2.1).

Theorem 1 Assume A1-A5, and A6(0). Also assume that

1. K is a bounded probability density;

2. $m \to \infty$, $h \to 0$, and $nh \to \infty$, as $c \to \infty$.

Then for all $x \in \mathbb{R}$,

$$\lim_{c \to \infty} \text{mse}(\hat{f}(x; m, n, h)) = 0.$$

A proof is given in Section 2.2.3.

2.2.2 Asymptotic expressions for mse and mise

We now turn to the asymptotic expressions of mse and mise. More restrictive Assumptions are needed to compute these asymptotic expansions. For one thing, it is assumed that the function $f(\cdot)$ and the set of functions $\{g(s|\cdot) : s \in \mathbb{R}\}$ are four times continuously differentiable.

Before stating the results let us introduce the notation o and O. For sequences of real numbers a_n and b_n, we say that

$$a_n = o(b_n) \text{ as } n \to \infty \text{ iff } \lim_{n\to\infty} a_n/b_n = 0.$$

For sequences of real numbers a_n and b_n, we say that

$$a_n = O(b_n) \text{ as } n \to \infty \text{ iff } \exists C \text{ s.t. } a_n \le C b_n \text{ for } n \text{ sufficiently large.}$$

Theorem 2 Assume A1-A5, and A6(4). Also assume

1. K is a bounded probability distribution function symmetric about zero with finite second moment;

2. $m \to \infty$, $n \to \infty$, $h \to 0$, and $nh \to \infty$ as $c \to \infty$.

Then

$$\begin{aligned} \mathrm{mse}(\hat{f}(x; m, n, h)) &= \left(h^2 \frac{1}{2} f''(x) \int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2} \int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s \right)^2 \\ &\quad + \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\left(h^2 + \frac{1}{m} \right)^2 + \frac{1}{nh} \right), \end{aligned} \tag{2.6}$$

where α is defined in (2.4) and (2.5).

The asymptotic expansion of mise follows.

Theorem 3 Assume A1-A5 and A6(4). Also assume

1. $f''(\cdot)$ is ultimately monotone;

2. $f^{(k)}(\cdot)$ is integrable for $k = 1, 2, 3, 4$;

3. K is a bounded probability density function symmetric about zero with finite second moment;

4. $m \to \infty$, $n \to \infty$, $h \to 0$, and $nh \to \infty$ as $c \to \infty$.

Then

$$\begin{aligned}\operatorname{mise}(\hat{f}(\cdot; m, n, h)) &= \int \left(h^2 \frac{1}{2} \left(\int u^2 K(u)\,\mathrm{d}u \right) f''(x) + \frac{1}{m}\frac{1}{2} \int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s \right)^2 \mathrm{d}x \\ &\quad + \frac{1}{nh} \int K^2(u)\,\mathrm{d}u \\ &\quad + o\left(\left(h^2 + \frac{1}{m} \right)^2 + \frac{1}{nh} \right), \qquad (2.7)\end{aligned}$$

where α is defined in (2.4) and (2.5).

Theorem 3 follows from Theorem 2 provided the o term in (2.6) is integrable. Proofs of Theorems 2 and 3 are presented in Section 2.2.3.

Compare (2.7) to the mise for standard kernel density estimation (e.g., Wand and Jones [1995]),

$$\begin{aligned}\operatorname{mise}(\hat{g}(\cdot; h)) &= \int \left(h^2 \frac{1}{2} \left(\int u^2 K(u)\,\mathrm{d}u \right) g''(x) \right)^2 \mathrm{d}x + \frac{1}{nh} \int K(u)^2\,\mathrm{d}u \\ &\quad + o\left(h^4 + \frac{1}{nh} \right). \qquad (2.8)\end{aligned}$$

To aid in the comparison, recall that mse can be decomposed into a squared bias term and a variance term:

$$\operatorname{mse}(\hat{f}(x; m, n, h)) = \operatorname{bias}^2(\hat{f}(x; m, n, h)) + \operatorname{var}(\hat{f}(x; m, n, h)) \quad \forall x \in \mathbb{R} \qquad (2.9)$$

where

$$\operatorname{bias}(\hat{f}(x; m, n, h)) = \mathrm{E}(\hat{f}(x; m, n, h)) - f(x).$$

We have a similar decomposition for mise:

$$\operatorname{mise}(\hat{f}(\cdot; m, n, h)) = \int \operatorname{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x + \int \operatorname{var}(\hat{f}(x; m, n, h))\,\mathrm{d}x. \quad (2.10)$$

We see that mise can be decomposed into integrated squared bias and integrated variance. We get similar formulas for the standard kernel density estimator $\hat{g}$. Note the $O(1/nh)$ terms in the mise expansions in (2.7) and (2.8) are the same for both estimators. In the proof of Theorem 3 we show that this term is the leading term for the integrated variance. The remaining leading terms in (2.7) and (2.8) are those of the integrated squared bias.

For our estimator $\hat{f}$, the bias itself can be further decomposed. Suppose that the density of an observation $\bar{X}_m(Z)$ exists and is given by $f_m(\cdot)$. Then

$$\text{bias}(\hat{f}(x;m,n,h)) \quad = \quad (\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x)) \tag{2.11}$$

The first component, $\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)$, is the expected distance from our kernel estimator to the density of the observations $\bar{X}_m(Z)$. We can think of this as the bias due to the kernel smoothing. The second component, $f_m(x) - f(x)$, is the distance from the density of the observations to the true target density. If the observations were measured with no error, then this distance would be zero. We can therefore think of this as the bias due to measurement error.

Both the standard kernel density estimator and our estimator are biased due to the kernel smoothing. The proof of Theorem 3 shows that the leading term of this bias is the $O(h^2)$ term seen in the mise expansion (2.7) for $\hat{f}$. This same $O(h^2)$ term is present in (2.8), the mise expansion of $\hat{g}$. This term shows that bias is dependent on the curvature of the target density through the second derivative of the target density. In addition, our estimator also has a bias component due to the measurement error. The leading term of this component is the $O(1/m)$ term in (2.7). As expected, the bias of our estimator due to the measurement error in the observations $\bar{X}_m(Z_i)$, $i = 1, \ldots, n$, decreases as the number of internal samples m increases. Also note that the bias of our estimator due to the measurement error depends on the distribution of the conditional variance function $\sigma^2(\cdot)$ through α.

The asymptotic mise for our estimator $\hat{f}$ is

$$\int\left(h^2\frac{1}{2}\left(\int u^2K(u)\mathrm{d}u\right)f''(x)+\frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\mathrm{d}s\right)^2\mathrm{d}x+\frac{1}{nh}\int K^2(u)\mathrm{d}u. \tag{2.12}$$

By choosing m, n, and h to minimize this asymptotic mise, we can achieve the optimal asymptotic convergence. Define

$$A=\sqrt{\sqrt{\frac{\int\beta_2(x)^2\,\mathrm{d}x}{2\int\beta_1(x)^2\,\mathrm{d}x}+\frac{(\int\beta_1(x)\beta_2(x)\,\mathrm{d}x)^2}{16(\int\beta_1(x)^2\,\mathrm{d}x)^2}}-\frac{\int\beta_1(x)\beta_2(x)\,\mathrm{d}x}{4\int\beta_1(x)^2\,\mathrm{d}x}},$$

where

$$\beta_1(x)=\frac{f''(x)}{2}\int u^2K(u)\,\mathrm{d}u$$

and

$$\beta_2(x)=\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s.$$

Then the optimal m, n, and h, denoted m^*, n^*, and h^*, are

$$m^* = \left(\frac{2A^3\int\beta_1(x)\beta_2(x)\,\mathrm{d}x+2A\int\beta_2(x)\,\mathrm{d}x}{\int K^2(u)\,\mathrm{d}u}\right)^{2/7}c^{2/7}, \tag{2.13}$$

$$n^* = \left(\frac{\int K^2(u)\,\mathrm{d}u}{2A^3\int\beta_1(x)\beta_2(x)\,\mathrm{d}x+2A\int\beta_2(x)\,\mathrm{d}x}\right)^{2/7}c^{5/7},\text{ and} \tag{2.14}$$

$$h^* = A\left(\frac{\int K^2(u)\,\mathrm{d}u}{2A^3\int\beta_1(x)\beta_2(x)\,\mathrm{d}x+2A\int\beta_2(x)\,\mathrm{d}x}\right)^{1/7}c^{-1/7}. \tag{2.15}$$

Substituting m^*, n^*, and h^* into (2.12) shows that the optimal rate of convergence is of the order $c^{-4/7}$. In fact, when m, n, and h are chosen such that m is of the order $c^{2/7}$, n is of the order $c^{5/7}$, and h is of the order $c^{-1/7}$ the optimal rate of convergence of mise is achieved. We note that for the case in which Z is assumed to be univariate, the optimal rate of convergence is also $c^{-4/7}$ (Steckley and Henderson [2003]).

In standard kernel density estimation, the optimal rate of convergence is $c^{-4/5}$ (Wand and Jones [1995]). The decrease in the rate of convergence is a consequence of the additional bias in our estimator due to measurement error. For each of the n observations $\bar{X}_m(Z_i)$, we must use m units of computer time to deal with the measurement error bias, and $m \to \infty$ as $c \to \infty$. In the standard kernel density estimation setting, each observation requires only one unit of computer time since there is no measurement error.

Although we phrased the optimal rate of convergence in terms of mise, the same applies to the mse. So the optimal rate of convergence of mse for our estimator $\hat{f}(x; m, n, h)$ is $c^{-4/7}$.

2.2.3 Proofs

In this section, we present the proofs of Theorems 1, 2, and 3.

The proof of the consistency in quadratic mean

The following lemma is useful in the proof of Theorem 1.

Lemma 2 Assume A1-A4 and A6(0). Also assume that

1. K is nonnegative and integrable;
2. $m \to \infty$ and $h \to 0$ as $c \to \infty$.

Then

$$\lim_{c \to \infty} \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y = f(x) \int K(u)\,\mathrm{d}u.$$

Proof: By (2.3),

$$
\begin{aligned}
&\int \frac{1}{h}K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y \\
&= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) \iint \frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|y-\frac{z}{\sqrt{m}})f(y-\frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}y \\
&= \iiint \frac{1}{h}K\left(\frac{x-y}{h}\right) \frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|y-\frac{z}{\sqrt{m}})f(y-\frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}y \\
&= \iiint K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}u. \qquad (2.16)
\end{aligned}
$$

We will show that there exists an integrable function $\tilde{f}$ such that for all $c > C$, for some nonnegative number C,

$$
K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}}) \le \tilde{f}(s,z,u). \qquad (2.17)
$$

Then by Lebesgue's dominated convergence theorem,

$$
\begin{aligned}
&\lim_{c\to\infty} \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y \\
&= \lim_{c\to\infty} \iiint K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}u \\
&= \iiint \lim_{c\to\infty} K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}u \\
&= \iiint K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) \lim_{c\to\infty} g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}})\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}u. \qquad (2.18)
\end{aligned}
$$

For any given $u \in \mathbb{R}$, $uh \to 0$ and for any given $z \in \mathbb{R}$, $z/\sqrt{m} \to 0$ as $c \to \infty$. Then, by the continuity of $f(\cdot)$ and $g(s|\cdot)$,

$$
\lim_{c\to\infty} g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}}) = g(s|x)f(x) \quad \forall s \in \mathbb{R}.
$$

Thus,

$$
\begin{aligned}
\lim_{c\to\infty} \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y &= \iiint K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x)f(x)\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}u \\
&= f(x) \iiint K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right) g(s|x)\,\mathrm{d}u\,\mathrm{d}z\,\mathrm{d}s \\
&= f(x) \int K(u)\,\mathrm{d}u \int g(s|x)\,\mathrm{d}s.
\end{aligned}
$$

The second equality follows from an application of Fubini's Theorem for a nonnegative integrand. For $x \in \mathbb{R}$ such that $f(x)$ is nonzero, $\int g(s|x)\,\mathrm{d}s = 1$. The result then follows, once we establish (2.17). By Assumptions A5 and A6(0),

$$\begin{aligned} K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right)g(s|x-uh-\frac{z}{\sqrt{m}})f(x-uh-\frac{z}{\sqrt{m}}) \\ \le K(u)\frac{1}{s}\,\phi\left(\frac{z}{s}\right)I(0<s\le B_S)B_gB_f \\ = \tilde{f}(s,z,u). \end{aligned}$$

$\square$

Proof of Theorem 1: Let x be arbitrary. Recall the decomposition of $\mathrm{mse}(\hat{f}(x;m,n,h))$ in (2.9). Since $(\bar{X}_m(Z_i) : 1 \le i \le n)$ are i.i.d. with common probability density f_m,

$$\begin{aligned} \mathrm{E}(\hat{f}(x;m,n,h)) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\ &= \mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\ &= \int\frac{1}{h}K\left(\frac{x-y}{h}\right)f_m(y)\,\mathrm{d}y. \end{aligned}$$

Since K is a probability density and thus nonnegative and integrable, all of the assumptions for Lemma 2 hold, and so

$$\begin{aligned} \lim_{c\to\infty}\mathrm{E}(\hat{f}(x;m,n,h)) &= f(x)\int K(u)\,\mathrm{d}u \\ &= f(x). \end{aligned}$$

It follows that $\mathrm{bias}^2(\hat{f}(x;m,n,h))$ tends to zero as $c \to \infty$.

Again since $(\bar{X}_m(Z_i) : 1 \leq i \leq n)$ are i.i.d. with common probability density f_m,

$$\begin{aligned}
\operatorname{var}(\hat{f}(x;m,n,h)) &= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\
&= \frac{1}{n}\operatorname{var}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\
&= \frac{1}{n}\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)^2 - \frac{1}{n}\left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2 \\
&= \frac{1}{nh}\mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) - \frac{1}{n}\left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2 \qquad (2.19) \\
&\leq \frac{1}{nh}\mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\
&= \frac{1}{nh}\int \frac{1}{h}K^2\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y.
\end{aligned}$$

Since K is a bounded probability density, K^2 is integrable. Therefore Lemma 2 holds and

$$\lim_{c\to\infty}\int \frac{1}{h}K^2\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y = f(x)\int K^2(u)\,\mathrm{d}u < \infty.$$

Now, $1/nh$ tends to zero as $c \to \infty$, and so

$$\lim_{c\to\infty}\operatorname{var}(\hat{f}(x;m,n,h)) = 0.$$

By the decomposition in (2.9),

$$\lim_{c\to\infty}\operatorname{mse}(\hat{f}(x;m,n,h)) = 0,$$

and since x was arbitrary, the result follows.

$\square$

The proofs of the asymptotic expressions for mse and mise

To compute the rates at which mse and mise converge to zero, we will make use of the decomposition of bias given in (2.11). As a reminder, the decomposition is

$$\operatorname{bias}(\hat{f}(x;m,n,h)) \quad = \quad (\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x)).$$

Computing the rates at which mse and mise converge to zero, requires additional smoothness in f and g. Recall that for all $x \in \mathbb{R}$ and $s \in \mathbb{R}$,

$$\alpha(x, s) = g(s|y)f(y),$$

and for nonnegative integer k,

$$\alpha^{(k+1)}(x, s) = \frac{d}{dy}\alpha^{(k)}(y, s)\big|_{y=x},$$

where $\alpha^{(0)}(x, s) = \alpha(x, s)$.

Lemmas 4 and 5 presented below are useful for both the mse result and the mise result. Both results use Taylor's theorem with integral remainder, presented here as a lemma.

Lemma 3 Let $k \geq 0$ be an integer. Assume f is k times continuously differentiable and $k+1$ times differentiable. Also assume that $f^{(k+1)}$ is integrable on $(x, x+h)$. Then

$$f(x+h) = f(x)+hf'(x)+\frac{h^2}{2!}f^{(2)}(x)+\cdots+\frac{h^k}{k!}f^{(k)}(x)+\frac{h^{k+1}}{k!}\int_0^1 (1-t)^k f^{(k+1)}(x+th)\,\mathrm{d}t.$$

The proof of Taylor's theorem with integral remainder involves repeated application of integration by parts which follows from the fundamental theorem of calculus (e.g., Rudin [1987, p.149]).

Lemma 4 Assume that for some $k \geq 1$,

1. $f(\cdot)$ is $2k$ times continuously differentiable, and

2. for all $s \in \mathbb{R}$, $g(s|\cdot)$ is $2k$ times continuously differentiable.

Then for all x,

$$\begin{aligned} f_m(x) &= \sum_{j=0}^{k-1} \frac{1}{m^j}\frac{1}{2^j j!}\int s^{2j}\alpha^{(2j)}(x, s)\,\mathrm{d}s \\ &\quad + \frac{1}{m^k}\frac{1}{(2k-1)!}\iiint_0^1 (1-t)^{2k-1}z^{2k}\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2k)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s. \end{aligned}$$

Remark 1 *Since* $\int \alpha(x,s)\,\mathrm{d}s = f(x)$*, if* $k=1$*,*

$$f_m(x) = f(x) + \frac{1}{m}\iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s,$$

and for $k=2$*,*

$$\begin{aligned} f_m(x) = f(x) \quad &+ \quad \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s \\ &+ \quad \frac{1}{m^2}\frac{1}{3!}\iiint_0^1 (1-t)^3 z^4\frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s. \end{aligned}$$

Proof of Lemma 4: By (2.3) and Taylor's theorem with integral remainder

$$\begin{aligned} f_m(x) &= \iint \frac{1}{s}\,\phi\left(\frac{z}{s}\right)\alpha(x - \frac{z}{\sqrt{m}}, s)\,\mathrm{d}z\,\mathrm{d}s \\ &= \iint \frac{1}{s}\,\phi\left(\frac{z}{s}\right)\sum_{j=0}^{2k-1}\frac{1}{j!}\left(\frac{-z}{\sqrt{m}}\right)^j \alpha^{(j)}(x,s)\,\mathrm{d}z\,\mathrm{d}s \\ &\quad + \iiint_0^1 \frac{1}{s}\phi\left(\frac{z}{s}\right)\frac{1}{(2k-1)!}\left(\frac{z}{\sqrt{m}}\right)^{2k}(1-t)^{2k-1}\alpha^{(2k)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\ &= \sum_{j=0}^{2k-1}\frac{1}{j!}\left(\frac{-1}{\sqrt{m}}\right)^j\int \alpha^{(j)}(x,s)\int z^j\frac{1}{s}\,\phi\left(\frac{z}{s}\right)\,\mathrm{d}z\,\mathrm{d}s \qquad (2.20) \\ &\quad + \frac{1}{m^k}\frac{1}{(2k-1)!}\iiint_0^1 (1-t)^{2k-1}z^{2k}\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2k)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s. \end{aligned}$$

Note that $(1/s)\,\phi(\cdot/s)$ is the density of a normal random variable with mean 0 and variance s^2. Then the odd moments of this distribution are zero and the jth moment for j even is

$$\frac{j!s^j}{(j/2)!2^{(j/2)}}.$$

Substituting the moments into (2.20) gives the result.

□

The next lemma gives conditions that ensure that f_m is differentiable and the jth derivative $f_m^{(j)}(x)$ is given by $f^{(j)}(x)$ plus bounded, lower order terms.

Lemma 5 Assume A1-A5 and A6($2k$) for some $k \geq 1$. Then $f_m(\cdot)$ is 2k-2 times differentiable everywhere and

1. for $j = 0, \ldots, 2k-2$, and for all x

$$f_m^{(j)}(x) = f^{(j)}(x) + \frac{1}{m} R_m^{(j)}(x),$$

where

$$R_m^{(j)}(x) = \iiint_0^1 (1-t) z^2 \frac{1}{s} \phi(\frac{z}{s}) \alpha^{(2+j)}(x - t\frac{z}{\sqrt{m}}, s) \, \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}s; \tag{2.21}$$

2. for $j = 0, \ldots, 2k-2$, and for all x and c, $\exists B_1 > 0$ such that

$$|R_m^{(j)}(x)| \leq B_1.$$

If, in addition to the above assumptions, assume

1. $f^{(j)}(\cdot)$ is integrable for $j = 1, \ldots, 2k$,

then for $j = 0, \ldots, 2k-2$ and for all c, $\exists B_2 > 0$ such that

$$\int |R_m^{(j)}(x)| \, \mathrm{d}x \leq B_2.$$

Proof: Note that the first result for the case in which $j = 0$ is Lemma 4, so

$$f_m(x) = f(x) + \frac{1}{m} \iiint_0^1 (1-t) z^2 \frac{1}{s} \phi(\frac{z}{s}) \alpha^{(2)}(x - t\frac{z}{\sqrt{m}}, s) \, \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}s.$$

Since $f(\cdot)$ is 2k times continuously differentiable, to prove the first result for the remaining j it is enough to show that

$$\frac{\partial}{\partial x} \iiint_0^1 (1-t) z^2 \frac{1}{s} \phi(\frac{z}{s}) \alpha^{(2+j)}(x - t\frac{z}{\sqrt{m}}, s) \, \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}s \tag{2.22}$$

exists and is equal to

$$\iiint_0^1 (1-t) z^2 \frac{1}{s} \phi(\frac{z}{s}) \alpha^{(3+j)}(x - t\frac{z}{\sqrt{m}}, s) \, \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}s$$

for $j = 0, 1, \ldots, 2k-3$. Assumption A6($2k$) imply that for any s, $\alpha(x, s)$ is 2k times continuously differentiable in its first argument, so all that needs to be shown is that the derivative and integration operations can be interchanged in (2.22). The integrand in question can be rewritten

$$I(0 < t < 1)I(0 < s < B_S)(1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2+j)}(x - t\frac{z}{\sqrt{m}}, s).$$

Note the integrand is continuous and differentiable in x. And for all x, the derivative of the integrand is

$$I(0 < t < 1)I(0 < s < B_S)(1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(3+j)}(x - t\frac{z}{\sqrt{m}}, s).$$

Now, by Assumption A6($2k$), $\alpha^{(3+j)}(x - t\frac{z}{\sqrt{m}}, s) \leq dB_f B_g$ for some constant d, so the derivative of the integrand is bounded in absolute value by the integrable function

$$I(0 < t < 1)I(0 < s < B_S)dB_f B_g z^2\frac{1}{s}\phi(\frac{z}{s}).$$

Then by Lemma 1, the interchange is valid. For $j = 0, 1, \ldots, 2k-2$, the second result follows from the fact that the above bound is uniform in x.

Turning to the final result of this lemma, for any $j = 0, 1, \ldots, 2k-2$,

$$\begin{aligned}
\int |R_m^{(j)}(x)|\,\mathrm{d}x &= \int \Big| \iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2+j)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\Big|\,\mathrm{d}x \\
&\leq \iiiint_0^1 z^2\frac{1}{s}\phi(\frac{z}{s})\big|\alpha^{(2+j)}(x - t\frac{z}{\sqrt{m}}, s)\big|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}x \\
&= \iiiint_0^1 z^2\frac{1}{s}\phi(\frac{z}{s})\big|f^{(2+j)}(x - t\frac{z}{\sqrt{m}})g(s|x - t\frac{z}{\sqrt{m}}) + \\
&\qquad \cdots + f(x - t\frac{z}{\sqrt{m}})g^{(2+j)}(s|x - t\frac{z}{\sqrt{m}})\big|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}x \\
&\leq \iiiint_0^1 z^2\frac{1}{s}\phi(\frac{z}{s})I(0 < s < B_S)B_g\big(\big|f^{(2+j)}(x - t\frac{z}{\sqrt{m}})\big| + \\
&\qquad \cdots + \big|f(x - t\frac{z}{\sqrt{m}})\big|\big)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\,\mathrm{d}x \\
&\leq \iiint_0^1 z^2\frac{1}{s}\phi(\frac{z}{s})I(0 < s < B_S)B_g \int \big(\big|f^{(2+j)}(x - t\frac{z}{\sqrt{m}})\big| + \\
&\qquad \cdots + \big|f(x - t\frac{z}{\sqrt{m}})\big|\big)\,\mathrm{d}x\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s.
\end{aligned}$$

Since $f^{(j)}(\cdot)$ is integrable for $j = 0, \ldots, 2k$, the above bound is finite and uniform in c.

□

The following two lemmas concern the two bias components in (2.11) and are useful for the mse result.

Lemma 6 Assume A1-A5 and A6(2). Also assume that $m \to \infty$ as $c \to \infty$. Then

$$f_m(x) = f(x) + \frac{1}{m}\frac{1}{2}\int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s + o(\frac{1}{m}).$$

If in addition A6(4) is assumed, then

$$f_m(x) = f(x) + \frac{1}{m}\frac{1}{2}\int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s + \frac{1}{m^2}\frac{1}{8}\int s^4 \alpha^{(4)}(x, s)\,\mathrm{d}s + o(\frac{1}{m^2}).$$

Proof: We present only the proof to the second result. The proof of the first result is similar. First note by Assumption A6(4),

$$\lim_{c\to\infty} \alpha^{(4)}(x - t\frac{z}{\sqrt{m}}, s) = \alpha^{(4)}(x, s).$$

By Lemma 4,

$$\begin{aligned} f_m(x) = f(x) \;&+\; \frac{1}{m}\frac{1}{2}\int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s \\ &+\; \frac{1}{m^2}\frac{1}{3!}\iiint_0^1 (1-t)^3 z^4 \frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s. \end{aligned}$$

The integrand in the last term can be written

$$I(0 < t < 1)I(0 < s < B_S)(1-t)^3 z^4 \frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x - t\frac{z}{\sqrt{m}}, s),$$

and is bounded in absolute value by the integrable function

$$I(0 < t < 1)I(0 < s < B_S) d B_f B_g z^4 \frac{1}{s}\phi(\frac{z}{s}),$$

where d is some constant. Then by Lebesgue's dominated convergence theorem,

$$\begin{aligned}
&\lim_{c\to\infty} \frac{1}{3!} \iiint_0^1 (1-t)^3 z^4 \frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&= \frac{1}{3!} \iiint_0^1 (1-t)^3 z^4 \frac{1}{s}\phi(\frac{z}{s}) \lim_{c\to\infty} \alpha^{(4)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&= \frac{1}{3!} \iiint_0^1 (1-t)^3 z^4 \frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&= \frac{1}{3!} \int \left(\int_0^1 (1-t)^3\,\mathrm{d}t\right)\left(\int z^4 \frac{1}{s}\phi(\frac{z}{s})\,\mathrm{d}z\right)\alpha^{(4)}(x,s)\,\mathrm{d}s \\
&= \frac{1}{3!}\frac{1}{4}\frac{4!}{2!4}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s \\
&= \frac{1}{8}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s.
\end{aligned}$$

$\square$

Lemma 7 Assume A1-A5 and A6(4). Also assume

1. K is a probability distribution function symmetric about zero with finite second moment;
2. $m \to \infty$ and $h \to 0$ as $c \to \infty$.

Then

$$\mathrm{E}\big(\hat{f}(x;m,n,h)\big) - f_m(x) = h^2\frac{1}{2}f''(x)\int u^2K(u)\,\mathrm{d}u + o(h^2).$$

If in addition A6(6) is assumed and K has a finite fourth moment, then

$$\begin{aligned}
&\mathrm{E}\big(\hat{f}(x;m,n,h)\big) - f_m(x) = h^2\frac{1}{2}f^{(2)}(x)\int u^2K(u)\,\mathrm{d}u \\
&+ \frac{h^2}{m}\frac{1}{4}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u + h^4\frac{1}{24}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u \\
&+ o\left(\frac{h^2}{m} + h^4\right).
\end{aligned}$$

Proof: We present only the proof to the second result. The proof of the first result is similar. As in the proof of Theorem 1,

$$
\begin{aligned}
\mathrm{E}(\hat{f}(x;m,n,h)) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\
&= \mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) \\
&= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y \\
&= \int K(u) f_m(x-uh)\,\mathrm{d}u.
\end{aligned}
$$

By Lemma 5, f_m is four times differentiable and the fourth derivative is bounded. Then by Taylor's theorem with integral remainder

$$
\begin{aligned}
\int K(u)f_m(x-uh)\,\mathrm{d}u &= f_m(x)\int K(u)\,\mathrm{d}u - hf_m^{(1)}(x)\int uK(u)\,\mathrm{d}u \\
&\quad + h^2\frac{1}{2}f_m^{(2)}(x)\int u^2K(u)\,\mathrm{d}u + h^3\frac{1}{6}f_m^{(3)}(x)\int u^3K(u)\,\mathrm{d}u \\
&\quad + h^4\frac{1}{6}\int\int_0^1 u^4K(u)(1-t)^3 f_m^{(4)}(x-tuh)\,\mathrm{d}t\,\mathrm{d}u.
\end{aligned}
$$

Since K is a probability distribution symmetric about zero, K integrates to one and its odd moments are zero. Then

$$
\begin{aligned}
\int K(u)f_m(x-uh)\,\mathrm{d}u = f_m(x) + h^2\frac{1}{2}f_m^{(2)}(x)\int u^2K(u)\,\mathrm{d}u \\
+ h^4\frac{1}{6}\int\int_0^1 u^4K(u)(1-t)^3 f_m^{(4)}(x-tuh)\,\mathrm{d}t\,\mathrm{d}u. \quad (2.23)
\end{aligned}
$$

Again by Lemma 5,

$$
f_m^{(2)}(x) = f^{(2)}(x) + \frac{1}{m}\int\int\int_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x-t\frac{z}{\sqrt{m}},s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s.
$$

As in the proof of Lemma 5, the integrand in the above expression can be bounded for all c in absolute value by an integrable function. Therefore by Lebesgue's dominated convergence theorem and by continuity

$$\begin{aligned}
&\lim_{c\to\infty}\iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x-t\frac{z}{\sqrt{m}},s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\\
&=\iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(4)}(x,s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\\
&=\frac{1}{2}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s.
\end{aligned}$$

It follows that the second term in the right hand side of (2.23) is

$$\begin{aligned}
h^2\frac{1}{2}f_m^{(2)}(x)\int u^2K(u)\,\mathrm{d}u &= h^2\frac{1}{2}f^{(2)}(x)\int u^2K(u)\,\mathrm{d}u\\
&\quad+\frac{h^2}{m}\frac{1}{4}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u+o\left(\frac{h^2}{m}\right).
\end{aligned}$$

Consider the integral in the last term in the right hand side of (2.23). By Lemma 5,

$$f_m^{(4)}(x-tuh)=f^{(4)}(x-tuh)+\frac{1}{m}\iiint_0^1(1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(6)}(x-tuh-t'\frac{z}{\sqrt{m}},s)\,\mathrm{d}t'\,\mathrm{d}z\,\mathrm{d}s,$$

and

$$\iiint_0^1(1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(6)}(x-tuh-t'\frac{z}{\sqrt{m}},s)\,\mathrm{d}t'\,\mathrm{d}z\,\mathrm{d}s,$$

is bounded for all c. Then

$$\lim_{c\to\infty}f_m^{(4)}(x-tuh)=f^{(4)}(x).$$

Since $f_m^{(4)}$ is bounded, the integrand in the last term in the right hand side of (2.23) is bounded in absolute value for all c by an integrable function. Lebesgue's dominated convergence theorem then gives

$$\begin{aligned}
\lim_{c\to\infty}\iint_0^1 u^4K(u)(1-t)^3f_m^{(4)}(x-tuh)\,\mathrm{d}t\,\mathrm{d}u &= \iint_0^1 u^4K(u)(1-t)^3f^{(4)}(x)\,\mathrm{d}t\,\mathrm{d}u\\
&=\frac{1}{4}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u.
\end{aligned}$$

It follows that the last term in the right hand side of (2.23) is

$$h^4\frac{1}{6}\iint_0^1 u^4K(u)(1-t)^3f_m^{(4)}(x-tuh)\,\mathrm{d}t\,\mathrm{d}u=h^4\frac{1}{24}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u+o(h^4).$$

The result follows.

□

Now we prove the result on the asymptotic expression of mse.

Proof of Theorem 2: Recall from (2.9) that

$$\mathrm{mse}(\hat{f}(x;m,n,h)) = \mathrm{bias}^2(\hat{f}(x;m,n,h)) + \mathrm{var}(\hat{f}(x;m,n,h)) \quad \forall x \in \mathbb{R}$$

Since K is a bounded probability distribution function, K and K^2 are integrable. Then by Lemma 2,

$$\lim_{c\to\infty} \int \frac{1}{h} K^2\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y = f(x) \int K^2(u)\,\mathrm{d}u,$$

and

$$\lim_{c\to\infty} \left(\int \frac{1}{h} K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y\right)^2 = f^2(x) \left(\int K(u)\,\mathrm{d}u\right)^2.$$

From (2.19) in the proof of Theorem 1

$$\begin{aligned}\mathrm{var}(\hat{f}(x;m,n,h)) &= \frac{1}{nh}\mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) - \frac{1}{n}\left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2 \\ &= \frac{1}{nh}\int \frac{1}{h}K^2\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y + \frac{1}{n}\left(\int \frac{1}{h}K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y\right)^2.\end{aligned}$$

Then

$$\mathrm{var}(\hat{f}(x;m,n,h)) = \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).$$

Recall from (2.11),

$$\mathrm{bias}(\hat{f}(x;m,n,h)) = (\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x))$$

Then by Lemmas 6 and 7,

$$\begin{aligned}\mathrm{bias}(\hat{f}(x;m,n,h)) &= h^2 \frac{1}{2} f''(x)\,\mathrm{d}x \int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2}\int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s \\ &+ o\left(h^2 + \frac{1}{m}\right).\end{aligned}$$

The result follows.

$\square$

A brief outline of the proof of Theorem 3 follows. Recall the decomposition of $\mathrm{mise}(\hat{f}(\cdot; m, n, h))$ in (2.10). In (2.11) the bias term was further decomposed. The bias term in $\mathrm{mise}(\hat{f}(\cdot; m, n, h))$ is then given by

$$\int \mathrm{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x = \int ((\mathrm{E}(\hat{f}(x; m, n, h)) - f_m(x)) + (f_m(x) - f(x)))^2\,\mathrm{d}x$$

$$= \int (\mathrm{E}(\hat{f}(x; m, n, h)) - f_m(x))^2\,\mathrm{d}x \tag{2.24}$$

$$+2 \int (\mathrm{E}(\hat{f}(x; m, n, h)) - f_m(x))(f_m(x) - f(x))\,\mathrm{d}x \tag{2.25}$$

$$+ \int (f_m(x) - f(x))^2\,\mathrm{d}x. \tag{2.26}$$

In Lemma 10, the normalized limit for the variance term in the decomposition of $\mathrm{mise}(\hat{f}(\cdot; m, n, h))$ in (2.10) is computed. In Lemmas 11, 12, and 13, normalized limits are computed for the terms (2.24), (2.25), and (2.26), respectively. Theorem 3 follows immediately from Lemmas 10, 11, 12, and 13. Before stating and proving Lemmas 10, 11, 12, and 13, a couple of useful lemmas are presented first including the following lemma in which the assumption that f'' is ultimately monotone is introduced. A function γ whose domain is the real line is said to ultimately monotone if there exists a $B > 0$ such that γ is monotone on $[B, \infty)$ and monotone on $(-\infty, -B)$. This assumption is useful in satisfying the assumptions for Lebesgue's dominated convergence theorem which is used in Lemmas 12 and 13.

Lemma 8 Assume

1. $f''(\cdot)$ is ultimately monotone;

2. $f'(\cdot)$ and $f''(\cdot)$ are integrable;

3. $f''(\cdot)$ is continuous.

Then

1. for $j = 1, 2$ and $k = 1, 2$, $\exists\, C_{jk} \geq 0$ and an integrable function $\tilde{f}_{jk}$ such that for all $c > C_{jk}$,

$$\iiint_0^1 I(0 < s \leq B_S) z^2 \frac{1}{s}\, \phi\left(\frac{z}{s}\right) |f^{(k)}(x - t\frac{z}{\sqrt{m}})^j|\, \mathrm{d}t\, \mathrm{d}z\, \mathrm{d}s \leq \tilde{f}_{jk}(x) \qquad \forall x \in \mathbb{R};$$

2. for $j = 1, 2$, $\exists\, C_j \geq 0$ and an integrable function $\tilde{h}_j$ such that for all $c > C_j$,

$$\iiint_0^1 I(0 < s \leq B_S) z^2 \frac{1}{s}\, \phi\left(\frac{z}{s}\right) f(x - \frac{z}{\sqrt{m}})^j\, \mathrm{d}t\, \mathrm{d}z\, \mathrm{d}s \leq \tilde{h}_j(x) \qquad \forall x \in \mathbb{R}.$$

Proof: Assumption 1 implies that $f''(\cdot)^2$ is ultimately monotone. Assumption 1 also implies that $f(\cdot)$ and $f'(\cdot)$ are ultimately monotone so that $f(\cdot)^2$ and $f'(\cdot)^2$ are ultimately monotone. By Assumption 2, $f''(\cdot)$ is integrable. It then follows that $\exists B_u > 0$ such that on the set $[B_u, \infty)$, $|f''(x)|$ is nonincreasing and on the set $(-\infty, -B_u]$, $|f''(x)|$ is nondecreasing. That is to say for all x_1 and x_2 such that $B_u \leq x_1 < x_2$, $|f''(x_1)| \geq |f''(x_2)|$ and for all x_1 and x_2 such that $x_1 < x_2 \leq -B_u$, $|f''(x_1)| \leq |f''(x_2)|$. It follows that the function $f''(\cdot)^2$ exhibits similar behavior. In the same way, it can be shown that the functions $f(\cdot)$, $|f'(\cdot)|$, $f(\cdot)^2$, and $f'(\cdot)^2$ behave similarly.

Note that this behavior together with the continuity of Assumption 3 implies that $f(\cdot)$, $f(\cdot)^2$, $f'(\cdot)$, $f'(\cdot)^2$, and $f''(\cdot)$, $f''(\cdot)^2$ are bounded. Also note that the above behavior along with the integrability in Assumption 2 implies that $f(\cdot)$, $f'(\cdot)$ and $f''(\cdot)$ are square integrable.

Consider the first result for $j = 1$ and $k = 2$. In this case

$$
\begin{aligned}
&\iiint_0^1 I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) |f''(x - t\frac{z}{\sqrt{m}})|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
=&\iiint_0^1 I(z \ge 0) I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) |f''(x - t\frac{z}{\sqrt{m}})|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \qquad (2.27)\\
&+\iiint_0^1 I(z < 0) I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) |f''(x - t\frac{z}{\sqrt{m}})|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \qquad (2.28)\\
=&\, A(x) + B(x),
\end{aligned}
$$

where $A(x)$ and $B(x)$ are given by (2.27) and (2.28), respectively. Since $t \in (0, 1)$ and $m \to \infty$ as $c \to \infty$, $\exists C$ such that for all $c > C$, $t/\sqrt{m} < 1/2$. Therefore, if $z \ge 0$ and $x > B_u + (1/2)z$, then

$$|f''(x - t\frac{z}{\sqrt{m}})| \le |f''(x - \frac{1}{2}z)|.$$

If $z \ge 0$ and $x < -B_u$, then

$$|f''(x - t\frac{z}{\sqrt{m}})| \le |f''(x)|.$$

Similarly, if $z < 0$ and $x > B_u$, then

$$|f''(x - t\frac{z}{\sqrt{m}})| \le |f''(x)|.$$

If $z < 0$ and $x < -B_u + (1/2)z$, then

$$|f''(x - t\frac{z}{\sqrt{m}})| \le |f''(x - \frac{1}{2}z)|.$$

It was established above that f'' is bounded. Let $B_{f''}$ be this bound. Then for all $c > C$ and for all $x \in \mathbb{R}$,

$$
\begin{aligned}
A(x) \le& \iiint_0^1 I(x < -B_u) I(z \ge 0) I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) |f''(x)|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&+ B_{f''} \iiint_0^1 I(-B_u \le x \le B_u + \frac{1}{2}z) I(z \ge 0) I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&+ \iiint_0^1 I(B_u + \frac{1}{2}z < x) I(z \ge 0) I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) |f''(x - \frac{1}{2}z)|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s,
\end{aligned}
\qquad (2.29)
$$

and

$$B(x) \leq \iiint_0^1 I(x < -B_u + \frac{1}{2}z)I(z < 0)I(0 < s \leq B_S)z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right)|f''(x - \frac{1}{2}z)|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s$$
$$+ B_{f''} \iiint_0^1 I(-B_u + \frac{1}{2}z \leq x \leq B_u)I(z < 0)I(0 < s \leq B_S)z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s$$
$$+ \iiint_0^1 I(B_u < x)I(z < 0)I(0 < s \leq B_S)z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right)|f''(x)|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s. \qquad (2.30)$$

Let $\hat{A}(x)$ be the upper bound of $A(x)$ given in (2.29). Let $\hat{B}(x)$ be the upper bound of $B(x)$ given in (2.30). For all $x \in \mathbb{R}$, let $\tilde{f}_{12}(x) = \hat{A}(x) + \hat{B}(x)$. Note that $f''(\cdot)$ is integrable. It follows that $\tilde{f}_{12}(\cdot)$ is integrable which proves this case for the first result. The other cases are similar.

□

In the following lemma, sufficient conditions are given for a limit used in Lemmas 12 and 13.

Lemma 9 Assume A1-A5 and A6(2). Also assume that $m \to \infty$ as $c \to \infty$. Then for all $x \in \mathbb{R}$

$$\lim_{c\to\infty} R_m^{(0)}(x) = \frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s,$$

where $R_m^{(0)}(x)$ is defined in (2.21).

Proof: Let $x \in \mathbb{R}$. From (2.21),

$$R_m^{(0)}(x) = \iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2)}(x - t\frac{z}{\sqrt{m}}, s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s.$$

Bounding the integrand on the left hand side as in the proof of Lemma 5 allows us to apply Lebesgue's dominated convergence theorem:

$$
\begin{aligned}
&\lim_{c\to\infty}\iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2)}(x-t\frac{z}{\sqrt{m}},s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&= \iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\lim_{c\to\infty}\alpha^{(2)}(x-t\frac{z}{\sqrt{m}},s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&= \iiint_0^1 (1-t)z^2\frac{1}{s}\phi(\frac{z}{s})\alpha^{(2)}(x,s)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\
&= \frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s.
\end{aligned}
$$

□

Recall the decomposition of $\mathrm{mise}(\hat{f}(\cdot;m,n,h))$ in (2.10) and the further decomposition of the bias component in (2.25). In the following lemmas, the normalized limits of the components of $\mathrm{mise}(\hat{f}(\cdot;m,n,h))$ are computed. In the first of these lemmas, the variance term is considered.

Lemma 10 Assume A1-A5 and A6(3). Also assume

1. $f^{(k)}(\cdot)$ is integrable for $k=1,2,3$;

2. K is a bounded probability distribution function with finite first moment;

3. $m\to\infty$, $n\to\infty$, $h\to 0$ and $nh\to\infty$ as $c\to\infty$.

Then

$$\lim_{c\to\infty} nh\int \mathrm{var}(\hat{f}(x;m,n,h))\,\mathrm{d}x = \int K(u)^2\,\mathrm{d}u$$

Proof: As in (2.19),

$$\mathrm{var}(\hat{f}(x;m,n,h)) = \frac{1}{nh}\mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) - \frac{1}{n}\left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2.$$

Therefore

$$
\begin{aligned}
\lim_{c\to\infty} nh\int \mathrm{var}(\hat{f}(x;m,n,h))\,\mathrm{d}x &= \lim_{c\to\infty}\int \mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\mathrm{d}x \\
&\quad -\lim_{c\to\infty} h\int\left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2\mathrm{d}x. \qquad (2.31)
\end{aligned}
$$

Observe that

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right) &= \int \frac{1}{h}K\left(\frac{x-y}{h}\right) f_m(y)\,\mathrm{d}y \\
&= \int K(u) f_m(x-uh)\,\mathrm{d}u.
\end{aligned}$$

Therefore, the above expectation is bounded, say by B, for all c since by Lemma 5, $f_m(\cdot)$ is bounded for all c and K is a probability density. Also, the integral

$$\begin{aligned}
\int \mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\mathrm{d}x &= \iint K(u) f_m(x-uh)\,\mathrm{d}u\,\mathrm{d}x \\
&= \int K(u) \int f_m(x-uh)\,\mathrm{d}x\,\mathrm{d}u \\
&= \int f_m(x)\,\mathrm{d}x \int K(u)\,\mathrm{d}u \\
&= 1
\end{aligned}$$

since f_m and K are probability densities. Therefore, for all c,

$$\int \left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2 \mathrm{d}x \leq \int B\,\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\mathrm{d}x = B.$$

Since $h \to 0$ as $c \to \infty$,

$$\lim_{c\to\infty} h \int \left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\right)^2 \mathrm{d}x = 0. \tag{2.32}$$

As for

$$\lim_{c\to\infty} \int \mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\mathrm{d}x,$$

similar to the above,

$$\int \mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\mathrm{d}x = \iint K^2(u) f_m(x-uh)\,\mathrm{d}u\,\mathrm{d}x.$$

By Lemma 5, f_m is differentiable and f'_m is bounded so that Taylor's theorem with integral remainder gives

$$\begin{aligned}
\iint K^2(u) f_m(x-uh)\,\mathrm{d}u\,\mathrm{d}x &= \int f_m(x)\,\mathrm{d}x \int K^2(u)\,\mathrm{d}u \\
&\quad - h \iiint_0^1 uK^2(u) f'_m(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}x
\end{aligned}$$

Note that

$$\begin{aligned}
\left|\iiint_0^1 uK^2(u)f_m'(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}x\right| &\leq \iiint_0^1 |u|K^2(u)|f_m'(x-vuh)|\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}x \\
&= \iint_0^1 |u|K^2(u)\int |f_m'(x-vuh)|\,\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}u \\
&= \int |f_m'(x)|\,\mathrm{d}x \int |u|K^2(u)\,\mathrm{d}u.
\end{aligned}$$

By Assumption 2, $\int |u|K^2(u)\,\mathrm{d}u$ is finite. It follows from Lemma 5 that $\int |f_m'(x)|\,\mathrm{d}x$ is bounded for all c. Therefore,

$$\lim_{c\to\infty} h\iiint_0^1 uK^2(u)f_m'(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}x = 0, \tag{2.33}$$

since $h \to 0$ as $c \to \infty$. Now consider

$$\lim_{c\to\infty}\int f_m(x)\,\mathrm{d}x\int K^2(u)\,\mathrm{d}u = \lim_{c\to\infty}\left(\int f(x)\,\mathrm{d}x + \frac{1}{m}\int R_m^{(0)}(x)\,\mathrm{d}x\right)\int K^2(u)\,\mathrm{d}u.$$

By Assumption 2, $\int K^2(u)\,\mathrm{d}u$ is finite. By Lemma 5, it was shown that $\int |R_m^{(0)}(x)|\,\mathrm{d}x$ is bounded for all c. Since $m \to \infty$ as $c \to \infty$,

$$\begin{aligned}
\lim_{c\to\infty}\int f_m(x)\,\mathrm{d}x\int K^2(u)\,\mathrm{d}u &= \int f(x)\,\mathrm{d}x\int K^2(u)\,\mathrm{d}u \\
&= \int K^2(u)\,\mathrm{d}u. \qquad (2.34)
\end{aligned}$$

Therefore by (2.33) and (2.34),

$$\lim_{c\to\infty}\int \mathrm{E}\left(\frac{1}{h}K^2\left(\frac{x-\bar{X}_m(Z_i)}{h}\right)\right)\mathrm{d}x = \int K^2(u)\,\mathrm{d}u. \tag{2.35}$$

Substituting (2.32) and (2.35) into (2.31) gives the desired result.

□

Now consider $\int \mathrm{bias}^2(\hat{f}(x;m,n,h))\,\mathrm{d}x$, which is the first term of the mise expression in (2.10). In the following three lemmas, the normalized limits of the components of this term given by (2.24), (2.25), and (2.26) are computed.

Lemma 11 Assume A1-A5 and A6(4). Also assume

1. $f^{(k)}(\cdot)$ is integrable for $k = 1, 2, 3, 4$;
2. K is a probability distribution function symmetric about zero with finite second moment;
3. $m \to \infty$ and $h \to 0$ as $c \to \infty$.

Then

$$\lim_{c\to\infty} \frac{1}{h^4} \int (\mathrm{E}\big(\hat{f}(x; m, n, h)\big) - f_m(x))^2 \,\mathrm{d}x = \int \left(\frac{1}{2} \left(\int u^2 K(u) \,\mathrm{d}u \right) f''(x) \right)^2 \mathrm{d}x.$$

Proof: From Lemma 5, f_m is twice differentiable and f''_m is bounded so that by Taylor's theorem with integral remainder, for all $x \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{E}\big(\hat{f}(x; m, n, h)\big) &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f_m(y) \,\mathrm{d}y \\
&= \int K(u) f_m(x - uh) \,\mathrm{d}u \\
&= f_m(x) \int K(u) \,\mathrm{d}u - h f'_m(x) \int u K(u) \,\mathrm{d}u \\
&\quad + h^2 \iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh) \,\mathrm{d}v \,\mathrm{d}u.
\end{aligned}$$

By Assumption 2,

$$\mathrm{E}\big(\hat{f}(x; m, n, h)\big) - f_m(x) = h^2 \iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh) \,\mathrm{d}v \,\mathrm{d}u.$$

Therefore

$$\begin{aligned}
&\lim_{c\to\infty} \frac{1}{h^4} \int (\mathrm{E}\big(\hat{f}(x; m, n, h)\big) - f_m(x))^2 \,\mathrm{d}x \\
&\qquad = \lim_{c\to\infty} \int \left(\iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh) \,\mathrm{d}v \,\mathrm{d}u \right)^2 \mathrm{d}x.
\end{aligned}$$

By Fatou's lemma

$$\liminf_{c\to\infty} \int \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u \right)^2 \mathrm{d}x$$
$$\geq \int \liminf_{c\to\infty} \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u \right)^2 \mathrm{d}x. \quad (2.36)$$

It follows from Lemma 5 that $|f_m''(\cdot)|$ is bounded for all c. Then by Lebesgue's dominated convergence theorem

$$\lim_{c\to\infty} \iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u$$
$$= \iint_0^1 \lim_{c\to\infty} u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u$$
$$= \iint_0^1 u^2 K(u)(1-v) \lim_{c\to\infty} f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u. \quad (2.37)$$

By Lemma 5,

$$f_m''(x - vuh) = f''(x - vuh) + \frac{1}{m} R_m^{(2)}(x - vuh).$$

By Assumption A6(4), f'' is continuous. By Lemma 5, $|R_m^{(2)}(\cdot)|$ is bounded for all c. By Assumption 8, $h \to 0$ and $1/m \to 0$ as $c \to \infty$. So

$$\lim_{c\to\infty} f_m''(x - vuh) = f''(x).$$

By substituting into (2.37),

$$\lim_{c\to\infty} \iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u = \frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u \right) f''(x).$$

It follows that

$$\lim_{c\to\infty} \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u \right)^2 = \left(\frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u \right) f''(x) \right)^2.$$

By substituting into (2.36),

$$\liminf_{c\to\infty} \int \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u \right)^2 \mathrm{d}x$$
$$\geq \int \left(\frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u \right) f''(x) \right)^2 \mathrm{d}x.$$

By Assumptions A6(4) and 1, f'' is square integrable.

If

$$\limsup_{c\to\infty} \int \left(\iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh)\, \mathrm{d}v\, \mathrm{d}u \right)^2 \mathrm{d}x \\ \leq \int \left(\frac{1}{2} \left(\int u^2 K(u)\, \mathrm{d}u \right) f''(x) \right)^2 \mathrm{d}x,$$

then the result follows. By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\int \left(\iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh)\, \mathrm{d}v\, \mathrm{d}u \right)^2 \mathrm{d}x \\
&\leq \int \left(\iint_0^1 u^2 K(u)(1-v)\, \mathrm{d}v\, \mathrm{d}u \iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh)^2\, \mathrm{d}v\, \mathrm{d}u \right) \mathrm{d}x \\
&= \frac{1}{2} \int u^2 K(u)\, \mathrm{d}u \iiint_0^1 u^2 K(u)(1-v) f''_m(x - vuh)^2\, \mathrm{d}v\, \mathrm{d}u\, \mathrm{d}x \\
&= \frac{1}{2} \int u^2 K(u)\, \mathrm{d}u \iint_0^1 u^2 K(u)(1-v) \int f''_m(x - vuh)^2\, \mathrm{d}x\, \mathrm{d}v\, \mathrm{d}u \\
&= \frac{1}{2} \int f''_m(x)^2\, \mathrm{d}x \int u^2 K(u)\, \mathrm{d}u \iint_0^1 u^2 K(u)(1-v)\, \mathrm{d}v\, \mathrm{d}u \\
&= \frac{1}{4} \int f''_m(x)^2\, \mathrm{d}x \left(\int u^2 K(u)\, \mathrm{d}u \right)^2.
\end{aligned}$$

Therefore

$$\limsup_{c\to\infty} \int \left(\iint_0^1 u^2 K(u)(1-v) f''_m(x - vuh)\, \mathrm{d}v\, \mathrm{d}u \right)^2 \mathrm{d}x \\ \leq \frac{1}{4} \left(\limsup_{c\to\infty} \int f''_m(x)^2\, \mathrm{d}x \right) \left(\int u^2 K(u)\, \mathrm{d}u \right)^2. \quad (2.38)$$

By Lemma 5,

$$\begin{aligned}
f''_m(x)^2 &= (f''(x) + \frac{1}{m} R_m^{(2)}(x))^2 \\
&= f''(x)^2 + \frac{2}{m} f''(x) R_m^{(2)}(x) + \frac{1}{m^2} (R_m^{(2)}(x))^2.
\end{aligned}$$

By Assumption 1, $f''(\cdot)$ is integrable. By Lemma 5, $|R_m^{(2)}(\cdot)|$ is bounded for all c and $\int (R_m^{(2)}(x))^2\, \mathrm{d}x$ is bounded for all c. By Assumption 3, $1/m \to 0$ as $c \to \infty$.

Then

$$\lim_{c\to\infty} \int f_m''(x)^2 \,\mathrm{d}x = \int f''(x)^2 \,\mathrm{d}x,$$

which implies that

$$\limsup_{c\to\infty} \int f_m''(x)^2 \,\mathrm{d}x = \int f''(x)^2 \,\mathrm{d}x.$$

By substitution into (2.38),

$$\limsup_{c\to\infty} \int \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh) \,\mathrm{d}v \,\mathrm{d}u \right)^2 \mathrm{d}x$$
$$\leq \int \left(\frac{1}{2} \left(\int u^2 K(u) \,\mathrm{d}u \right) f''(x) \right)^2 \mathrm{d}x.$$

□

Lemma 12 Assume A1-A5 and A6(4). Also assume

1. $f''(\cdot)$ is ultimately monotone;
2. $f^{(k)}(\cdot)$ is integrable for $k = 1, 2, 3, 4$;
3. K is a probability distribution function symmetric about zero with finite second moment;
4. $m \to \infty$ and $h \to 0$ as $c \to \infty$.

Then

$$\lim_{c\to\infty} \frac{m}{h^2} \int 2(\mathrm{E}\big(\hat{f}(x; m, n, h)\big) - f_m(x))(f_m(x) - f(x)) \,\mathrm{d}x$$
$$= \int \frac{1}{2} \left(\int u^2 K(u) \,\mathrm{d}u \right) f''(x) \int s^2 \alpha^{(2)}(x, s) \,\mathrm{d}s \,\mathrm{d}x.$$

Proof: By Lemma 4,

$$f_m(x) - f(x) = \frac{1}{m} R_m^{(0)}(x).$$

As in the proof of Lemma 11,

$$\mathrm{E}\big(\hat{f}(x;m,n,h)\big) - f_m(x) = h^2 \iint_0^1 u^2 K(u)(1-v) f_m''(x - vuh)\,\mathrm{d}v\,\mathrm{d}u.$$

Then

$$\begin{aligned}&\frac{m}{h^2}\int 2(\mathrm{E}\big(\hat{f}(x;m,n,h)\big) - f_m(x))(f_m(x) - f(x))\,\mathrm{d}x\\ &\qquad = 2\int \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\right) R_m^{(0)}(x)\,\mathrm{d}x.\end{aligned}$$

It will be shown that $\exists \tilde{f}(\cdot)$ such that $\tilde{f}(\cdot)$ is integrable and for all $x \in \mathbb{R}$ and $c > C$ where C is some nonnegative number,

$$|\left(\iint_0^1 u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\right) R_m^{(0)}(x)| \leq \tilde{f}(x), \tag{2.39}$$

then by Lebesgue's dominated convergence theorem

$$\begin{aligned}&\lim_{c\to\infty} \frac{m}{h^2}\int 2(\mathrm{E}\big(\hat{f}(x;m,n,h)\big) - f_m(x))(f_m(x) - f(x))\,\mathrm{d}x\\ &\qquad = 2\lim_{c\to\infty}\int \left(\iint_0^1 u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\right) R_m^{(0)}(x)\,\mathrm{d}x\\ &\qquad = 2\int \lim_{c\to\infty}\left(\iint_0^1 u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\right) R_m^{(0)}(x)\,\mathrm{d}x.\end{aligned}$$

From Lemma 9, $\lim_{c\to\infty} R_m^{(0)}(x) = (1/2)\int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s$. By Lemma 5, $\exists B_{fR} > 0$ such that $|f_m''(\cdot)| \leq B_{fR}$ for all c. Then by Lebesgue's dominated convergence theorem,

$$\lim_{c\to\infty}\iint_0^1 u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u = \iint_0^1 \lim_{c\to\infty} u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u.$$

As in the proof of Lemma 11, $\lim_{c\to\infty} f_m''(x - vuh) = f''(x)$ so that

$$\lim_{c\to\infty}\iint_0^1 u^2 K(u)(1-v) f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u = \frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x).$$

It follows that

$$\begin{aligned}&\lim_{c\to\infty} \frac{m}{h^2}\int 2(\mathrm{E}\big(\hat{f}(x;m,n,h)\big) - f_m(x))(f_m(x) - f(x))\,\mathrm{d}x\\ &\qquad = \int \frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x) \int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s\,\mathrm{d}x.\end{aligned}$$

As for the validity of (2.39), observe that

$$
\begin{aligned}
&|\left(\iint_0^1 u^2K(u)(1-v)f_m''(x-vuh)\,\mathrm{d}v\,\mathrm{d}u\right)\ R_m^{(0)}(x)| \\
&\qquad\qquad\qquad \le B_{fR}\left(\iint_0^1 u^2K(u)(1-v)\,\mathrm{d}v\,\mathrm{d}u\right)\ |R_m^{(0)}(x)| \\
&\qquad\qquad\qquad = \frac{1}{2}B_{fR}\left(\int u^2K(u)\,\mathrm{d}u\right)\ |R_m^{(0)}(x)|.
\end{aligned}
$$

As in the proof of the last part of Lemma 5,

$$
\begin{aligned}
|R_m^{(0)}(x)| \ \le\ & \iiint_0^1 z^2\frac{1}{s}\phi(\frac{z}{s})I(0<s<B_S)B_g\big(\big|f''(x-t\frac{z}{\sqrt{m}})\big|+ \\
& \cdots+\big|f(x-t\frac{z}{\sqrt{m}})\big|\big)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s.
\end{aligned}
$$

The validity of (2.39) then follows from Lemma 8.

□

Lemma 13 Assume A1-A5 and A6(2). Also assume

1. $f''(\cdot)$ is ultimately monotone;
2. $f'(\cdot)$, $f''(\cdot)$ are integrable;
3. $m \to \infty$ as $c \to \infty$;

Then

$$
\lim_{c\to\infty} m^2\int(f_m(x)-f(x))^2\,\mathrm{d}x = \int\left(\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s\right)^2\,\mathrm{d}x.
$$

Proof: Observe that

$$
m^2\int(f_m(x)-f(x))^2\,\mathrm{d}x = \int(R_m^{(0)}(x))^2\,\mathrm{d}x
$$

where $R_m^{(0)}(\cdot)$ is defined in Lemma 5. It will be shown that $\exists\tilde{f}(\cdot)$ such that $\tilde{f}(\cdot)$ is integrable and for all $x\in\mathbb{R}$ and $c>C$ where C is some nonnegative number,

$$
(R_m^{(0)}(x))^2 \le \tilde{f}(x), \tag{2.40}
$$

then by Lebesgue's dominated convergence theorem

$$\begin{aligned}\lim_{c\to\infty} m^2 \int (f_m(x) - f(x))^2 \,\mathrm{d}x &= \lim_{c\to\infty} \int (R_m^{(0)}(x))^2 \,\mathrm{d}x \\ &= \int \lim_{c\to\infty} (R_m^{(0)}(x))^2 \,\mathrm{d}x \\ &= \int \left(\frac{1}{2}\int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s\right)^2 \mathrm{d}x.\end{aligned}$$

As for the validity of (2.40), from the proof of the last part of Lemma 5 observe that

$$\begin{aligned}(R_m^{(0)})^2 &\le \Big(\iiint_0^1 z^2 \frac{1}{s}\phi(\frac{z}{s}) I(0 < s < B_S) B_g \big(\big|f''(x - t\frac{z}{\sqrt{m}})\big| + \\ &\quad \cdots + \big|f(x - t\frac{z}{\sqrt{m}})\big|\big)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\Big)^2 \\ &\le d\Big(\iiint_0^1 z^2 \frac{1}{s}\phi(\frac{z}{s}) I(0 < s < B_S) B_g \big|f''(x - t\frac{z}{\sqrt{m}})\big|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\Big)^2 \\ &\quad \cdots + d\Big(\iiint_0^1 z^2 \frac{1}{s}\phi(\frac{z}{s}) I(0 < s < B_S) B_g \big|f(x - t\frac{z}{\sqrt{m}})\big|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\Big)^2,\end{aligned}$$

where d is some positive constant. Consider the first term on the right hand side of the inequality. By the Cauchy-Schwarz inequality,

$$\begin{aligned}&d\Big(\iiint_0^1 z^2 \frac{1}{s}\phi(\frac{z}{s}) I(0 < s < B_S) B_g \big|f''(x - t\frac{z}{\sqrt{m}})\big|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\Big)^2 \\ &\le dB_g^2 \iiint_0^1 I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right)\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\ &\qquad \times \iiint_0^1 I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) f''(x - t\frac{z}{\sqrt{m}})^2\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s \\ = \;& dB_g^2 \int I(0 < s \le B_S) s^2\,\mathrm{d}s \\ &\qquad \times \iiint_0^1 I(0 < s \le B_S) z^2 \frac{1}{s}\,\phi\left(\frac{z}{s}\right) f''(x - t\frac{z}{\sqrt{m}})^2\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s.\end{aligned}$$

The integral $\int I(0 < s \le B_S)s^2\,\mathrm{d}s$ is finite. Then by Lemma 8, $\exists\ C_1 \ge 0$ and an integrable function $\tilde{g}_1$ such that for all $c > C_1$,

$$d\Big(\iiint_0^1 z^2 \frac{1}{s}\phi(\frac{z}{s}) I(0 < s < B_S) B_g \big|f''(x - t\frac{z}{\sqrt{m}})\big|\,\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}s\Big)^2 \le \tilde{g}_1(x) \qquad \forall x \in \mathbb{R}.$$

The other terms are similar and it follows that (2.40) holds.

□

2.3 A Local Kernel Density Estimate

In this section, we introduce a local kernel estimate for the density of a conditional expectation. We first motivate and give some background on the local estimator in the standard density estimation setting. We then present a local estimator for our setting and give some results on the convergence of the estimator's mse.

Quite often a density will exhibit very different levels of curvature over mutually exclusive convex sets in its domain. Consider the normal mixture

$$(1/2)N(-1/2, 4^{-2}) + (1/2)N(1/2, 1).$$

The density is

$$f(x) = \frac{1}{2}\left(\frac{4}{\sqrt{2\pi}}\exp\left(\frac{-16(x+1/2)^2}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(x-1/2)^2}{2}\right)\right).$$

The density is plotted in Figure 2.1. For the interval containing the mode corresponding to the normal component with low variance, the curvature is quite high. On the other hand, for the interval out in the right tail of the normal component with low variance, the curvature is very low. Suppose we have data generated from this normal mixture and we apply the naive kernel density estimator (1.2) discussed in Section 1.2. To distinguish this estimator from the local estimator, let us call this estimator the global kernel density estimator. Because of the differences in curvature, whatever bandwidth we choose we will likely either

1. oversmooth the interval with high curvature;

2. undersmooth the interval with low curvature; or

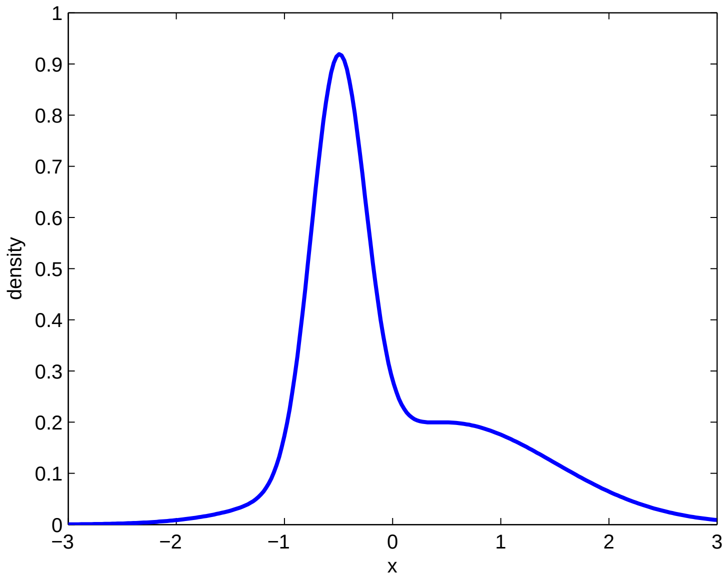

Figure 2.1: The density of the normal mixture $(1/2)N(-1/2, 4^{-2}) + (1/2)N(1/2, 1)$.

3. both oversmooth the interval with high curvature and undersmooth the interval with low curvature.

In this case we would like to choose different bandwidths for different locations in which we estimate the density. This is the idea of the local kernel density estimator denoted $\hat{g}_L$. Recall from (1.3) that the estimator has the form

$$\hat{g}_L(x; h(x)) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x)} K\left(\frac{x - Y_i}{h(x)}\right).$$

For the local kernel density estimator, bandwidth is a function of the point x where the target density g is being estimated, whereas for the global estimator given in (1.2), the bandwidth is constant. Viewed pointwise, the local estimator in (1.3) is just a standard kernel density estimator. But from a global perspective, the local kernel density estimator can be thought of as a continuum of individual global kernel density estimators with different bandwidths (Jones [1990]). Note that there is no guarantee that the local kernel density estimator for a finite sample will integrate to one so that the estimator itself may not be a proper probability density.

The intuition above that the bandwidth should be inversely proportional to curvature is reinforced theoretically. It turns out that the asymptotically optimal bandwidth $h(x)$ is proportional to $[g(x)/(g''(x))^2]^{1/5}$ (e.g., Jones [1990]). But note that even with the asymptotically optimal bandwidth the rate of mse and mise are no better than for the global kernel density estimator in (1.2). There is, however, an improvement (i.e., a decrease) in the multiplier of the rate (Jones [1990]).

A local kernel density estimator for $f(x)$, the density of $\mathrm{E}(X|Z)$ evaluated at x, is

$$\hat{f}_L(x; m, n, h(x)) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x)} K\left(\frac{x - \bar{X}_m(Z_i)}{h(x)} \right), \tag{2.41}$$

where, again,

$$\bar{X}_m(Z_i) = \frac{1}{m} \sum_{j=1}^{m} X_j(Z_i) \quad \text{for} \quad i = 1, \ldots, n.$$

Compare the local estimator (2.41) with the global estimator (2.1) introduced in Section 2.1. Considering that pointwise, the local kernel density estimator is the same as the global density estimator, the following mse results for $\hat{f}_L$ are immediate.

Theorem 4 Assume A1-A5 and A6(0). Also assume

1. K is a bounded probability density;

2. $m \to \infty$, $h(x) \to 0$, and $nh(x) \to \infty$, as $c \to \infty$ for all $x \in \mathbb{R}$.

Then for all $x \in \mathbb{R}$,

$$\lim_{c \to \infty} \mathrm{mse}(\hat{f}_L(x; m, n, h(x))) = 0.$$

Theorem 5 Assume A1-A5 and A6(4). Also assume

1. K is a bounded probability distribution function symmetric about zero with finite second moment;

2. $m \to \infty$, $n \to \infty$, $h(x) \to 0$, and $nh(x) \to \infty$ as $c \to \infty$ for all $x \in \mathbb{R}$.

Then for any $x \in \mathbb{R}$

$$\begin{aligned} \mathrm{mse}(\hat{f}_L(x,m,n,h(x))) &= \left(h(x)^2 \frac{1}{2} f''(x) \int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2} \int s^2 \alpha^{(2)}(x,s)\,\mathrm{d}s\right)^2 \\ &\quad + \frac{1}{nh(x)} f(x) \int K^2(u)\,\mathrm{d}u \\ &\quad + o\left(\left(h(x)^2 + \frac{1}{m}\right)^2 + \frac{1}{nh(x)}\right). \end{aligned} \tag{2.42}$$

Theorem 5 implies that, similar to the standard kernel density estimation setting, the optimal *rate* of mse convergence is the same for local and global estimators.

2.4 A Bias-Corrected Estimate

In this section, we introduce a bias-corrected estimate of the density of the conditional expectation. We motivate the estimator with a discussion of the jackknife bias-corrected estimator. For an introduction to the jackknife bias-corrected estimate see Efron and Tibshirani [1993]. Finally, we present some results on the asymptotic bias and variance of the bias-corrected estimate and show that the optimal rate of mse convergence is faster than for the naive, global estimator.

The jackknife estimator can be thought of as an extrapolation from one estimate back to another estimate that has nearly zero bias (e.g., Stefanski and Cook [1995]). To understand this interpretation of the jackknife estimator, we turn to an example. A similar example was presented in Stefanski and Cook [1995]. Suppose we want to estimate $\theta = g(\mu)$ where g is nonlinear. We are given i.i.d. data $\{X_1, \ldots, X_m\}$ drawn from a $N(\mu, \sigma^2)$ distribution. We take our estimate, denoted $\hat{\theta}_m$, to be

$g(\bar{X}_m)$ where $\bar{X}_m$ is the sample mean of the data. A Taylor expansion shows that for an estimate based on any sample size m,

$$\mathrm{E}(\hat{\theta}_m) \approx \theta + \frac{1}{m}\beta. \tag{2.43}$$

We actually know that $\beta = \sigma^2 g''(\mu)/2$, but that is not needed for our discussion. The point is that the bias, $\mathrm{E}(\hat{\theta}_m) - \theta$ is approximately linear in the inverse sample size m. Then if we know β and $\mathrm{E}(\hat{\theta}_m)$ for some m, by extrapolating on the line given in (2.43) back to $1/m = 0$, we have a nearly unbiased estimate of θ. The remaining bias is from the lower order terms in the Taylor expansion of $\mathrm{E}(\hat{\theta}_m)$.

If we have an estimate of $\mathrm{E}(\hat{\theta}_m)$, all we need is another estimate $\mathrm{E}(\hat{\theta}_{\tilde{m}})$ for $\tilde{m} \neq m$ to estimate β. For the standard jackknife estimator, $\mathrm{E}(\hat{\theta}_m)$ is estimated with $\hat{\theta}_m$ and $\mathrm{E}(\hat{\theta}_{m-1})$ is estimated with $\hat{\theta}_{(\cdot)} = \sum_{k=1}^{m} \hat{\theta}_{(k)}/m$ where for $k = 1, \ldots, m$, $\hat{\theta}_{(k)}$, the leave-out-one estimator, is the estimator based on all the data less X_k. The jackknife bias-corrected estimator $\dot{\theta}$ is then

$$\begin{aligned} \dot{\theta} &= \hat{\theta}_m - (m-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_m) \\ &= m\hat{\theta}_m - (m-1)\hat{\theta}_{(\cdot)}. \end{aligned}$$

For our global estimator (2.1), we know that from Theorem 2,

$$\mathrm{E}(\hat{f}(x; m, n, h)) \approx f(x) + h^2\beta_1 + \frac{1}{m}\beta_2, \tag{2.44}$$

where

$$\beta_1 = \frac{1}{2} f''(x)\,\mathrm{d}x \int u^2 K(u)\,\mathrm{d}u$$

and

$$\beta_2 = \frac{1}{2}\int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s.$$

Here the bias is approximately linear in the square of the bandwidth (h^2) and the inverse of the internal sample size ($1/m$). Given an estimate of $\mathrm{E}(\hat{f}(x; m, n, h))$ for some m and h, we would like to extrapolate back to $1/m = 0$ and $h^2 = 0$ on the plane specified in (2.44).

Similar to the typical jackknife estimator, we take the global estimate $\hat{f}(x; m, n, h)$ as an approximation of $\mathrm{E}(\hat{f}(x; m, n, h))$. To determine β_1 and β_2 and thus extrapolate back to $1/m = 0$ and $h^2 = 0$, we need to estimate $\mathrm{E}(\hat{f}(x; m, n, h))$ at two other pairs of (m, h). Alternatively, we can save ourselves a bit of work by choosing only one other pair $(\tilde{m}, \tilde{h})$ such that $(1/\tilde{m}, \tilde{h^2})$ lies on the line determined by $(0, 0)$ and $(1/m, h^2)$.

We could estimate $\mathrm{E}(\hat{f}(x; \tilde{m}, n, \tilde{h}))$ as the average of the leave-out-one estimators as is done for the typical jackknife estimator. This will require m computations of the density estimator. As a computationally friendly alternative, consider instead taking $\tilde{m} = m/2$ and $\tilde{h} = \sqrt{2}h$ and take the estimate $\hat{f}(x; \tilde{m}, n, \tilde{h})$ as an approximation of $\mathrm{E}(\hat{f}(x; \tilde{m}, n, \tilde{h}))$. Note that $(1/\tilde{m}, \tilde{h^2})$ lies on the line determined by $(0, 0)$ and $(1/m, h^2)$.

Using the data points $\hat{f}(x; m, n, h)$ and $\hat{f}(x; m/2, n, \sqrt{2}h)$ and extrapolating back to $1/m = 0$ and $h^2 = 0$ gives the bias-corrected estimator

$$\dot{f}(x; m, n, h) = 2\hat{f}(x; m, n, h) - \hat{f}(x; m/2, n, \sqrt{2}h). \tag{2.45}$$

We emphasize that just like the leave-out-one jackknife estimator, the data can be reused to estimate $\hat{f}(x; m/2, n, \sqrt{2}h)$. That is to say, the estimator $\hat{f}(x;m/2,n,\sqrt{2}h)$ can be computed with the same data set with which $\hat{f}(x; m, n, h)$ is computed less half of the internal samples. However in some cases, it would be possible to generate a new data set to estimate $\hat{f}(x; m/2, n, \sqrt{2}h)$. For the remainder of this section, we consider the asymptotic bias and variance of the bias-corrected estimator given in (2.45). The results cover both the case where the data is reused in computing $\hat{f}(x; m/2, n, \sqrt{2}h)$ and the case where a new data set is generated.

Recall from (2.11),

$$\text{bias}(\hat{f}(x;m,n,h)) = (\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x))$$

Then the bias of the estimate $\dot{f}(x;m,n,h)$ can be expressed as

$$\begin{aligned}
\text{bias}(\dot{f}(x;m,n,h)) &= \mathrm{E}(\dot{f}(x;m,n,h)) - f(x) \\
&= 2\left[\mathrm{E}(\hat{f}(x;m,n,h)) - f(x)\right] - \left[\mathrm{E}\hat{f}(x;m/2,n_0,\sqrt{2}h) - f(x)\right] \\
&= 2\left[(\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x)) + (f_m(x) - f(x))\right] \\
&\quad - \left[(\mathrm{E}(\hat{f}(x;m/2,n_0,\sqrt{2}h)) - f_{m/2}(x)) + (f_{m/2}(x) - f(x))\right]. \quad (2.46)
\end{aligned}$$

From Lemma 7,

$$\begin{aligned}
&\mathrm{E}(\hat{f}(x;m,n,h)) - f_m(x) = h^2\frac{1}{2}f^{(2)}(x)\int u^2K(u)\,\mathrm{d}u \\
&+ \frac{h^2}{m}\frac{1}{4}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u + h^4\frac{1}{24}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u \\
&+ o\left(\frac{h^2}{m} + h^4\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}(\hat{f}(x;m/2,n_0,\sqrt{2}h)) = 2h^2\frac{1}{2}f^{(2)}(x)\int u^2K(u)\,\mathrm{d}u \\
&+ 4\frac{h^2}{m}\frac{1}{4}\int s^2\alpha^{(4)}(x,s)\,\mathrm{d}s\int u^2K(u)\,\mathrm{d}u + 4h^4\frac{1}{24}f^{(4)}(x)\int u^4K(u)\,\mathrm{d}u \\
&+ o\left(\frac{h^2}{m} + h^4\right).
\end{aligned}$$

From Lemma 6,

$$f_m(x) - f(x) = \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s + \frac{1}{m^2}\frac{1}{8}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s + o(\frac{1}{m^2})$$

and

$$f_{m/2}(x) - f(x) = 2\frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s + 4\frac{1}{m^2}\frac{1}{8}\int s^4\alpha^{(4)}(x,s)\,\mathrm{d}s + o(\frac{1}{m^2}).$$

Substituting into (2.46) proves the following theorem.

Theorem 6 Assume A1-A5 and A6(6). Also assume

1. K is a bounded probability distribution function symmetric about zero with finite fourth moment;

2. $m \to \infty$ and $h \to 0$ as $c \to \infty$.

Then

$$\begin{aligned} \operatorname{bias}(\dot{f}(x;m,n,h)) &= -h^4 \frac{1}{12} f^{(4)}(x) \int u^4 K(u)\,\mathrm{d}u \\ &\quad - \frac{h^2}{m}\frac{1}{2} \int s^2 \alpha^{(4)}(x,s)\,\mathrm{d}s \int u^2 K(u)\,\mathrm{d}u \\ &\quad - \frac{1}{m^2}\frac{1}{4} \int s^4 \alpha^{(4)}(x,s)\,\mathrm{d}s \\ &\quad + o\left(\left(h^2 + \frac{1}{m}\right)^2\right). \end{aligned}$$

As for the variance of the $\dot{f}(x;m,n,h)$ note that from the proof of Theorem 2,

$$\operatorname{var}(\hat{f}(x;m,n,h)) = \frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right)$$

and

$$\operatorname{var}(\hat{f}(x;m/2,n,\sqrt{2}h)) = \frac{1}{\sqrt{2}nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).$$

Also,

$$\begin{aligned} |\operatorname{cov}(\hat{f}(x;m,n,h), \hat{f}(x;m/2,n,\sqrt{2}h))| &\le \sqrt{\operatorname{var}(\hat{f}(x;m,n,h))\operatorname{var}(\hat{f}(x;m/2,n,\sqrt{2}h))} \\ &\le \frac{1}{2^{1/4}}\frac{1}{nh} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right). \end{aligned}$$

Then

$$
\begin{aligned}
\operatorname{var}(\dot{f}(x;m,n,h)) &= \operatorname{var}(2\hat{f}(x;m,n,h) - \hat{f}(x;m/2,n,\sqrt{2}h) \\
&= 4\operatorname{var}(\hat{f}(x;m,n,h)) + \operatorname{var}(\hat{f}(x;m/2,n,\sqrt{2}h)) \\
&\quad -4\operatorname{cov}(\hat{f}(x;m,n,h),\hat{f}(x;m/2,n,\sqrt{2}h)) \\
&\leq 4\frac{1}{nh}f(x)\int K^2(u)\,\mathrm{d}u + \frac{1}{\sqrt{2}nh}f(x)\int K^2(u)\,\mathrm{d}u \\
&\quad +4\frac{1}{2^{1/4}}\frac{1}{nh}f(x)\int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right) \\
&= \left(4+\frac{1}{2^{1/2}}+\frac{4}{2^{1/4}}\right)\frac{1}{nh}f(x)\int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).
\end{aligned}
$$

This shows that $\operatorname{var}(\dot{f}(x;m,n,h))$ is $O(\frac{1}{nh})$. Similarly,

$$
\operatorname{var}(\dot{f}(x;m,n,h)) \geq \left(4+\frac{1}{2^{1/2}}-\frac{4}{2^{1/4}}\right)\frac{1}{nh}f(x)\int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh}\right).
$$

Since

$$
4+\frac{1}{2^{1/2}}-\frac{4}{2^{1/4}} \approx 1.34,
$$

we conclude that the asymptotic variance of $\dot{f}(x;m,n,h)$ is greater than the asymptotic variance of the global estimator $\hat{f}(x;m,n,h)$. Therefore, it is likely the actual variance of the bias-corrected estimate is greater than the variance for the global estimate. This is a common theme for bias-corrected estimates (Efron and Tibshirani [1993]).

The above asymptotic bias and variance results for $\dot{f}(x;m,n,h)$ imply that if m, n, and h are chosen such that m is of the order $c^{2/11}$, n is of the order $c^{9/11}$, and h is of the order $c^{-1/11}$ the optimal rate of convergence of mse is obtained and that optimal rate is $c^{-8/11}$. Recall the optimal rate of mse for the global estimator $\hat{f}(x;m,n,h)$ was $c^{-4/7}$. Thus, the bias-correction leads to improved convergence. But as we noted above, the variance is greater for the bias-corrected estimate and this can adversely affect performance, especially for modest sample sizes.

The bias-corrected estimate (2.45) is based on the global estimator $\hat{f}(x; m, n, h)$. We can formulate a local version $\dot{f}_L$ of the bias-corrected estimator as follows:

$$\dot{f}_L(x; m, n, h(x)) = 2\hat{f}_L(x; m, n, h(x)) - \hat{f}_L(x; m/2, n, \sqrt{2}h(x)), \tag{2.47}$$

where $\hat{f}_L(x; m, n, h(x))$ is the local estimator defined in (2.41).

We immediately have the following result for the bias of the local estimator:

Theorem 7 Assume A1-A5 and A6(6). Also assume

1. K is a bounded probability distribution function symmetric about zero with finite fourth moment;

2. $m \to \infty$ and $h(x) \to 0$ as $c \to \infty$ for all $x \in \mathbb{R}$.

Then for any $x \in \mathbb{R}$,

$$\begin{aligned} \text{bias}(\dot{f}_L(x; m, n, h(x))) &= -h(x)^4 \frac{1}{12} f^{(4)}(x) \int u^4 K(u)\, \mathrm{d}u \\ &\quad - \frac{h(x)^2}{m} \frac{1}{2} \int s^2 \alpha^{(4)}(x, s)\, \mathrm{d}s \int u^2 K(u)\, \mathrm{d}u \\ &\quad - \frac{1}{m^2} \frac{1}{4} \int s^4 \alpha^{(4)}(x, s)\, \mathrm{d}s \\ &\quad + o\left(\left(h(x)^2 + \frac{1}{m}\right)^2\right). \end{aligned}$$

The following inequality involving variance is also immediate:

$$\begin{aligned} &\left(4 + \frac{1}{2^{1/2}} - \frac{4}{2^{1/4}}\right) \frac{1}{nh(x)} f(x) \int K^2(u)\, \mathrm{d}u + o\left(\frac{1}{nh(x)}\right) \\ &\quad \leq \text{var}(\dot{f}_L(x; m, n, h(x))) \\ &\quad \leq \left(4 + \frac{1}{2^{1/2}} + \frac{4}{2^{1/4}}\right) \frac{1}{nh(x)} f(x) \int K^2(u)\, \mathrm{d}u + o\left(\frac{1}{nh(x)}\right). \end{aligned} \tag{2.48}$$

Chapter 3

Implementation and Simulation

In this chapter, we address the implementation of our estimators for the density of the conditional expectation discussed in Chapter 2 and study their performance. Section 3.1 addresses the implementation which requires specifying a number of inputs. For some of these inputs, we develop a data-based selection method based on the ideas of empirical-bias bandwidth selection (EBBS) used in local polynomial regression (Ruppert [1997]). In Section 3.2, we then compare the performance of the estimators for some simulated test cases.

3.1 Implementation

Implementation requires the specification of a number of inputs. For the standard kernel density estimator presented in (1.2), one must choose the kernel K and the bandwidth h. For the estimators of the density of the conditional expectation including the global kernel density estimator (2.1), the local kernel density estimator (2.41), and the bias-corrected estimator (2.45), one must choose K, h, as well as the number of external samples n and the number of internal samples m.

We choose K to be the Epanechnikov kernel which is

$$K(x) = \frac{3}{4}(1 - x^2)I(|x| < 1).$$

Epanechnikov (Epanechnikov [1967]) showed this kernel was optimal in terms of minimizing the mise for the standard kernel density estimator (1.2). For more discussion on this topic see Wand and Jones [1995].

The rest of this section deals with the choice of the parameters m, n, and h.

In Section 3.1.1 we consider the selection of these parameters for the global kernel density estimator (2.1). We present a data-based method to select these parameters based on EBBS developed by Ruppert [1997]. We present the algorithm and briefly discuss why we chose this method. In Section 3.1.2 we apply a similar method to selecting the parameters for the local kernel density estimator (2.41). For this estimator we must specify a function for the bandwidth h so the algorithm is more complicated than the one given for the global estimator. Finally, in Section 3.1.3, the data-based parameter selection method is applied to the bias-corrected estimator (2.45).

3.1.1 Global Kernel Density Estimate

In Chapter 2 we saw how to choose the bandwidth h, the number of internal samples m, and the number of external samples n for the global estimator $\hat{f}(x; m, n, h)$ to obtain optimal convergence (see (2.13)). However the expressions for m, n, and h given in (2.13) involve unknowns such as $f''(x)$, the second derivative of the target density, and $\int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s$ where $\alpha^{(2)}$ is defined in (2.4) and (2.5) as the second derivative with respect to the first argument of the function

$$\alpha(y, s) = g(s|y) f(y).$$

To implement the estimator $\hat{f}(x; m, n, h)$ in an optimal way, one could attempt to estimate these unknown quantities and plug these estimates into the expressions given in (2.13). This type of estimator is known as a plug-in estimator (Wand and Jones [1995]). In fact it is quite doable to estimate the unknowns f and f'' needed for the plug-in estimator. Other needed estimates, including an estimate of the second derivative of α, appear very difficult to obtain.

To choose the parameters m, n, and h needed to implement the estimator $\hat{f}(x; m, n, h)$ we turn from optimizing the asymptotic mise to optimizing an approximation of mise. From (2.10), mise can be decomposed as

$$\operatorname{mise}(\hat{f}(\cdot; m, n, h)) = \int \operatorname{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x + \int \operatorname{var}(\hat{f}(x; m, n, h))\,\mathrm{d}x.$$

It was shown in the proof of Theorem 3 that

$$\int \operatorname{var}(\hat{f}(x; m, n, h))\,\mathrm{d}x = \frac{1}{nh}\int K^2(u)\,\mathrm{d}u + o(\frac{1}{nh}).$$

An approximation for the variance component in mise is the asymptotic approximation,

$$\frac{1}{nh}\int K^2(u)\,\mathrm{d}u,$$

which is readily available. Also in the proof of Theorem 3, it was shown that

$$\begin{aligned}
&\int \operatorname{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x \\
&\quad = \int \left(h^2\frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x) + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s\right)^2 \mathrm{d}x \\
&\qquad + o\left(\left(h^2 + \frac{1}{m}\right)^2\right).
\end{aligned}$$

As explained above, the asymptotic approximation

$$\int \left(h^2\frac{1}{2}\left(\int u^2 K(u)\,\mathrm{d}u\right) f''(x) + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x, s)\,\mathrm{d}s\right)^2 \mathrm{d}x$$

is not immediately useful given the unknowns in the approximation. To approximate the bias component in mise we will instead build and estimate a model of bias for each x. Squaring the bias and numerically integrating will then provide an empirical model of integrated squared bias. Adding the integrated variance approximation to this gives an empirical model of mise which can then be optimized with respect to m, n, and h.

The idea of building and empirically estimating a model of bias to be used in the selection of an estimator's parameters was introduced in Ruppert [1997]. In this paper, the method, called empirical-bias bandwith selection (EBBS), was applied to bandwidth selection in local polynomial regression. Schulman [1998] established convergence results for the bandwidth selector in the context of local polynomial regression. Staudenmayer [2000] applied EBBS to local polynomial regression in which the covariates are measured with error.

EBBS uses a model of bias suggested by the asymptotic expression of the expected value of the estimator. In our case, by Lemmas 6 and 7,

$$\begin{aligned} \mathrm{E}\big(\hat{f}(x;m,n,h)\big) &= f(x) + h^2\frac{1}{2}f''(x)\int u^2K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s \\ &\qquad + o(h^2 + \frac{1}{m}). \end{aligned}$$

The asymptotic expression

$$\mathrm{E}\big(\hat{f}(x;m,n,h)\big) = f(x) + h^2\frac{1}{2}f''(x)\int u^2K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2}\int s^2\alpha^{(2)}(x,s)\,\mathrm{d}s, \quad (3.1)$$

suggests the following model:

$$\mathrm{E}\big(\hat{f}(x;m,n,h)\big) = \beta_0(x) + \beta_1(x)h^2 + \beta_2(x)\frac{1}{m}. \quad (3.2)$$

Here $\beta_0(x)$ approximately corresponds to $f(x)$, the target density evaluated at x. The bias of $\hat{f}(x;m,n,h)$ is then approximately given by

$$\beta_1(x)h^2 + \beta_2(x)\frac{1}{m}. \quad (3.3)$$

The EBBS model of bias used in local polynomial regression is a polynomial in h (Ruppert [1997], Staudenmayer [2000]). In our case the model of bias is polynomial in h as well as $1/m$. Lemmas 6 and 7 allow for more terms used in the asymptotic expression of $\mathrm{E}\big(\hat{f}(x;m,n,h)\big)$ given in (3.1) which would give more terms in model

(3.2). Such a model would be a better approximation of $\mathrm{E}\big(\hat{f}(x; m, n, h)\big)$ but would require the estimation of additional parameters. In this thesis, we use the model (3.2).

Though approximate, notice that the model of bias does capture the fact that as $h \to 0$ and $1/m \to 0$, bias tends to zero. Suppose that we can estimate the model (3.2). This not only gives us an empirical model of bias that can be used in selecting the needed parameters m, n, and h but also gives another estimator which will be of some use. Extrapolating the estimated model to $h = 1/m = 0$ gives an approximately unbiased estimate of $f(x)$. This approximately unbiased estimate of $f(x)$ is of course $\hat{\beta}_0$, the estimate of β_0. Based on the discussion of jackknife bias-correction, one can argue $\hat{\beta}_0$ is essentially a jackknife estimate. For more on this see Staudenmayer [2000].

The estimation of the model (3.2) at x_0 for a given computer budget c is outlined in the following algorithm.

1. Generate a sample of the data using half of the computer budget. To do this fix n_0 and m_0 such that $n_0 m_0 = c/2$.

2. Establish a grid of pairs of bandwidths h and internal samples m given by the Cartesian product $(h_1, \ldots, h_{I_1}) \times (m_1, \ldots, m_{I_2})$. The largest internal sample, m_{I_2}, is equal to m_0 so that only half the computer budget is used. Ruppert [1997] suggests evenly spacing the bandwidths on a log scale. We follow this suggestion for the bandwidths and the number of internal samples.

3. For each pair in the grid of bandwidths and internal samples compute the kernel density estimator. This gives the data:

$$[(h_i, m_j), \hat{f}(x_0; m_j, n, h_i)] \qquad i = 1, \ldots, I_1, \quad j = 1, \ldots, I_2.$$

4. Take $\hat{f}(x_0; m_j, n, h_i)$ as an approximation of $\mathrm{E}\big(\hat{f}(x_0; m_j, n, h_i)\big)$ for each i and j. Estimate (3.2) with the data computed in step 3. We use global least squares regression. Note that in the context of local polynomial regression, Ruppert [1997], Schulman [1998], and Staudenmayer [2000] use local least squares to estimate the model. Local least squares may provide a better estimate but it requires the specification of additional tuning parameters to be discussed below. We are content with global least squares as it gives good performance for the test cases considered in Section 3.2.

The estimation procedure above is repeated on an equally spaced grid of x values over the range of the observations $X_j(Z_i)$, $j = 1, \ldots, m_0$, $i = 1, \ldots, n_0$. Following Ruppert's suggestion (Ruppert [1997]), we smooth the estimates $\hat{\beta}_1(x)$ and $\hat{\beta}_2(x)$ over x. The result is an approximation of bias

$$\hat{\beta}_1(x)h^2 + \hat{\beta}_2(x)\frac{1}{m}.$$

at each x in the grid. Squaring the bias at each x on the grid and numerically integrating gives us an approximation of the bias component in mise as a function of h and m. Adding to this the variance component approximation

$$\frac{1}{nh}\int K^2(u)\,\mathrm{d}u,$$

gives an approximation of mise as a function of m, n, and h.

To compute the optimal m, n, and h for the computer budget c, we minimize the approximation of mise with respect to m, n, and h given the constraints

1. $mn = c$;

2. $n_0 \leq n \leq 2n_0$ and $m_0 \leq m \leq 2m_0$.

The first constraint is simply the computer budget constraint. The second constraint arises because we have already used half of the computer budget to generating n_0 external samples and m_0 internal samples.

Note that it is implicitly assumed that our approximation for mise is appropriate for the computer budget c. The approximation for bias was estimated under the constraint that the computer budget not exceed $c/2$. As Ruppert [1997] points out, the EBBS bias approximation captures the bias for the given finite sample. Asymptotics are used only to suggest a model for the bias. This implies that the bias coefficients β_1 and β_2 in (3.3) should be different for different sample sizes corresponding to different computer budgets. We will assume that the change in these coefficients is small enough such that the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ computed for the budget $c/2$ are reasonably good estimates for β_1 and β_2 given a computer budget of c.

In the above algorithm there are a number of tuning parameters that must be selected including n_0, m_0, h_1, h_{I_1}, m_1, I_1, and I_2. Ideally, we would like to establish values for these tuning parameters that work for most instances of the problem. This could be done with an experiment involving Monte Carlo simulation as in Ruppert [1997], Schulman [1998], and Staudenmayer [2000]. However, for this thesis we simply offer some guidelines and report the values that worked well in the test cases presented below.

For choosing n_0, the initial allocation of external samples given half of the computer budget, we note that it is asymptotically optimal to set $n_0 = b(c/2)^{5/7}$ for some positive constant b. We found a constant of $b = 1/3$ worked well for the cases in which Z is univariate. For Z multivariate, it was better to take more external samples (b=2). We then took m_0, the initial allocation of internal samples given half the computer budget, to be $(c/2)/n_0$.

We must also choose the lower and upper bounds for the bandwidth grid, h_1 and h_{I_1}, respectively. We note that if h_{I_1} is chosen too large, (3.3) is not a good model for the larger bandwidths h. But if h_1 is chosen too small the variance approximation will not be very good for the smaller bandwidths h. We found that $h_1 = 0.1$ and $h_{I_1} = 0.5$ worked well. Similar considerations need to be made in choosing m_1, the lower bound for the internal samples grid. If m_1 is too small, (3.3) is not a good model for the smaller numbers of internal samples. We found that $0.05m_0$ worked well. Finally, for the number of bandwidths I_1 and the number of internal samples I_2, we found that $I_1 = I_2 = 5$ was adequate.

3.1.2 Local Kernel Density Estimate

In this section, we again apply the ideas of EBBS to the selection of inputs for our local kernel density estimator (2.41), which has the form

$$\hat{f}_L(x; m, n, h(x)) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x)} K\left(\frac{x - \bar{X}_m(Z_i)}{h(x)}\right).$$

The inputs that we must choose are again m, n, and h but in this case h is a function. As a result the data-based algorithm for choosing the inputs is a bit more complicated.

In Theorem 5, we established an expression for the local estimate's asymptotic mse. Unfortunately, at this time we don't have an equivalent result for the mise. But let us assume that we are able to integrate $\mathrm{mse}(\hat{f}_L(x, m, n, h(x)))$ so that we obtain the following asymptotic mise

$$\int \left((h(x)^2 \frac{1}{2} \left(\int u^2 K(u)\,\mathrm{d}u \right) f''(x) + \frac{1}{m}\frac{1}{2} \int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s \right)^2 \mathrm{d}x$$
$$+ \frac{1}{n} \int \frac{f(x)}{h(x)}\,\mathrm{d}x \int K^2(u)\,\mathrm{d}u.$$

The goal here is to select m, n, and $h(\cdot)$ to minimize mise. As was discussed for the global kernel density estimator, we could try the plug-in estimator: optimize asymptotic mise with respect to m, n, and $h(\cdot)$ and then estimate the unknowns in the resulting expressions for m, n, and $h(\cdot)$. This has the same difficulty in that some of the unknowns will be very difficult to estimate. And actually in this case it seems impossible to analytically compute the optimal expressions.

We then turn to building an approximation of mise as was done in Section 3.1.1. We again approximate the variance component with its asymptotic expression

$$\frac{1}{n}\int \frac{f(x)}{h(x)}\,\mathrm{d}x \int K^2(u)\,\mathrm{d}u.$$

To approximate the bias component in mise, as was done in Section 3.1.1 we first build a model for the bias of the estimator at a particular x. The asymptotic expression for expectation of $\hat{f}_L(x, m, n, h)$ is used as a guide. This expression is identical to the asymptotic expression for the expectation of $\hat{f}(x, m, n, h)$ except that the h is replaced with $h(x)$. The model for the expectation of $\hat{f}_L(x, m, n, h)$ is

$$\mathrm{E}\big(\hat{f}_L(x; m, n, h)\big) = \beta_0(x) + \beta_1(x)h(x)^2 + \beta_2(x)\frac{1}{m}. \tag{3.4}$$

Once again, $\beta_0(x)$ approximately corresponds to $f(x)$, the target density evaluated at x. An estimate $\hat{\beta}_0(x)$ of $\beta_0(x)$ is an approximately unbiased estimator of $f(x)$. The bias of $\hat{f}_L(x; m, n, h)$ is then approximately given by

$$\beta_1(x)h(x)^2 + \beta_2(x)\frac{1}{m}. \tag{3.5}$$

To estimate the model, we can use the same algorithm presented in Section 3.1.1. We again repeat the estimation for x on an equally spaced grid over the range of observations $X_j(Z_i)$, $j = 1, \ldots, m_0$, $i = 1, \ldots, n_0$.

Assume for the moment that we have estimates $\hat{\beta}_1(x)$ and $\hat{\beta}_2(x)$ for all x. Then our approximation of the bias component of mise is

$$\int \left(\hat{\beta}_1(x)h(x)^2 + \hat{\beta}_2(x)\frac{1}{m}\right)^2 \mathrm{d}x,$$

and our approximation of mise is

$$\int \left(\hat{\beta}_1(x)h(x)^2 + \hat{\beta}_2(x)\frac{1}{m}\right)^2 \mathrm{d}x + \frac{1}{n}\int \frac{f(x)}{h(x)}\,\mathrm{d}x \int K^2(u)\,\mathrm{d}u. \tag{3.6}$$

Once again, we want to optimize (3.6) with respect to m, n, and $h(\cdot)$ given the constraints

1. $mn = c$;

2. $n_0 \leq n \leq 2n_0$ and $m_0 \leq m \leq 2m_0$.

This is a very difficult constrained optimization problem. We found that applying numerical algorithms to this constrained optimization problem gives poor results for the test cases discussed in Section 3.2.

We simplified the optimization in two ways. First we ignore the second constraint that $n_0 \leq n \leq 2n_0$ and $m_0 \leq m \leq 2m_0$. This means that in the end, our chosen m and n may fall outside of these constraints in which case we choose the m and n that are closest to the optimal but still obey the constraints. Second, we assumed that $h(x) = \eta(x)^2 m^{-1/2}$ where $\eta(\cdot)$ is some nonzero function. Such a choice results, after optimizing over n and m, in a rate of convergence for mise of $c^{-4/7}$ for any choice of η. This is the optimal rate for the global kernel density estimator and we suspect it is the optimal rate for the local kernel density estimator since local kernel estimates don't improve the optimal rate in the standard density estimation setting.

If $h(x) = \eta(x)^2 m^{-1/2}$ and we ignore the second constraint, then for any $\eta(\cdot)$, we can analytically compute the optimal m and n. We plug the optimal m and n back into the approximation for mise. All that remains is to optimize the approximation with respect to $\eta(\cdot)$. This is not possible to do analytically so we do a numerical search. We restrict $\eta(\cdot)$ to be a continuous piecewise linear function in which the slope changes are allowed at a set of knots $\xi_1, \xi_2, \ldots, \xi_{K_n}$. Then $\eta(x)$ can be written

$$\eta(x) = \sum_{i=1}^{K_n+2} \gamma_i b_i(x),$$

where $b_1(x) = 1, b_2(x) = x, b_3(x) = (x - \xi_1)_+, \ldots, b_{K_n+2}(x) = (x - \xi_{K_n})_+$ in which t_+ denotes the positive part and $(\gamma_1, \ldots, \gamma_{K_n+2})$ are the slope coefficients (Hastie et al. [2001]). With $\eta(\cdot)$ restricted to this form, the optimization problem that remains is simply to choose the slope coefficients to minimize the approximation of mise. This is done numerically in Matlab using the fmincon function. Note that computing the approximation of mise for any given $\eta(\cdot)$ requires numerical integration involving $\hat{\beta}_1(x)$, $\hat{\beta}_2(x)$, and $f(x)$ for x on the grid. The function $f(x)$ is unknown but we can use the approximately unbiased estimate $\hat{\beta}_0(x)$ in its place. The function fmincon requires an initial value for the slope coefficient vector. We use Latin hypercube sampling to search for the initial value.

We found that often for some x the chosen $h(x)$ is too small, leaving the density excessively rough in areas. This may be a result of estimation error. To deal with this issue, we introduce bounds on h. The bounds of h evaluated at a particular x are determined by spans. A span is defined as follows. For a real number y, the function $\lceil y \rceil$ returns the smallest integer that is greater than or equal to y. Let ρ be a positive number less than or equal to one and suppose there are n data points $(\bar{X}_m(Z_1), \ldots, \bar{X}_m(Z_n))$. The span at x is the Euclidean distance from x to the $\lceil \rho n \rceil$th nearest data point. This definition will do for our purposes but for a more general definition see Staudenmayer [2000]. We will take the lower bound of $h(x)$ to be a span for some ρ_1 and the upper bound to be a span for some ρ_2, where $\rho_1 < \rho_2$.

The bounds correspond to the bandwidth function used in the nearest neighbor density estimator (Loftsgaarden and Quesenberry [1965]). Suppose the target density is g. Wand and Jones [1995] point out that this bandwidth evaluated at a particular x is essentially proportional to $g(x)^{-1}$. They go on to note that this bandwidth is not an ideal surrogate for the optimal choice, which was reported above to be $[g(x)/(g''(x))^2]^{1/5}$. Even so, we found that the bounds at a particular x, which are essentially proportional to $f(x)^{-1}$ and not ideal, work well at smoothing out the overly rough regions.

The tuning parameters discussed in Section 3.1.1 must be specified for this algorithm as well. For the test cases presented below the same values specified in Section 3.1.1 work for this algorithm. There are some additional tuning parameters including the number of knots K_n, the knots themselves $(\xi_1, \xi_2, \ldots, \xi_{K_n})$, ρ_1 and ρ_2 for the bounds on the bandwidth, and bounds on the slope coefficients $(\gamma_1, \ldots, \gamma_{K_n+2})$. For $K_n \geq 5$, the results were very similar so we stuck with $K_n = 5$. We spaced the knots evenly over the range of the data and found that taking $\rho_1 = 0.005$, $\rho_2 = 0.2$ and bounding the slope coefficients between -1 and 1 worked well.

3.1.3 Bias-Corrected Density Estimate

Now we turn to the implementation of the bias-corrected estimator presented in Section 2.4,

$$\dot{f}(x; m, n, h) = 2\hat{f}(x; m, n, h) - \hat{f}(x; m/2, n, \sqrt{2}h).$$

We also have the local version of this estimator which is

$$\dot{f}_L(x; m, n, h(x)) = 2\hat{f}_L(x; m, n, h(x)) - \hat{f}_L(x; m/2, n, \sqrt{2}h(x)).$$

We use the same data to compute the estimators on the RHS for both $\dot{f}(x; m, n, h)$ and $\dot{f}_L(x; m, n, h(x))$. For the sake of generality, let us focus the discussion of implementation on the local estimator.

We again would like to use an expression for asymptotic mise to guide the modeling of mise. Recalling the decomposition of mise, we thus need asymptotic expressions for integrated, squared bias and integrated variance. Theorem 3.1.3 gives an asymptotic expression for bias. Let us assume that we can integrate squared bias so that we have the asymptotic expression of integrated, squared bias

$$\int \Bigg(- h(x)^4 \frac{1}{12} f^{(4)}(x) \int u^4 K(u)\,\mathrm{d}u - \frac{h(x)^2}{m}\frac{1}{2} \int s^2 \alpha^{(4)}(x, s)\,\mathrm{d}s \int u^2 K(u)\,\mathrm{d}u \\ - \frac{1}{m^2}\frac{1}{4} \int s^4 \alpha^{(4)}(x, s)\,\mathrm{d}s \Bigg)^2 \mathrm{d}x.$$

Let us also assume that the upper and lower bounds on variance given in (3.1.3) integrate. Moreover, since we are reusing the data, assume that the covariance of $\hat{f}_L(x; m, n, h(x))$ and $\hat{f}_L(x; m/2, n, \sqrt{2}h(x))$ is approximately equal to its upper bound

$$\frac{1}{2^{1/4}} \frac{1}{nh(x)} f(x) \int K^2(u)\,\mathrm{d}u + o\left(\frac{1}{nh(x)}\right),$$

so that we can approximate the variance component in mise with the integrated asymptotic expression from the lower bound of $\mathrm{var}(\dot{f}_L(x; m, n, h(x))$. This approximation is

$$\int \left(4 + \frac{1}{2^{1/2}} - \frac{4}{2^{1/4}}\right) \frac{1}{nh(x)} f(x) \int K^2(u)\,\mathrm{d}u\,\mathrm{d}x. \tag{3.7}$$

Returning to the bias, Theorem 3.1.3 suggests that we model the expectation of $\dot{f}_L(x; m, n, h(x))$ as

$$\mathrm{E}\big(\dot{f}_L(x; m, n, h)\big) = \beta_0(x) + \beta_1(x)h(x)^4 + \beta_2(x)\frac{h(x)^2}{m} + \beta_3(x)\frac{1}{m^2}.$$

The bias of $\dot{f}_L(x; m, n, h(x))$ is then approximately given by

$$\beta_1(x)h(x)^4 + \beta_2(x)\frac{h(x)^2}{m} + \beta_3(x)\frac{1}{m^2}. \tag{3.8}$$

We thus have an approximation for the variance component of mise (3.7) and a model for the bias (3.8). If implementing the local version $\dot{f}_L(x; m, n, h(x))$, proceed as in Section 3.1.2. If implementing the global version $\dot{f}(x; m, n, h)$, proceed as in Section 3.1.1. The tuning parameter values from the previous sections work well here.

3.2 Simulation Results

In this section we examine the performance of the implementations discussed in the previous section on three test cases. To assess performance we consider representative plots and the behavior of estimated mise.

In the first test case, $Z = (Z_1, Z_2) =_d N(\mu, \Sigma)$ where

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Conditional on Z,

$$X(Z) =_d N\left(Z_1 + Z_2, \left(1 - \frac{1}{1 + 2^{-1/2}|Z_1 - Z_2|}\right)^2\right)$$

Then the random variable $\mathrm{E}(X|Z) = Z_1 + Z_2$ is normally distributed with mean 0 and variance 2. This is a straightforward example in which Z is multivariate and all the assumptions for Theorem 3 which gives an asymptotic expansion of mise for the global estimator are satisfied. We consider this example mainly to numerically verify that the rate of mise convergence for the global estimator is $c^{-4/7}$ as suggested by Theorem 3.

In the second and third test cases, we consider more interesting target densities. In the second test case, Z is a bimodal normal mixture, $(1/2)N(-1, 3^{-2}) + (1/2)N(1, 3^{-2})$. Conditional on Z, the random variable $X(Z) =_d N(Z, (1 + Z^2)^2)$. Then the random variable $\mathrm{E}(X|Z) = Z$ and it is thus a bimodal normal mixture, $(1/2)N(-1, 3^{-2}) + (1/2)N(1, 3^{-2})$. The density f of $\mathrm{E}(X|Z)$ is

$$f(x) = \frac{1}{2}\left(\frac{3}{\sqrt{2\pi}}\exp\left(\frac{-9(x+1)^2}{2}\right) + \frac{3}{\sqrt{2\pi}}\exp\left(\frac{-9(x-1)^2}{2}\right)\right).$$

This density is plotted in Figure 3.1. Note that conditional on Z, $\mathrm{var}(X|Z) = (1 + Z^2)^2$ so that the variability in the observations $\bar{X}_m(Z)$ increases as Z moves further from 0. For this test case, we will compare the performance for each of the estimators introduced in Chapter 2.

Note that in the second test case, Z is univariate. Since the $\mathrm{var}(X|Z)$ is unbounded, this example does not satisfy the assumptions for the result in Steckley and Henderson [2003] nor does it satisfy the assumptions of any of the results in this thesis. This case then also serves as a test of the robustness of the estimators presented in Chapter 2. The same is true for the third case.

For the third test case, Z is a normal mixture $(1/2)N(-1/2, 4^{-2}) + (1/2)N(1/2, 1)$ and conditional on Z, the random variable $X(Z) =_d N(Z, (1 + Z^2)^2)$. Again, the random variable $\mathrm{E}(X|Z) = Z$ so its distribution is the bimodal normal mixture, $(1/2)N(-1/2, 4^{-2}) + (1/2)N(1/2, 1)$. The density f is

$$f(x) = \frac{1}{2}\left(\frac{4}{\sqrt{2\pi}}\exp\left(\frac{-16(x+1/2)^2}{2}\right) + \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(x-1/2)^2}{2}\right)\right).$$

This is the density discussed in Section 2.3. See Figure 2.1 for a plot of this

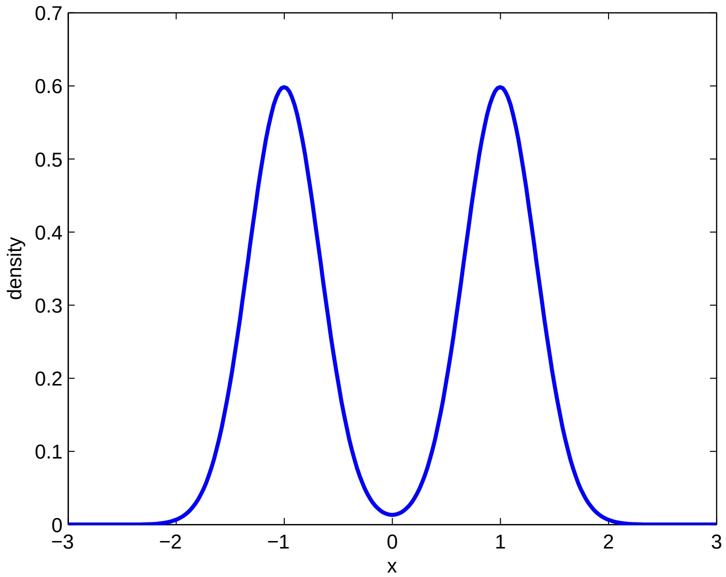

Figure 3.1: The density of the normal mixture $(1/2)N(-1, 3^{-2}) + (1/2)N(1, 3^{-2})$.

density. As discussed in Section 2.3, this target density exhibits very different levels of curvature. We might then expect the local kernel density estimator to outperform the global kernel density estimator for this test case and we focus on this comparison in Section 3.2.3.

3.2.1 Test Case 1

In Figure 3.2, the naive global density estimator is plotted for two different computer budgets along with the target density for the first test case. The figure shows that, as expected, the performance of the estimator improves as the computer budget increases.

We now turn to mise convergence. For clarity, we no longer suppress the dependence of the various estimators and parameters on the computer budget c.

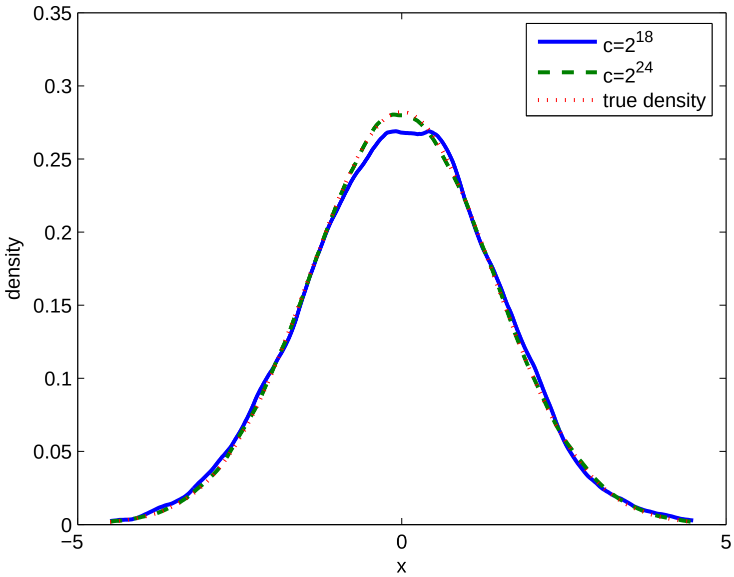

Figure 3.2: The global kernel density estimator for two different computer budgets along with the target density.

To estimate mise(c), mise at a given computer budget c, we first replicate the density estimator 50 times:

$$\{\hat{f}(\cdot; m(c), n(c), h(c))_k : k = 1, \ldots, 50\}.$$

We define integrated squared error (ise) as follows:

$$\text{ise}(c) = \int [\hat{f}(x; m(c), n(c), h(c)) - f(x)]^2 \, \mathrm{d}x.$$

For each k=1,. . . , 50, we use numerical integration to compute

$$\text{ise}_k(c) = \int [\hat{f}(x; m(c), n(c), h(c))_k - f(x)]^2 \, \mathrm{d}x.$$

Our estimate for mise(c) is then

$$\hat{\text{mise}}(c) = \frac{1}{50} \sum_{k=1}^{50} \text{ise}_k(c).$$

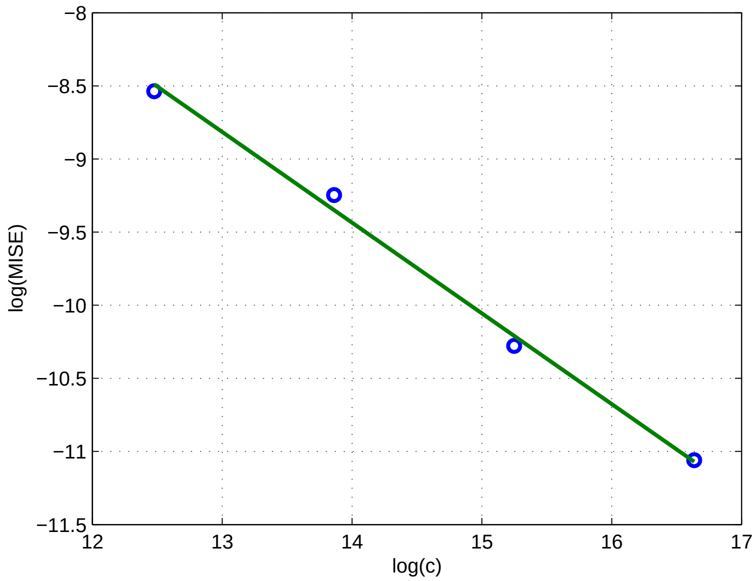

Figure 3.3: Plot of $\log(\text{mise}(c))$ vs. $\log(c)$ at $c = 2^{18}$, 2^{20}, 2^{22} for the global kernel density estimator.

In Figure 3.3, we plot $\log(\text{mise}(c))$ vs. $\log(c)$ at $c = 2^{18}$, 2^{20}, 2^{22}, 2^{24} and the least squares regression line for the global estimator. The linearity of the plot suggests that over the particular range of computer budgets c, the estimator's $\text{mise}(c)$ has the form

$$\text{mise}(c) = Vc^{\gamma}$$

for some constants V and γ. Suppose that $\hat{\delta}_0$ and $\hat{\delta}_1$ are the estimated intercept and slope of the regression line plotted in the figures. Then $\hat{\delta}_1$ estimates γ and $\exp(\hat{\delta}_0/\hat{\delta}_1)$ estimates V. Given that the optimal mise convergence rate is $c^{-4/7}$ we expect that, asymptotically, $\gamma = -4/7 \approx -0.57$. The estimated intercept and slope in Figure 3.3 are -7.51 and -0.62, respectively. So it appears that the estimator performs as expected. Of course, we can never be sure that c is large enough over the range we have considered so that the comparison is valid.

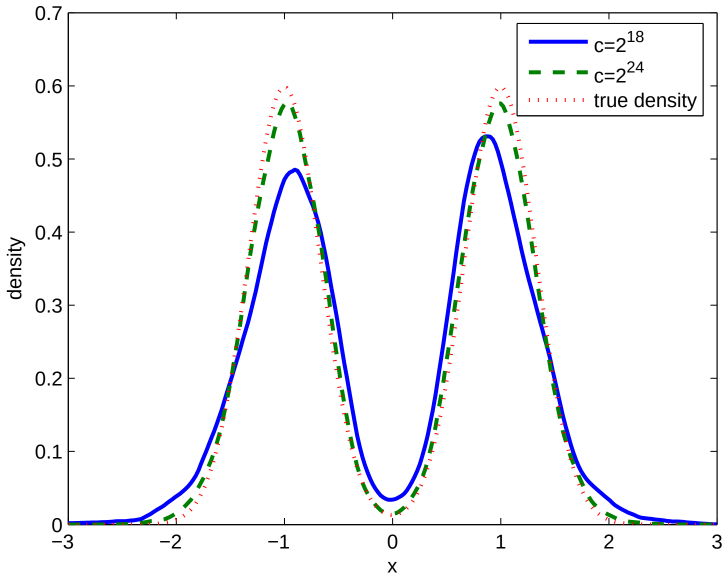

Figure 3.4: The global kernel density estimator for two different computer budgets along with the target density.

3.2.2 Test Case 2

Now we consider test case 2 in which the target density is bimodal. In Figure 3.4, the naive global density estimator is plotted for two different computer budgets along with the target density for the second test case. Figure 3.5 is a similar plot for the local kernel density estimator. It seems from the plots that the performance of these two estimators is very similar for this test case. It is also clear from the figures that for each estimator, performance improves as the computer budget increases. Finally we note that for both estimators and for both computer budgets the estimators are generally closer to the actual density for values of x closer to zero.

This final observation is likely a result of the double smoothing discussed in Section 2.1. For this test case the variability of the observations $\bar{X}_m(Z)$ increase

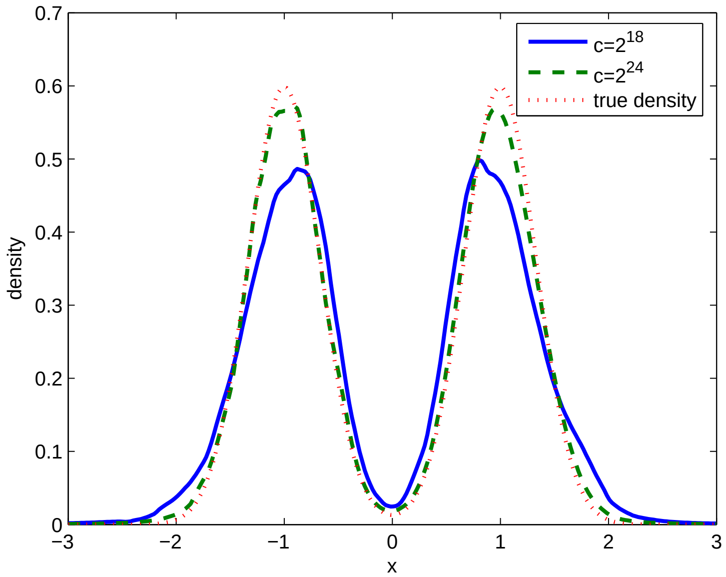

Figure 3.5: The local kernel density estimator for two different computer budgets along with the target density.

as Z moves further from 0. Since $\mathrm{E}(X|Z) = Z$, the observations $\bar{X}_m(Z)$ tend to become more variable as they increase in absolute value. So the measurement error in the observations is greater for observations further from 0. With increased measurement error comes increased smoothing of the observations. Hence the observations further from 0 are oversmoothed.

To better compare the local kernel density estimator and the global naive kernel density estimator, we attempt to study the mise convergence. For each of the estimators we estimate mise(c) over a range c as was done for the first test case. In Figure 3.6 and Figure 3.7, we plot log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22}, 2^{24} and the least squares regression line for the global estimator and the local estimator, respectively.

We again see linearity in the plots. For the global estimator in Figure 3.6, the estimated intercept and slope are 4.14 and -0.66, respectively. For the local

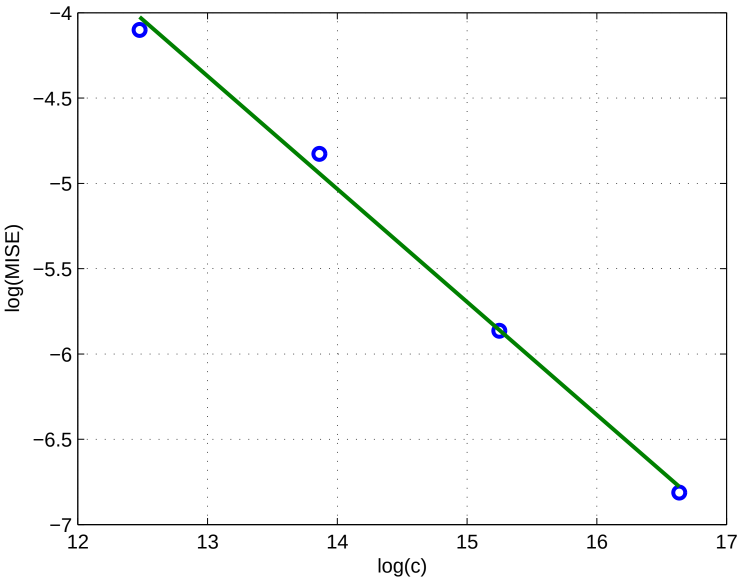

Figure 3.6: Plot of log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22} for the global kernel density estimator.

estimator in Figure 3.7 the estimated intercept and slope are 3.98 and -0.65, respectively. So it appears that the estimators perform equally well. Also comparing the estimated convergence rate with the optimal rate $c^{-4/7}$ suggested by the results in Chapter 2 and Steckley and Henderson [2003], it seems that the estimators perform a bit better than expected.

For this test case, the estimators built on the EBBS idea of empirically estimating a model of bias, perform quite well. It is interesting to look at the estimated bias model itself. Recall the model of bias from (3.3)

$$\beta_1(x)h^2 + \beta_2(x)\frac{1}{m}.$$

This model was suggested by the asymptotic expression for the expectation (3.1)

$$\mathrm{E}\big(\hat{f}(x; m, n, h)\big) = f(x) + h^2 \frac{1}{2} f''(x) \int u^2 K(u)\,\mathrm{d}u + \frac{1}{m}\frac{1}{2} \int s^2 \alpha^{(2)}(x, s)\,\mathrm{d}s.$$

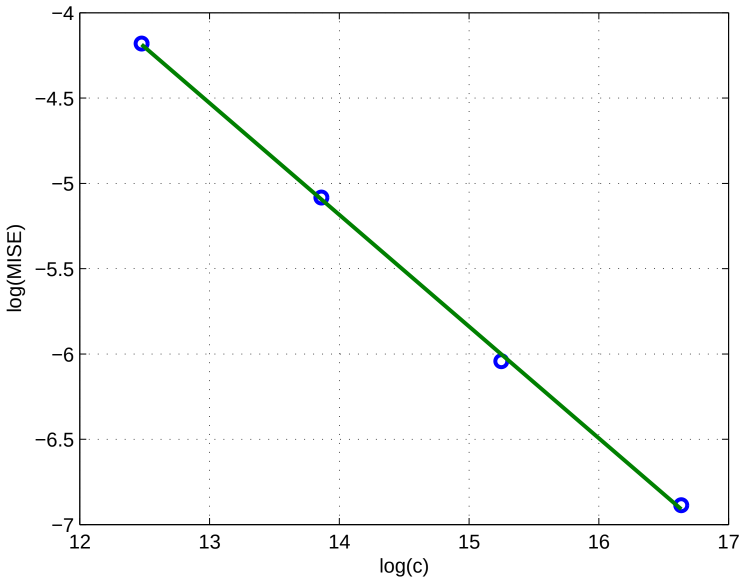

Figure 3.7: Plot of log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22} for the local kernel density estimator.

Consider $\beta_1(x)$. This term asymptotically corresponds to the coefficient $(1/2)f''(x)\int u^2K(u)\mathrm{d}u$ in (3.1) but it was noted earlier that $\beta_1(x)$ captures the effect of h^2 on the bias for the given finite computer budget c (or rather $c/2$, as discussed in Section 3.1.1). It is not an estimate of $(1/2)f''(x)\int u^2K(u)\mathrm{d}u$. However for c very large, we might expect $\beta_1(x)$ to look somewhat like $(1/2)f''(x)\int u^2K(u)\mathrm{d}u$. In Figure 3.8, we plot $(1/2)f''(x)\int u^2K(u)\mathrm{d}u$ and $\hat{\beta}_1(\cdot)$ which was estimated in computing the global kernel density estimator for $c = 2^{24}$. It is interesting to see that the $\hat{\beta}_1(\cdot)$ does in fact follow the shape of $(1/2)f''(x)\int u^2K(u)\mathrm{d}u$.

In Figure 3.9, the bias-corrected local density estimator is plotted for two different computer budgets along with the target density. Comparing this plot with those for the global and local kernel density estimators in Figures 3.4 and 3.5, respectively, indicates that, especially at the smaller computer budget, the bias-corrected estimator tends to be more variable. Given the discussion in Section 2.4,

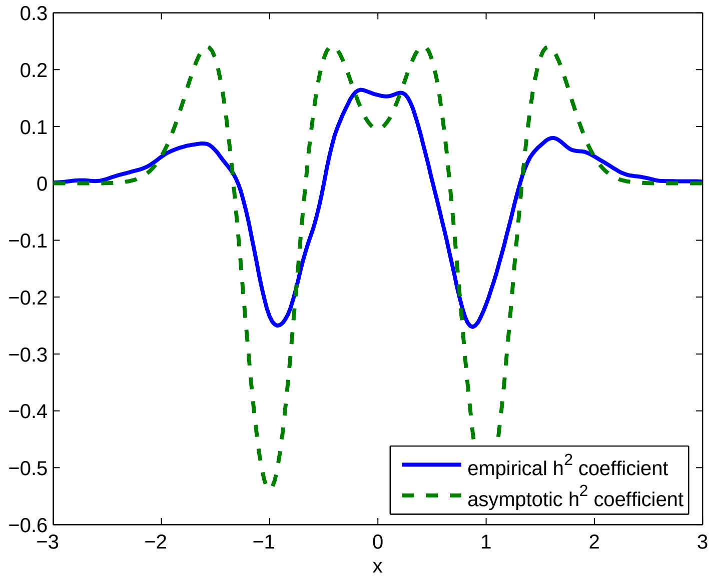

Figure 3.8: The empirical coefficient of h^2, $\hat{\beta}_1(\cdot)$, and the asymptotic coefficient of h^2 which is $(1/2)f''(x)\int u^2K(u)\,\mathrm{d}u$.

this is expected. Comparing the figures also indicates that for the larger computer budget, the bias-corrected estimator outperforms the other two estimators.

To further test this last observation, we estimate mise(c) at $c = 2^{18}$, 2^{20}, 2^{22}, 2^{24} for the bias-corrected estimator as was done for the global and local estimators above. Figure 3.10 is a plot of log(mise(c)) vs. log(c) and the least squares regression line. Again, the linearity of the plot indicates

$$\mathrm{mise}(c) = Vc^{\gamma}$$

for some constants V and γ over the specified range of c. The estimated intercept and slope of the regression line in the plot are 4.74 and -0.77, respectively. Recall that the slope estimates γ, which we expect, asymptotically to be $-8/11 \approx -0.73$ based on the mse result in Section 2.4. The estimated mise convergence here is nearly exactly what we would expect asymptotically. Also note that the perfor-

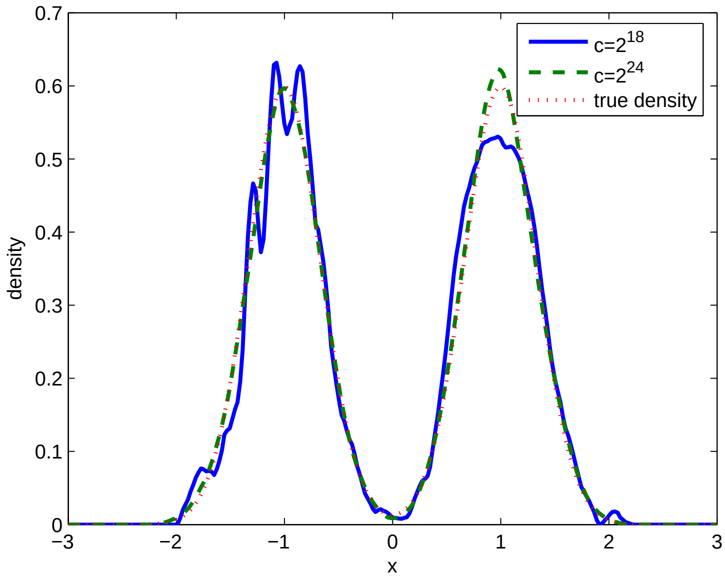

Figure 3.9: The local bias-corrected density estimator for two different computer budgets along with the target density.

mance of the bias-corrected estimator in terms of its convergence rate is superior to the global and local kernel density estimators. But we point out that this is for very large values of c. The representative plots in Figure 3.9 indicate that for modest values of c the bias-corrected estimator is highly variable and may not be a good choice.

3.2.3 Test Case 3

In test case 3 the target density exhibits contrasting levels of curvature. We focus on comparing the performance of the naive global kernel density estimator to the local kernel density estimator. In Figure 3.11 and Figure 3.12, the respective estimators are plotted for two different computer budgets along with the target density. At the larger computer budget, the estimators are very similar. At the

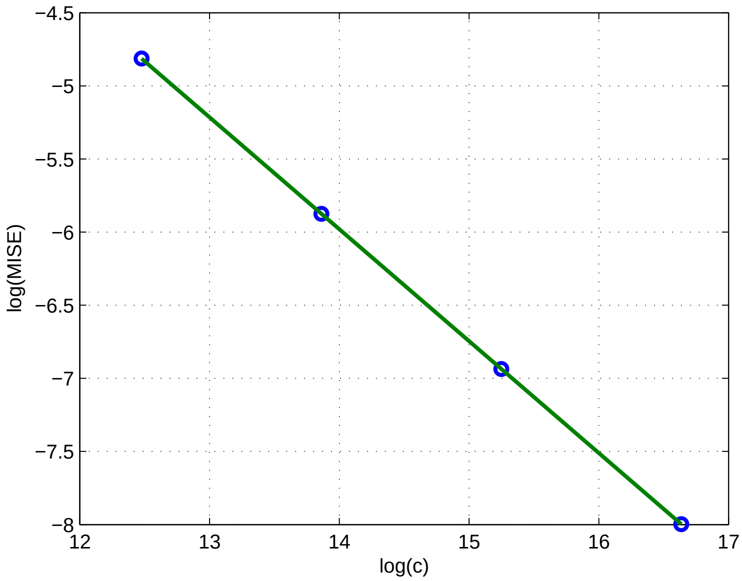

Figure 3.10: Plot of log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22} for the local bias-corrected kernel density estimator.

lower computer budget, the local estimator appears to be more variable.

Based on Figures 3.11 and 3.12 alone, it is difficult to distinguish the performance of the two estimators. In Figure 3.13 and Figure 3.14 we plot log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22}, 2^{24} and the least squares regression line for the global estimator and the local estimator, respectively.

We again see linearity in the plots. For the global estimator in Figure 3.13, the estimated intercept and slope are 2.31 and -0.57, respectively. For the local estimator in Figure 3.14 the estimated intercept and slope are 1.84 and -0.55. The slopes indicate that the rate of convergence for both estimators is very close to the expected rate which is again $c^{-4/7}$. We do however see a smaller intercept for the local estimator. This indicates the constant V is smaller for the local estimator which falls in line with the result in the standard density estimation setting in which the local estimator's optimal mise has a smaller constant multiplier of $c^{-4/7}$

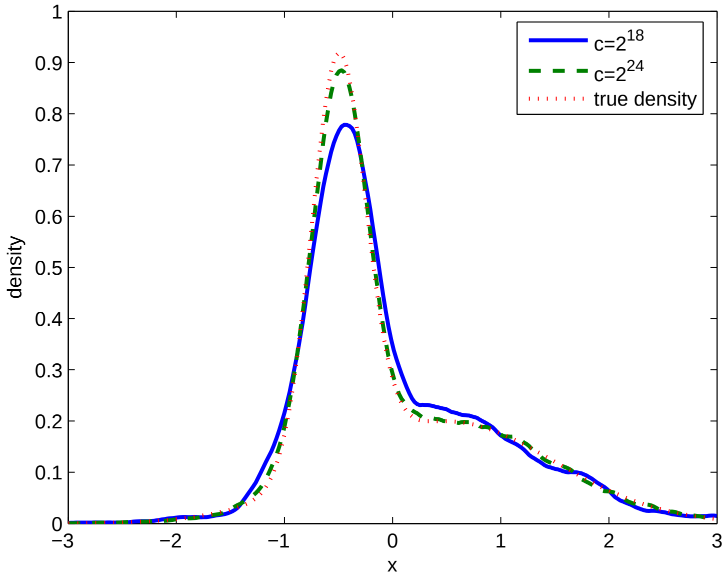

Figure 3.11: The global kernel density estimator for two different computer budgets along with the target density.

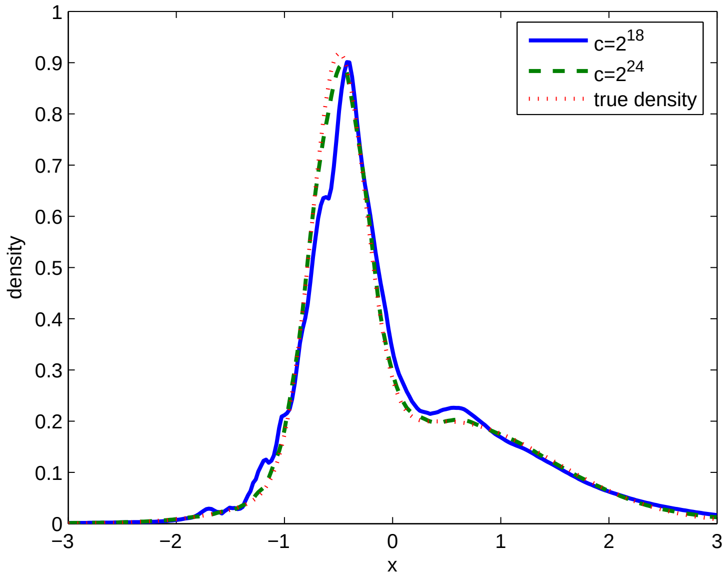

Figure 3.12: The local kernel density estimator for two different computer budgets along with the target density.

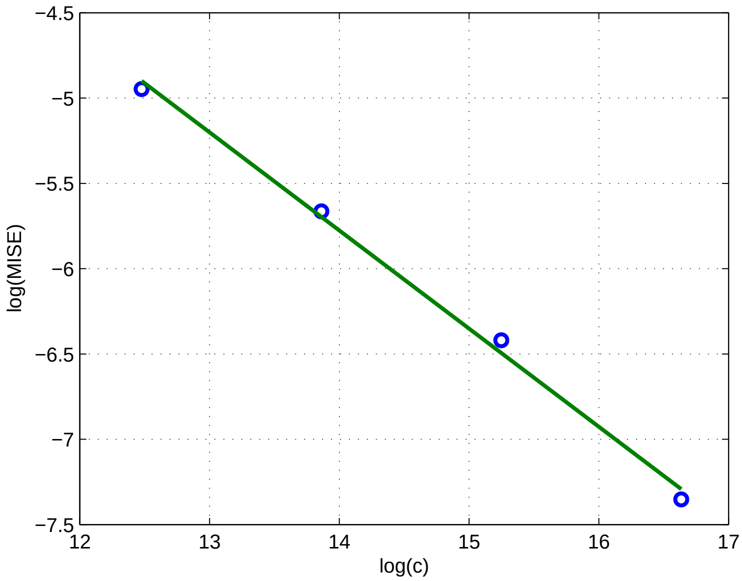

Figure 3.13: Plot of log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22} for the global kernel density estimator.

than the global estimator's constant multiplier (Jones [1990]).

Figure 3.15 plots the bandwidth for the local estimator and the target density. We see that the EBBS implementation performs as we might hope. The bandwidth is smallest for the interval on which the curvature of the density is the greatest.

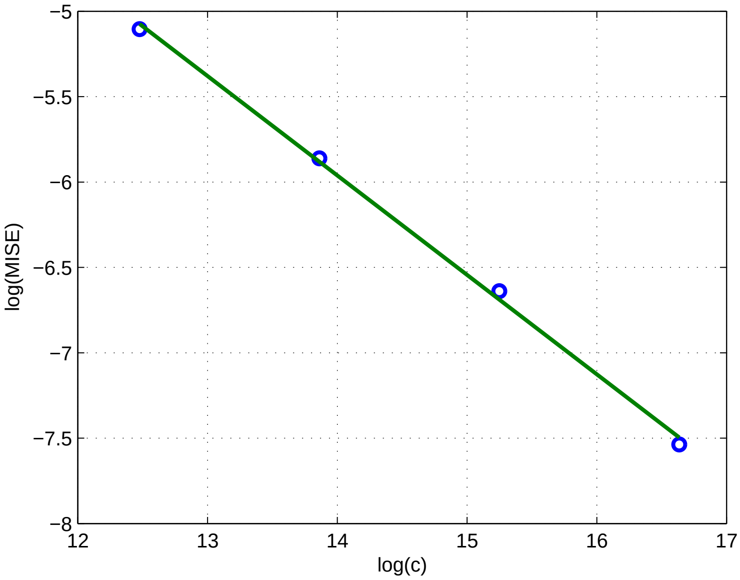

Figure 3.14: Plot of log(mise(c)) vs. log(c) at $c = 2^{18}$, 2^{20}, 2^{22} for the local kernel density estimator.

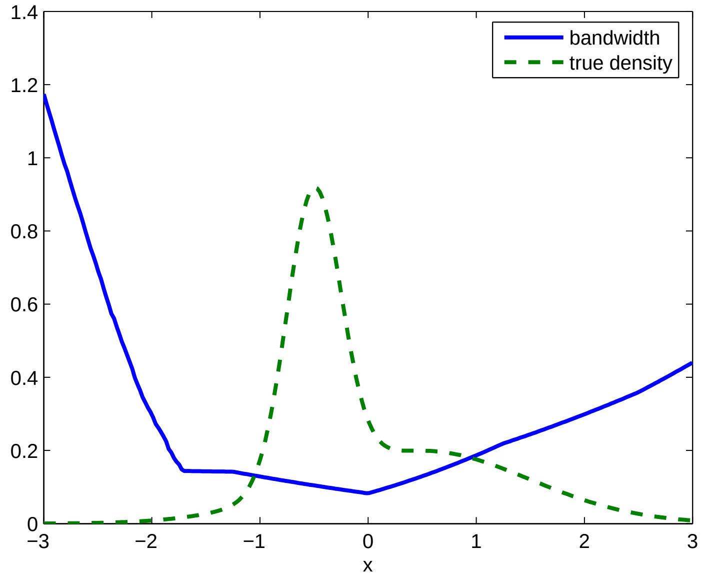

Figure 3.15: The bandwidth for the local kernel density estimator along with the target density.

Chapter 4

Service System Performance in the Presence of an Uncertain Arrival Rate

In this chapter we explore performance for a service system in which the arrival process cannot be determined with certainty. We focus on performance related to service level. This is the fraction of customers that wait in lines for less than a prescribed amount of time before receiving service and is a commonly used metric. We consider two possible interpretations of uncertainty, the RVAR case and the UAR case. These cases were discussed in Section 1.3. Each of these cases, we claim, requires different measures to gauge performance. We identify what performance measures should be computed and discuss how they can be computed for the RVAR and UAR cases. We also consider the implications of ignoring uncertainty associated with the arrival process.

The appropriate long-run performance measures differ in the RVAR and UAR cases in terms of how one should weight performance conditional on a given realized arrival rate function. In the RVAR case there are more customers expected on days when the arrival rate Λ is large, so more customers experience the performance associated with a large arrival rate. In the UAR case weighting by the arrival rate may be inappropriate. These long-run performance measures can be viewed as "customer-focussed" since they indicate what a customer can expect in terms of performance.

We also look at "manager-focused" performance measures. These are short-run performance measures, i.e., "what might happen tomorrow." This kind of information is valuable because it can help to explain variability in daily performance.

They are “manager-focused” because they indicate what a manager could see on any particular day. But of course, long-run and short-run performance measures are relevant to both managers and customers.

Given that we can choose appropriate performance measures, we then look at how to compute them. A common approach is to use closed-form expressions based on steady-state results for simple queueing models. When such approximations are inaccurate or infeasible, simulation provides an alternative way to compute performance. We discuss both steady-state approximations and simulation-based estimates.

The remainder of this chapter is organized as follows. In Section 4.1 we consider the RVAR case and the performance measure giving the long-run fraction of customers that wait less than a prescribed amount of time in queue before receiving service. We give an expression for this quantity, and then consider approximations given by steady-state expectations. We also show that performance will typically be overestimated if a randomly-varying arrival rate is ignored. We then turn to short-run performance, which is the distribution of the fraction of calls answered in the given time limit for a single instance of a period. We give a steady-state approximation based on a central limit theorem. The section concludes by discussing how one can use simulation to estimate both short-run and long-run performance measures efficiently. In Section 4.2 we turn to the UAR case and again suggest appropriate performance measures for the short-run and long-run. We again consider approximations based on steady-state expectations. The section concludes with a discussion of simulation procedures to estimate the performance measures. In Section 4.3 we describe a set of experiments designed to examine performance for both cases. Specifically, we wanted to determine which factors impact the performance measures and assess the quality of the approximations as compared to the simulation-based estimates.

4.1 Randomly Varying Arrival Rates

In order to make the RVAR model more concrete, we begin this section with an example of an RVAR model adapted from a model given by Whitt [1999]. In this model, the arrival process on a given day is Poisson with arrival rate function $B(\lambda(s) : s \geq 0)$, where $(\lambda(s) : s \geq 0)$ is a deterministic "profile" describing the relative intensities of arrivals, and B is a random "busyness" parameter indicating how busy the day is. To simplify the analysis we assume that the day can be divided into periods so that $\lambda(\cdot)$ is constant within each period. The analysis that we present in this section generalizes beyond this particular model, but we return to this model for the RVAR experiments in Section 4.3.1.

The key long-run performance measure is the long-run fraction of customers that receive satisfactory service in a given period. A customer receives satisfactory service if her delay in queue is at most τ seconds. Common choices for τ are 20 seconds (a moderate delay) and 0 seconds (no delay).

For much of what follows we focus on a single period (e.g., 10am - 10.15am) in the day, arbitrarily representing this time period as time 0 through time t. With an abuse of notation, let Λ_i denote the real-valued random arrival rate within this period on day i. We assume that once the random arrival rate Λ_i is realized for the period on day i, it is constant throughout the period (i.e., from time 0 to time t).

Let S_i denote the number of satisfactory calls (calls that are answered within the time limit τ) in the period on day i out of a total of N_i calls that are received.

Notice that here we consider any call that abandons to be unsatisfactory. Some planners prefer to ignore calls that abandon within very short time frames. There is a difference, but it is not important for our discussion.

Over n days, the fraction of satisfactory calls is

$$\frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} N_i}.$$

Assume that days are i.i.d., the staffing level is fixed throughout, and $\mathrm{E}N_1 < \infty$. (Assuming days are i.i.d. ignores the inter-day correlations seen in Brown et al. [2005] and Steckley et al. [2005]. More general dependence structures can be captured in essentially the same framework.) The last assumption holds if $\mathrm{E}\Lambda_1 < \infty$. Dividing both the numerator and denominator by n and taking the limit as $n \to \infty$, the strong law then implies that the long-run fraction of satisfactory calls is

$$\frac{\mathrm{E}S_1}{\mathrm{E}N_1}. \tag{4.1}$$

This ratio gives performance as a function of staffing level. But how do we compute it?

First note that

$$\begin{aligned} \mathrm{E}N_1 &= \mathrm{E}\mathrm{E}[N_1|\Lambda_1] \\ &= \mathrm{E}[\Lambda_1 t] \\ &= t\mathrm{E}\Lambda_1, \end{aligned} \tag{4.2}$$

so that $\mathrm{E}N_1$ is easily computed. Computing $\mathrm{E}S_1$ is more difficult. We again condition on Λ_1 to obtain $\mathrm{E}S_1 = \mathrm{E}s(\Lambda)$, where $s(\lambda)$ is the conditional expected number of satisfactory calls in the period, conditional on $\Lambda_1 = \lambda$. Our initial goal is an expression for $s(\lambda)$.

Fix the arrival rate to be deterministic and equal to λ (for now). Let $X(\cdot;\lambda) = (X(s;\lambda) : s \geq 0)$ be a Markov process used to model the call center when there is a fixed arrival rate λ. In specialized cases one can take X to be the process giving the number of customers in the system, but it may be more complicated. Suppose that a customer arriving at time s will receive satisfactory service if and only if $X(s;\lambda) \in B$ for some distinguished set of states B.

Example 1 *A common model of a call center is an $M/M/c+M$ queue, i.e., the Erlang-A model. There are c servers, service times are exponentially distributed, and the arrival process is Poisson. Customers are willing to wait an exponentially-distributed amount of time (the "patience time") in the queue, and abandon if they do not reach a server by that time. Here we take $X(s;\lambda)$ to be the number of customers in the system at time s. Then X is a continuous-time Markov chain (CTMC). Suppose that a service is considered satisfactory if and only if the customer immediately reaches a server. Then we can take $B = \{0, 1, 2, ..., c-1\}$, i.e., a service is satisfactory if and only if the number of customers in the system is $c-1$ or less when the customer arrives.*

Example 2 *Consider the same model as in the previous example, but now define a service to be satisfactory if and only if the customer reaches a server in at most $\tau > 0$ seconds so long as she doesn't abandon. The state space of the CTMC defined in the previous example is no longer rich enough to determine, upon a customer arrival, whether that customer will receive satisfactory service or not. We turn to a different Markov process in such a case. Without loss of generality, suppose that as soon as a customer arrives, the patience and service times for that customer are sampled and therefore known. Since customers are served in FIFO order we can determine, for every customer that has arrived by time s, whether that customer* *will abandon or not, and if not which agent the customer will be served by. Let* $V_i(s;\lambda)$ *denote the virtual work load, i.e., the "work in process" for agent* i *at time* s, $i = 1, \ldots, c$. *The quantity* $V_i(s;\lambda)$ *gives the time required for agent* i *to complete the service of all customers in the system at time* s *that are, or will be, served by agent* i. *Let* $X(s;\lambda)$ *be the vector* $(V_i(s;\lambda) : 1 \leq i \leq c)$. *The process* $X(\cdot;\lambda) = (X(s;\lambda) : s \geq 0)$ *is a Markov process, albeit a rather complicated one, and we can take* $B = \{v : \min_{i=1}^{c} v_i \leq \tau\}$, *so that a service is satisfactory if and only if at least one server will be available to answer a call within* τ *seconds of a customer's arrival.*

Let $P_\varphi(\cdot)$ denote the probability measure when the Markov process has initial distribution φ. Let ν and π be, respectively, the distribution of the Markov process at time 0 and the stationary distribution (assumed to exist and be unique). Proposition 8 serves as a foundation for the use of steady-state approximations for performance measures in both the deterministic and random arrival rate contexts.

Proposition 8 *Under the conditions above,*

$$s(\lambda) = \lambda \int_0^t P_\nu(X(s;\lambda) \in B)\, ds.$$

If $\nu = \pi$, *so that the Markov process is in steady-state at time 0, then*

$$s(\lambda) = \lambda t f(\lambda),$$

where $f(\lambda) = P_\pi(X(0;\lambda) \in B)$ *is the steady-state probability that the system is in state* B. *We can interpret* $f(\lambda)$ *as the long-run fraction of customers that receive satisfactory service.*

Proof: For notational simplicity we suppress the dependence on λ. For $s \geq 0$, let $U(s) = I(X(s) \in B)$, where $I(\cdot)$ is the indicator function that is 1 if its argument is true and 0 otherwise. Note that X can be defined such that U is left continuous and has right hand limits. Let $L = (L(s) : s \geq 0)$ be the arrival process. Then L is a Poisson process with rate λ. For arbitrary $v \geq 0$, $(L(v+u) - L(v) : u \geq 0)$ is independent of $(U(s) : 0 \leq s \leq v)$ and $(L(s) : 0 \leq s \leq v)$. Then $s(\lambda) = \lambda \mathrm{E}_\nu \int_0^t U(s)\,\mathrm{d}s$ by the PASTA result (e.g., [Wolff, 1989, Section 5.16]). By Fubini's theorem, for arbitrary $v \geq 0$, $\mathrm{E}_\nu \int_0^v U(s)\,\mathrm{d}s = \int_0^v \mathrm{E}_\nu U(s)\,\mathrm{d}s$. Therefore

$$\mathrm{E}_\nu \int_0^v U(s)\,ds = \int_0^v P_\nu(X(s) \in B)\,ds. \tag{4.3}$$

Taking $v = t$, it follows that $s(\lambda) = \lambda \int_0^t P_\nu(X(s) \in B)\,ds$.

For the second result the system is in steady state at time 0 so that $\nu = \pi$. But $P_\pi(X(s) \in B) = P_\pi(X(0) \in B)$ for all $s \geq 0$. Defining $f(\lambda) = P_\pi(X(0) \in B)$, it follows from (4.3) that

$$\mathrm{E}_\pi \int_0^v U(s)\,ds = v f(\lambda), \tag{4.4}$$

and so $s(\lambda) = \lambda t f(\lambda)$.

To see that $f(\lambda)$ can be interpreted as the long-run fraction of customers that receive satisfactory service, define the stochastic process $A = (A(s) : s \geq 0)$, where $A(s) = \int_0^s U(u)\,dL(u)$. Then the fraction of customers that have received satisfactory service up to time v is given by $A(v)/L(v)$. It is assumed that as $v \to \infty$, $A(v)/L(v)$ converges to some constant p, where p is the long-run fraction of customers that receive satisfactory service. We show that $f(\lambda) = p$. From the PASTA result (e.g., [Wolff, 1989, Section 5.16]), since $A(v)/L(v)$ converges to p, $\int_0^v U(s)\,ds/v$ also converges to p as $v \to \infty$. But $p = \mathrm{E}_\nu p = \mathrm{E}_\nu \lim_{v\to\infty}(1/v) \int_0^v U(s)\,ds$. By the bounded convergence theorem,

$$\mathrm{E}_\nu \lim_{v\to\infty} \frac{1}{v} \int_0^v U(s)\,ds = \lim_{v\to\infty} \frac{1}{v} \mathrm{E}_\nu \int_0^v U(s)\,ds.$$

By (4.4), $\lim_{v\to\infty}(1/v)\mathrm{E}_\nu \int_0^v U(s)\,ds = f(\lambda)$. Therefore $f(\lambda) = p$. $\square$

4.1.1 Steady-State Approximations

Suppose that we adopt the steady-state approximation $s(\lambda) \approx \lambda t f(\lambda)$. Here λt is the expected number of customer arrivals in the period and $f(\lambda)$ is the long-run fraction of customers that receive satisfactory service. From (4.1) and (4.2) we see that

$$\frac{\mathrm{E}S_1}{\mathrm{E}N_1} = \frac{\mathrm{E}s(\Lambda_1)}{t\mathrm{E}\Lambda_1} \approx \frac{\mathrm{E}[\Lambda_1 f(\Lambda_1)]}{\mathrm{E}\Lambda_1}. \tag{4.5}$$

The fact that one should weight $f(\Lambda)$ by the arrival rate in (4.5) is well known. It is implicit (and at times explicit) in the work of Harrison and Zeevi [2005] and Whitt [2004] for example. Chen and Henderson [2001] did *not* perform this weighting in their analysis. So their results do not directly apply to the RVAR case, in contrast to what is claimed there. (But their results may apply in the UAR case considered in Section 4.2.)

What are the consequences of ignoring a randomly-varying arrival rate when predicting performance in a call center? In that case we would first estimate a deterministic arrival rate. The most commonly used estimates converge to $\mathrm{E}\Lambda_1$ as the data size increases. We then estimate performance as $f(\mathrm{E}\Lambda_1)$.

Together with (4.5), Proposition 9 below establishes that if f is decreasing and concave over the range of Λ_1, then we will overestimate performance if a random arrival rate is ignored. The function f is, in great generality, decreasing in λ. For many models it is also concave, at least in the region of interest; see Chen and Henderson [2001].

Proposition 9 *Suppose that f is decreasing and concave on the range of Λ_1. Then*

$$\frac{E[\Lambda_1 f(\Lambda_1)]}{E\Lambda_1} \leq f(E\Lambda_1).$$

Proof: We have that

$$\begin{aligned} \mathrm{E}[\Lambda_1 f(\Lambda_1)] &\leq (\mathrm{E}\Lambda_1)(\mathrm{E}f(\Lambda_1)) && (4.6) \\ &\leq (\mathrm{E}\Lambda_1)f(\mathrm{E}\Lambda_1) && (4.7) \end{aligned}$$

establishing the result. The inequality (4.6) follows since f is decreasing (see, e.g., Whitt [1976]), and (4.7) uses Jensen's inequality. □

For certain models and distributions of Λ_1, we may be able to compute (4.5) exactly. In general though, this will not be possible. In such a case we can use some numerical integration technique. The problem is quite straightforward since f is typically easily computed and the integral $\mathrm{E}[\Lambda_1 f(\Lambda_1)]$ is one-dimensional.

We now turn from long-run performance to short-run performance. We want to determine the distribution of S_1/N_1, the fraction of satisfactory calls in a single period $[0, t]$ of a single day. (We define $0/0 = 1$.) Our approach is to condition on Λ, the arrival rate for the period.

Suppose that conditional on Λ, the period is long enough that the fraction of calls answered on time is close to its steady-state mean $f(\Lambda)$. This transformation of the random variable Λ is our first approximation. It ignores the "process variability" that arises even for a fixed arrival rate.

We can refine this approximation to take into account process variability. The key to the refinement is a central limit theorem (CLT) for S_1/N_1 assuming a fixed λ. We first show how to establish the CLT under special conditions, obtaining an expression for the variance $\sigma^2(\cdot)$ in the process, and then argue that it should hold in much greater generality (albeit with a difficult-to-compute variance).

Let the arrival rate λ be fixed. Suppose that our goal is to answer calls immediately. Suppose further that the number-in-system process $X = (X(s) : s \geq 0)$ can be modeled as an irreducible continuous-time Markov chain on the finite state space $\{0, 1, \ldots, d\}$, where $d > c$. (It is not essential that the state space be finite, but it allows us to avoid verifying technical conditions.) Let $M(s)$ be the number of transitions by time s, and let $Y = (Y_n : n \geq 0)$ be the embedded discrete-time Markov chain. Then we can write

$$\frac{S_1}{N_1} \approx \frac{U_{M(t)}}{V_{M(t)}}, \tag{4.8}$$

where

$$\begin{aligned} U_n &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1}+1, Y_{i-1} \leq c-1) \text{ and} \\ V_n &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1}+1). \end{aligned}$$

Here U_n gives the fraction of the first n transitions that correspond to an arriving customer finding a server available. Similarly, V_n gives the fraction of the first n transitions that correspond to an arrival joining the system. Notice that V_n does not count blocked customers. This is why the relation in (4.8) is not an equality. When d is large enough that few customers are turned away, the approximation should be very good.

Theorem 10 *Under the assumptions given above,*

$$\sqrt{\lambda s}\left(\frac{U_{M(s)}}{V_{M(s)}} - \frac{u}{v}\right) \Rightarrow N(0, \sigma^2(\lambda))$$

as $s \to \infty$, where u, v and $\sigma^2(\lambda)$ are specified in the proof below.

Proof: The proof has 3 steps. The key step is to establish the joint CLT

$$\sqrt{n}\left(\begin{pmatrix} U_n \\ V_n \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow N(0, \Sigma) \tag{4.9}$$

as $n \to \infty$, where $N(0, \Sigma)$ denotes a Gaussian random vector with mean 0 and covariance matrix Σ, and u, v and Σ are specified below. The final 2 steps consist of applying a random time change and then the delta method.

To establish (4.9) we apply a Markov chain CLT (see, e.g., [Meyn and Tweedie, 1993, Theorem 17.4.4]). That result applies only to univariate processes, but the result easily extends to multivariate processes through an application of the Cramér-Wold device (see, e.g., [Billingsley, 1968, Theorem 7.7]). Consider the (irreducible, finite-state-space) Markov chain $\tilde{Y} = (\tilde{Y}_i : i \geq 0)$, where $\tilde{Y}_i = (Y_i, Y_{i+1})$. We can write

$$\begin{aligned} U_n - u &= \frac{1}{n} \sum_{i=0}^{n-1} \tilde{h}_1(\tilde{Y}_i) \text{ and} \\ V_n - v &= \frac{1}{n} \sum_{i=0}^{n-1} \tilde{h}_2(\tilde{Y}_i), \end{aligned}$$

where

$$\begin{aligned} \tilde{h}_1(x, y) &= I(y = x + 1, x \leq c - 1) - u \text{ and} \\ \tilde{h}_2(x, y) &= I(y = x + 1) - v. \end{aligned}$$

Let $\tilde{\pi}$ be the stationary distribution of $\tilde{Y}$. We choose u and v to be steady-state means, so that $\tilde{\pi}\tilde{h}_i = \sum_{(x,y)} \tilde{\pi}(x, y)\tilde{h}_i(x, y) = 0$ for $i = 1, 2$. Let $\tilde{P}$ be the transition matrix of $\tilde{Y}$, and let $\tilde{g}_1$ and $\tilde{g}_2$ solve Poisson's equation

$$\tilde{P}\tilde{g}_i(x, y) = \tilde{g}_i(x, y) - \tilde{h}_i(x, y),$$

for $i = 1, 2$ and all (x, y). We then obtain (4.9), where

$$\begin{aligned} \Sigma_{ij} &= \mathrm{E}_{\tilde{\pi}}[(\tilde{g}_i(\tilde{Y}_1) - \tilde{P}\tilde{g}_i(\tilde{Y}_0))(\tilde{g}_j(\tilde{Y}_1) - \tilde{P}\tilde{g}_j(\tilde{Y}_0))] \\ &= \mathrm{E}_{\tilde{\pi}}[\tilde{g}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0) + \tilde{h}_i(\tilde{Y}_0)\tilde{g}_j(\tilde{Y}_0) - \tilde{h}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0)], \end{aligned}$$

where the second equality follows as in Meyn and Tweedie [1993], Equation 17.47.

In fact, we obtain a stronger result, namely a functional CLT. This observation, together with the random-time-change result [Billingsley, 1968, Theorem 17.1] allows us to conclude that

$$\sqrt{M(s)}\left(\begin{pmatrix} U_{M(s)} \\ V_{M(s)} \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow N(0,\Sigma)$$

as $s \to \infty$. Now, $M(s)/s \to \gamma$ as $s \to \infty$ a.s., where γ is the long-run rate of transitions in the continuous-time Markov chain X. The converging-together lemma [Billingsley, 1968, Problem 1, p. 28] then implies that

$$\sqrt{\gamma s}\left(\begin{pmatrix} U_{M(s)} \\ V_{M(s)} \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow N(0,\Sigma)$$

as $s \to \infty$.

The final step applies the delta method (e.g., [Billingsley, 1968, Problem 2, p. 34], using the function $\phi(x,y) = x/y$, to conclude that

$$\sqrt{\gamma s}\left(\frac{U_{M(s)}}{V_{M(s)}} - \frac{u}{v}\right) \Rightarrow N(0,\eta^2),$$

where

$$\begin{aligned} \eta^2 &= \nabla\phi(u,v)^T \Sigma \nabla\phi(u,v) \\ &= \frac{\Sigma_{11} - 2(u/v)\Sigma_{12} + (u/v)^2\Sigma_{22}}{v^2}. \end{aligned}$$

Setting $\sigma^2(\lambda) = \lambda\eta^2/\gamma$ yields the result. □

Equation (4.8) and Theorem 10 establish that conditional on Λ, the fraction S_1/N_1 is approximately normally distributed with mean $u/v \approx f(\Lambda)$ and variance $\sigma^2(\Lambda)/\Lambda t$. So we can approximate the distribution of S_1/N_1 by the normal mixture $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$.

Remark 2 The variance of this normal mixture is

$$\mathrm{var} f(\Lambda) + \mathrm{E}\frac{\sigma^2(\Lambda)}{\Lambda t},$$

which can be viewed as a decomposition of the variance into contributions from arrival rate uncertainty and process uncertainty respectively.

To compute the distribution of this normal mixture we need to be able to compute the constant $\sigma^2(\lambda)$, which in turn depends on γ and η^2 (which also depend on λ). The following formulae are useful in this regard. They exploit the strong relationships between the 2-step Markov chain $\tilde{Y}$ and the single-step Markov chain Y, and between the continuous-time Markov chain X and its embedded chain Y. Let $\beta(i)$ denote the rate at which the CTMC X leaves state i, and let π_X and π_Y denote the steady-state distributions associated with X and Y respectively. Since

$$\pi_X(y) = \frac{\pi_Y(y)/\beta(y)}{\sum_z \pi_Y(z)/\beta(z)},$$

it follows that

$$\gamma = \sum_{y=0}^{d} \pi_X(y)\beta(y) = \left(\sum_{z=0}^{d} \pi_Y(z)/\beta(z)\right)^{-1}.$$

Note that π_X or π_Y are easily computed, and therefore so is γ.

We also need to compute u and v. These are given by

$$\begin{aligned} u &= \sum_{i=0}^{c-1} \pi_Y(i) P_Y(i, i+1) \text{ and} \\ v &= \sum_{i=0}^{d-1} \pi_Y(i) P_Y(i, i+1), \end{aligned}$$

where P_Y is the transition matrix of Y.

Finally, recall that for $1 \leq i, j \leq 2$

$$\begin{aligned} \Sigma_{ij} &= \mathrm{E}_{\tilde{\pi}}[\tilde{g}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0) + \tilde{h}_i(\tilde{Y}_0)\tilde{g}_j(\tilde{Y}_0) - \tilde{h}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0)] \\ &= \sum_{x,y} \pi_Y(x) P_Y(x,y)[\tilde{g}_i(x,y)\tilde{h}_j(x,y) + \tilde{h}_i(x,y)\tilde{g}_j(x,y) - \tilde{h}_i(x,y)\tilde{h}_j(x,y)]. \end{aligned}$$

It remains to specify how to compute $\tilde{g}_i(x,y)$. Define

$$h_i(x) = \mathrm{E}_x \tilde{h}_i(x, Y_1) = \sum_{y=0}^{d} \tilde{h}_i(x,y) P_Y(x,y)$$

to be the "smoothed" version of $\tilde{h}_i$, for $i = 1,2$ and $x = 0,\ldots,d$. There are multiple solutions to the equations defining $\tilde{g}_i$, all of which differ by an additive constant. In what follows we use one such solution for $\tilde{g}_i$, which is

$$\begin{aligned}
\tilde{g}_i(x,y) &= \sum_{k=0}^{\infty} \mathrm{E}_{(x,y)} \tilde{h}_i(Y_k, Y_{k+1}) \\
&= \tilde{h}_i(x,y) + \sum_{k=1}^{\infty} \mathrm{E}_{(x,y)} \tilde{h}_i(Y_k, Y_{k+1}) \\
&= \tilde{h}_i(x,y) + \sum_{k=1}^{\infty} \mathrm{E}_{(x,y)} h_i(Y_k) \\
&= \tilde{h}_i(x,y) + g_i(y),
\end{aligned}$$

where

$$g_i(y) = \sum_{k=0}^{\infty} \mathrm{E}_y h_i(Y_k)$$

solves $(P_Y - I)g_i(y) = -h_i(y)$ for all y, and has the property that $\pi_Y g_i = 0$. It is therefore possible to compute g_i from these latter relations, and then substitute back to obtain $\tilde{g}_i$.

We believe that a central limit theorem result holds in greater generality. Given a fixed arrival rate λ, now assume only that the call center can be modeled with a Markov process $X = (X(s) : s \geq 0)$ and that a customer arriving at time s will receive satisfactory service if and only if $X(s) \in B$ for some distinguished set of states. For Theorem 10, we required that X be the number-in-system process and $B = \{0, 1, \ldots, c\}$ which corresponds to the case in which a call is satisfactory if and only if it is handled immediately. In assuming only that X is a Markov process, we now allow for the case in which a call is satisfactory if and only if it is answered within $\tau > 0$ seconds so long as the call doesn't abandon. See Example 2 for more on this case.

For fixed arrival rate λ and assuming $X = (X(s) : s \geq 0)$ is a Markov process, we argue heuristically (non-rigorously) as follows. Let T_i denote the time of the ith customer arrival. Define $Z_i = X(T_i)$ to be the state of the Markov process at the time of the ith customer arrival. The ith customer receives satisfactory service if and only if $Z_i \in B$. So assuming a fixed arrival rate λ, S_1/N_1 has the same distribution as

$$\frac{1}{N(t)} \sum_{i=1}^{N(t)} I(Z_i \in B),$$

where $N(s)$ is a Poisson random variable with mean λs giving the number of arrivals in $[0, s]$.

The strong Markov property for $X(\cdot)$ ensures that $(Z_i : i \geq 1)$ is a Markov chain. We can then apply a central limit theorem (e.g., Meyn and Tweedie [1993, Chapter 17]) to assert that under appropriate conditions

$$\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} I(Z_i \in B) - f(\lambda) \right] \Rightarrow \sigma(\lambda) N(0,1),$$

as $n \to \infty$, where $\sigma^2(\lambda)$ is a variance constant. Again under appropriate conditions a random-time-change theorem ensures that

$$N^{1/2}(s) \left[\frac{1}{N(s)} \sum_{i=1}^{N(s)} I(Z_i \in B) - f(\lambda) \right] \Rightarrow \sigma(\lambda) N(0,1)$$

as $s \to \infty$. A converging-together argument then ensures that

$$(\lambda s)^{1/2} \left[\frac{1}{N(s)} \sum_{i=1}^{N(s)} I(Z_i \in B) - f(\lambda) \right] \Rightarrow \sigma(\lambda) N(0,1). \tag{4.10}$$

The limit result (4.10) then ensures that, so long as t is "large enough", conditional on $\Lambda = \lambda$,

$$\frac{S_1}{N_1} =_d \frac{1}{N(t)} \sum_{i=1}^{N(t)} I(Z_i \in B) \approx_d N\left(f(\lambda), \frac{\sigma^2(\lambda)}{\lambda t} \right),$$

where $\approx_d$ denotes approximate equality in distribution. Unconditioning, we then assert that

$$\frac{S_1}{N_1} \approx_d N\left(f(\Lambda), \frac{\sigma^2(\Lambda)}{\Lambda t}\right), \tag{4.11}$$

so that the realized fraction of acceptable calls is approximately a mixture of normal random variables.

This argument follows a similar path to the one in the proof of Theorem 10. First, we established a central limit theorem for the fraction of the first n customers that receive satisfactory service, then we applied a random time change and finally we used the converging-together lemma. However this more general result is limited in that we do not always know the variance constant $\sigma^2(\cdot)$ and in some cases we do not know the long-run fraction of satisfactory calls $f(\cdot)$ either.

4.1.2 Simulation-Based Estimates

The approximations for long-run and short-run performance described above may be inappropriate, either because the steady-state approximations for time-dependent quantities may be inaccurate for a non-negligible set of arrival rates, or because the true system is not well modeled by simple models for which steady-state results are readily computed. It is natural to then turn to simulation to compute performance measures.

In terms of long-run performance, we have already noted that the problem reduces to computing $\mathrm{E}S_1$, the expected number of satisfactory calls in a particular period. This is straightforward using simulation. One can simply generate the arrival rate process, Λ say, and then conditional on the realized value, simulate the call center for the day, giving a realization of S_1. Repeating this process in i.i.d. fashion gives $S_1, \ldots, S_n$ say, which can be averaged to give an estimate of $\mathrm{E}S_1$.

But we can develop more efficient (in the sense of lower variance) estimators of $\mathrm{E}S_1$ by taking advantage of structure.

For definiteness, suppose we adopt the model that the arrival rate is given by $B(\lambda(s) : s \geq 0)$, where $\lambda(\cdot)$ is constant in each period, and B is a random "busyness" factor. If we know $\mathrm{E}B$ then we can use $B - \mathrm{E}B$ as a control variate, i.e., we use

$$\frac{1}{n}\sum_{i=1}^{n}(S_i - \beta(B_i - \mathrm{E}B))$$

to estimate $\mathrm{E}S_1$, where $((S_i, B_i) : i = 1, \ldots, n)$ are i.i.d. and distributed as (S_1, B), and β is a constant that is chosen to maximize the variance reduction; see, e.g., Law and Kelton [2000]. However, we will typically know much more than just the mean of B. If we know its distribution then, as discussed in Glasserman [2004], p. 220, stratifying on B should yield larger variance reduction than using $B - EB$ as a control variate. See Glasserman [2004], Section 4.3, for details on how to implement stratification.

For short-run performance we wish to compute the distribution of S_1/N_1. This random variable does not have a (Lebesgue) density since it is supported on the rationals. Its probability mass function is also uninformative. Therefore, we would probably estimate a moderately coarse histogram (say, with bins of width $\Delta x = 0.01$). The height of the bin $[x, x + \Delta x]$ is proportional to $F(x + \Delta x) - F(x)$, where F is the distribution function of S_1/N_1. Hence, estimating this histogram is equivalent to estimating the distribution function at the fixed set of points $\Delta x, 2\Delta x, \ldots, 1$. This estimation is straightforward based on i.i.d. observations of (S_1, N_1), and one can apply standard results (e.g., Ross [1996], pp. 360–363) to compute tolerance bounds for F. As with estimating $\mathrm{E}S_1$, one can stratify on the busyness parameter B to reduce variance in the estimation of the quantities

$F(x) = P(S_1 - xN_1 \leq 0)$.

4.2 Uncertain Arrival Rates

Suppose that the arrival rate function does not vary from day to day and is given by the fixed function $(\lambda(s) : s \geq 0)$ say, but we do not know this function with certainty. This situation can arise, for example, when a call center is opening for the first time, when a new product is added to an existing portfolio of products, or when a new marketing promotion comes into effect. It corresponds to what we usually interpret as "forecast uncertainty," and commonly arises in dynamic business environments.

Just as in the RVAR case, in the long run we are interested in $\mathrm{E}S_1/\mathrm{E}N_1$, the long-run fraction of satisfactory calls in a given period, and in the short run we are interested in the distribution of S_1/N_1, the fraction of satisfactory calls in a single period in a single day. In the long-run we will eventually learn the true arrival rate through observation, but decisions need to be made before that eventuates, which helps to explain our interest in this case.

We focus on a single period $[0, t]$ of the day and assume that the true arrival rate takes on the constant value λ^* in this period. Let Λ denote a random variable representing our knowledge of the value λ^*. In the RVAR case we obtained a new sample from the distribution of Λ every day. In contrast, in the UAR case, although we cannot directly observe it, Λ takes on the true value λ^* on the first day and then remains constant.

Arguing as in the previous section, conditional on $\Lambda = \lambda$ the long-run fraction of satisfactory calls is

$$\frac{\mathrm{E}[S_1|\Lambda = \lambda]}{\mathrm{E}[N_1|\Lambda = \lambda]} = \frac{s(\lambda)}{\lambda t}, \tag{4.12}$$

where we have used the notation $s(\lambda)$ for the conditional expectation of S_1 given $\Lambda = \lambda$. The unconditional long-run performance is therefore $s(\Lambda)/(\Lambda t)$, which is random because it depends on the unknown Λ. We might then select the staffing level so that, with high probability, this fraction is larger than some specified level. But how do we compute $s(\Lambda)$?

4.2.1 Steady-State Approximations

In this section we employ the steady-state approximation $s(\lambda) = \lambda t f(\lambda)$. In that case we see from (4.12) that long-run performance is simply $f(\Lambda)$. The expected long-run performance is then $\mathrm{E}f(\Lambda)$, which differs from the RVAR case in that it does not weight the function f by Λ, as noted in the introduction. The random variable $f(\Lambda)$ can be analyzed reasonably easily once f and the distribution of Λ are known, as noted earlier.

Turning to short-run performance, the argument leading to Theorem 10 is directly relevant, and so we can approximate the distribution of S_1/N_1 as

$$N(f(\Lambda), \sigma^2(\Lambda)/(\Lambda t)).$$

This distribution is an amalgam of *parameter uncertainty* in that the true arrival rate is unknown, and *process uncertainty* that is exhibited through the normal distribution for any given Λ.

4.2.2 Simulation-Based Estimates

If steady-state approximations are deemed inappropriate then we may turn to simulation. Recall from (4.12) that long-run performance is given by the random quantity $s(\Lambda)/(\Lambda t)$. We can write this as

$$\mathrm{E}\left[\left.\frac{S_1}{\Lambda t}\right|\Lambda\right]. \tag{4.13}$$

The distribution function of the conditional expectation (4.13) is relevant for computing the probability that long-run performance is satisfactory. The density is of interest in understanding how the uncertainty modeled by Λ translates into uncertainty about performance. Methods for estimating the distribution function of a conditional expectation can be found in Lee [1998]. Methods for estimating the density of a conditional expectation can be found in Chapters 2 and 3. These simulation methods involve a combination of "macro replications" that sample observations of Λ, and "micro-replications" that estimate the conditional expectation for a sampled value of Λ.

One may prefer to simply determine summary statistics of (4.13) such as the mean. In this case, the discussion given in Section 4.1.2 about the use of stratified sampling is directly relevant.

Recall that the short-run performance measures in the UAR case coincide exactly with those for the RVAR case, and so the methods sketched in Section 4.1.2 are directly relevant.

4.3 Experimental Insights

We conducted experiments to examine performance given a randomly varying arrival rate. In Section 4.3.1 we focus on the RVAR case, using data from four call centers studied in Steckley et al. [2005] to guide our experiments. Specifically, we wanted to determine which factors impact the performance measures discussed in Section 4.1, assess the quality of the approximations as compared to the simulation-based estimates of performance, and learn more about the behavior of systems with a random arrival rate. The factors we chose to examine included (a) the level of variability in the (Poisson) arrival rate; (b) the duration of the (exponential) service times; and (c) the (exponential) rate at which customers abandon the system.

In Section 4.3.2 we assess the quality of the UAR measures discussed in Section 4.2 paying special attention to the long-run, simulation-based density estimate which is an application of the work from Chapters 2 and 3.

4.3.1 RVAR Experiments

In this section we consider the RVAR performance measures. The design of the experiment is discussed in Section 4.3.1 and the results are presented in Section 4.3.1. We continue to focus our analysis on a single period.

Experimental Design

For our experiments, we model the call center as an $M/M/c+M$ queue (i.e., the Erlang-A model) with a random arrival rate Λ. We adopt the model adapted from Whitt [1999] discussed earlier in which the arrival rate in the ith instance of the period is given by $B_i\lambda$, where the B_is are i.i.d. We model B_i as uniform with mean 1 so that Λ is uniform with mean λ. We chose the uniform distribution because it is simple and it effectively illustrates the essential ideas. One could easily substitute a more realistic distribution. The choice of the endpoints of the uniform distribution are discussed below.

For these experiments, we have set the length of the period at one hour. A call is defined to have received satisfactory service if it is answered immediately, i.e.,

$\tau = 0$.

Using both the analytic approximations discussed above and the corresponding simulation models, we estimate the performance measures discussed in Section 4.1.1 and Section 4.1.2 for a number of scenarios. The simulations were modeled and run using software developed by Eric Buist and Pierre L'Ecuyer (Buist and L'Ecuyer [2005]), which was chosen for its ease of modeling call center operations and capturing the desired performance statistics, as well as its very fast simulation run times.

The scenarios are summarized in Table 4.1. We vary the expected number of calls per hour (λ). We also vary the variability in the arrival rate in terms of a quantity we call the variance factor. The variance factor is defined as the ratio of the variance of the number of calls per hour under the random arrival rate Λ and the variance of the number of calls per hour given a deterministic arrival rate λ. The level of the variance factor then determines the endpoints of the uniform distribution for Λ and thus determines the variability of Λ. Finally, we allow the mean service time and mean abandonment time to vary.

The range of variance factors (as well as arrival rates and average handle times) included in these experiments is based on the actual historical data from four diverse call centers that we have studied; additional details and examples from this dataset are presented in Steckley et al. [2005].

In Table 4.1, a variance factor of one corresponds to the case in which the arrival rate is deterministic and equal to λ. An abandonment rate of 0 corresponds to the case in which there is no abandonment, in which case the call center is modeled as an $M/M/c$ queue.

For each scenario, we selected the number of servers c to be the minimum value

Table 4.1: Experimental Design

Factor	Levels
Mean number of calls per hour (λ)	250 1000 4000
Variance factor	1 3 6
Service rate per hour (μ)	12 6
Abandonment rate per hour(θ)	0 6 12

so that the long-run fraction of calls that are served immediately for a system with a deterministic arrival rate λ is at least 90%.

For the simulations, we used an extensive warm-up period. The parameter settings (arrival rate, service time distribution, abandonment time distribution) for the warm-up period were identical to those used in the simulation of the actual period for which data was captured. Therefore, our data reflects steady-state performance.

Results

Both the simulation-based estimates and steady-state approximations for long-run performance (long-run fraction of satisfactory calls) are reported in Table 4.2. The simulation results are accurate to approximately 2 decimal places, and so are reported only to that accuracy. Due to space considerations we present only selected scenarios. This selection illustrates the essential characteristics and trends seen in the results as a whole.

The approximations and simulation-based estimates are very similar. We ex-

Table 4.2: Simulation-based estimates and approximations (in parentheses) of long-run performance

			Variance factor		
λ	μ	θ	1	3	6
250	12	0	0.91 (0.91)	0.87 (0.87)	0.82 (0.81)
1000	12	0	0.90 (0.90)	0.87 (0.87)	0.82 (0.82)
4000	12	0	0.91 (0.90)	0.88 (0.87)	0.83 (0.82)
1000	6	0	0.91 (0.91)	0.85 (0.84)	0.76 (0.76)
1000	12	6	0.89 (0.90)	0.87 (0.87)	0.84 (0.84)
1000	12	12	0.90 (0.91)	0.88 (0.89)	0.86 (0.86)

pect such agreement since the simulated period should exhibit steady-state behavior after the extensive warm-up we used. In fact, assuming that the simulated periods are in steady-state, we have equality in (4.5). Then as the number of simulated periods tends to infinity, the simulation-based estimate tends to the steady-state approximation.

When the variance factor is one so that there is no variability in the arrival rate, the long-run fraction of satisfactory calls is very close to 0.9. This is because the number of servers c is specifically chosen so that the long-run fraction of satisfactory calls will be at least 0.9 in this case. When the variance factor is strictly greater than one, so that there is variability in the arrival rate, the long-run fraction of satisfactory calls is less than 0.9 as suggested by Proposition 9. We also see that the more variable the arrival rate, the worse the performance. We see that the degradation can be significant. It is on the order of 5% - 10% for some of the cases.

The results also indicate that abandonment reduces the negative impacts of variability in the arrival rate. To understand this, note that in a no-abandonment model, customers with long waiting times remain in the system, creating a "chain reaction" of waiting for future customers. In contrast, with abandonment, these customers leave the system quickly, thereby avoiding the chain reaction encountered in a no-abandonment model. This reasoning suggests that the same trend would be observed if we had instead defined a call to have received satisfactory service if the call does not abandon and is answered within $\tau > 0$ seconds. Although we believe this trend holds in general, in some cases in which τ is very large and the rate of abandonment θ is also very large, the abandoning calls may actually drive down the long-run fraction of satisfactory calls.

For short-run performance, we turn to the distribution of S_1/N_1, the fraction of satisfactory calls in a single instance of the period. We have two possible approximations for this distribution. The first is given by the distribution of $f(\Lambda)$. The second is given by the distribution of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$. Figure 4.1 plots the simulation-based estimate of the distribution (histogram) along with the density of the two approximations for a particular case. The final bar of the histogram corresponds to the observed S_1/N_1 ratios that were exactly one. The density of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$ has been truncated at one and the probability of the truncated region has been plotted as a "histogram" bar just to the right of one. The estimate of the density of $f(\Lambda)$ is a kernel estimate.

The simulation-based histogram shows that the distribution of S_1/N_1 has a spike around one and a skewed left tail for the given staffing level. We saw the same general shape for all the scenarios in which there is variability in the arrival rate.

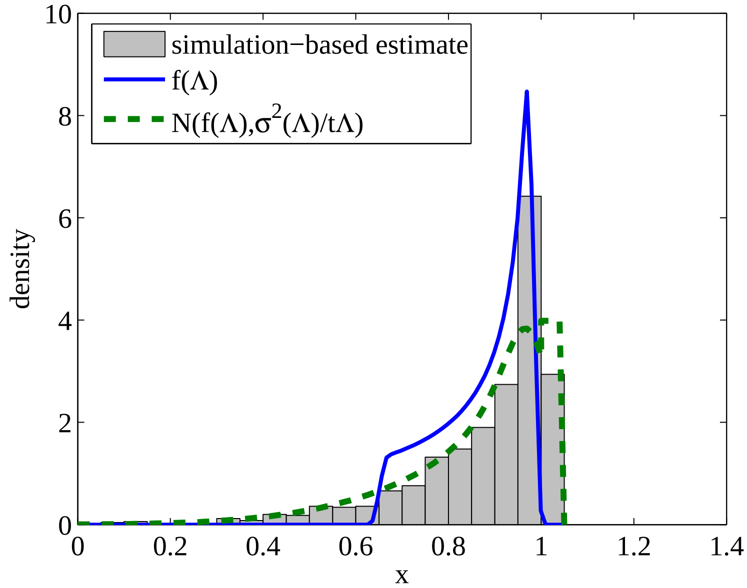

Figure 4.1: Distribution estimates when $\lambda = 1000$, $\mu = 12$, $c = 97$, $\theta = 0$, and the variance factor $= 3$.

The shape indicates that it is quite likely that performance for a single instance of the period will be excellent, with the fraction of satisfactory calls greater than 0.9. But with a significant probability, the fraction of satisfactory calls will be less than 0.9 and can be as bad as 0.5 with nontrivial probability.

The approximations in Figure 4.1 track the simulation-based results fairly well. The normal mixture approximation is a much better estimate in the left tail.

To better understand the general shape of the distribution when there is variability in the arrival rate, consider Figure 4.2 which plots the mean $f(\cdot)$ and variance $\sigma^2(\cdot)/(\cdot)t$ of the normal mixture over the support of the arrival rate distribution for the case plotted in Figure 4.1. When the arrival rate is small, the mean is very close to one and the variance is very small. This corresponds to the situation in which the call center is comfortably overstaffed and nearly all calls receive satisfactory service. For such λ, $N(f(\lambda), \sigma^2(\lambda)/\lambda t)$ has a very concentrated density in

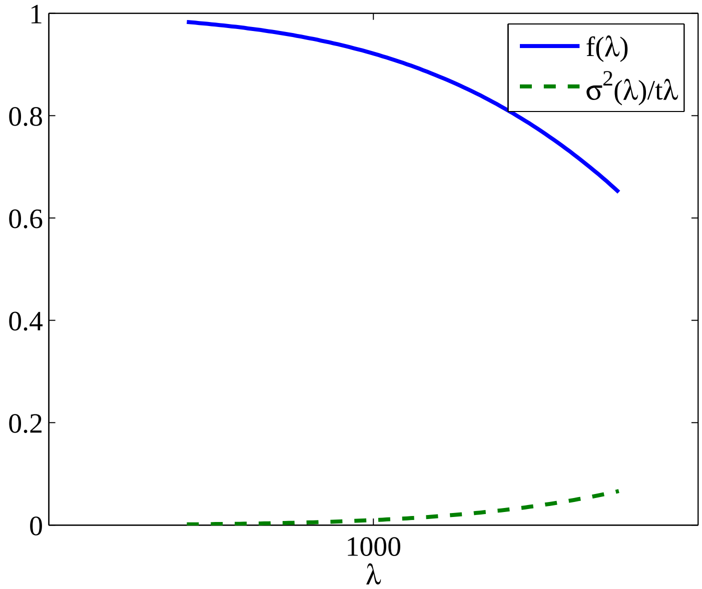

Figure 4.2: Plot of $f(\cdot)$ and $\sigma^2(\cdot)/t(\cdot)$ for the scenario of Figure 4.1.

the neighborhood of one. The larger arrival rates result in lower means and higher variances. This corresponds to a situation in which the call center is understaffed and performance becomes more variable. In such cases, $N(f(\lambda), \sigma^2(\lambda)/\lambda t)$ takes on small values and is more dispersed.

In Figure 4.3, we present a plot of the various estimates for the case in which all parameters are the same, except the variance factor which has increased to 6. There is now an even greater skew in the left tail, which means that there is higher probability of disastrous performance for a single instance of a period. In fact, as variability in the arrival rate becomes extremely large (variance factor ≥ 50), the distribution of S_1/N_1 becomes bimodal with one mode at 1 and the other at 0. Intuitively, the arrival rate distribution is so spread out that it rarely takes on values that our staffing level is designed to handle, instead taking values that are either very large, or very small relative to the staffing level. Therefore, performance

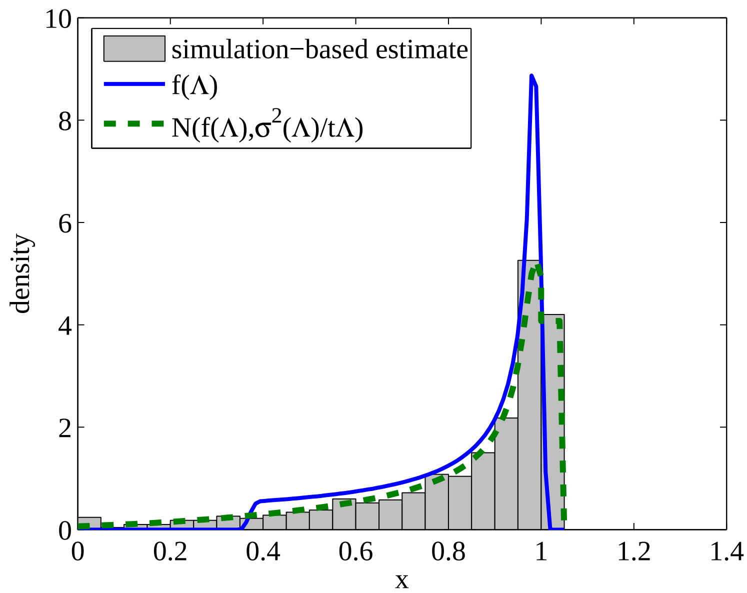

Figure 4.3: Plots of the distribution estimates when $\lambda = 1000$, $\mu = 12$, $c = 97$, $\theta = 0$, and the variance factor $= 6$.

is either very poor, or very good, with little chance of moderate performance.

Further examination of Figures 4.1 and 4.3 suggests that the approximations improve as variability in the arrival rate increases. Indeed, we saw this trend in the other scenarios in our experimental design. To understand this trend, first note that the normal approximation for S_1/N_1 is provably good when the periods are long, but deteriorates as the periods become shorter. For shorter periods, N_1 can be small with high probability. As a consequence, the actual distribution of S_1/N_1 will exhibit a right skew. Note that the right skew will be less for small λ since S_1/N_1 then clusters around one. But for any deterministic λ, there will be a discrepancy in the symmetric normal approximation and the right-skewed actual distribution. When the arrival rate Λ is random we smooth the normal approximation over the possible values of Λ to get our approximation $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$. The approximation is essentially a kernel density estimate with local bandwidth

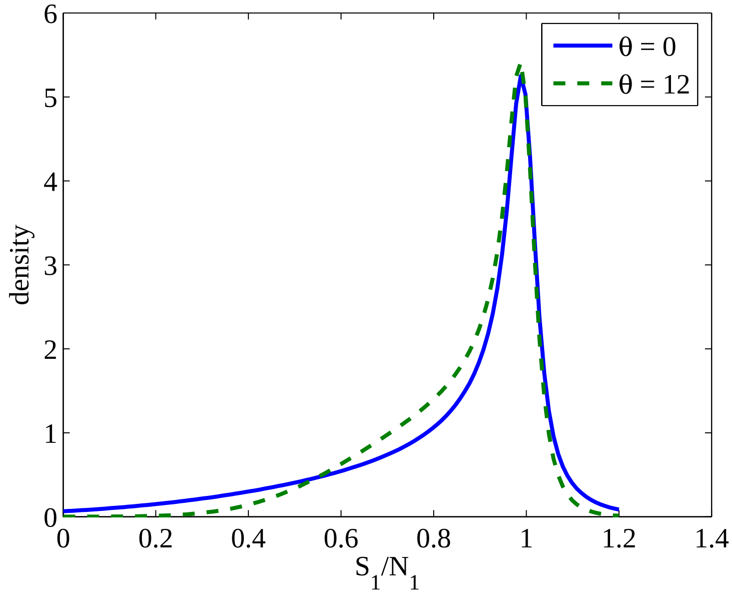

Figure 4.4: Plot of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$ when $\theta = 0$ and $\theta = 12$, with $\lambda = 1000$, $\mu = 12$, $c = 97$, and the variance factor $= 6$.

$\sigma^2(\cdot)/t(\cdot)$. Figure 4.2 shows that for large λ, where the discrepancy between the normal approximation and actual distribution is significant, $\sigma^2(\lambda)/t(\lambda)$ is relatively large and we smooth more heavily. For smaller λ when the discrepancy is less significant, we do less smoothing. As a result, the approximation gets visually tighter.

To examine the effect of abandonment on short-run performance, we plot the density of $N(f(\Lambda), \sigma^2(\Lambda)/\Lambda t)$ for a particular scenario with, and without, abandonment in Figure 4.4. The densities are very similar around one but the density corresponding to abandonment is less skewed to the left. Similar characteristics are seen in the simulation-based histogram and the distribution of $f(\Lambda)$. The intuition here is the same as for the effect of abandonment on long-run performance.

4.3.2 UAR Experiments

For the UAR experiments, we again model the call center as an $M/M/c + M$ queue (i.e., the Erlang-A model). Recall that in the UAR case, the arrival rate λ for the given period is assumed to be deterministic but unknown. We continue to assume that the arrival rate is constant throughout the period. Here we model our uncertainty in the arrival rate with Λ. In particular, let $\Lambda = B\lambda$, where B is uniformly distributed with mean one, and where the endpoints of the distribution are determined by a variance factor, as was done in Section 4.3.1. Then Λ is uniform with mean λ, i.e., our uncertainty in the arrival rate is uniformly distributed about the true arrival rate λ. The level of uncertainty is measured by the variance factor. Similar to the RVAR experiments, a more realistic distribution could have been chosen for Λ but we chose the uniform distribution because it is simple and effectively illustrates the essential ideas.

We have set the length of the period to one hour. Similar to Section 4.3.1, a call is defined to have received satisfactory service if it is answered immediately, and we selected the number of servers to be the minimum value so that the long-run fraction of calls that are served immediately for a system with a deterministic arrival rate λ is at least 90%. To avoid confusing the number of servers with the computer budget, we let $\tilde{c}$ denote the number of servers and c denote the computer budget in this section.

Recall that the short-run measures for the RVAR case and the UAR case have the same form, although we note that the interpretations are slightly different. To get a sense of how short-run UAR measures behave, consult the plots of the short-run RVAR measures in Section 4.3.1. In this section we focus on the long-run measures presented in Section 4.2, which estimate the distribution of

$$\mathrm{E}\left[\left.\frac{S_1}{\Lambda t}\right|\Lambda\right].$$

For the simulation-based estimate, we estimate the density of this random variable using the global estimator and implementation introduced in Chapters 2 and 3.

In order to assess the quality of this estimator we did the following. For each simulation of the call center with a given arrival rate, service rate and abandonment rate, the initial conditions (number-in-system) were sampled from the steady-state distribution with the same parameter settings. Therefore, the simulated data reflects steady-state behavior. As a result, the steady-state approximation which is the density of $f(\Lambda)$ is the target density of the simulation-based estimates.

In Figure 4.5 we plot the simulation-based estimate at two different computer budgets along with the steady-state density. Here, $\lambda = 1000$, $\mu = 12$, $\tilde{c} = 97$, $\theta = 0$, and the variance factor $= 3$.

It appears that the simulation-based estimates are converging to the steady-state density as expected. But even after a computer budget of $c = 2^{19}$ (which results in an EBBS optimal allocation of $n = 2473$ external samples), there is still appreciable error. This problem is particularly difficult for kernel density estimation given the boundary at 1 (for more on the boundary issue, see Wand and Jones [1995]).

The general shapes of the long-run UAR density estimates presented in Figure 4.5 are similar to the short-run RVAR density estimates in Figure 4.1. In fact, we see that the figures share the same parameter settings and both plot the density of $f(\Lambda)$. But we stress that the interpretations are different. For the UAR case, the interpretation is as follows. Given the uncertainty in the arrival rate modeled by Λ, with high probability the long-run fraction of calls answered immediately

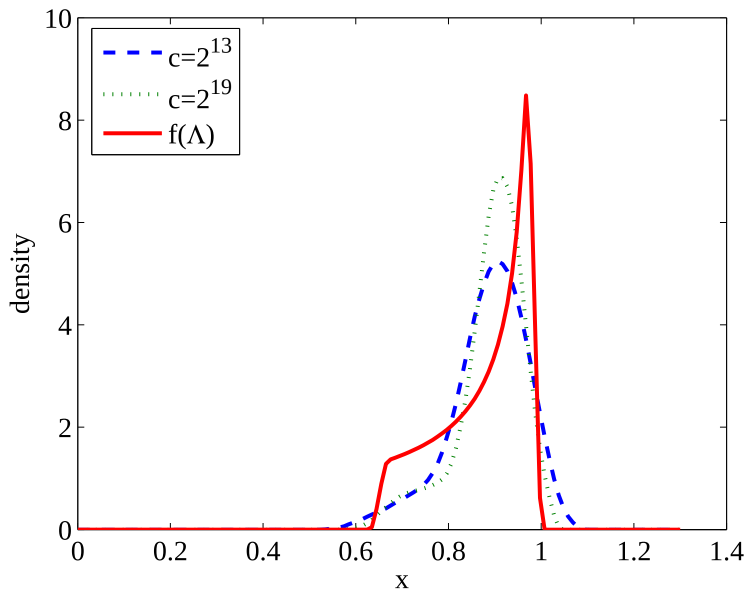

Figure 4.5: Distribution estimates when $\lambda = 1000$, $\mu = 12$, $\tilde{c} = 97$, $\theta = 0$, and the variance factor $= 3$.

will be greater than 0.9. But with nontrivial probability, the long-fraction of calls answered immediately could be worse than 0.7.

The general agreement in shape of the densities in Figures 4.1 and 4.5 can be traced to the fact that the random variables whose densities are being estimated are very similar. For the UAR long-run case, the random variable, $\mathrm{E}[(S_1/\Lambda t)|\Lambda]$, is essentially the random variable S_1/N_1 for the RVAR short-run case with the process variability integrated out. The difference arises because

$$\mathrm{E}\left[\frac{S_1}{\Lambda t}\middle|\Lambda\right] \neq_d \mathrm{E}\left[\frac{S_1}{N_1}\middle|\Lambda\right].$$

But the random variables are similar enough that the impact of the variance factor and abandonment on the long-run UAR measures are well described by Figures 4.3 and 4.4, respectively.

BIBLIOGRAPHY

S. Andradóttir and P. W. Glynn. Computing Bayesian means using simulation. Submitted for publication, 2003.

A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

A. Bassamboo, J. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. 2004. Submitted.

P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.

P. Billingsley. *Probability and Measure*. Wiley, New York, 3rd edition, 1995.

L. Breiman. *Probability*. Addison-Wesley, Reading, MA, 1968.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

E. Buist and P. L'Ecuyer. A java library for simulating contact centers. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, Piscataway NJ, 2005. IEEE.

R. J. Carrol and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, Dec 1988.

B. P. K. Chen and S. G. Henderson. Two issues in setting call center staffing levels. *Annals of Operations Research*, 108:175–192, 2001.

R. C. H. Cheng. Selecting input models. In J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, editors, *Proceedings of the 1994 Winter Simulation Conference*, pages 184–191, Piscataway, NJ, 1994. IEEE.

R. C. H. Cheng and W. Holland. Calculation of confidence intervals for simulation output. Submitted for publication, 2003.

R. C. H. Cheng and W. Holland. Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation*, 57: 219–241, 1997.

R. C. H. Cheng and W. Holland. Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation*, 60: 183–205, 1998.

S. E. Chick. Input distribution selection for simulation experiments: accounting for input uncertainty. *Operations Research*, 49:744–758, 2001.

A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A. N. Avramidis. Markov chain models of a telephone call center in blend mode. Submitted, 2004.

L. Devroye and G. Lúgosi. *Combinatorial Methods in Density Estimation.* Springer, New York, 2001.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14(1):153–158, 1967.

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: tutorial, review and research prospects. *Manufacturing and Service Operations Management*, 5: 79–141, 2003.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.

P. Glasserman. Performance continuity and differentiability in Monte Carlo optimization. In M. Abrams, P. Haigh, and J. Comfort, editors, *Proceedings of the 1988 Winter Simulation Conference*, pages 518–524, Piscataway, NJ, 1988. IEEE.

W. K. Grassmann. Finding the right number of servers in real-world queuing systems. *Interfaces*, 18(2):94–104, 1988.

J. M. Harrison and A. Zeevi. A method for staffing large call centers using stochastic fluid models. *Manufacturing & Service Operations Management*, 2005. To appear.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

S. G. Henderson. Input model uncertainty: why do we care and what should we do about it? In S. E. Chick, P. J. Sánchez, D. J. Morrice, and D. Ferrin, editors, *Proceedings of the 2003 Winter Simulation Conference*, page To appear, Piscataway, NJ, 2003. IEEE.

M. C. Jones. Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, 32(3):361–371, 1990.

M. C. Jones and D. F. Signorini. A comparison of higher-order bias kernel density estimators. *Journal of the American Statistical Association*, 92(439):1063–1073, September 1997.

G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, 3rd edition, 2000.

P. L'Ecuyer. A unified view of the IPA, SF and LR gradient estimation techniques. *Management Science*, 36:1364–1383, 1990.

P. L'Ecuyer. On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41:738–748, 1995.

S. H. Lee. *Monte Carlo Computation of Conditional Expectation Quantiles*. PhD thesis, Stanford University, Stanford, CA, 1998.

S. H. Lee and P. W. Glynn. Computing the distribution function of a conditional expectation via Monte Carlo: discrete conditioning spaces. *ACM Transactions on Modeling and Computer Simulation*, 13(3):238–258, July 2003.

S. H. Lee and P. W. Glynn. Computing the distribution function of a conditional expectation via Monte Carlo: discrete conditioning spaces. In P. A. Farrington, H. Black Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 1654–1663, Piscataway, NJ, 1999. IEEE.

D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36:1049–1051, 1965.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

E. Parzen. On the estimation of a probability density function and the mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, September 1962.

B. L. S. P. Rao. *Nonparametric Functional Estimation*. Academic Press, New York, 1983.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, September 1956.

A. M. Ross. *Queueing Systems with Daily Cycles and Stochastic Demand with Uncertain Parameters*. PhD thesis, University of California, Berkeley, Berkeley, California, USA, 2001.

S. M. Ross. *Stochastic Processes*. Wiley, New York, 2nd edition, 1996.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1987.

D. Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, 92 (439):1049–1062, September 1997.

A. E. Schulman. *A Comparison of Local Bandwidth Selectors for Local Polynomial Regression*. PhD thesis, Cornell University, Ithaca, NY, 1998.

J. W. Staudenmayer. *Local Polynomial Regression in the Presence of Covariate Measurement Error: An Improved SIMEX Estimator*. PhD thesis, Cornell University, Ithaca, NY, 2000.

S. G. Steckley and S. G. Henderson. A kernel approach to estimating the density of a conditional expectation. In S. E. Chick, P. J. Sánchez, D. J. Morrice, and D. Ferrin, editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 383–391, Piscataway, NJ, 2003. IEEE.

S. G. Steckley, S. G. Henderson, and V. Mehrotra. Service system planning in the presence of a random arrival rate. 2005. Working paper.

L. A. Stefanski and J. R. Cook. Simulation-extrapolation: The measurement error jacknife. *Journal of the American Statistical Association*, 90(432):1247–1256, Dec 1995.

G. M. Thompson. Server staffing levels in pure service environments when the true mean daily customer arrival rate is a normal random variate. Unpublished manuscript, 1999.

M. P. Wand and M. C. Jones. *Kernal Smoothing*. Chapman & Hall, London, 1995.

W. Whitt. *Stochastic-Process Limits*. Springer Series in Operations Research. Springer, New York, 2002.

W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. 2004. Submitted.

W. Whitt. Bivariate distributions with given marginals. *The Annals of Statistics*, 4:1280–1289, 1976.

W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212, 1999.

R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs NJ, 1989.

F. Zouaoui and J. R. Wilson. Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions*, 35(3):781–792, 2003.

F. Zouaoui and J. R. Wilson. Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions*, 36(11):1135–1151, 2004.