Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

A KERNEL APPROACH TO ESTIMATING THE DENSITY OF A CONDITIONAL EXPECTATION

Samuel G. Steckley
Shane G. Henderson

School of Operations Research and Industrial Engineering
Cornell Universtiy
Ithaca, NY 14853, U.S.A.

ABSTRACT

Given uncertainty in the input model and parameters of a simulation study, the goal of the simulation study often becomes the estimation of a conditional expectation. The conditional expectation is expected performance conditional 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 paper we estimate the density of a conditional expectation using ideas from the field of kernel density estimation. We present a result on asymptotically optimal rates of convergence and examine a number of numerical examples.

1 INTRODUCTION

Let X be a real-valued random variable with $\mathrm{E}|X| < \infty$. Let Z be some other random variable. The conditional expectation $\mathrm{E}(X|Z)$ is a random variable that represents one's best guess (in a certain sense) as to the value of X given only the value of the random variable Z. In this paper 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 variable Z, and
2. 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.

In this paper we confine our attention to the case where Z is a real-valued random variable, but we are working on establishing analogous results when Z is a more complicated random object.

Such a generalization is important, because our primary motivation for studying this problem stems from the issue of input model uncertainty. This form of uncertainty arises when one is not completely certain what input distributions and associated parameters should be used in a 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 input distributions and parameters. For example, this is the case in the papers Cheng (1994), Cheng and Holland (1997), Cheng and Holland (1998), Cheng and Holland (2003), Chick (2001), Zouaoui and Wilson (2001b), Zouaoui and Wilson (2001a). See Henderson (2003) for further discussion.

The input model uncertainty problem maps to the setting in this paper as follows. The random object Z corresponds to a selection of input distributions and associated parameters for a simulation experiment. The random variable X represents an estimate of a performance measure from the simulation model. Its distribution is dependent on the choice Z of input distributions and parameters. The conditional expectation $\mathrm{E}(X|Z)$ represents the expected value of the performance measure as a function of the input distributions and parameters. It is essentially what one would compute from the simulation if the simulation were allowed to run for an infinite amount of time. Notice that it is still a random variable owing to the uncertainty in the input distributions and parameters. A density of $\mathrm{E}(X|Z)$ gives a sense of the uncertainty in the estimate of the performance measure due to the uncertainty in the values of the input distributions and parameters. Example 3 of Henderson (2003) discusses this density in the setting of a queueing simulation, and describes how it may be interpreted.

Very 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. We use kernel density estimation methods to estimate the required

density. The analysis of our estimator draws from methods that are used in variable-bandwidth kernel density estimation methods (Hall 1990).

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.

This paper is organized as follows. In §2 we describe our estimation methodology, and show that under fairly general conditions the error in our density estimator converges at rate $c^{-2/7}$, where c is the overall computational budget. We then present some numerical examples in §3. Some brief conclusions and directions for future research appear in §4.

2 ESTIMATION METHODOLOGY

Our problem is very similar in structure to that of Lee and Glynn (1999). 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 $\sigma^2(\cdot) \equiv \mathrm{Var}(X|Z = \cdot)$.

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 at the value x is of the form

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

where 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 (Wand and Jones 1995, p. 11). 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

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

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.

We wish to analyze the asymptotics of the estimator. For a given computer budget c, let $m = m(c)$ and $n = n(c)$ be chosen so that the total computational effort required to generate $\hat{f}(x; m, n, h)$ is approximately c. Following Lee and Glynn (1999), the computational effort required to compute $\hat{f}(x; m, n, h)$ is

$$\alpha_1 n(c) + \alpha_2 n(c) m(c)$$

where α_1 and α_2 are the average computational effort used to generate Z_i and $X_j(Z_i)$, respectively. We need $m(c) \to \infty$ as $c \to \infty$ to ensure that $\bar{X}_{m(c)}(Z_i) \to \mathrm{E}(X|Z = Z_i)$ and we can assume $\alpha_2 = 1$ without loss of generality. It follows that $m(c)$ 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.

The error criterion that we choose to use in analyzing the convergence is mean integrated squared error (mise), defined as

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

This error criterion is not without its drawbacks (see Devroye and Györfi 1985) but its mathematical simplicity is appealing. Switching the order of integration yields

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

Note that the integrand is mean squared error (MSE) which decomposes into squared bias and variance. We thus have

$$\begin{aligned}\mathrm{mise}(\hat{f}(\cdot; m, n, h)) \quad &= \quad \int \mathrm{bias}^2(\hat{f}(x; m, n, h))\,\mathrm{d}x \\ &\quad + \int \mathrm{Var}(\hat{f}(x; m, n, h))\,\mathrm{d}x.\end{aligned}$$

In order to simplify the bias and variance calculations we make the following additional assumptions. Let $N(a_1, a_2)$ denote a normally distributed random variable with mean a_1 and variance a_2.

- A1. Conditional on $(Z_i : 1 \leq i \leq n)$, $\bar{X}_m(Z_i) \sim N(\mu(Z_i), m^{-1}\sigma^2(Z_i))$ for $i = 1, \ldots, n$ and are (conditionally) independent;
- A2. The kernel K is the density of a $N(0, 1)$ random variable;
- A3. The function $\mu(\cdot)$ is strictly monotonic on its domain;
- A4. The bandwidth h is defined by $m = ah^{-\delta}$ where $\delta > 0$ and $a > 0$ are constants independent of c.

If the central limit theorem holds, then for large m assumption A1 is approximately true. This assumption will be further examined in §3. In A2, we specify the kernel K to be normal. Together, the normality of A1 and A2 allow us to derive compact expressions for bias and variance for the estimator $\hat{f}(x; m, n, h)$. This will be illustrated in the proof presented in §2.1.

Assumption A3 is a simplifying assumption that ensures that a change of variable that we employ later is valid. Since we have $m \to \infty$ as $c \to \infty$, A4 ensures that $h \to 0$ as $c \to \infty$ which is necessary for convergence in the standard kernel density estimation setting. Also note that given A4, h is completely determined by m and δ. Now the density estimator is a function of x, m, n, and δ so $\hat{f}(x; m, n, h)$ and $\text{mise}(\hat{f}(\cdot; m, n, h))$ can now be written, with an abuse of notation, as $\hat{f}(x; m, n, \delta)$ and $\text{mise}(\hat{f}(\cdot; m, n, \delta))$ respectively.

In §2.1 the case in which the variance function $\sigma^2(\cdot)$ is constant is examined. In §2.2, the general case is considered.

2.1 Constant Variance Function $\sigma^2(\cdot)$

The analysis for the case in which the variance function $\sigma^2(\cdot)$ is constant is similar to that in the standard kernel density estimation setting. Hence the following assumptions, together with A2 and A4, are similar to those found in Prakasa Rao (1983), p. 44:

- A5. f'' is a bounded, continuous, square-integrable function;
- A6. $nh \to \infty$ and $m \to \infty$ as $c \to \infty$.

Finally, the constant variance assumption is noted:

- A7. $\sigma^2(\cdot) \equiv \sigma^2 > 0$.

In Propositions 1 and 2 we make use of o ("small oh") notation. For sequences of real numbers a_n and b_n, we say that

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

Taylor's Theorem with integral remainder is useful in the proof of Proposition 1. We state it here as a lemma. A proof can be found on p. 278 of Apostol (1967).

Lemma 1 Assume f is twice continuously differentiable. Then

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

Proposition 1 Assume A1-A7. Then

$$\begin{aligned}\text{mise}(\hat{f}(\cdot; m, n, \delta)) &= \left(h^2 + \frac{\sigma^2}{m}\right)^2 b_1 + \frac{b_2}{nh} \\ &\quad + o\left((h^2 + \frac{1}{m})^2 + \frac{1}{nh}\right), \quad (1)\end{aligned}$$

where

$$b_1 = \frac{1}{4}\int f''(x)^2\,\mathrm{d}x$$

and

$$b_2 = \frac{1}{2\sqrt{\pi}}.$$

Since $h = (a/m)^{1/\delta}$, (1) is equivalent to the following: for $0 < \delta \leq 2$,

$$\text{mise}(\hat{f}(\cdot; m, n, \delta)) = \frac{b_1^c}{m^2} + \frac{b_2}{nh} + o\left(\frac{1}{m^2} + \frac{1}{nh}\right),$$

and for $\delta > 2$,

$$\text{mise}(\hat{f}(\cdot; m, n, \delta)) = \left(\frac{a}{m}\right)^{4/\delta} b_1 + \frac{b_2}{nh} + o\left(\frac{1}{m^{4/\delta}} + \frac{1}{nh}\right),$$

where

$$b_1^c = \frac{1}{4}(aI(\delta = 2) + \sigma^2)^2 \int f''(x)^2\,\mathrm{d}x.$$

Proof:

$$\begin{aligned}&\mathrm{E}(\hat{f}(x; m, n, \delta)) \\ &= \mathrm{E}\left(\frac{1}{h}K\left(\frac{x - \bar{X}_m(Z_1)}{h}\right)\right) \\ &= \mathrm{E}\left(\mathrm{E}\left[\frac{1}{h}K\left(\frac{x - \bar{X}_m(Z_1)}{h}\right)\,\middle|\, Z_1\right]\right). \quad (2)\end{aligned}$$

Conditional on Z_1,

$$\bar{X}_m(Z_1) \sim N(\mu(Z_1), \frac{\sigma^2}{m}).$$

Since K is the density of a $N(0,1)$ random variable, $h^{-1}K(\cdot/h)$ is the density of a $N(0,h^2)$ random variable. Then the conditional expectation above is a convolution of the densities of two normal random variables and so

$$\mathrm{E}(\hat{f}(x;m,n,\delta)) = \mathrm{E}\left(\frac{1}{\sqrt{h^2+\frac{\sigma^2}{m}}}K\left(\frac{x-\mu(Z_1)}{\sqrt{h^2+\frac{\sigma^2}{m}}}\right)\right).$$

Define

$$\eta = \sqrt{h^2+\frac{\sigma^2}{m}},$$

so that $E\,\hat{f}(x;m,n,h)$ can be expressed as

$$\mathrm{E}\left(\frac{1}{\eta}K\left(\frac{x-\mu(Z_1)}{\eta}\right)\right). \tag{3}$$

A change of variable using A3 shows that (3) is given by

$$\int \frac{1}{\eta}K\left(\frac{x-y}{\eta}\right) f(y)\,\mathrm{d}y, \tag{4}$$

and another change of variable gives

$$\int f(x-u\eta)K(u)\,\mathrm{d}u.$$

Since f is twice continuously differentiable, we have by Lemma 1 that

$$\begin{aligned}
&\mathrm{E}\,\hat{f}(x;m,n,h)\\
&\quad= f(x)\int K(u)\,\mathrm{d}u - \eta f'(x)\int uK(u)\,\mathrm{d}u\\
&\qquad+ \eta^2\iint_0^1 (1-t)\,f''(x-tu\eta)u^2K(u)\,\mathrm{d}t\mathrm{d}u\\
&\quad= f(x)+\eta^2\iint_0^1 (1-t)\,f''(x-tu\eta)u^2K(u)\,\mathrm{d}t\mathrm{d}u.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\int \mathrm{bias}^2(\hat{f}(x;m,n,\delta))\,\mathrm{d}x\\
&\quad= \eta^4\int\left[\iint_0^1 (1-t)\,f''(x-tu\eta)u^2K(u)\,\mathrm{d}t\mathrm{d}u\right]^2\mathrm{d}x.
\end{aligned}$$

A rather involved argument that uses Lebesgue's dominated convergence theorem (see, e.g., p. 45 of Prakasa Rao 1983) establishes that

$$\int\left[\iint_0^1 (1-t)\,f''(x-tu\eta)u^2K(u)\,\mathrm{d}t\mathrm{d}u\right]^2\mathrm{d}x \;\to\; b_1$$

as $c\to\infty$. It follows that

$$\begin{aligned}
&\int \mathrm{bias}^2(\hat{f}(x;m,n,\delta))\,\mathrm{d}x\\
&\quad= \left(h^2+\frac{\sigma^2}{m}\right)^2 b_1 + o\left(\left(h^2+\frac{1}{m}\right)^2\right).
\end{aligned} \tag{5}$$

Similarly,

$$\begin{aligned}
&\mathrm{Var}(\hat{f}(x;m,n,\delta))\\
&\quad= \frac{1}{n}\left[\mathrm{var}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_1)}{h}\right)\right)\right]\\
&\quad= \frac{1}{n}\left[\mathrm{E}\left(\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_1)}{h}\right)\right)^2\right)\right.\\
&\qquad\left.-\left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_1)}{h}\right)\right)\right)^2\right] \qquad (6)\\
&\quad= \frac{1}{nh}\frac{1}{2\sqrt{\pi}}\mathrm{E}\left(\mathrm{E}\left[\frac{1}{h/\sqrt{2}}K\left(\frac{x-\bar{X}_m(Z_1)}{h/\sqrt{2}}\right)\,\middle|\,Z_1\right]\right)\\
&\qquad+o\left(\frac{1}{nh}\right). \qquad (7)
\end{aligned}$$

To obtain (7) we have used the fact that the second term in (6) is $f^2(x)$ plus error terms, as shown in the bias calculations above. Therefore, the second term in (6) is of order n^{-1} and therefore $o((nh)^{-1})$. We then use the fact that when $K(\cdot;h^2)$ is the density of a normal random variable with mean 0 and variance h^2,

$$K^2(\cdot;h^2) = \frac{K(\cdot;h^2/2)}{\sqrt{2\pi}}$$

to obtain (7).

Applying the same steps to

$$\mathrm{E}\left(\mathrm{E}\left[\frac{1}{h/\sqrt{2}}K\left(\frac{x-\bar{X}_m(Z_1)}{h/\sqrt{2}}\right)\,\middle|\,Z_1\right]\right) \tag{8}$$

as were applied to

$$\mathrm{E}\left(\mathrm{E}\left[\frac{1}{h}K\left(\frac{x-\bar{X}_m(Z_1)}{h}\right)\,\middle|\,Z_1\right]\right)$$

gives

$$\mathrm{var}(\hat{f}(x;m,n,\delta)) = \frac{1}{nh}\frac{1}{2\sqrt{\pi}}f(x)+o\left(\frac{1}{nh}\right).$$

Then

$$\int \mathrm{var}(\hat{f}(x;m,n,\delta))\,\mathrm{d}x = \frac{1}{nh}\frac{1}{2\sqrt{\pi}}+o\left(\frac{1}{nh}\right), \tag{9}$$

and (1) follows from (5) and (9).

The normality given by A1 and A2 assures that the convolutions in (2) and (8) have well defined forms. Specifically, for each convolution we get a normal density evaluated at x with expectation and variance resulting from the sum of the convolved distributions' expectations and variances, respectively. Consider (4) which is immediate from the convolution performed on the expectation. This shows that we are effectively doing standard kernel density estimation using kernels that have squared bandwidth $\eta^2 = h^2 + m^{-1}\sigma^2$ rather than just h^2. Note that the extra component $m^{-1}\sigma^2$ is the variance of the error of the estimate $\bar{X}_m(Z_1)$ about $\mathrm{E}(X|Z_1)$ conditional on Z_1. So under the the stated assumptions, the effect of using estimates $\bar{X}_m(Z)$ of realizations of $\mathrm{E}(X|Z)$ rather than actual realizations of $\mathrm{E}(X|Z)$ is an effective bandwidth whose square is wider than h^2 by the variance of the error of the estimate.

Compare (1) to the mise in standard kernel density estimation (Wand and Jones 1995),

$$\begin{aligned}\mathrm{mise}(\hat{g}(\cdot;h)) &= h^4 b_1 \int u^2 K(u)\,\mathrm{d}u + \frac{1}{nh}\int K(u)^2\,\mathrm{d}u \\ &\quad + o(h^4 + \frac{1}{nh}). \qquad (10)\end{aligned}$$

Note that when K is the density of a $N(0,1)$ random variable,

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

and

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

in which case (10) is precisely the same as (1) except for the order of the leading term on the $\int \mathrm{bias}^2\,\mathrm{d}x$ term. For (10) it is h^4 whereas for (1) it is $(h^2 + m^{-1}\sigma^2)^2$. Recalling that $h = (a/m)^{1/\delta}$, we see the order on the $\int \mathrm{bias}^2\,\mathrm{d}x$ term in (1) is actually larger for $0 < \delta < 2$ as compared to (10). For $\delta \geq 2$, the order is the same. So the wider bandwidth arising from using estimates rather than realizations of $\mathrm{E}(X|Z)$ translates into diminished convergence on the $\int \mathrm{bias}^2\,\mathrm{d}x$ term in mise for $0 < \delta < 2$.

We wish to choose n, m, and δ to optimize the rate of convergence of $\mathrm{mise}(\hat{f}(\cdot;m,n,\delta))$ in terms of the computer budget c and compare it to the optimal rate of convergence for standard kernel density estimation. To do this, we drop the low order terms and consider the large sample approximation of $\mathrm{mise}(\hat{f}(\cdot;m,n,\delta))$:

$$\mathrm{amise}(\hat{f}(\cdot;m,n,\delta)) = \begin{cases} \frac{b_1^c}{m^2} + \frac{b_2}{nh} & \text{if } 0 < \delta \leq 2 \\ \left(\frac{a}{m}\right)^{4/\delta} b_1 + \frac{b_2}{nh} & \text{if } \delta > 2. \end{cases}$$

Here, amise stands for asymptotic mean integrated squared error. We now solve for the n^*, m^*, and δ^* that give the optimal rate of convergence of $\mathrm{amise}(\hat{f}(\cdot;m,n,\delta))$, which in turn give the optimal rate of convergence of $\mathrm{mise}(\hat{f}(\cdot;m,n,\delta))$.

As mentioned above, we must satisfy the asymptotic relationship $m(c)n(c)/c \to 1$ as $c \to \infty$. For the large sample approximation, take $n = c/m$. It turns out that for any given δ, we can solve for m^* and thus $n^* = c/m^*$:

$$m^*(\delta) = \begin{cases} d_1(\delta)c^{\delta/(3\delta+1)} & \text{if } 0 < \delta \leq 2 \\ d_2(\delta)c^{\delta/(\delta+5)} & \text{if } \delta > 2, \end{cases}$$

where

$$d_1(\delta) = a^{1/(3\delta+1)}\left(\frac{2b_1^c\delta}{b_2(\delta+1)}\right)^{\delta/(3\delta+1)},$$

and

$$d_2(\delta) = a^{5/(\delta+5)}\left(\frac{4b_1}{b_2(\delta+1)}\right)^{\delta/(\delta+5)}.$$

We note that m^* and n^* are such that assumption A6 holds for any δ. Substituting into our expressions for $\mathrm{amise}(\hat{f}(\cdot;m,n,\delta))$ gives

$$\mathrm{amise}(\hat{f}(\cdot;m^*;n^*;\delta)) = \begin{cases} \bar{d}_1(\delta)c^{-2\delta/(3\delta+1)} & \text{if } 0 < \delta \leq 2 \\ \bar{d}_2(\delta)c^{-4/(\delta+5)} & \text{if } \delta > 2, \end{cases}$$

where

$$\begin{aligned}\bar{d}_1(\delta) &= a^{-2/(3\delta+1)} b_2^{2\delta/(3\delta+1)} \\ &\quad \times \left(\frac{3\delta+1}{2\delta}\right)\left(\frac{2b_1^c\delta}{(\delta+1)}\right)^{(\delta+1)/(3\delta+1)},\end{aligned}$$

and

$$\bar{d}_2(\delta) = (ab_2)^{4/(\delta+5)}\left(\frac{5+\delta}{4}\right)\left(\frac{4b_1}{\delta+1}\right)^{(\delta+1)/(\delta+5)}.$$

Thus the best rate of convergence is attained at $\delta^* = 2$. Then the optimal choice of m is

$$m^* = dc^{2/7},$$

where

$$d = a^{1/7}\left(\frac{4b_1^c}{3b_2}\right)^{2/7},$$

and the optimal amise is

$$\mathrm{amise}(\hat{f}(\cdot;m^*;n^*;\delta^*)) = \bar{d}c^{-4/7},$$

where

$$\bar{d} = a^{-2/7} b_2^{4/7}\left(\frac{7}{4}\right)\left(\frac{4b_1^c}{3}\right)^{3/7}.$$

The optimal rate of convergence of mise is thus $c^{-4/7}$. Although m^* given above is the optimal choice of m for this rate of convergence, it is possible to achieve the rate $c^{-4/7}$ so long as asymptotically

$$m(c) = d_3 c^{\frac{2}{7}}$$

for any positive constant d_3 and of course $n = cm^{-1}$.

In standard kernel density estimation, the optimal rate of convergence is $c^{-4/5}$ (Wand and Jones 1995, p. 23). The decrease in rate of convergence is expected in that for each of the n observations $\bar{X}_m(Z_i)$, m units of computer time are required and $m \to \infty$ as $c \to \infty$, whereas in the standard kernel density estimation setting, each observation requires only one unit of computer time. In addition, as we noted above, for $0 < \delta < 2$ the convergence of $\int \text{bias}^2 \, dx$ in the expression for mise is slower in this setting as compared to standard kernel density estimation. The slower convergence for $0 < \delta < 2$ induces choosing $\delta^* = 2$ so that it does play a role in determining the optimal rate of convergence.

2.2 General Variance Function $\sigma^2(\cdot)$

In this section, assumption A7 is relaxed. Define the function $\rho(\cdot) \equiv \sigma^2(\mu^{-1}(\cdot))$. We make the following additional assumptions:

- A8. The function $\rho(\cdot)$ is bounded above and also away from zero, is twice differentiable and its derivatives are continuous and bounded;
- A9. f, f', and f'' are square integrable.

Proposition 2 Assume A1-A6, A8 and A9. Then for $0 < \delta \leq 2$,

$$\text{mise}(\hat{f}(\cdot; m, n, \delta)) = \frac{b_1^v}{m^2} + \frac{b_2}{nh} + o\left(\frac{1}{m^2} + \frac{1}{nh}\right),$$

and for $\delta > 2$

$$\text{mise}(\hat{f}(\cdot; m, n, \delta)) = \left(\frac{a}{m}\right)^{4/\delta} b_1 + \frac{b_2}{nh} + o\left(\frac{1}{m^{\frac{4}{\delta}}} + \frac{1}{nh}\right),$$

where

$$\begin{aligned} b_1^v &= \frac{1}{4}\int [f''(x)(aI(\delta = 2) + \rho(x)) \\ &\quad + 2f'(x)\rho'(x) + f(x)\rho''(x)]^2 dx, \end{aligned}$$

and b_1 and b_2 are the same as above.

The proof of Proposition 2, which uses techniques from the variable bandwidth literature (Hall 1990) but is otherwise similar to the proof of Proposition 1, will be given elsewhere.

We remark that the condition that $\rho(\cdot)$ be bounded away from zero is used only for the case $\delta \in (0, 2)$.

The only difference in mise in the constant variance function case and the general variance function case is in the coefficient of the $\int \text{bias}^2 \, dx$ term when $0 < \delta \leq 2$. In the general case, we have b_1^v whereas in the constant variance function case we have b_1^c. We first note that when we have a constant variance function ($\sigma^2(\cdot) \equiv \sigma^2$), $\rho(\cdot) \equiv \sigma^2$ and ρ' and ρ'' are zero so that b_1^v simplifies to b_1^c. Secondly we note that in the general case, $\sigma^2(\cdot)$ plays a significant role in the constant b_1^v through the functions ρ, ρ', and ρ''. And finally we note that if we define the function $\beta(\cdot) \equiv f(\cdot)\rho(\cdot)$, b_1^v is very similar to

$$\frac{1}{4}\int (\beta''(x))^2 \, dx.$$

So we expect b_1^v to be large when the function $\beta''(\cdot)$ tends to be large in magnitude.

Following the same line of reasoning as in the constant variance case, the best rate of convergence is attained at $\delta^* = 2$ and

$$m^* = d^v c^{2/7},$$

where

$$d^v = a^{1/7}\left(\frac{4b_1^v}{3b_2}\right)^{2/7}.$$

The optimal amise is

$$\text{amise}(\hat{f}(\cdot; m^*; n^*; \delta^*)) = \bar{d}^v c^{-4/7},$$

where

$$\bar{d}^v = a^{-2/7} b_2^{4/7}\left(\frac{7}{4}\right)\left(\frac{4b_1^v}{3}\right)^{3/7}.$$

So the same rate is achieved as in the constant variance case but the coefficient is different. We again note that the optimal rate $c^{-4/7}$ is attainable provided that asymptotically

$$m = d_4 c^{2/7}$$

for any positive constant d_4 and $n = cm^{-1}$.

3 EXAMPLES

In this section we examine the convergence of a few basic examples and compare to the theoretical results presented in §2. Specifically, for each example we look at mise as a function of the computer budget c. For clarity, we no longer suppress the dependence of our functions on c. For example, we now write mise(c), $m(c)$, and $n(c)$.

To estimate mise(c), we first replicate the density estimator 50 independent times:

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

We define integrated squared error (ise) as follows:

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

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

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

Our estimate for mise(c) is then

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

In calculating $\hat{f}_k(\cdot; m(c), n(c), \delta)$ we take $\delta = 2$, and $m(c) = \lfloor rc^{2/7} \rfloor$ as suggested in §2, where the constant r was chosen in brief preliminary experiments to be 30 for Example 1 and 1 for the other examples. We took $h = m^{-1/\delta}$ (so that $a = 1$). We estimate mise(c) for the following values of c:

$$\{c = 1024 \times 2^l : l = 1, \ldots, 8\}.$$

Example 1: In the first example we let $Z \sim \text{Beta}(4, 4)$ (a Beta(a_1, a_2) random variable has density on $(0, 1)$ proportional to $x^{a_1-1}(1-x)^{a_2-1}$) and conditional on $Z = z$, $X \sim N(z, 0.5)$. Then the true density of the conditional expectation f is just the density of the Beta(4, 4) distribution. In Figure 1, we plot log(mise(c)) vs. log(c).

The linearity of the plot suggests that asymptotically,

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

for some constants V and γ. Theoretically we expect $\gamma = -4/7 \approx -0.57$. Note that γ is the slope of the $(\log(c), \log(\text{mise}(c)))$ plot and the estimated slope of the plot in Figure 1 is -0.54. This is very close to the expected rate of convergence.

Example 2: In this example we consider a non-constant variance function $\sigma^2(\cdot)$. Once again, let $Z \sim \text{Beta}(4, 4)$. Conditional on $Z = z$, we take $X \sim N(z, z^2)$. The target density f is again the density of the Beta(4, 4) distribution. We present the log(mise(c)) vs. log(c) plot in Figure 2. The plot is linear and the slope is estimated to be -0.45, indicating poorer convergence as compared to Example 1. This is likely the result of the variance function $\sigma^2(z) = z^2$ on the interval $(0, 1)$ and zero elsewhere. We will further discuss the impact of this variance function in Example 3

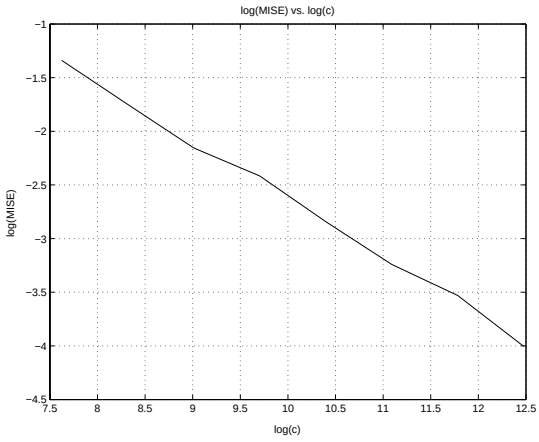

Figure 1: Mean Integrated Squared Error as a Function of Computational Budget for Example 1

but we note here that this variance function does not satisfy the smoothness assumption A8 at $z = 1$.

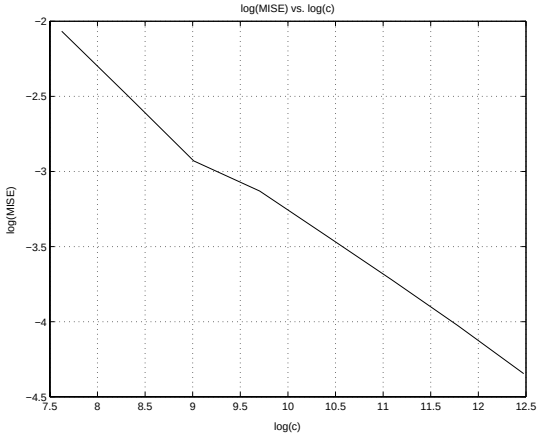

Figure 2: Mean Integrated Squared Error as a Function of Computational Budget for Example 2

Example 3: In this example we study the impact of violating the assumption A1 in which we assume that conditional on $(Z_i : 1 \leq i \leq n)$, $\bar{X}_m(Z_i)$ is normally distributed for $i = 1, \ldots, n$. We take Z to have a Beta(4, 4) distribution shifted to the right by one unit so the support of Z is the interval $(1, 2)$. Conditional on $Z = z$, we now suppose $X \sim \exp(1/z)$, i.e., conditional on $Z = z$, X is exponentially distributed with mean $1/z$. Note that conditional on $(Z_i : 1 \leq i \leq n)$, $\bar{X}_m(Z_i) \sim \text{Gamma}(m, Z_i/m)$ for $i = 1, \ldots, n$ (a Gamma(a_1, a_2) random variable has density on $(0, \infty)$ proportional to $x^{a_1-1}e^{-x/a_2}$), so that assumption A1 is violated. The target density f is the Beta(4, 4) density shifted to the right by one unit.

In Figure 3, we give the log(mise(c)) vs. log(c) plot. The slope is estimated to be -0.44. The rates of convergence in Examples 2 and 3 are quite similar, suggesting that the normality of $\bar{X}_m(Z_i)$, $i = 1, \ldots, n$, is not crucial to the rate of convergence. This is to be expected since the central limit theorem (CLT) tells us that, conditional on $(Z_i : 1 \leq i \leq n)$, for large m, $\bar{X}_m(Z_i)$ behaves approximately like a random variable with a $N(Z_i, Z_i^2/m)$ distribution as in Example 2.

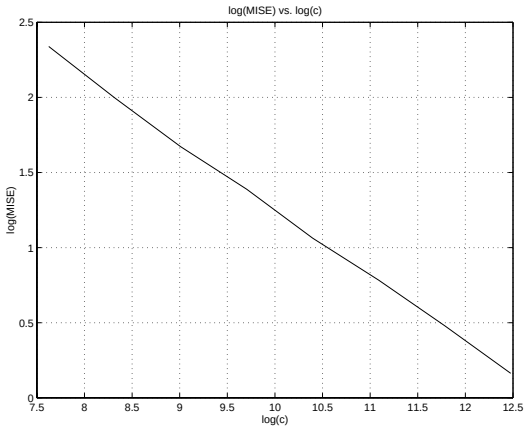

Figure 3: Mean Integrated Squared Error as a Function of Computational Budget for Example 3

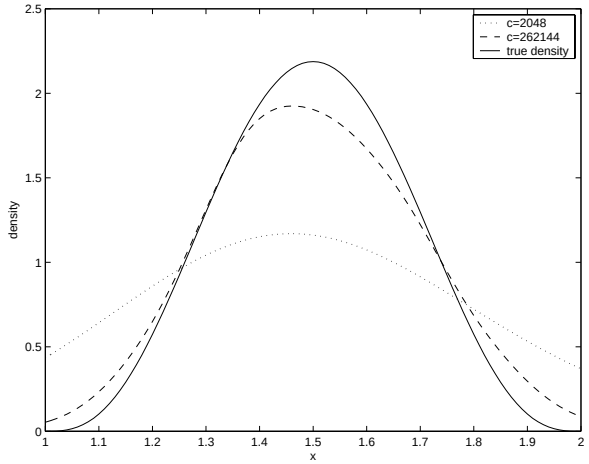

Figure 4: The True Density and Two Estimators for Example 3

The rate -0.44 is not as good as that seen in Example 1. We suspect that this reduction in rate, and that seen in Example 2, is due to the function $\rho(\cdot)$ being discontinuous at $z = 1, 2$. This "boundary effect" is perhaps evident in Figure 4 where the performance of the density estimates deteriorates near the boundaries of the plot. The density estimate for $c = 2048$ is quite poor, but improves when $c = 262144$. For $c = 262144$, the density estimate is slightly skewed to the left. This same skewness was evident in other independent replications of the experiment. We believe that the skewness is a natural result of the form of the variance function – $\sigma^2(z) = z^2$ on the interval $(1, 2)$ and zero elsewhere. Recall that to generate the estimated density our observations $\bar{X}_m(Z_i)$, $i = 1, \ldots, n$, are smoothed by normal kernels with bandwidth $(h^2 + \sigma^2(Z_i)/m)^{1/2}$. For large m, $\bar{X}_m(Z_i) \approx Z_i$. Then the observations $\bar{X}_m(Z_i)$, $i = 1, \ldots, n$ that are larger in value are smoothed more than the observations with smaller value resulting in the skewness seen in the plot.

This "nonuniform smoothing" will probably be typical in examples with nonconstant variance, but the theory presented in this paper shows that the convergence rate will not be affected. Of course, while the *rate* may not be affected, the magnitude of the error may be significantly affected through multiplicative constants.

4 CONCLUSIONS AND FUTURE RESEARCH

We have shown how to share a computational budget between external sampling of Z and internal sampling conditional on values of Z so as to minimize the amise of the density estimator. The amise can converge to 0 at rate $c^{-4/7}$, where c is the computational budget. This is slower than the $c^{-4/5}$ rate exhibited in the standard density estimation context, and both of these rates are slower than the standard Monte Carlo rate c^{-1} when one is estimating an expectation. Nevertheless, we believe that the insight one obtains from the estimated density justifies the additional computational effort involved. Furthermore, one does not need an especially accurate estimate of the density in order to get some idea of the extent of the effect of input uncertainty.

Clearly much remains to be done.

- We need to generalize our results beyond the case where Z is real-valued, so as to capture multiple input parameters and/or distributions.
- The rate of convergence of the estimator seems to strongly depend on the smoothness of ρ and, as observed in experiments not reported here, smoothness of the target density. We need to understand this better.
- In view of the relatively slow convergence of our estimators, confidence intervals for estimates of $f(x)$, or more generally, confidence bands for the entire density f would be of great value.
- A key assumption is that $\bar{X}_m(z)$ is exactly normally distributed. This assumption often holds approximately due to the central limit theorem, since we require m to grow with the computational budget. The results for Example 3 suggest that non-normality *may* not severely impact the rate of convergence, but we need to better under-

stand this impact. It is also of interest to consider problems that do not fit the framework here, such as steady-state simulation and quantile estimation.

- We have shown how to choose the bandwidth only up to a multiplicative constant. This constant can have a strong impact on the performance of the estimator, even though it doesn't change the asymptotic *rate* of convergence. So just as in the standard kernel-density estimation case, bandwidth selection remains an issue.
- In view of the popularity of histogram estimators, it would be interesting to explore their asymptotic performance in our setting. They are known to converge at a slower rate than kernel-based estimators in the i.i.d. setting (Freedman and Diaconis 1981).

We are pursuing, or plan to pursue, all of these topics.

ACKNOWLEDGMENTS

The first author was supported by a National Defense Science and Engineering Graduate Fellowship. The work of the second author was partially supported by National Science Foundation grants DMI 0224884 and DMI 0230528.

REFERENCES

Andradóttir, S., and P. W. Glynn. 2003. Computing Bayesian means using simulation. Submitted for publication.

Apostol, T. M. 1967. *Calculus, Volume I*. 2nd ed. New York: Wiley.

Cheng, R. C. H. 1994. Selecting input models. In *Proceedings of the 1994 Winter Simulation Conference*, ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 184–191. Piscataway, NJ: IEEE.

Cheng, R. C. H., and W. Holland. 1997. Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* 57:219–241.

Cheng, R. C. H., and W. Holland. 1998. Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation* 60:183–205.

Cheng, R. C. H., and W. Holland. 2003. Calculation of confidence intervals for simulation output. Submitted for publication.

Chick, S. E. 2001. Input distribution selection for simulation experiments: accounting for input uncertainty. *Operations Research* 49:744–758.

Devroye, L., and L. Györfi. 1985. *Nonparametric Density Estimation : The L_1 View*. New York: Wiley.

Freedman, D., and P. Diaconis. 1981. On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheorie verw. Gebeite* 57:453–476.

Hall, P. 1990. On the bias of variable bandwidth curve estimators. *Biometrika* 77 (3): 529–535.

Henderson, S. G. 2003. Input model uncertainty: why do we care and what should we do about it? In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. E. Chick, P. J. Sánchez, D. J. Morrice, and D. Ferrin, To appear. Piscataway, NJ: IEEE.

Lee, S. H. 1998. *Monte Carlo Computation of Conditional Expectation Quantiles*. Ph.D. thesis, Stanford University, Stanford, CA.

Lee, S. H., and P. W. Glynn. 1999. Computing the distribution function of a conditional expectation via Monte Carlo: discrete conditioning spaces. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. Black Nembhard, D. T. Sturrock, and G. W. Evans, 1654–1663. Piscataway, NJ: IEEE.

Prakasa Rao, B. L. S. 1983. *Nonparametric Functional Estimation*. New York: Academic Press.

Wand, M., and M. Jones. 1995. *Kernel Smoothing*. London: Chapman & Hall.

Zouaoui, F., and J. R. Wilson. 2001a. Accounting for input model and parameter uncertainty in simulation. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 290–299. Piscataway, NJ: IEEE.

Zouaoui, F., and J. R. Wilson. 2001b. Accounting for parameter uncertainty in simulation input modeling. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 354–363. Piscataway, NJ: IEEE.

AUTHOR BIOGRAPHIES

SAMUEL G. STECKLEY is a Ph.D. candidate in the School of Operations Research and Industrial Engineering at Cornell University. His primary field of interest is input model uncertainty in discrete-event simulation. He is the recipient of an NDSEG fellowship. His e-mail address is <steckley@orie.cornell.edu>.

SHANE G. HENDERSON is an assistant professor in the School of Operations Research and Industrial Engineering at Cornell University. He has previously held positions at the University of Michigan and the University of Auckland. He is an associate editor for the *ACM Transactions on Modeling and Computer Simulation*, *Operations Research Letters*, and *Mathematics of Operations Research*, and the newsletter editor for the INFORMS College on Simulation. His research interests include discrete-event simulation, queueing theory and scheduling problems. His e-mail address is <sgh9@cornell.edu>, and his web page URL is <www.orie.cornell.edu/~shane>.