Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

ASYMPTOTIC PROPERTIES OF KERNEL DENSITY ESTIMATORS WHEN APPLYING IMPORTANCE SAMPLING

Marvin K. Nakayama

Computer Science Department
New Jersey Institute of Technology
Newark, NJ 07102, USA

ABSTRACT

We study asymptotic properties of kernel estimators of an unknown density when applying importance sampling (IS). In particular, we provide conditions under which the estimators are consistent, both pointwise and uniformly, and are asymptotically normal. We also study the optimal bandwidth for minimizing the asymptotic mean square error (MSE) at a single point and the asymptotic mean integrated square error (MISE). We show that IS can improve the asymptotic MSE at a single point, but IS always increases the asymptotic MISE. We also give conditions ensuring the consistency of an IS kernel estimator of the sparsity function, which is the inverse of the density evaluated at a quantile. This is useful for constructing a confidence interval for a quantile when applying IS. We also provide conditions under which the IS kernel estimator of the sparsity function is asymptotically normal. We provide some empirical results from experiments with a small model.

1 INTRODUCTION

Estimation (without importance sampling) of an unknown density function f has been studied extensively in the statistics literature, starting with Rosenblatt (1956) and Parzen (1962). In addition to the density being of interest in its own right, which motivates much of the work in this area, the problem also arises in the context of estimating a quantile. Specifically, for $0 < p < 1$, the p-quantile of a distribution F is defined as $\xi_p \equiv \inf\{x : F(x) \geq p\}$. The standard estimator of ξ_p satisfies a central limit theorem (CLT), with asymptotic variance constant $p(1-p)/f^2(\xi_p)$, where f denotes the density of F; e.g., see Section 2.3.3 of Serfling (1980). Thus, it is also of interest to estimate $1/f(\xi_p)$, which is known as the *sparsity function*, as it allows construction of a confidence interval for a quantile.

Kernel estimation is a well-known technique of "smoothing," which is accomplished by taking the convolution of an estimator with a given kernel. The kernel is often (but not always) a symmetric, unimodal density function, and the user also needs to specify a parameter h, known as the bandwidth, which needs to shrink to 0 at an appropriate rate as the sample size grows to establish consistency of the kernel estimator. For an overview of kernel estimation without importance sampling, see Wand and Jones (1995).

Obtaining a good estimate of the density in the tails of the distribution requires large sample sizes when simulating with crude Monte Carlo (CMC), i.e., no variance reduction is applied. To address this issue, we now consider kernel estimation of f when using importance sampling (IS), which can be especially effective for studying such rare events; e.g., see Chapters V.1 and VI of Asmussen and Glynn (2007). In this paper, we establish some asymptotic properties of IS kernel density estimators. We provide conditions under which they are consistent at a single point and under which they are uniformly consistent. We also study the asymptotic properties of their mean square error (MSE) at a single point and mean integrated square error (MISE), which we exploit to determine the optimal bandwidth to minimize the asymptotic MSE or MISE. We show that the asymptotic MSE at a single point can be reduced by applying IS, but IS always increases the asymptotic MISE. We further give conditions when the IS kernel density estimator

satisfies various CLTs, which we use to construct an asymptotically valid confidence interval for the density at a fixed point. We also develop a consistent kernel estimator of the sparsity function when applying IS and give conditions under which it satisfies a CLT. While many of our results on the IS kernel density estimator generalize the work on non-IS kernel estimators of Parzen (1962), that paper does not consider the estimation of the sparsity function.

The rest of our paper has the following layout. Section 2 reviews background material on naive (i.e., non-kernel) estimation of a density (Section 2.1), kernel density estimation without IS (Section 2.2), and IS (Section 2.3). Section 3 presents the IS kernel density estimator and studies its asymptotic properties. Section 4 contains empirical results from experiments with a small example. We provide some concluding remarks in Section 5. The proofs are given in Nakayama (2011).

2 BACKGROUND

We start by providing some background material on naive density estimation with CMC, kernel estimators using CMC, and importance sampling.

2.1 CMC Naive Density Estimator

Suppose that F is an absolutely continuous distribution function with density function f. We assume F and f are unknown, and the goal is to use simulation to estimate f, either at a fixed value y or the entire function. One (naive) approach for doing this is as follows. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) samples having density f. Then define the empirical distribution function

$$F_n(y) = \frac{1}{n}\sum_{j=1}^{n} I(X_j \le y),$$

where $I(\cdot)$ denotes the indicator function, which is 1 (resp., 0) if the argument is true (resp., false). Since

$$f(y) = \frac{d}{dy}F(y) = \lim_{h\to 0}\frac{F(y+h)-F(y-h)}{2h},$$

a natural estimator of $f(y)$ is

$$f_n(y) = \frac{F_n(y+h)-F_n(y-h)}{2h} \tag{1}$$

for a constant $h > 0$, which is known as the *bandwidth* or *smoothing parameter*. We call f_n the *CMC naive density estimator*, which we note is a (central) finite-difference estimator of $f(y)$; e.g., see Section VII.1 of Asmussen and Glynn (2007). In general, we want h to be small, and we assume that $h = h_n \to 0$ as the sample size $n \to \infty$. To simplify notation, we sometimes write h rather than h_n.

2.2 CMC Kernel Density Estimator

Let

$$k(u) = \frac{1}{2}I(-1 < u \le 1), \tag{2}$$

which is the density function of a unif$[-1,1]$ distribution, and note that we can rewrite $f_n(y)$ in (1) as

$$f_n(y) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}k\left(\frac{y-X_j}{h}\right). \tag{3}$$

This suggests that we can replace k in (3) with another density function or more generally a function k that integrates to 1. We call k a *kernel*, and we then define the *CMC kernel density estimator* as

$$f_n^*(y) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}k\left(\frac{y-X_j}{h}\right) = \frac{1}{n}\sum_{j=1}^{n}k_h(y-X_j),$$

where $k_h(u) = \frac{1}{h}k(\frac{u}{h})$. Throughout this paper, all estimators with an asterisk are kernel estimators. When k (and so also k_h) is a density function, for a fixed sample $X_1, X_2, \ldots, X_n$, we see that $f_n^*(y) \geq 0$ for all y and

$$\int f_n^*(y)\,dy = \frac{1}{n}\sum_{j=1}^{n}\int k_h(y - X_j)\,dy = 1, \tag{4}$$

so f_n^* is a density function. (All integrals are taken over the entire real line.) However, when k is not a density function, then f_n^* may not be a density.

We assume the following:

Assumption A1 The kernel k satisfies the following conditions:

$$\int k(x)\,dx = 1, \tag{5}$$

$$\sup_x |k(x)| < \infty, \quad \int |k(x)|\,dx < \infty, \quad \lim_{|x|\to\infty} |xk(x)| = 0. \tag{6}$$

Some examples of kernels satisfying Assumption A1 include

$$k(x) = \frac{3}{4}(1 - x^2)I(-1 \leq x \leq 1), \tag{7}$$

$$k(x) = \frac{1}{(2\pi)^{1/2}}e^{-x^2/2}, \tag{8}$$

$$k(x) = \frac{3 - x^2}{2(2\pi)^{1/2}}e^{-x^2/2}. \tag{9}$$

The functions in (7) and (8) are known as the *Epanechnikov* and *Gaussian* kernels, respectively. While (7)–(8) are nonnegative and therefore densities, the function in (9) is negative at some points. For more details on these and other kernels, see Chapter 2 of Wand and Jones (1995).

When Assumption A1 holds and f is continuous at y, Parzen (1962) shows that if $h = h_n \to 0$ as $n \to \infty$, then

$$\begin{aligned} E_f[f_n^*(y)] &\to f(y), \\ nh_n V_f[f_n^*(y)] &\to f(y)\rho_k, \end{aligned} \tag{10}$$

as $n \to \infty$, where E_f and V_f denote expectation and variance, respectively, under density f, and

$$\rho_k = \int k^2(z)\,dz, \tag{11}$$

which is finite under Assumption A1. If in addition $nh_n \to \infty$, then the mean square error of $f_n^*(y)$ satisfies

$$\mathrm{MSE}[f_n^*(y)] \to 0 \tag{12}$$

as $n \to \infty$, so $f_n^*(y) \Rightarrow f(y)$ as $n \to \infty$, where "$\Rightarrow$" denotes convergence in distribution, which is equivalent to convergence in probability when the limit is deterministic; e.g., see Section 25 of Billingsley (1999).

We can further give the rate of convergence in (12) under additional conditions:

Assumption A2 The kernel k satisfies $\int xk(x)\,dx = 0$ and $\int x^2|k(x)|\,dx < \infty$.

When we further assume A2 holds and the second derivative f'' of f is continuous and bounded in a neighborhood of y, then $nh_n \to \infty$ implies

$$\mathrm{MSE}[f_n^*(y)] = \frac{1}{nh_n}f(y)\rho_k + o\left(\frac{1}{nh_n}\right) + h_n^4\left(\frac{f''(y)}{2}\eta_k\right)^2 + o(h_n^4)$$

as $n \to \infty$, where

$$\eta_k = \int x^2 k(x)\,dx. \tag{13}$$

For details, see Parzen (1962) or Theorem 2.2 of Pagan and Ullah (1999).

An alternative way of looking at f_n^* is that it is the convolution of k_h and F_n:

$$f_n^*(y) = \int k_h(y-x)\,dF_n(x). \tag{14}$$

The empirical distribution function F_n assigns mass $1/n$ to each sample point X_j, so we see that f_n^* is a "smoothing" that spreads out each point mass over the support of the scaled kernel k_h. Also, if k is a density with mean 0, then f_n^* is the equally weighted mixture of n scaled densities k_h with means $X_1, X_2, \ldots, X_n$.

2.3 Importance Sampling

Now suppose that rather than generating samples using the original density f, we apply importance sampling using a change of measure having density function g (with respect to Lebesgue measure). Specifically, assume that f is absolutely continuous with respect to g, i.e., $g(x) = 0$ implies $f(x) = 0$; see, e.g., p. 422 of Billingsley (1999) for more details. Let $L(x) = f(x)/g(x)$ denote the *likelihood ratio* evaluated at x. Then

$$F(y) = \int I(x \le y) f(x)\,dx = \int I(x \le y) \frac{f(x)}{g(x)} g(x)\,dx = E_g[I(X \le y)L(X)], \tag{15}$$

where E_g denotes expectation under density g.

The representation in (15) suggests estimating $F(y)$ as follows. First generate $X_1, X_2, \ldots, X_n$ as i.i.d. samples having density g. Then we can define an estimator of the distribution function F as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{j=1}^{n} I(X_j \le y) L(X_j), \tag{16}$$

which we call the IS estimator of F. Throughout this paper, all estimators with a hat are IS estimators. We can then define an estimator of $f(y)$ by taking a finite difference of $\hat{F}_n$:

$$\hat{f}_n(y) = \frac{\hat{F}_n(y+h) - \hat{F}_n(y-h)}{2h},$$

which we call the *IS naive density estimator*.

3 IS KERNEL DENSITY ESTIMATOR

As in (14) a smoothed IS density estimator results from convolving a scaled kernel k_h and $\hat{F}_n$:

$$\hat{f}_n^*(y) = \int k_h(y-x)\,d\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} L(X_j) k_h(y - X_j),$$

which we call the *IS kernel density estimator*. Suppose for the moment that k is a density. Then since $L(\cdot) \ge 0$, we have $\hat{f}_n^*(y) \ge 0$ for all y. But in contrast to (4), we see that

$$\int \hat{f}_n^*(y)dy = \frac{1}{n} \sum_{j=1}^{n} L(X_j) \int k_h(y - X_j)\,dy = \frac{1}{n} \sum_{j=1}^{n} L(X_j) \neq 1$$

in general, although $E_g[L(X)] = 1$, so $\hat{f}_n^*$ is not a density function. We now study its asymptotic properties.

3.1 Pointwise and Uniform Consistency

Let V_g denote variance under density g, and let

$$\mathrm{MSE}[\hat{f}_n^*(y)] = E_g[(\hat{f}_n^*(y) - f(y))^2] = V_g[\hat{f}_n^*(y)] + (E_g[\hat{f}_n^*(y)] - f(y))^2$$

be the mean square error of $\hat{f}_n^*(y)$. The following result shows the MSE at a fixed point y asymptotically vanishes when the bandwidth h is chosen appropriately.

Theorem 1 Suppose f is continuous at y, and k satisfies Assumption A1. If $h = h_n \to 0$ as $n \to \infty$, then

$$E_g[\hat{f}_n^*(y)] \to f(y) \tag{17}$$

as $n \to \infty$. If in addition L is continuous at y and $E_g[L^2(X)] < \infty$, then

$$nh_n V_g[\hat{f}_n^*(y)] \to L(y) f(y) \rho_k \tag{18}$$

as $n \to \infty$, where ρ_k is defined in (11). If in addition $nh_n \to \infty$, then

$$\mathrm{MSE}[\hat{f}_n^*(y)] \to 0 \tag{19}$$

as $n \to \infty$, so $\hat{f}_n^*(y) \Rightarrow f(y)$ as $n \to \infty$.

We next provide conditions under which $\hat{f}_n^*$ is a uniformly consistent estimator of f, in the sense given in the theorem below. To do this, let $i = \sqrt{-1}$, and define the Fourier transform of the kernel k to be $\psi(t) = \int e^{-itx} k(x)\, dx$, which exists for all t when k satisfies Assumption A1.

Theorem 2 Suppose that f is uniformly continuous and $E_g[L^2(X)] < \infty$. Also, suppose k satisfies Assumption A1,

$$k(x) = k(-x) \text{ for all } x,\ k \text{ is continuous everywhere, and } \int |\psi(t)|\, dt < \infty. \tag{20}$$

Assume that $h_n \to 0$ and $nh_n^2 \to \infty$ as $n \to \infty$. Then $\sup_{-\infty < y < \infty} |\hat{f}_n^*(y) - f(y)| \Rightarrow 0$ as $n \to \infty$.

The kernels in (7)–(9) satisfy (20). But the uniform kernel (2) is not continuous.

3.2 Asymptotic MSE

We now want to study the rate of convergence of (19), which we will use in Section 3.3 to determine the optimal bandwidth h_n to minimize the asymptotic MSE. To do this we require the additional assumptions on the kernel given in A2 so that we can determine the rate of convergence in (17).

Theorem 3 Suppose the conditions of Theorem 1 hold, and further suppose the kernel k satisfies Assumption A2 and the second derivative f'' of f is continuous and bounded in a neighborhood of y. Then

$$E_g[\hat{f}_n^*(y)] - f(y) = h_n^2 \frac{f''(y)}{2} \eta_k + o(h_n^2), \tag{21}$$

$$\mathrm{MSE}[\hat{f}_n^*(y)] = \frac{1}{nh_n} L(y) f(y) \rho_k + o\left(\frac{1}{nh_n}\right) + h_n^4 \left(\frac{f''(y)}{2} \eta_k\right)^2 + o(h_n^4), \tag{22}$$

as $n \to \infty$ when $h_n \to 0$ and $nh_n \to \infty$, where η_k is given in (13).

Since $E_g[\hat{f}_n^*(y)] = E_f[f_n^*(y)]$, the bias of the IS kernel density estimator $\hat{f}_n^*(y)$ is independent of the change of measure applied. However, the variance of $\hat{f}_n^*(y)$ does depend on the IS density g. We now study the asymptotic MSE (AMSE), which is defined as just the highest-order terms of the MSE:

$$\mathrm{AMSE}[\hat{f}_n^*(y)] = \frac{1}{nh_n} L(y) f(y) \rho_k + h_n^4 \left(\frac{f''(y)}{2} \eta_k\right)^2. \tag{23}$$

For the case of CMC, AMSE$[f_n^*(y)]$ is the same as (23) but with $L(y) \equiv 1$. Thus, for a fixed kernel k and bandwidth rate h_n, the IS kernel estimator $\hat{f}_n^*(y)$ has smaller AMSE than the CMC kernel estimator $f_n^*(y)$ if and only if $L(y) < 1$, or equivalently, $f(y) < g(y)$.

Note that the AMSE in (23) at the point y is affected by the choice of the IS density g by only the value of the likelihood ratio at y. Thus, there is the potential to significantly decrease the AMSE at y by choosing g so that $g(y) \gg f(y)$, subject to maintaining $E_g[L^2(X)] < \infty$ so that our theorems remain valid.

Hong and Liu (2010) study a related problem of estimating the derivative of a distribution function with respect to a model parameter θ, e.g., a customer arrival rate in a queueing system. Their estimator is essentially a finite difference, and they use importance sampling to obtain only samples that lie in the difference, which leads to a faster convergence rate. But the applicability of the approach seems limited.

Instead of studying the MSE of our estimators of f at only a single point y, we now examine a measure of overall quality of the estimators of the entire density function. One such metric is the mean integrated square error (MISE). For our IS kernel estimator, this is given by

$$\text{MISE}[\hat{f}_n^*] = E_g\left[\int (\hat{f}_n^*(y) - f(y))^2\, dy\right] = \int E_g\left[(\hat{f}_n^*(y) - f(y))^2\right] dy = \int \text{MSE}[\hat{f}_n^*(y)]\, dy,$$

where the interchange of expectation and integral is justified by Fubini's theorem (Theorem 18.3 of Billingsley 1999) since the integrand is nonnegative; thus, the MISE is the integrated MSE. We thus define the asymptotic MISE (AMISE) as the integrated AMSE:

$$\begin{aligned}\text{AMISE}[\hat{f}_n^*] &= \int \text{AMSE}[\hat{f}_n^*(y)]\, dy = \frac{\rho_k}{nh_n}\int L(y)f(y)\, dy + \frac{h_n^4\eta_k^2}{4}\int (f''(y))^2\, dy \\ &= \frac{\rho_k}{nh_n}E_g[L^2(X)] + \frac{h_n^4\eta_k^2}{4}\tau_f, \end{aligned} \tag{24}$$

where $\tau_f = \int (f''(y))^2\, dy$, which we assume is finite. When instead applying CMC, we get AMISE$[f_n^*]$ is the same as (24) but with $E_g[L^2(X)]$ replaced with 1. But the Cauchy-Schwarz inequality implies

$$E_g[L^2(X)] \geq E_g^2[L(X)] = E_f^2[1] = 1, \tag{25}$$

where the inequality becomes equality if and only if $g = f$ almost everywhere (a.e.), the latter condition meaning that no IS is used. Thus, AMISE$[\hat{f}_n^*] \geq$ AMISE$[f_n^*]$ with equality if and only if $g = f$ a.e. *We then conclude that IS always does worse when estimating the entire density function, but IS can do better when estimating the density at only a single point.*

3.3 Optimal Bandwidth

We now want to determine the rate of the bandwidth h_n that will minimize the AMSE of $\hat{f}_n^*(y)$. Assume that $f(y) > 0$ and $f''(y)\eta_k \neq 0$, so the two terms in (23) are nonzero. Differentiating (23) with respect to h_n and equating this to zero give the asymptotically optimal bandwidth

$$h_n^\star = \left(\frac{L(y)f(y)\rho_k}{(f''(y)\eta_k)^2}\right)^{1/5} n^{-1/5}. \tag{26}$$

Substituting this into (23) then gives the optimal AMSE as

$$\text{AMSE}_\star[\hat{f}_n^*(y)] = \left[\frac{5}{4}\left(L(y)f(y)\rho_k\right)^{4/5}\left(f''(y)\eta_k\right)^{2/5}\right] n^{-4/5},$$

so at the optimal rate, the MSE of $\hat{f}_n^*(y)$ decreases as $n^{-4/5}$, which is strictly slower than the canonical rate of n^{-1} typically arising for unbiased estimators. Note that this conclusion is obtained under the assumption

that $f''(y) \neq 0$ and $\eta_k \neq 0$, the latter of which holds when the kernel is a density so $\eta_k > 0$. But by allowing the kernel to take on negative values, as in (9), the rate at which the AMSE shrinks can be improved (and can be made to be arbitrarily close to n^{-1}); e.g., see Section 2.8 of Wand and Jones (1995).

We can similarly minimize the AMISE with respect to h_n, although the analysis in the previous section suggests that applying IS to estimate the entire density is not efficient. The optimal bandwidth to minimize AMISE in (24) is

$$h_n^\star = \left(\frac{\rho_k E_g[L^2(X)]}{\tau_f \eta_k^2} \right)^{1/5} n^{-1/5}, \tag{27}$$

which corresponds to the optimal AMISE as

$$\text{AMISE}_\star[\hat{f}_n^*] = \left[\frac{5}{4} \left(E_g[L^2(X)] \rho_k \right)^{4/5} (\tau_f \eta_k^2)^{1/5} \right] n^{-4/5}.$$

When CMC instead is applied, the optimal AMISE is the same except the $E_g[L^2(X)]$ is replaced with 1.

Of course, the asymptotically optimal bandwidths $h_n^\star$ in (26) and (27) are not directly implementable since they depend on $L(y)$, $f(y)$ and $f''(y)$, which are unknown. In the case of CMC, others have suggested data-based methods to estimate the unknown quantities; e.g., see Section 3.6 of Wand and Jones (1995).

We now compare the AMSE and AMISE of kernel density estimators with IS and CMC when using the optimal bandwidths. Define the ratios

$$R_\star(y) = \frac{\text{AMSE}_\star[\hat{f}_n^*(y)]}{\text{AMSE}_\star[f_n^*(y)]} = L^{4/5}(y) \quad \text{and} \quad \bar{R}_\star = \frac{\text{AMISE}_\star[\hat{f}_n^*]}{\text{AMISE}_\star[f_n^*]} = E_g^{4/5}[L^2(X)],$$

which we note do not depend on the kernel k. Thus, when estimating the density at a single point y using the optimal bandwidths, the value of the likelihood ratio at y determines by how much the AMSE changes when applying IS. If we instead focus on estimating the entire density function using the optimal bandwidths, then the amount by which IS degrades the AMISE is determined by the second moment of the likelihood ratio, which is never less than 1 by (25).

We now compute the values of $R_\star(y)$ and $\bar{R}_\star$ for a simple example.

Example 1 Suppose that f is the $N(0,1)$ density, where $N(a,b^2)$ denotes a normal random variable with mean a and variance b^2. Also, suppose that for IS, g is the $N(\nu,1)$ density for some constant ν, so $L(x) = f(x)/g(x) = \exp(-\nu x + \nu^2/2)$ and $E_g[L^2(X)] = \exp(\nu^2) < \infty$ for any ν. Table 1 presents the values of the ratios $R_\star(y)$ and $\bar{R}_\star$ for different values of y and ν. A ratio less than 1 means that the IS kernel density estimator outperforms the CMC kernel density estimator. Note that $R_\star(y)$ is sometimes much smaller than 1, so when estimating the density at a single point y, it is possible to obtain significant improvement using IS. We also have $R_\star(y) > 1$ for some combinations of y and ν, so IS can also do worse at a single point. However, $\bar{R}_\star$ is always greater than 1, so IS always does worse when estimating the entire density function. For example, when estimating $f(3)$, we can get a $1/0.027 \approx 36$-fold reduction in AMSE by applying IS with $\nu = 3$, but if we are interested in estimating the entire density function f, then applying IS with the same ν is more than 1300 times worse than not applying it.

Table 1: IS can significantly reduce $\text{AMSE}_\star$, but it can also increase. $\text{AMISE}_\star$ always increases with IS.

	$R_\star(y)$				
ν	$y=0$	$y=1$	$y=2$	$y=3$	$\bar{R}_\star$
1	1.49	0.670	0.301	0.135	2.23
2	4.95	1	0.202	0.041	24.5
3	36.6	3.32	0.301	0.027	1339

3.4 Asymptotic Normality and Confidence Intervals

We now give conditions under which the IS kernel density estimator $\hat{f}_n^*(y)$ satisfies central limit theorems.

Theorem 4 Suppose the conditions of Theorem 1 hold and that $E_g[L^{2+\delta}(X)] < \infty$ for some $\delta > 0$. Then

$$(nh_n)^{1/2}\left(\hat{f}_n^*(y) - E_g[\hat{f}_n^*(y)]\right) \Rightarrow N(0, L(y)f(y)\rho_k) \tag{28}$$

as $n \to \infty$. If in addition the conditions of Theorem 3 hold and $nh_n^5 \to 0$, then as $n \to \infty$,

$$(nh_n)^{1/2}\left(\hat{f}_n^*(y) - f(y)\right) \Rightarrow N(0, L(y)f(y)\rho_k). \tag{29}$$

The CLT in (28) is centered at the mean of $\hat{f}_n^*(y)$, which in general is not equal to $f(y)$ because of the bias. Hence, (29) is more practically useful. Also, the asymptotically optimal rates in (26) and (27) do not satisfy $nh_n^5 \to 0$, so we are not guaranteed that the CLT in (29) holds for the optimal bandwidths.

We now describe how to construct a confidence interval for $f(y)$ based on the CLT (29). When f is the density of the output of a complicated simulation and IS is employed using density g, it is likely the case that $g(y)$ is unknown (or difficult to compute). Thus, $g(y)$ and correspondingly $L(y) = f(y)/g(y)$, which appears in the asymptotic variance in (29), need to be estimated to construct confidence intervals based on the CLT. We can estimate $g(y)$ using $g_n^*(y) = \frac{1}{n}\sum_{j=1}^n k'_{h'_n}(y - X_j)$, which is the standard kernel density estimator of g at y with kernel k' and bandwidth h'_n. Here, $X_1, X_2, \ldots, X_n$ are the same samples from density g used to construct $\hat{f}_n^*(y)$, and k' and h'_n could be the same as or different from the kernel k and bandwidth h_n used in $\hat{f}_n^*(y)$. Standard kernel theory (Parzen 1962) shows that $g_n^*(y) \Rightarrow g(y)$ as $n \to \infty$ when g is continuous at y, k' satisfies (5)–(6) and $h'_n \to 0$ and $nh'_n \to \infty$ as $n \to \infty$. Thus, Theorem 1 and Slutsky's theorem (e.g., p. 19 of Serfling 1980) imply $L_n(y) \equiv \hat{f}_n^*(y)/g_n^*(y) \Rightarrow f(y)/g(y) = L(y)$ as $n \to \infty$. Consequently, using the CLT (29), we can form an asymptotically valid $100(1-\alpha)\%$ confidence interval for $f(y)$ as

$$\left(\hat{f}_n^*(y) \pm z_{1-\alpha/2}\left(\frac{L_n(y)\hat{f}_n^*(y)\rho_k}{nh_n}\right)^{1/2}\right),$$

where $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$ and Φ is the distribution function of a $N(0,1)$.

3.5 Estimating the Sparsity Function

Suppose that we are interested in estimating the p-quantile $\xi_p = F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$ of F for a fixed $0 < p < 1$. Glynn (1996) develops an IS quantile estimator $\hat{\xi}_{p,n,1} = \hat{F}_{n,1}^{-1}(p)$, where $\hat{F}_{n,1} = \hat{F}_n$ defined in (16). He also considers another IS quantile estimator $\hat{\xi}_{p,n,2} = \hat{F}_{n,2}^{-1}(p)$, where $\hat{F}_{n,2}(y) = 1 - n^{-1}\sum_{j=1}^n I(X_j > y)L(X_j)$. The first (resp., second) quantile estimator $\hat{\xi}_{p,n,1}$ (resp., $\hat{\xi}_{p,n,2}$) is more appropriate to use when $p \approx 0$ (resp., $p \approx 1$). Glynn shows that if $E_g[L^3(X)] < \infty$ and $f(\xi_p) > 0$, then the following CLTs hold:

$$\sqrt{n}[\hat{\xi}_{p,n,\ell} - \xi_p] \Rightarrow N(0, \beta_{p,\ell}^2), \tag{30}$$

as $n \to \infty$ for $\ell = 1, 2$, where

$$\beta_{p,1}^2 = \frac{E_g[I(X \leq \xi_p)L^2(X)] - p^2}{f^2(\xi_p)}, \tag{31}$$

$$\beta_{p,2}^2 = \frac{E_g[I(X > \xi_p)L^2(X)] - (1-p)^2}{f^2(\xi_p)}. \tag{32}$$

Chu and Nakayama (2011) prove (30) holds for $\ell = 1$ (resp., $\ell = 2$) when the moment condition on the likelihood ratio is relaxed to $E_g[I(X < \xi_p + \gamma)L^{2+\delta}(X)] < \infty$ (resp., $E_g[I(X > \xi_p - \gamma)L^{2+\delta}(X)] < \infty$) for

some $\gamma > 0$ and $\delta > 0$. If we have consistent estimators of the numerator and denominator in (31) (resp., (32)), then we can construct a confidence interval for ξ_p based on (30) for $\ell = 1$ (resp., $\ell = 2$). Tukey (1965) calls $1/f(\xi_p)$ the *sparsity function* at p, which Parzen (1979) names the *quantile density function*.

Chu and Nakayama (2011) develop consistent estimators of the numerators and denominators in (31) and (32). To handle the denominator, note that $\frac{d}{dp}F^{-1}(p) = 1/f(\xi_p)$, and their estimator of the sparsity function is the finite difference

$$\frac{\hat{F}_{n,\ell}^{-1}(p+h_n) - \hat{F}_{n,\ell}^{-1}(p-h_n)}{2h_n}, \tag{33}$$

for $\ell = 1,2$, which they show is consistent for bandwidth $h_n = cn^{-1/2}$ for any constant $c \neq 0$. Moreover, under the further assumption that f is continuous at ξ_p, then (33) is consistent when $h_n \to 0$ and nh_n^2 goes to a positive constant or ∞ as $n \to \infty$. We now consider IS kernel estimators of $f(\xi_p)$ and $1/f(\xi_p)$.

Theorem 5 Suppose $f(\xi_p) > 0$ and f is continuous in a neighborhood of ξ_p. Also, assume $E_g[L^{2+\delta}(X)] < \infty$ for some $\delta > 0$, and k satisfies Assumption A1 and (20). If $h_n \to 0$ and $nh_n^2 \to \infty$ as $n \to \infty$, then

$$\hat{f}_n^*(\hat{\xi}_{p,n,\ell}) \Rightarrow f(\xi_p) \quad \text{and} \quad \frac{1}{\hat{f}_n^*(\hat{\xi}_{p,n,\ell})} \Rightarrow \frac{1}{f(\xi_p)},$$

as $n \to \infty$ for $\ell = 1,2$. In addition, suppose the assumptions of Theorem 3 hold for $y = \xi_p$, the characteristic function $\phi(t) = E_f[e^{itX}]$ of f satisfies $\int |t\phi(t)|\,dt < \infty$, and $\int |t\psi(t)|\,dt < \infty$. Then if $nh_n^4 \to \infty$ and $nh_n^5 \to 0$,

$$(nh_n)^{1/2}\left(\hat{f}_n^*(\hat{\xi}_{p,n,\ell}) - f(\xi_p)\right) \Rightarrow N(0, L(\xi_p)f(\xi_p)\rho_k),$$
$$(nh_n)^{1/2}\left(\frac{1}{\hat{f}_n^*(\hat{\xi}_{p,n,\ell})} - \frac{1}{f(\xi_p)}\right) \Rightarrow N\left(0, \frac{L(\xi_p)\rho_k}{f^3(\xi_p)}\right),$$

as $n \to \infty$ for $\ell = 1,2$.

Csörgő and Révész (1981), Theorem 5.5.2, establish the almost sure (uniform) consistency of the analogous kernel estimator for $f(\xi_p)$ for CMC. Falk (1986) proves a CLT for a different type of kernel estimator for $1/f(\xi_p)$ when applying CMC; his proof technique expresses F_n^{-1} as F^{-1} evaluated at the empirical CDF of n i.i.d. uniform$(0,1)$ samples, but this approach does not generalize when applying IS. Liu and Yang (2011) develop a bootstrap estimator of the IS asymptotic variance $\beta_{p,1}^2$ in (31), and they show under alternative hypotheses (e.g., $E_g[L^{4+\delta_1}(X)] < \infty$ and $E_g[|X|^{3+\delta_2}] < \infty$ for some $\delta_1, \delta_2 > 0$) that their estimator satisfies a CLT with rate $n^{-1/4}$, which is slower than the rate $(nh_n)^{-1/2}$ in the CLTs in Theorem 5.

4 EMPIRICAL STUDY

We now present some results from running experiments to construct confidence intervals (CIs) for a quantile of a small stochastic model when applying IS using the methods in Section 3.5. The model we consider is a stochastic activity network (SAN), previously studied by Hsu and Nelson (1990) and Chu and Nakayama (2011). The SAN consists of $d = 5$ independent activities, where each activity i has lifetime A_i, which is exponential with mean 1. The SAN has $q = 3$ paths, and B_j denotes the set of activities on path j, where $B_1 = \{1,2\}$, $B_2 = \{1,3,5\}$, and $B_3 = \{4,5\}$. Let $T_j = \sum_{i \in B_j} A_i$ be the length of the path j, and let $X = \max_{j=1,\ldots,q} T_j$ be the length of the longest path, which is the time to complete the project modeled by the SAN. Our goal is to estimate and construct a 90% CI for ξ_p, the p-quantile of X. The CDF of X is $F(x) = 1 + (3 - 3x - x^2/2)e^{-x} + (-3 - 3x + x^2/2)e^{-2x} - e^{-3x}$ for $x \geq 0$. In our experiments, we used sample sizes $n = 100 \times 4^r$, $r = 0,1,2,3$, and we estimated coverage (and average half widths) of the constructed CIs from 10^4 independent replications.

We experimented with $p \approx 1$, so we used the IS quantile estimator $\hat{\xi}_{p,n,2}$. We applied IS via an approach described in Chu and Nakayama (2011). Based on an idea in Juneja, Karandikar, and Shahabuddin (2007) for estimating tail probabilities in SANs, the IS measure is a mixture of q measures, where the jth measure in the mixture exponentially tilts the activities on path j, and activities not on path j retain their original distribution. We determine the tilting parameter for each path by applying an idea suggested by Glynn (1996). For further details, see Chu and Nakayama (2011).

To construct CIs for ξ_p we employed various methods to estimate $\lambda_p \equiv 1/f(\xi_p)$ from (32). We used IS kernel estimators $1/\hat{f}_n^*(\hat{\xi}_{p,n,2})$ with the uniform kernel (2), the Epanechnikov kernel (7), and the Gaussian kernel (8). (In contrast to (7) and (8), kernel (2) does not satisfy (20), so Theorem 5 does not assure its consistency.) In our tables below, these correspond to columns labeled "Unif. kernel", "Epan. kernel", and "Gauss. kernel", respectively. Also we applied the central finite-difference (CFD) estimator in (33) of Chu and Nakayama (2011). Columns labeled "Exact λ_p" are for CIs constructed using the exact value of λ_p. In all our IS CIs, we used the same estimator of the numerator of (32). We also ran the same experiments with CMC for comparison.

For CFD, issues arise when p is close to 1, the bandwidth h_n shrinks slowly, and the sample size n is small. In this case, we can have $p + h_n \geq 1$, but then the CFD estimator would evaluate the inverse of the estimated CDF at a point outside of its domain $(0,1)$. Therefore, when this occurs, rather than evaluate $\hat{F}_{n,2}^{-1}$ at $q_{1,n} \equiv p + h_n$ and $q_{2,n} \equiv p - h_n$, we instead evaluate at $q_{1,n} = 1 - (1-p)/10$ and $q_{2,n} = 2p - 1 + (1-p)/10$, the second point chosen so that $q_{1,n}$ and $q_{2,n}$ are symmetric about p. Some adjustment of this type must be done to ensure that the inverted estimated CDF is evaluated at points within its domain, but it can lead to poor estimates of the sparsity function for reasonably large sample sizes n. This seems to be due to the fact that when applying the adjustment, $q_{1,n}$ and $q_{2,n}$ may not be approaching p for the range of n with which we are experimenting. But the validity of the CFD requires that $q_{1,n} - q_{2,n} \to 0$, which may only occur when n is very large. See Chu and Nakayama (2011) for more details.

Table 2 gives the results from experiments for $p = 0.99$. Comparing the columns for exact λ_p (which do not depend on the bandwidth) for CMC and IS shows that IS reduces average half widths by more than a factor of 4, demonstrating the effectiveness of our IS scheme. We experimented with CFD and kernel estimators having bandwidths $h_n = 0.5n^{-v}$ for $v = 1/2$, $1/3$ and $1/5$. For CFD, the coverage for $v = 1/2$ does better than the other values of v. Also, $v = 1/2$ gives better estimates on average of λ_p for CFD, as seen by comparing the average half widths for CFD with those for exact λ_p. But for the kernel estimators of λ_p, $v = 1/5$ (which is the asymptotically optimal value in (26) for minimizing MSE of the kernel density estimator) outperforms $v = 1/2$ and $1/3$ when n is smaller. When $v = 1/2$, the kernel methods have undercoverage even for $n = 6400$ (and is especially poor for CMC), but our consistency theory for the kernel estimator of λ_p in Theorem 5 does not cover the case when $v = 1/2$. (The asymptotic theory for CFD in Chu and Nakayama 2011 allows for $0 < v \leq 1/2$.) For small n, the Gaussian and uniform kernels do better at estimating λ_p than the Epanechnikov kernel, but all of the kernel estimators perform about equally for large n when $v = 1/3$ and $1/5$. Additional experiments (not shown) estimating more extreme quantiles with $p = 1 - 10^{-e}$ for $e > 2$ show that CFD gives as bad or even higher overcoverage, while kernel estimators for $v = 1/3$ and $1/5$ give close to nominal coverage when $n \geq 1600$. Thus, it seems that for extreme quantiles, kernel estimators can produce CIs with closer to nominal coverage than CFD.

5 CONCLUDING REMARKS

This paper analyzed kernel density estimators when importance sampling is applied. We provided conditions under which IS kernel density estimators are pointwise and uniformly consistent. The estimators are also asymptotically normal, and we developed asymptotically valid confidence intervals for $f(y)$. We further provided expressions for the asymptotic MSE and MISE, which we used to determine the optimal bandwidths to minimize these asymptotic measures for a given IS density. One conclusion is that IS can improve

Table 2: For CIs for the 0.99-quantile when using CMC and IS, coverage levels are closer to nominal (0.9) for kernel estimators than for CFD, and average half widths (given in parentheses) for kernel estimators are closer than CFD to those for exact λ_p.

	$h_n = 0.5n^{-1/2}$									
	CMC					IS				
n	CFD	Unif. kernel	Epan. kernel	Gauss. kernel	Exact λ_p	CFD	Unif. kernel	Epan. kernel	Gauss. kernel	Exact λ_p
100	0.507	0.092	0.060	0.116	0.945	0.981	0.544	0.425	0.579	0.873
	(1.109)	(0.154)	(0.104)	(0.188)	(2.025)	(0.712)	(0.220)	(0.158)	(0.239)	(0.445)
400	0.924	0.182	0.122	0.222	0.915	0.989	0.700	0.617	0.721	0.897
	(1.430)	(0.150)	(0.102)	(0.180)	(1.012)	(0.372)	(0.174)	(0.133)	(0.174)	(0.232)
1600	0.981	0.346	0.250	0.400	0.906	0.991	0.778	0.733	0.802	0.901
	(0.824)	(0.139)	(0.096)	(0.161)	(0.506)	(0.188)	(0.110)	(0.093)	(0.105)	(0.117)
6400	0.936	0.545	0.420	0.587	0.898	0.941	0.829	0.809	0.841	0.897
	(0.291)	(0.121)	(0.086)	(0.132)	(0.253)	(0.068)	(0.059)	(0.054)	(0.056)	(0.059)
	$h_n = 0.5n^{-1/3}$									
	CMC					IS				
n	CFD	Unif. kernel	Epan. kernel	Gauss. kernel	Exact λ_p	CFD	Unif. kernel	Epan. kernel	Gauss. kernel	Exact λ_p
100	0.507	0.196	0.132	0.230	0.945	0.981	0.688	0.612	0.715	0.873
	(1.109)	(0.312)	(0.214)	(0.367)	(2.025)	(0.712)	(0.342)	(0.265)	(0.340)	(0.445)
400	0.924	0.429	0.309	0.474	0.915	0.989	0.808	0.770	0.827	0.897
	(1.430)	(0.352)	(0.248)	(0.395)	(1.012)	(0.372)	(0.226)	(0.200)	(0.215)	(0.232)
1600	0.981	0.660	0.569	0.684	0.906	0.991	0.859	0.841	0.867	0.901
	(0.824)	(0.333)	(0.251)	(0.343)	(0.506)	(0.188)	(0.118)	(0.113)	(0.114)	(0.117)
6400	0.989	0.775	0.727	0.795	0.898	0.991	0.879	0.872	0.882	0.897
	(0.403)	(0.236)	(0.201)	(0.226)	(0.253)	(0.094)	(0.059)	(0.058)	(0.058)	(0.059)
	$h_n = 0.5n^{-1/5}$									
	CMC					IS				
n	CFD	Unif. kernel	Epan. kernel	Gauss. kernel	Exact λ_p	CFD	Unif. kernel	Epan. kernel	Gauss. kernel	Exact λ_p
100	0.507	0.331	0.231	0.373	0.945	0.981	0.764	0.713	0.777	0.873
	(1.109)	(0.520)	(0.367)	(0.585)	(2.025)	(0.712)	(0.415)	(0.348)	(0.390)	(0.445)
400	0.924	0.622	0.525	0.648	0.915	0.989	0.859	0.838	0.864	0.897
	(1.430)	(0.601)	(0.450)	(0.620)	(1.012)	(0.372)	(0.232)	(0.222)	(0.223)	(0.232)
1600	0.981	0.776	0.735	0.796	0.906	0.991	0.881	0.873	0.883	0.901
	(0.824)	(0.466)	(0.401)	(0.441)	(0.506)	(0.188)	(0.117)	(0.115)	(0.115)	(0.117)
6400	0.989	0.851	0.831	0.860	0.898	0.991	0.890	0.888	0.890	0.897
	(0.403)	(0.252)	(0.242)	(0.244)	(0.253)	(0.094)	(0.059)	(0.058)	(0.058)	(0.059)

(relative to CMC) the AMSE when estimating the density at only a single point y, with $L^{4/5}(y)$ being the factor by which the AMSE changes when applying IS with the optimal bandwidth. But IS always does worse (in terms of AMISE) when estimating the entire density function, and $E_g^{4/5}[L^2(X)] \geq 1$ is the factor by which the AMISE increases when using IS with the optimal bandwidth. We also developed a consistent kernel estimator of the sparsity function when applying IS, which is useful for constructing a confidence interval for a quantile when using IS, and also gave conditions under which this estimator satisfies a CLT. We included some empirical results from experimenting with a small model. The results suggest that kernel estimators of the sparsity function λ_p may lead to better CIs for extreme quantiles than the CFD estimator of Chu and Nakayama (2011).

We are currently investigating kernel density estimation when applying other variance-reduction techniques, such as control variates; e.g., see Chapter V of Asmussen and Glynn (2007). Section 3.5 considers a "plug-in" kernel estimator of the sparsity function λ_p, and we are now studying other kernel estimators of λ_p, such as those considered in Falk (1986) for CMC.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation under Grant No. CMMI-0926949. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Asmussen, S., and P. Glynn. 2007. *Stochastic simulation: Algorithms and analysis*. New York: Springer.

Billingsley, P. 1999. *Convergence of probability measures*. Second ed. New York: John Wiley & Sons.

Chu, F., and M. K. Nakayama. 2011. "Confidence Intervals for Quantiles When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation*, to appear.

Csörgő, M., and P. Révész. 1981. *Strong approximations in probability and statistics*. New York: Academic Press.

Falk, M. 1986. "On the estimation of the quantile density function". *Statistics & Probability Letters* 4:69–73.

Glynn, P. W. 1996. "Importance sampling for Monte Carlo estimation of quantiles". In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, 180–185: Publishing House of St. Petersburg University, St. Petersburg, Russia.

Hong, L. J., and G. Liu. 2010. "Pathwise estimation of probability sensitivities through terminating and steady-state simulations". *Operations Research* 58:357–370.

Hsu, J. C., and B. L. Nelson. 1990. "Control variates for quantile estimation". *Management Science* 36:835–851.

Juneja, S., R. Karandikar, and P. Shahabuddin. 2007. "Asymptotics and Fast Simulation for Tail Probabilities of Maximum of Sums of Few Random Variables". *ACM Transactions on Modeling and Computer Simulation* 17:article 2, 35 pages.

Liu, J., and X. Yang. 2011. "The convergence rate and asymptotic distribution of bootstrap quantile variance estimator for importance sampling". Preprint.

Nakayama, M. K. 2011. "Kernel density estimators when applying importance sampling". In preparation.

Pagan, A., and A. Ullah. 1999. *Nonparametric econometrics*. England: Cambridge University Press.

Parzen, E. 1962. "On estimation of a probability density and mode". *Annals of Mathematical Statistics* 33:1065–1076.

Parzen, E. 1979. "Density Quantile Estimation Approach to Statistical Data Modelling". In *Smoothing Techniques for Curve Estimation*. Berlin: Springer.

Rosenblatt, M. 1956. "Remarks on some non-parametric estimates of a density function". *Annals of Mathematical Statistics* 27:832–837.

Serfling, R. J. 1980. *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons.
Tukey, J. W. 1965. "Which Part of the Sample Contains the Information?". *Proc Natl Acad Sci USA* 53:127–134.
Wand, M. P., and M. C. Jones. 1995. *Kernel smoothing*. London: Chapman & Hall.

AUTHOR BIOGRAPHY

MARVIN K. NAKAYAMA is a professor in the Department of Computer Science at the New Jersey Institute of Technology. He received his Ph.D. in operations research from Stanford University. He won second prize in the 1992 George E. Nicholson Student Paper Competition sponsored by INFORMS and is a recipient of a CAREER Award from the National Science Foundation. He is the Stochastic Models Area Editor for *ACM Transactions on Modeling and Computer Simulation* and the Simulation Area Editor for *INFORMS Journal on Computing*.