Monte Carlo and Quasi-Monte Carlo Density Estimation via Conditioning

Pierre L'Ecuyer
Département d'Informatique et de Recherche Opérationnelle, Pavillon Aisenstadt, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7, lecuyer@iro.umontreal.ca

Florian Puchhammer
Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Basque Country, Spain; and Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, fpuchhammer@bcamath.org

Amal Ben Abdellah
Département d'Informatique et de Recherche Opérationnelle, Pavillon Aisenstadt, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7, amal.ben.abdellah@umontreal.ca

Estimating the unknown density from which a given independent sample originates is more difficult than estimating the mean, in the sense that for the best popular non-parametric density estimators, the mean integrated square error converges more slowly than at the canonical rate of $\mathcal{O}(1/n)$. When the sample is generated from a simulation model and we have control over how this is done, we can do better. We examine an approach in which conditional Monte Carlo yields, under certain conditions, a random conditional density which is an unbiased estimator of the true density at any point. By averaging independent replications, we obtain a density estimator that converges at a faster rate than the usual ones. Moreover, combining this new type of estimator with randomized quasi-Monte Carlo to generate the samples typically brings a larger improvement on the error and convergence rate than for the usual estimators, because the new estimator is smoother as a function of the underlying uniform random numbers.

Key words: density estimation; conditional Monte Carlo; quasi-Monte Carlo

1. Introduction

Simulation is commonly used to generate n realizations of a random variable X that may represent a payoff, a cost, or a performance of some kind, and then to estimate from this sample the unknown expectation of X together with a confidence interval on this expectation (Asmussen and Glynn 2007, Law 2014). Simulation books focus primarily on how to improve the quality of the estimator of $\mathbb{E}[X]$ and of the confidence interval. Estimating a given quantile of the distribution of X, or the sensitivity of $\mathbb{E}[X]$ with respect to some parameter in the model, also with a confidence interval, are other well-studied topics in the literature.

However, large simulation experiments can provide a lot more information than just point estimates with confidence intervals. Running simulations of a complex system for hours, with thousands of runs, only to report confidence intervals on a few single numbers is poor data valorization. A

simulation experiment can give much more useful information than this. In particular, it can provide an estimate of the entire distribution of X, and not only its expectation or a specific quantile. Moreover, and perhaps more importantly, users are typically more interested in the whole distribution than on a confidence interval on the mean. The following examples of typical simulation models show why.

In many real-life stochastic simulation models, the prime focus of interested is the distribution of certain random *delays*. These delays can be for example the waiting times of calls in a telephone call center, the waiting times of patients at a walk-in medical clinic or at the emergency, the waiting time of passengers at an airport checking counter, the delivery time of an order, the travel time in some transportation network, etc. In all these situations, a user will be mostly interested in *what his own waiting time is likely to be*. That is, she/he is much more interested in the *probability distribution* of the waiting time than in having a good estimate of the expectation (or global mean) (Nelson 2008, Smith and Nelson 2015).

When one makes a call to a call center and all agents are busy, a good forecast of the waiting time is certainly appreciated. Based on this forecast, the caller may decide that there is enough time to engage in another activity before getting an answer. The expected waiting time alone is not sufficient to make such as decision, because for example it does not tell the probability of missing the call when going out for x minutes. A distributional forecast, which provides a density of the waiting time distribution (perhaps conditional on the current time and system state) is much more informative and helpful (Thiongane et al. 2021). This applies to waiting times in many other types of service systems. Users are interested in the density (or distribution) of their waiting time, not just the expectation. When a manufacturer orders parts from a supplier, or a retail store orders items from the manufacture or distributor, an estimate of the density of the time until delivery (and not just its expectation) gives them good idea of what can happen, including the probability that the parts or items arrive on time, and the distribution of the delay if there is one. In a large construction project that involves many activities of random durations and precedence constraints, the total time to complete the project is a random variable X usually modeled by a stochastic activity network (see Section 4.3). Knowing the density of X permits one to assess the risks in signing contracts that impose various types of penalties when X is too large. In many other situations, X is a cost or a profit and estimating the density of X is again more interesting and useful than just the expectation. In a finance application, for example, X may represent an investment loss over a given month, and the density of X provides much more information on the possibilities of large losses (and perhaps bankruptcy) than just the mean. We will report numerical experiments for a finance-type example in Section 4.6.

So, simulation users are interested in the whole distribution of the output and not only the mean. One way to visualize the *entire distribution of* X is to look at the empirical *cumulative distribution function* (cdf) of the observations. But density estimators (including histograms) are preferred because they give a better visual insight on the distribution than the cdf. For this reason, leading simulation software routinely provides histograms and enhanced boxplots that give a rough idea of the distribution of the output random variables of interest. For example, the standard output display in Simio$^{\text{TM}}$ is a histogram enhanced with a boxplot, named SMORE (Simio Measure Of Risk and Error) (Sturrock and Pegden 2010, Smith and Nelson 2015). When X has a continuous distribution, these histograms and boxplots are in fact just primitive forms of density estimators. So why are better density estimators not routinely offered? Mainly because non-parametric density estimation is difficult.

If the density of X is *assumed* to have a known *parametric form*, e.g., a normal or gamma distribution, then one can estimate the parameters from data in the usual way (e.g., by maximum likelihood) and things are simple. But in typical complex models, X does not have a known and simple form of distribution. There are *semi-parametric* procedures in which the density is assumed to belong to a Hilbert space of functions which are linear combinations a finite number of fixed basis functions, and the coefficients are estimated by penalized regression. These are known as smoothing spline models (Gu and Qiu 1993, Yu et al. 2020). But it is often difficult to select basis functions that capture the unknown density and a good choice depends on the problem. In this paper, we focus on *non-parametric* methods, in the sense that we assume no particular form for the density of X. On the other hand, the input distributions in the simulation model may be parametric (they often are). The most widely used non-parametric density estimation methods are the *histogram* and the *kernel density estimator* (KDE) (Parzen 1962, Silverman 1986, Wand and Jones 1995, Scott 2015). Given n independent realizations of X, the mean integrated square error (MISE) between the true density and a histogram with optimally selected divisions converges only as $\mathcal{O}(n^{-2/3})$. With the KDE, the MISE converges as $\mathcal{O}(n^{-4/5})$ in the best case. These rates are slower than the canonical $\mathcal{O}(n^{-1})$ rate for the variance of the sample average as an unbiased estimator of the mean. The slower rates stem from the presence of bias. See for example Scott (2015) for the details. For a histogram, taking wider rectangles reduces the variance but increases the bias by flattening out the short-range density variations. A compromise must be made to minimize the MISE. The same happens with the KDE, with the rectangle width replaced by the bandwidth of the kernel. Selecting a good bandwidth for the KDE is particularly difficult. The bandwidth should ideally vary over the interval in which we estimate the density; it should be smaller where the density is larger and/or smoother, and vice-versa. This is complicated to implement. Handling discontinuities

in the density is also problematic. These difficulties have discouraged the use of KDEs (instead of histograms) to report simulation results in general-purpose software.

The KDE and other related density estimation methods were developed mainly for the situation where n independent realizations of X are given and nothing else is known, as traditionally assumed in classical non-parametric statistics, and one wishes to estimate the density from them (Scott 2015). But in a Monte Carlo setting in which the n observations are generated by simulation, there are opportunities to do better by controlling the way we generate the realizations and by exploiting the fact that we know the underlying stochastic model. This is the subject of the present paper.

Our approach combines two general ideas. The first one is to build a smooth estimator of the cdf via conditional Monte Carlo (CMC), and take the corresponding conditional density to estimate the unknown density. We call it a *conditional density estimator* (CDE). Under appropriate conditions, the CDE is unbiased and has uniformly-bounded variance, so its MISE is $\mathcal{O}(n^{-1})$ for n samples. This idea of using CMC was mentioned by Asmussen and Glynn (2007), page 146, Example 4.3, and further studied in Asmussen (2018), but only for the special case of estimating the density of a sum of i.i.d. continuous random variables having a known density. Asmussen (2018) simply "hides" the last term of the sum, meaning that the last random variable is not generated, and he takes a shifted version of the known density of this last variable to estimate the density, the value at risk, and the conditional value at risk of the sum. His setting is equivalent to a sum of two independent random variables: the first one is the partial sum which is generated and on which we condition, and the second one is the last variable which is not generated. Fu (2006) mentioned this same idea in one of his examples.

Smoothing by CMC before taking a stochastic derivative has been studied earlier for estimating the derivative of an expectation (Gong and Ho 1987, L'Ecuyer and Perron 1994, Fu and Hu 1997) and the derivative of a quantile (Fu et al. 2009) with respect to a *model parameter*. This is known as *smoothed perturbation analysis* (SPA). In retrospect, one can say that the CDE at a given point x is an SPA estimator obtained by viewing the cdf $F(x)$ as the expectation and x as the model parameter. However, nobody studied this idea for density estimation until Asmussen (2018) did it for his special case.

The main contribution of this paper is to show how this CDE approach can be used to estimate the density in a much more general setting than Asmussen (2018), to give conditions under which it provides an unbiased density estimator, and to examine how effective it is via experiments on several types of examples. In most of these examples, X is *not* defined as a sum of random variables, and we often have to hide more than just one random variable to do the conditioning. A key unbiasedness condition is that the conditional cdf must be a continuous function of the point x at which we estimate the density. In other words, the conditional distribution of X under the

selected conditioning must have a density with respect to the Lebesgue measure. The variance of the density estimator may depend strongly on which variables we hide, i.e., on what we are conditioning. We illustrate this with several examples and we provide guidelines for the choice of conditioning. Interestingly, while the KDE is defined as an average of n randomly-shifted copies of the (fixed) kernel density, the CDE is an average of n conditional densities which are generally different and random.

In addition to being unbiased, the CDE often has less variation than the KDE as a function of the underlying uniform random numbers. As a result, its combination with *randomized quasi-Monte Carlo* (RQMC) tends to bring much more improvement than for the KDE. We have observed this in all our experiments. Under appropriate conditions, it can be proved that combining the CDE with RQMC provides a density estimator whose MISE converges at a faster rate than $\mathcal{O}(n^{-1})$, for instance $\mathcal{O}(n^{-2+\epsilon})$ for any $\epsilon > 0$ in some situations. We observe this fast rate empirically on numerical examples. This happens essentially when the CDE is a smooth function of the underlying uniforms. To our knowledge, this type of convergence rate has never been proved or observed for non-parametric density estimation.

The combination of RQMC with an ordinary KDE was studied by Ben Abdellah et al. (2021), who were able to prove a faster rate than $\mathcal{O}(n^{-4/5})$ for the MISE when the RQMC points have a small number of dimensions. They observed this faster rate empirically on examples. They also showed that the MISE reduction from RQMC degrades rapidly when the bandwidth is reduced (to reduce the bias) or when the dimension increases. The CDE+RQMC approach studied in the present paper avoids this problem (there is no bias and no bandwidth) and is generally much more effective than the KDE+RQMC combination. We provide numerical comparisons in our examples.

Other Monte Carlo density estimators were proposed very recently, also based on the idea of estimating the derivative of the cdf, but using a likelihood ratio (LR) method instead. The LR method was originally designed to estimate the derivative of the expectation with respect to parameters of the distribution of the underlying input random variables (Glynn 1987, L'Ecuyer 1990). Laub et al. (2019) proposed an estimator that combines a clever change of variable with the LR method, to estimate the density of a sum of random variables as in Asmussen (2018), but in a setting where the random variables can be dependent. Peng et al. (2018) proposed a generalized version of the LR gradient estimator method, named GLR, to estimate the derivative of an expectation with respect to a more general model parameter. Lei et al. (2018) sketched out how GLR could be used to estimate a density. Formulas for these GLR density estimators are given in Theorem 1 of Peng et al. (2020). We compare them with the CDE estimators in our numerical illustrations.

Density estimation has other applications than just visualizing the distribution of an output random variable (Van der Vaart 2000, Scott 2015). For instance when computing a confidence

interval for a quantile using the central-limit theorem (CLT), one needs a density estimator at the quantile to estimate the variance (Serfling 1980, Asmussen and Glynn 2007, Nakayama 2014a,b). See Section B.4 in the Supplement. Another application is for maximum likelihood estimation when the likelihood does not have a closed-form expression, so to maximize it with respect to some parameter θ, the likelihood function (which in the continuous case is a density at any value of θ) must be estimated (Van der Vaart 2000, Peng et al. 2020). A related application is the estimation of the posterior density of θ given some data, in a Bayesian model (Efron and Hastie 2016).

The remainder is organized as follows. In Section 2, we define our general setting, recall key facts about density estimators, introduce the general CDEs considered in this paper, prove some of their properties, and give small examples to provide insight on the key ideas. We also briefly recall GLR density estimators. In Section 3, we explain how to combine the CDE with RQMC and discuss the convergence properties for this combination. Section 4 reports experimental results with various examples. Some of the examples feature creative ways of conditioning to improve the effectiveness of the method. Additional examples are examined in the Online Supplement. Section 5 summarizes the key issues and guidelines on the construction and applications of the CDE. A conclusion is given in Section 6. The main ideas of this paper were presented at a SAMSI workshop on QMC methods in North Carolina, and at a RICAM workshop in Linz, Austria, both in 2018.

2. Model and conditional density estimator

2.1. Density estimation setting

We have a real-valued random variable X that can be simulated from its exact distribution, but we do not know the cdf F and density f of X. Typically, X will be an easily computable function of several other random variables with known densities. Our goal is to estimate f over a finite interval $[a, b]$. Let $\hat{f}_n$ denote an estimator of f based on a sample of size n. We measure the quality of $\hat{f}_n$ by the *mean integrated square error* (MISE), defined as

$$\mathrm{MISE} = \mathrm{MISE}(\hat{f}_n) = \int_a^b \mathbb{E}[(\hat{f}_n(x) - f(x))^2]\mathrm{d}x. \tag{1}$$

The MISE is the sum of the *integrated variance* (IV) and the *integrated square bias* (ISB):

$$\mathrm{MISE} = \mathrm{IV} + \mathrm{ISB} = \int_a^b \mathbb{E}(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)])^2\mathrm{d}x + \int_a^b (\mathbb{E}[\hat{f}_n(x)] - f(x))^2\mathrm{d}x.$$

A standard way of constructing $\hat{f}_n$ when $X_1, \ldots, X_n$ are n independent realizations of X is via a KDE, defined as follows (Parzen 1962, Scott 2015):

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right),$$

where the *kernel* k is a probability density over $\mathbb{R}$, usually symmetric about 0 and non-increasing over $[0,\infty)$, and the constant $h>0$ is the *bandwidth*, whose role is to stretch [or compress] the kernel horizontally to smooth out [or unsmooth] the estimator $\hat{f}_n$. The KDE was developed for the setting in which $X_1,\dots,X_n$ are given a priori, and it is the most popular estimator for this situation. It can be used as well when $X_1,\dots,X_n$ are independent observations produced by simulation from a generative model, but then there is an opportunity to do better, as we now explain.

2.2. Conditioning and the stochastic derivative as an unbiased density estimator

Since the density of X is the derivative of its cdf, $f(x)=F'(x)$, a natural idea would be to take the derivative of an estimator of the cdf as a density estimator. The simplest candidate for a cdf estimator is the *empirical cdf*

$$\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[X_i\le x],$$

but $\mathrm{d}\hat{F}_n(x)/\mathrm{d}x=0$ almost everywhere, so this one *cannot* be a useful density estimator. Here, $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$ at each x, but its derivative is a biased estimator of $F'(x)$. That is, because of the discontinuity of $\hat{F}_n$, we cannot exchange the derivative and expectation:

$$0=\mathbb{E}\left[\frac{\mathrm{d}\hat{F}_n(x)}{\mathrm{d}x}\right]\neq\frac{\mathrm{d}\mathbb{E}[\hat{F}_n(x)]}{\mathrm{d}x}=F'(x).$$

A general framework to construct a continuous estimator of F via CMC is the following. Replace the indicator $\mathbb{I}[X\le x]$ by its *conditional cdf* given filtered (reduced) information $\mathcal{G}$: $F(x\mid\mathcal{G})\stackrel{\text{def}}{=}\mathbb{P}[X\le x\mid\mathcal{G}]$, where $\mathcal{G}$ is a sigma-field that contains not enough information to reveal X but enough to compute $F(x\mid\mathcal{G})$. Here, knowing the realization of $\mathcal{G}$ means knowing the realizations of all $\mathcal{G}$-measurable random variables. Our CDE to estimate $f(x)$ will be the *conditional density* $f(x\mid\mathcal{G})\stackrel{\text{def}}{=}F'(x\mid\mathcal{G})=\mathrm{d}F(x\mid\mathcal{G})/\mathrm{d}x$, when it exists. We assume that this estimator can be computed (or approximated) for (almost) all realizations of $\mathcal{G}$. Under the following assumption, we prove that $f(x\mid\mathcal{G})$ exists almost surely and is an unbiased estimator of $f(x)$ whose variance is bounded uniformly in x. Since $F(\cdot\mid\mathcal{G})$ cannot decrease, $f(\cdot\mid\mathcal{G})$ is never negative.

Assumption 1. *For all realizations of $\mathcal{G}$, $F(x\mid\mathcal{G})$ is a continuous function of x over the interval $[a,b]$, and is differentiable except perhaps at a countable set of points $D(\mathcal{G})\subset[a,b]$. For all $x\in[a,b]$, $F(x\mid\mathcal{G})$ is differentiable at x w.p.1. There is also a random variable Γ defined over the same probability space as $F(x\mid\mathcal{G})$, such that $\mathbb{E}[\Gamma^2]\le K_\gamma$ for some constant $K_\gamma<\infty$, and for which $\sup_{x\in[a,b]\setminus D(\mathcal{G})}F'(x\mid\mathcal{G})\le\Gamma$.*

Proposition 1. *Under Assumption 1, $\mathbb{E}[f(x\mid\mathcal{G})]=f(x)$ and $\mathrm{Var}[f(x\mid\mathcal{G})]\le K_\gamma$ for all $x\in[a,b]$.*

PROOF. We adapt the proof of Theorem 1 of L'Ecuyer (1990). By Theorem 8.5.3 of Dieudonné (1969), which is a form of mean value inequality theorem for non-differentiable functions, for every $x \in [a,b]$ and $\delta > 0$, with probability 1, we have

$$0 \le \frac{\Delta(x,\delta,\mathcal{G})}{\delta} \stackrel{\text{def}}{=} \frac{F(x+\delta \mid \mathcal{G}) - F(x \mid \mathcal{G})}{\delta} \le \sup_{y \in [x,x+\delta] \setminus D(\mathcal{G})} F'(y \mid \mathcal{G}) \le \Gamma.$$

Then, by the dominated convergence theorem,

$$\mathbb{E}\left[\lim_{\delta \to 0} \frac{\Delta(x,\delta,\mathcal{G})}{\delta}\right] = \lim_{\delta \to 0} \mathbb{E}\left[\frac{\Delta(x,\delta,\mathcal{G})}{\delta}\right],$$

which shows the unbiasedness. Moreover, $\mathrm{Var}[f(x \mid \mathcal{G})] = \mathrm{Var}[F'(x \mid \mathcal{G})] \le \mathbb{E}[\Gamma^2] \le K_\gamma$. □

Suppose now that $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(n)}$ are n independent realizations of $\mathcal{G}$, so $F(x \mid \mathcal{G}^{(1)}), \ldots, F(x \mid \mathcal{G}^{(n)})$ are independent realizations of $F(x \mid \mathcal{G})$, and consider the CDE

$$\hat{f}_{\mathrm{cde},n}(x) = \frac{1}{n} \sum_{i=1}^{n} f(x \mid \mathcal{G}^{(i)}). \tag{2}$$

Under Assumption 1, it follows from Proposition 1 that $\mathrm{ISB}(\hat{f}_{\mathrm{cde},n}) = 0$ and $\mathrm{MISE}(\hat{f}_{\mathrm{cde},n}) = \mathrm{IV}(\hat{f}_{\mathrm{cde},n}) \le (b-a)K_\gamma/n$. An unbiased estimator of this IV is given by

$$\widehat{\mathrm{IV}} = \widehat{\mathrm{IV}}(\hat{f}_{\mathrm{cde},n}) = \frac{1}{n-1} \int_a^b \sum_{i=1}^{n} \left[f(x \mid \mathcal{G}^{(i)}) - \hat{f}_{\mathrm{cde},n}(x)\right]^2 \mathrm{d}x. \tag{3}$$

In practice, this integral can be approximated by evaluating the integrand at a finite number of points over $[a,b]$ and taking the average, multiplied by $(b-a)$.

The variance of the CDE estimator at x is $\mathrm{Var}[f(x \mid \mathcal{G})]$, where x is fixed and $\mathcal{G}$ is random. This differs from the variance associated with the conditional density $f(\cdot \mid \mathcal{G})$, which is $\mathrm{Var}[X \mid \mathcal{G}]$. It is well known that in general, when estimating $\mathbb{E}[X]$, a CMC estimator never has a larger variance than X itself, and the more information we hide, the smaller the variance. That is, if $\mathcal{G} \subset \tilde{\mathcal{G}}$ are two sigma-fields such that $\mathcal{G}$ contains only a subset of the information of $\tilde{\mathcal{G}}$, then

$$\mathrm{Var}[\mathbb{E}[X \mid \mathcal{G}]] \le \mathrm{Var}[\mathbb{E}[X \mid \tilde{\mathcal{G}}]] \le \mathrm{Var}[X]. \tag{4}$$

Noting that $F(x \mid \mathcal{G}) = \mathbb{E}[\mathbb{I}[X \le x] \mid \mathcal{G}]$, we also have

$$\mathrm{Var}[F(x \mid \mathcal{G})] \le \mathrm{Var}[F(x \mid \tilde{\mathcal{G}})] \le \mathrm{Var}[\mathbb{I}[X \le x]] = F(x)(1-F(x)).$$

Thus, (4) applies as well to the (conditional) cdf estimator. However, applying it to the CDE is less straightforward. It is obviously not true that $\mathrm{Var}[F'(x \mid \mathcal{G})] \le \mathrm{Var}[\mathrm{d}\mathbb{I}[X \le x]/\mathrm{d}x]$ because the latter is zero almost everywhere. Nevertheless, we can prove the following.

LEMMA 1. *If $\mathcal{G} \subset \tilde{\mathcal{G}}$ both satisfy Assumption 1, then for all $x \in [a,b]$, we have* $\mathrm{Var}[f(x \mid \mathcal{G})] \le \mathrm{Var}[f(x \mid \tilde{\mathcal{G}})]$.

PROOF. The result does not follow directly from (4) because F' is not an expectation; this is why our proof does a little detour. For an arbitrary $x \in [a,b]$ and a small $\delta > 0$, define the random variable $I = I(x,\delta) = \mathbb{I}[x < X \le x+\delta]$. We have $\mathbb{E}[I \mid \mathcal{G}] = F(x+\delta \mid \mathcal{G}) - F(x \mid \mathcal{G})$, as in the proof of Proposition 1, and similarly for $\tilde{\mathcal{G}}$. Using (4) with I in place of X gives

$$\mathrm{Var}[\mathbb{E}[I \mid \mathcal{G}]] \le \mathrm{Var}[\mathbb{E}[I \mid \tilde{\mathcal{G}}]]. \tag{5}$$

We have

$$f(x \mid \mathcal{G}) = \lim_{\delta \to 0} \frac{F(x+\delta \mid \mathcal{G}) - F(x \mid \mathcal{G})}{\delta} = \lim_{\delta \to 0} \mathbb{E}[I(x,\delta)/\delta \mid \mathcal{G}]$$

and similarly for $\tilde{\mathcal{G}}$. Combining this with (5), we obtain

$$\begin{aligned} \mathrm{Var}[f(x \mid \mathcal{G})] &= \mathrm{Var}[\lim_{\delta \to 0} \mathbb{E}[I(x,\delta)/\delta \mid \mathcal{G}]] \;=\; \lim_{\delta \to 0} \mathrm{Var}[\mathbb{E}[I(x,\delta)/\delta \mid \mathcal{G}]] \\ &\le \lim_{\delta \to 0} \mathrm{Var}[\mathbb{E}[I(x,\delta)/\delta \mid \tilde{\mathcal{G}}]] \;=\; \mathrm{Var}[\lim_{\delta \to 0} \mathbb{E}[I(x,\delta)/\delta \mid \tilde{\mathcal{G}}]] \;=\; \mathrm{Var}[f(x \mid \tilde{\mathcal{G}})], \end{aligned}$$

in which the exchange of "Var" with the limit (at two places) can be justified by a similar argument as in Proposition 1. More specifically, we need to apply the dominated convergence theorem to $\mathbb{E}[I(x,\delta)/\delta \mid \mathcal{G}]$, which is just the same as in Proposition 1, and also to its square, which is also valid because the square is bounded uniformly by Γ^2. This completes the proof. □

This lemma tells us that conditioning on less information (hiding more) always reduces the variance of the CDE (or keep it the same). But if we hide more, the CDE may be harder or more costly to compute, so a compromise must be made to minimize the work-normalized MISE (which is the MISE multiplied by the expected time to compute the estimator), and the best compromise is generally problem-dependent. When none of $\mathcal{G}$ or $\tilde{\mathcal{G}}$ is a subset of the other, the variances of the corresponding conditional density estimators may differ significantly, and Lemma 1 does not apply, so other strategies must be used to select $\mathcal{G}$ when there are multiple possibilities.

In our setting, the most important condition is that $\mathcal{G}$ must satisfy Assumption 1. Any such $\mathcal{G}$ provides an unbiased density estimator with finite variance. When there are multiple choices, in general we want to choose $\mathcal{G}$ so that the conditional density tends to be *spread out* as opposed to being concentrated in a narrow peak. We give concrete examples of this in Section 4. This criterion is heuristic. If f is very spiky itself, then the CDE must be spiky as well, because $\mathrm{Var}[X \mid \mathcal{G}] \le \mathrm{Var}[X]$, and yet $\mathrm{Var}[f(x \mid \mathcal{G})]$ can be very small, even zero in degenerate cases. Also, a large $\mathrm{Var}[X \mid \mathcal{G}]$ for all $\mathcal{G}$ is not sufficient, because the large variance may come from two or more separate spikes, and this is why we write "spread out" instead of "large variance". Roughly, we want the CDE $f(\cdot \mid \mathcal{G})$ to be spread out relative to f, for all realizations of $\mathcal{G}$.

A more elaborate selection criterion should take into account the IV of the CDE, its computing cost, and also the variation of of the resulting CDE as a function of the underlying uniform random

numbers, in case we want to use RQMC to generate those random numbers (see Section 3). For real-life models, it is usually much too hard to precompute such measures, so the best practice would be to identify a few promising candidates and either: (1) perform pilot runs to compare their effectiveness and select one or (2) take a convex combination of the corresponding CDEs, as explained in Section 2.4. We believe that finding a good $\mathcal{G}$ will always remain largely problem-dependent and it sometimes requires creativity. No simple selection method works universally. On the other hand, to make good selections, it is useful to understand certain basic principles. We illustrate this with a variety of examples in the next subsection and in Section 4.

2.3. Small examples to provide insight

To illustrate some key ideas, this subsection provides simple examples formulated in the special setting in which $X = h(Y_1, \ldots, Y_d)$ where $Y_1, \ldots, Y_d$ are independent continuous random variables, each Y_j has cdf F_j and density f_j, and we condition on $\mathcal{G} = \mathcal{G}_{-k}$ defined as the information that remains after erasing the value taken by the single input variable Y_k. We can write $\mathcal{G}_{-k} = (Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_d)$. The CDE $f(x \mid \mathcal{G}_{-k})$ will be related to the density f_k and will depend on the form of h. Checking for the continuity of the conditional cdf is usually easy in this case. Note that this setting is only a particular case of our framework. In many applications, X is not defined like this in a way that $\mathcal{G}_{-k}$ would satisfy Assumption 1 for some k. In Section 4, we examine examples that do not fit this setting and we provide more elaborate forms of conditioning.

Our first example is a sum of random variables, similar to Asmussen (2018). It conveys the CDE idea in a simple setting. It also shows that selecting which variable to hide is not straightforward even in this very simple setting, and that the optimal choice may depend on the value of x at which we estimate the density. The second example shows how the choice of $\mathcal{G}$ can make a significant difference in performance, and that it is usually better to hide variables having a larger variance contribution. The third example illustrates what we have to do to verify Assumption 1 for a given application. The fourth example shows that we cannot always obtain an unbiased CDE by hiding a single variable. The fifth example shows that it is not always easy to know what is the optimal information to hide. On the other hand, the CDE can still work well even if we do not use the optimal $\mathcal{G}$.

Example 1. A very simple situation is when $X = h(Y_1, \ldots, Y_d) = Y_1 + \cdots + Y_d$, a sum of d independent continuous random variables. By hiding Y_k for an arbitrary k, we get

$$F(x \mid \mathcal{G}_{-k}) = \mathbb{P}[X \le x \mid S_{-k}] = \mathbb{P}[Y_k \le x - S_{-k}] = F_k(x - S_{-k}),$$

where $S_{-k} \stackrel{\text{def}}{=} \sum_{j=1,\, j\neq k}^{d} Y_j$, and the density estimator becomes $f(x \mid \mathcal{G}_{-k}) = f_k(x - S_{-k})$. This form also works when the Y_j's are not independent if we are able to compute the density of Y_k conditional

on $\mathcal{G}_{-k}$. It then suffices to replace f_k by this conditional density. Asmussen (2018) studied exactly this model, with independent variables and $k = d$.

When the Y_j's have different distributions and we want to hide one, which one should we hide? Intuition may suggest to hide the one having the largest variance. This simple rule works well in a majority of cases, although it is not always optimal. In particular, the optimal choice of variable Y_k may depend on the value of x at which we estimate the density. To illustrate this, let $d = 2$, $X = Y_1 + Y_2$, $f_1(y) = 2y$, and $f_2(y) = 2(1-y)$, for $y \in (0,1)$. Then, $f(x) > 0$ for $0 < x < 2$. If we hide Y_2, the density estimator at x is $f_2(x - Y_1)$ and its second moment is $\mathbb{E}[f_2^2(x - Y_1)] = \int_0^1 f_2^2(x-y_1) f_1(y_1) \mathrm{d}y_1$ whereas if we hide Y_1, the density estimator at x is $f_1(x - Y_2)$ and its second moment is $\mathbb{E}[f_1^2(x - Y_2)] = \int_0^1 f_1^2(x-y_2) f_2(y_2) \mathrm{d}y_2$. One can easily verify that when x is close to 0, these integrands are nonzero only when both y_1 and y_2 are also close to 0, and then the second integral is smallest, so it is better to hide Y_1. When x is close to 2, the opposite is true and it is better to hide Y_2. In applications, changing the conditioning as a function of x adds complications and is normally not necessary. Using the same conditioning for all x, even when not optimal, is usually preferable because of its simplicity.

Example 2. The following small example provides further insight into the choice of $\mathcal{G}$. Suppose X is the sum of two independent uniform random variables: $X = Y_1 + Y_2$ where $Y_1 \sim \mathcal{U}(0,1)$ and $Y_2 \sim \mathcal{U}(0,\epsilon)$ where $0 < \epsilon < 1$. The exact density of X here is $f(x) = x/\epsilon$ for $0 \le x \le \epsilon$, $f(x) = 1$ for $\epsilon \le x \le 1$, and $f(x) = (1 + \epsilon - x)/\epsilon$ for $1 \le x \le 1 + \epsilon$. Figure 1 illustrates this density.

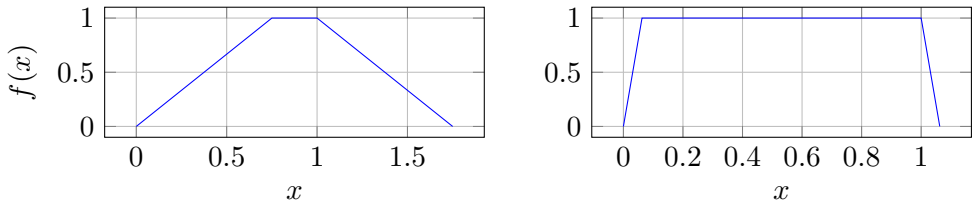

Figure 1 Exact density of X for the model in Example 2 with $\epsilon = 3/4$ (left) and $\epsilon = 1/16$ (right).

With $\mathcal{G} = \mathcal{G}_{-1}$, we have $F(x \mid \mathcal{G}_{-1}) = \mathbb{P}[X \le x \mid Y_2] = \mathbb{P}[Y_1 \le x - Y_2 \mid Y_2] = x - Y_2$ and the density estimator is $f(x \mid \mathcal{G}_{-1}) = 1$ for $Y_2 \le x \le 1 + Y_2$, and 0 elsewhere. If $\mathcal{G} = \mathcal{G}_{-2}$ instead, then $F(x \mid \mathcal{G}_{-2}) = \mathbb{P}[Y_2 \le x - Y_1 \mid Y_1] = (x - Y_1)/\epsilon$ and the density estimator is $f(x \mid \mathcal{G}_{-2}) = 1/\epsilon$ for $Y_1 \le x \le \epsilon + Y_1$, and 0 elsewhere. In both cases, Assumption 1 holds and the density estimator with one sample is a uniform density, but the second one is over a narrow interval if ϵ is small. When ϵ is small, $\mathcal{G} = \mathcal{G}_{-2}$ gives a density estimator $\hat{f}_{\mathrm{cde},n}$ which is a sum of high narrow peaks and has much larger variance. For this simple example, we can also derive exact formulas for the IV of the CDE under MC. For $\mathcal{G} = \mathcal{G}_{-1}$, $f(x \mid \mathcal{G}_{-1}) = \mathbb{I}[Y_2 \le x \le 1 + Y_2]$ is a Bernoulli random variable with mean $\mathbb{P}[x - 1 \le Y_2 \le x] = f(x)$, so its variance is $f(x)(1 - f(x))$. Integrating this over $[0, 1 + \epsilon]$ gives IV $= \epsilon/3$ for one

sample. For a sample of size n, this gives $\mathrm{IV} = \epsilon/(3n)$. For $\mathcal{G} = \mathcal{G}_{-2}$, $f(x \mid \mathcal{G}_{-2}) = \mathbb{I}[Y_1 \le x \le \epsilon + Y_1]/\epsilon$ has also mean $f(x)$, but its variance is $\epsilon^{-1} f(x)(1 - \epsilon f(x))$, which is much larger than $f(x)(1-f(x))$ when ϵ is small. Integrating over $[0,\, 1+\epsilon]$ gives $\mathrm{IV} = 1/\epsilon - 1 + \epsilon/3$ for one sample, which is also much larger than $\epsilon/3$ when ϵ is small. The take-away: It is usually better to condition on lower-variance information and hide variables having a large variance contribution.

Example 3. In this example, we illustrate how Assumption 1 can be verified. Let X be the sum of two independent normal random variables, $X = Y_1 + Y_2$, where $Y_1 \sim \mathcal{N}(0, \sigma_1^2)$, $Y_2 \sim \mathcal{N}(0, \sigma_2^2)$, and $\sigma_1^2 + \sigma_2^2 = 1$, so $X \sim \mathcal{N}(0,1)$. Let Φ and ϕ denote the cdf and density of the standard normal distribution. With $\mathcal{G} = \mathcal{G}_{-2}$, we have $F(x \mid \mathcal{G}_{-2}) = \mathbb{P}[Y_2 \le x - Y_1] = \Phi((x - Y_1)/\sigma_2)$ and the CDE is $f(x \mid \mathcal{G}_{-2}) = \phi((x - Y_1)/\sigma_2)/\sigma_2$. Assumption 1 holds with $\Gamma = \phi(0)/\sigma_2$ and $K_\gamma = \Gamma^2$, so this estimator is unbiased for $f(x) = \phi(x)$. Its variance is

$$\begin{aligned}
\mathrm{Var}[\phi((x - Y_1)/\sigma_2)/\sigma_2] &= \mathbb{E}[\exp[-(x - Y_1)^2/\sigma_2^2]/(2\pi\sigma_2^2)] - \phi^2(x) \\
&= \frac{1}{\sigma_2^2\sqrt{2\pi}} \mathbb{E}[\phi(\sqrt{2}(x - Y_1)/\sigma_2)] - \phi^2(x) \\
&= \frac{1}{\sigma_2\sqrt{2\pi(1+\sigma_1^2)}} \phi\left(\sqrt{2}x/\sqrt{1+\sigma_1^2}\right) - \phi^2(x). \qquad (6)
\end{aligned}$$

Example 4. If X is the min or max of two or more continuous random variables, then in general $F(\cdot \mid \mathcal{G}_{-k})$ is not continuous, so if we hide only one variable, Assumption 1 does not hold. Indeed, if $X = \max(Y_1, Y_2)$ where Y_1 and Y_2 are independent, with $\mathcal{G} = \mathcal{G}_{-2}$ (we hide Y_2), we have

$$\mathbb{P}[X \le x \mid Y_1 = y] = \begin{cases} \mathbb{P}[Y_2 \le x \mid Y_1 = y] = F_2(x) & \text{if } x \ge y; \\ 0 & \text{if } x < y. \end{cases}$$

If $F_2(y) > 0$, this function is discontinuous at $x = y$. The same holds for the maximum of more than two variables. One way to handle this is to generate all the variables, then hide the maximum and compute its conditional density given the other ones. Without loss of generality, suppose Y_1 is the maximum and $Y_2 = y_2$ the second largest. Then the CDE of the max is $f(x \mid \mathcal{G}) = f_1(x \mid Y_1 > y_2)$. Note that for independent random variables whose cdf's and densities have an analytical form, the cdf and density of the max can often be computed analytically. See Section 4.3 for more on this. A very similar story holds if we replace the max by the min.

Example 5. Suppose $X = Z \cdot C$ where $Z \sim N(0,1)$ and C is continuous with support over $(0, \infty)$. We can hide Z and generate $X \sim N(0, C^2)$ conditional on C, or do the opposite. Which one is best depends on the distribution of C. Here we have $\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}[X \mid C]] = \mathbb{E}[C^2]$ while $\mathrm{Var}[\mathbb{E}[X \mid C]] = 0$. So the usual variance decomposition tells us nothing about what to hide. This illustrates the fact that there is rarely a simple rule to find the optimal $\mathcal{G}$.

2.4. Convex combination of conditional density estimators

When there are many possible choices of $\mathcal{G}$ for a given problem, one can select more than one and take a convex linear combination of the corresponding CDEs as the final density estimator. This idea is well known for general mean estimators (Bratley et al. 1987). More specifically, suppose $\hat{f}_{0,n}, \ldots, \hat{f}_{q,n}$ are $q+1$ distinct unbiased density estimators. Typically, these estimators are dependent and based on the same simulations. They could be all CDEs based on different choices of $\mathcal{G}$ (so they will not hide the same information), but there could be non-CDEs as well. A convex combination can take the form

$$\hat{f}_n(x) = \beta_0 \hat{f}_{0,n}(x) + \cdots + \beta_q \hat{f}_{q,n}(x) = \hat{f}_{0,n}(x) - \sum_{\ell=1}^{q} \beta_\ell (\hat{f}_{0,n}(x) - \hat{f}_{\ell,n}(x)) \tag{7}$$

for all $x \in \mathbb{R}$, where $\beta_0 + \cdots + \beta_q = 1$. This is equivalent to choosing $\hat{f}_{0,n}(x)$ as the main estimator, and taking the q differences $\hat{f}_{0,n}(x) - \hat{f}_{\ell,n}(x)$ as control variables (Bratley et al. 1987), Problem 2.3.9. With this interpretation, the optimal coefficients β_ℓ can be estimated via standard control variate theory (Asmussen and Glynn 2007) by trying to minimize the IV of $\hat{f}_n(x)$ w.r.t. the β_ℓ's. More precisely, if we denote $\mathrm{IV}_\ell = \mathrm{IV}(\hat{f}_{\ell,n}(x))$ and $\mathrm{IC}_{\ell,k} = \int_a^b \mathrm{Cov}[\hat{f}_{\ell,n}(x), \hat{f}_{k,n}(x)]\mathrm{d}x$, we obtain

$$\mathrm{IV} = \mathrm{IV}\left(\hat{f}_n(x)\right) = \sum_{\ell=0}^{q} \beta_\ell^2 \mathrm{IV}_\ell + 2 \sum_{0 \le \ell < k \le q} \beta_\ell \beta_k \mathrm{IC}_{\ell,k}.$$

Given the IV_ℓ's and $\mathrm{IC}_{\ell,k}$'s (or good estimates of them), this IV is a quadratic function of the β_ℓ's, which can be minimized exactly as in standard least-squares linear regression. That is, the optimal coefficients β_j obey the standard linear regression formula. Estimating the density and coefficients from the same data yields biased but consistent density estimators, and the bias is rarely a problem. We followed this approach for some of the examples in Section 4. Cui et al. (2020) obtained an equivalent formula from a slightly different but equivalent reasoning.

Given that the best choice of $\mathcal{G}$ generally depends on x, one may also adopt a more refined approach which allows the coefficients β_j to depend on x:

$$\hat{f}_n(x) = \beta_0(x) \hat{f}_{0,n}(x) + \cdots + \beta_q(x) \hat{f}_{q,n}(x) = \hat{f}_{0,n}(x) - \sum_{\ell=1}^{q} \beta_\ell(x) (\hat{f}_{0,n}(x) - \hat{f}_{\ell,n}(x)), \tag{8}$$

where $\beta_0(x) + \cdots + \beta_q(x) = 1$ for all $x \in \mathbb{R}$. The optimal coefficients can be estimated by standard control variate theory at selected values of x, then for each $\ell \ge 1$, one can fit a smoothing spline to these estimated values, by least squares. This provides estimated optimal coefficients that are smooth functions of x, which can be used to obtain a final CDE. This type of strategy was used in L'Ecuyer and Buist (2008) to estimate varying control variate coefficients. The additional flexibility can improve the variance reduction in some situations.

2.5. A GLR density estimator (GLRDE)

The *generalized likelihood ratio* (GLR) method, originally developed by Peng et al. (2018) to estimate the derivative of an expectation with respect to some model parameter, can be adapted to density estimation, as shown in Peng et al. (2020). We summarize briefly here how this method estimates the density $f(x)$ in our general setting, so we can apply it in our examples and make numerical comparisons. The assumptions stated below differ slightly from those in Peng et al. (2020). In particular, here we do not have a parameter θ, the conditions on the estimator are required only in the area where $X \le x$, and we add a condition to ensure finite variance. As in Section 2.3, we assume here that $X = h(\mathbf{Y}) = h(Y_1, \ldots, Y_d)$ where $Y_1, \ldots, Y_d$ are independent continuous random variables, and Y_j has cdf F_j and density f_j. Let $P(x) = \{\mathbf{y} \in \mathbb{R}^d : h(\mathbf{y}) \le x\}$. For $j = 1, \ldots, d$, let $h_j(\mathbf{y}) := \partial h(\mathbf{y})/\partial y_j$, $h_{jj}(\mathbf{y}) := \partial^2 h(\mathbf{y})/\partial y_j^2$, and

$$\Psi_j(\mathbf{y}) = \frac{\partial \log f_j(y_j)/\partial y_j - h_{jj}(\mathbf{y})/h_j(\mathbf{y})}{h_j(\mathbf{y})}. \tag{9}$$

Assumption 2. *The Lebesgue measure of $h^{-1}((x-\epsilon, x+\epsilon))$ in $\mathbb{R}^d$ goes to 0 when $\epsilon \to 0$ (this means essentially that the density is bounded around x).*

Assumption 3. *The set $P(x)$ is measurable, the functions h_j, h_{jj}, and Ψ_j are well defined over it, and $\mathbb{E}[\mathbb{I}[X \le x] \cdot \Psi_j^2(\mathbf{Y})] < \infty$.*

Proposition 2. *Under Assumptions 2 and 3, the GLRDE $\mathbb{I}[X \le x] \cdot \Psi_j(\mathbf{Y})$ is an unbiased and finite-variance estimator of the density $f(x)$ at x.*

For the proof of Proposition 2 and additional details, see Peng et al. (2020).

3. Combining RQMC with the CMC density estimator

We now discuss how RQMC can be used with the CDE, and under what conditions it can provide a convergence rate faster than $\mathcal{O}(n^{-1})$ for the IV of the resulting unbiased estimator. For this, we first recall some basic facts about QMC and RQMC. More detailed coverages can be found in Niederreiter (1992), Dick and Pillichshammer (2010), and L'Ecuyer (2009, 2018), for example.

For a function $g : [0,1)^s \to \mathbb{R}$, the integration error by the average over a point set $P_n = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subset [0,1]^s$ is defined by

$$E_n = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{u}_i) - \int_{[0,1]^s} g(\mathbf{u})\mathrm{d}\mathbf{u}. \tag{10}$$

Classical QMC theory bounds this error as follows. Let $\mathfrak{v} \subseteq \mathcal{S} := \{1, \ldots, s\}$ denote an arbitrary subset of coordinates. For any point $\mathbf{u} = (u_1, \ldots, u_s) \in [0,1]^s$, $\mathbf{u}_\mathfrak{v}$ denotes the projection of $\mathbf{u}$ on the coordinates in $\mathfrak{v}$ and $(\mathbf{u}_\mathfrak{v}, \mathbf{1})$ is the point $\mathbf{u}$ in which u_j is replaced by 1 for each $j \notin \mathfrak{v}$. Let

$g_{\mathfrak{v}} := \partial^{|\mathfrak{v}|} g / \partial \mathbf{u}_{\mathfrak{v}}$ denote the partial derivative of g with respect to all the coordinates in $\mathfrak{v}$. When $g_{\mathfrak{v}}$ exists and is continuous for $\mathfrak{v} = \mathcal{S}$ (i.e., for all $\mathfrak{v} \subseteq \mathcal{S}$), the *Hardy-Krause (HK) variation* of g can be written as

$$V_{\mathrm{HK}}(g) = \sum_{\emptyset \neq \mathfrak{v} \subseteq \mathcal{S}} \int_{[0,1]^{|\mathfrak{v}|}} |g_{\mathfrak{v}}(\mathbf{u}_{\mathfrak{v}}, \mathbf{1})| \, \mathrm{d}\mathbf{u}_{\mathfrak{v}}. \tag{11}$$

On the other hand, the *star-discrepancy* of P_n is

$$D^*(P_n) = \sup_{\mathbf{u} \in [0,1]^s} \left| \frac{|P_n \cap [\mathbf{0}, \mathbf{u})|}{n} - \mathrm{vol}[\mathbf{0}, \mathbf{u}) \right|$$

where $\mathrm{vol}[\mathbf{0}, \mathbf{u})$ is the volume of the box $[\mathbf{0}, \mathbf{u})$. The classical *Koksma-Hlawka (KH) inequality* bounds the absolute error by the product of these two quantities, one that involves only the function g and the other that involves only the point set P_n:

$$|E_n| \leq V_{\mathrm{HK}}(g) \cdot D^*(P_n). \tag{12}$$

There are explicit construction methods (e.g., digital nets, lattice rules, and polynomial lattice rules) of deterministic point sets P_n for which $D^*(P_n) = \mathcal{O}((\log n)^{s-1}/n) = \mathcal{O}(n^{-1+\epsilon})$ for all $\epsilon > 0$. This means that functions g for which $V_{\mathrm{HK}}(g) < \infty$ can be integrated by QMC with a worst-case error that satisfies $|E_n| = \mathcal{O}(n^{-1+\epsilon})$. There are also known methods to randomize these point sets P_n in a way that each randomized point $\mathbf{u}_i$ has the uniform distribution over $[0,1)^s$, so $\mathbb{E}[E_n] = 0$, and the $\mathcal{O}(n^{-1+\epsilon})$ discrepancy bound is preserved, which gives

$$\mathrm{Var}[E_n] = \mathbb{E}[E_n^2] = \mathcal{O}(n^{-2+\epsilon}). \tag{13}$$

The classical definitions of variation and discrepancy given above are only one pair among an infinite collection of possibilities. There are other versions of (12), with different definitions of the discrepancy and the variation, such that there are known point set constructions for which the discrepancy converges as $\mathcal{O}(n^{-\alpha+\epsilon})$ for $\alpha > 1$, but the conditions on g to have finite variation are more restrictive (more smoothness is required) (Dick and Pillichshammer 2010).

From a practical viewpoint, getting a good estimate or an upper bound on the variation of g that can be useful to bound the RQMC variance is a notoriously difficult problem. Even just showing that the variation is finite is not always easy. However, finite variation is not a necessary condition. In many realistic applications in which variation is known to be infinite, RQMC can nevertheless reduce the variance by a large factor (L'Ecuyer 2009, L'Ecuyer and Munger 2012, He and Wang 2015). The appropriate explanation for this depends on the application. In many cases, part of the explanation is that the integrand g can be written as a sum of orthogonal functions (as in an ANOVA decomposition) and a set of terms in that sum have a large variance contribution and are smooth low-dimensional functions for which RQMC is very effective (L'Ecuyer and Lemieux

2000, L'Ecuyer 2009, Lemieux 2009). Making such a decomposition and finding the important terms is difficult for realistic problems, but to apply RQMC in practice, this is not needed. The usual approach in applications is to try it and compare the RQMC variance with the MC variance empirically. We will do that in Section 4.

To combine the CDE with RQMC, we must be able to write $F(x \mid \mathcal{G}) = \tilde{g}(x, \mathbf{u})$ and $f(x \mid \mathcal{G}) = \tilde{g}'(x, \mathbf{u}) = \mathrm{d}\tilde{g}(x, \mathbf{u})/\mathrm{d}x$ for some function $\tilde{g} : [a, b] \times [0, 1)^s$. The function $\tilde{g}'(x, \cdot)$ will act as g in (10). The combined CDE+RQMC estimator $\hat{f}_{\text{cde-rqmc},n}(x)$ will be defined by

$$\hat{f}_{\text{cde-rqmc},n}(x) = \frac{1}{n} \sum_{i=1}^{n} \tilde{g}'(x, \mathbf{U}_i), \tag{14}$$

which is the RQMC version of (2). To estimate the RQMC variance, we can perform n_r independent randomizations to obtain n_r independent realizations of $\hat{f}_{\text{cde-rqmc},n}$ in (14) with RQMC, and compute the empirical IV. By putting together the previous results, we obtain:

PROPOSITION 3. *If* $\sup_{x\in[a,b]} V_{\text{HK}}(\tilde{g}'(x, \cdot)) < \infty$*, then with RQMC points sets* P_n *with* $D^*(P_n) = \mathcal{O}((\log n)^{s-1}/n)$*, for any* $\epsilon > 0$*, we have* $\sup_{x\in[a,b]} \mathrm{Var}[\hat{f}_{\text{cde-rqmc},n}(x)] = \mathcal{O}(n^{-2+\epsilon})$*, so the MISE of the CDE+RQMC estimator converges as* $\mathcal{O}(n^{-2+\epsilon})$*.*

Although this is rarely done in practice, it is instructive to see how the HK variation of $\tilde{g}'(x, \cdot)$ can be bounded in our CDE setting, so that Proposition 3 applies. For this, we need to show that the integral of the partial derivative of $\tilde{g}'(x, \mathbf{u})$ with respect to each subset of coordinates of $\mathbf{u}$ is finite. In Section A of the Supplement, we do it for Examples 1 to 3. When the variation is unbounded, RQMC may still reduce the IV, but there is no guarantee. The GLRDE in Proposition 2 is typically discontinuous because of the indicator function, and therefore its HK variation is usually infinite.

4. Examples and numerical experiments

We now examine larger instructive examples for which we show how to construct a CDE, summarize the results of numerical experiments with the CDE and CDE+RQMC, and make comparisons with the GLRDE and KDE, with MC and RQMC. Section 4.1 gives the experimental framework used for all the numerical experiments. In Section 4.2, we use a three-dimensional real-life example to provide further insight on the choice of conditioning and make comparisons between methods. In Section 4.3, we estimate the density of the length X of the longest path between the source and destination in a stochastic network. This length may represent the total time to execute a project, the arrival time of a train at a given station, etc. The length of the shortest path can be handled in a similar way. In Section 4.4, X is the waiting time of a customer in a queuing system. We consider a single queue in the example, but a similar conditioning would apply for larger queueing systems as well. In Section 4.6, X is the payoff of a financial option. We show that by using a

clever conditioning with CDE+RQMC, the MISE can be reduced by huge factors. More examples are given in the Online Supplement. In all these examples, estimating the density of X has high practical relevance. Larger problem instances can also be handled with the same methods.

4.1. Experimental setting

Since the CDE is unbiased, we measure its performance by the IV, which equals the MISE in this case. To approximate the IV estimator (3) for a given n, we first take a stratified sample $e_1, \ldots, e_{n_e}$ of n_e *evaluation points* at which the empirical variance will be computed. We sample e_j uniformly in $[a+(j-1)(b-a)/n_e,\, a+j(b-a)/n_e)$ for $j = 1, \ldots, n_e$. Then we use the unbiased IV estimator

$$\widehat{\mathrm{IV}} = \frac{(b-a)}{n_e} \sum_{j=1}^{n_e} \widehat{\mathrm{Var}}[\hat{f}_n(e_j)],$$

where $\widehat{\mathrm{Var}}[\hat{f}_n(e_j)]$ is the empirical variance of the CDE at e_j, obtained as follows. We repeat the following n_r times, independently: Generate n observations of X from the density f with the given method (MC or RQMC), and compute the CDE at each evaluation point e_j. We then compute $\widehat{\mathrm{Var}}[\hat{f}_n(e_j)]$ as the empirical variance of the n_r density estimates at e_j, for each j. In all our examples, we used $n_r = 100$ and $n_e = 128$.

To estimate the convergence rate of the IV as a function of n with the different methods, we fit a model of the form $\mathrm{IV} \approx K n^{-\nu}$. For the CDE with independent points (no RQMC), this model holds exactly with $\nu = 1$. We hope to observe $\nu > 1$ with RQMC. The parameters K and ν are estimated by linear regression in log-log scale, i.e., by fitting the model $\log \mathrm{IV} \approx \log K - \nu \log n$ to data. Since n is always taken as a power of 2, we report the logarithms in base 2. We estimated the IV for $n = 2^{14}, \ldots, 2^{19}$ (6 values) to fit the regression model. We also report the observed $-\log_2 \mathrm{IV}$ for $n = 2^{19}$ and use $e19$ as a shorthand for this value in the tables. We use exactly the same procedure for the GLRDE. For the KDE, these values are for the MISE instead of the IV. In all cases, we used a normal kernel and a bandwidth h selected by the methodology described in Ben Abdellah et al. (2021). For some examples, we tried CDEs based on different choices of $\mathcal{G}$ and a convex combination as in Section 2.4.

We report results with the following types of point sets:

(1) independent points (MC);

(2) a randomly-shifted lattice rule (Lat+s);

(3) a randomly-shifted lattice rule with a baker's transformation (Lat+s+b);

(4) Sobol' points with a left random matrix scramble and random digital shift (Sob+LMS).

The short names in parentheses are used in the plots and tables. For the definitions and properties of these RQMC point sets, see L'Ecuyer and Lemieux (2000), Owen (2003), L'Ecuyer (2009, 2018). They are implemented in SSJ (L'Ecuyer 2016), which we used for our experiments. The parameters

of the lattice rules were found with the Lattice Builder software of L'Ecuyer and Munger (2016), using a fast-CBC construction method with the $\mathcal{P}_2$ criterion and order dependent weights $\gamma_{\mathfrak{v}} = \rho^{|\mathfrak{v}|}$, with ρ ranging from 0.05 to 0.8, depending on the example (a larger ρ was used when the dimension s was smaller). The baker's transformation sometimes improves the convergence rate by making the integrand periodic (Hickernell 2002), but it can also increase the variation of the integrand, so its impact on the variance can go either way.

4.2. Displacement of a cantilever beam

We consider the following (real-life) model for the displacement X of a cantilever beam with horizontal and vertical loads, taken from Bingham (2017):

$$X = h(Y_1, Y_2, Y_3) = \frac{4\ell^3}{Y_1 wt}\sqrt{\frac{Y_2^2}{w^4} + \frac{Y_3^2}{t^4}} \tag{15}$$

in which $\ell = 100$, $w = 4$ and $t = 2$ are constants (in inches), while Y_1 (Young's modulus), Y_2 (the horizontal load), and Y_3 (the vertical load), are independent normal random variables, $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, i.e., normal with mean μ_j and variance σ_j^2. The parameter values are $\mu_1 = 2.9 \times 10^7$, $\sigma_1 = 1.45 \times 10^6$, $\mu_2 = 500$, $\sigma_2 = 100$, $\mu_3 = 1000$, $\sigma_3 = 100$. We will denote $\kappa = 4\ell^3/(wt) = 5 \times 10^5$. The goal is to estimate the density of X over the interval [3.1707, 5.6675], which covers about 99% of the density (it clips 0.5% on each side). It is possible to have $X < 0$ in this model, but the probability is $\mathbb{P}[Y_1 < 0] = \Phi(-20) = 2.8 \times 10^{-89}$, which is negligible, so we can assume that $Y_1 > 0$. This example fits the framework of Section 2.3, with $d = 3$. We can hide any of the three random variables for the conditioning, and we will examine each case.

Conditioning on $\mathcal{G}_{-1}$ means hiding Y_1. We have

$$X = \frac{\kappa}{Y_1}\sqrt{\frac{Y_2^2}{w^4} + \frac{Y_3^2}{t^4}} \le x \quad \text{if and only if} \quad Y_1 \ge \frac{\kappa}{x}\sqrt{\frac{Y_2^2}{w^4} + \frac{Y_3^2}{t^4}} \stackrel{\text{def}}{=} W_1(x).$$

Note that $W_1(x) > 0$ if and only if $x > 0$. For $x > 0$,

$$F(x \mid \mathcal{G}_{-1}) = \mathbb{P}[Y_1 \ge W_1(x) \mid W_1(x)] = 1 - \Phi((W_1(x) - \mu_1)/\sigma_1)$$

which is continuous and differentiable in x, and

$$f(x \mid \mathcal{G}_{-1}) = -\phi((W_1(x) - \mu_1)/\sigma_1)W_1'(x)/\sigma_1 = \phi((W_1(x) - \mu_1)/\sigma_1)W_1(x)/(x\sigma_1).$$

If we condition on $\mathcal{G}_{-2}$ instead, i.e., we hide Y_2, we have $X \le x$ if and only if

$$Y_2^2 \le w^4\left((xY_1/\kappa)^2 - Y_3^2/t^4\right) \stackrel{\text{def}}{=} W_2(x).$$

If $W_2(x) \le 0$, then $f(x \mid \mathcal{G}_{-2}) = F(x \mid \mathcal{G}_{-2}) = \mathbb{P}[X \le x \mid W_2(x)] = 0$. For $W_2(x) > 0$, we have

$$\begin{aligned} F(x \mid \mathcal{G}_{-2}) &= \mathbb{P}[X \le x \mid W_2(x)] = \mathbb{P}\left[-\sqrt{W_2(x)} \le Y_2 \le \sqrt{W_2(x)} \mid W_2(x)\right] \\ &= \Phi((\sqrt{W_2(x)} - \mu_2)/\sigma_2) - \Phi(-(\sqrt{W_2(x)} + \mu_2)/\sigma_2), \end{aligned}$$

which is again continuous and differentiable in x, and

$$f(x \mid \mathcal{G}_{-2}) = \frac{\phi((\sqrt{W_2(x)} - \mu_2)/\sigma_2) + \phi(-(\sqrt{W_2(x)} + \mu_2)/\sigma_2)}{(\sigma_2\sqrt{W_2(x)})/(w^4 x (Y_1/\kappa)^2)} > 0.$$

If we condition on $\mathcal{G}_{-3}$, the analysis is the same as for $\mathcal{G}_{-2}$, by symmetry, and we get

$$f(x \mid \mathcal{G}_{-3}) = \frac{\phi((\sqrt{W_3(x)} - \mu_3)/\sigma_3) + \phi(-(\sqrt{W_3(x)} + \mu_3)/\sigma_3)}{(\sigma_3\sqrt{W_3(x)})/(t^4 x (Y_1/\kappa)^2)} > 0$$

for $W_3(x) > 0$, where $W_3(x)$ is defined in a similar way as $W_2(x)$.

For the GLRDE, we write $h(\mathbf{Y}) = (\kappa/Y_1)S^{1/2}$ where $S = Y_2^2/w^4 + Y_3^2/t^4$, and denote $Z_j = (Y_j - \mu_j)/\sigma_j^2 = -\partial \log f_j(Y_j)/\partial Y_j$ for $j = 1, 2, 3$. With this notation, we obtain $h_1(\mathbf{Y}) = -h(\mathbf{Y})/Y_1$, $h_{11}(\mathbf{Y}) = 2h(\mathbf{Y})/Y_1^2$, $h_2(\mathbf{Y}) = (\kappa/Y_1)(Y_2/w^4)S^{-1/2} = h(\mathbf{Y})Y_2/(Sw^4)$, $h_{22}(\mathbf{Y}) = (\kappa/(Y_1 w^4))(S^{-1/2} - S^{-3/2}Y_2^2/w^4)$, $h_3(\mathbf{Y}) = (\kappa/Y_1)(Y_3/t^4)S^{-1/2} = h(\mathbf{Y})Y_3/(St^4)$, $h_{33}(\mathbf{Y}) = (\kappa/(Y_1 t^4))(S^{-1/2} - S^{-3/2}Y_3^2/t^4)$. With a little calculation, this gives

$$\Psi_1(\mathbf{Y}) = \frac{Y_1 Z_1 - 2}{h(\mathbf{Y})}, \qquad \Psi_2(\mathbf{Y}) = -\frac{Y_2 Z_2 S + Y_3^2/t^4}{h(\mathbf{Y})\, Y_2^2/w^4}, \qquad \Psi_3(\mathbf{Y}) = -\frac{Y_3 Z_3 S + Y_2^2/w^4}{h(\mathbf{Y})\, Y_3^2/t^4}.$$

In addition to testing the individual estimators derived above, we also tested convex combinations of the three CDEs and of the three GLRDEs, as explained in Section 2.4, with coefficients β_ℓ that do not depend on x.

Table 1 Values of $\hat{\nu}$ and e19 with a CDE for each choice of $\mathcal{G}_{-k}$ and for the best convex combination (CDE-c), for the GLRDE with each Ψ_j and for the best convex combination (GLRDE-c), and for the KDE, for the cantilever beam model.

	$\mathcal{G}_{-1}$	$\mathcal{G}_{-2}$	$\mathcal{G}_{-3}$	CDE-c	Ψ_1	Ψ_2	Ψ_3	GLRDE-c	KDE
	e19								
MC	19.3	14.5	22.8	22.5	14.1	4.5	15.8	16.3	15.8
Lat+s	39.8	25.2	41.6	41.9	23.4	-2.5	26.4	26.5	21.9
Lat+s+b	44.5	23.7	46.8	47.0	23.3	5.7	24.7	25.1	21.0
Sob+LMS	44.0	23.6	45.7	46.1	23.4	2.8	25.5	25.9	21.5
	$\hat{\nu}$								
MC	0.97	0.98	0.99	0.98	1.02	0.55	0.94	0.95	0.76
Lat+s	1.99	1.95	2.06	2.04	1.38	—	1.51	1.52	1.03
Lat+s+b	2.24	2.08	2.27	2.25	1.37	—	1.24	1.25	0.93
Sob+LMS	2.21	2.03	2.21	2.21	1.32	—	1.31	1.32	0.97

Table 1 summarizes the results. The MISE is about 2^{-47} for the best CDE+RQMC compared with $2^{-15.8}$ for the usual KDE+MC, a gain by a factor of over $2^{31} \approx 2$ billions. This is probably much better accuracy than required in practice for this particular application. With RQMC, the convergence rate $\hat{\nu}$ is around 2 in all cases with the CDE methods, and much less for GLRDE and KDE. The GLRDE using Ψ_2 behaves very badly (the estimates $\hat{\nu}$ with RQMC are meaningless),

but with Ψ_1 and Ψ_3 (the best choice), it performs better that the KDE. Note that the denominator of Ψ_2 takes much smaller values on average than that of Ψ_3, and this can explain its larger variance.

For the CDE with lattice rules, the baker's transformation helps significantly for the CDE. Conditioning on $\mathcal{G}_{-2}$ does not give as much reduction as for the other choices. To provide visual insight, Figure 2 shows plots of five realizations of the conditional density for $\mathcal{G}_{-1}$, $\mathcal{G}_{-2}$, and $\mathcal{G}_{-3}$. The realizations of $f(\cdot \mid \mathcal{G}_{-2})$ have high narrow peaks, which explains the larger variance. The average of the five realizations is shown in red and the true density in black. In Figure 3, we zoom in on part of the estimated densities to show the difference between MC and RQMC. In each panel one can see the CDE using MC (in red), RQMC (in green), and the "true density" (black, dashed) estimated with RQMC using a large number of samples. We have $\mathcal{G}_{-1}$ with $n = 2^{10}$ on the left and $\mathcal{G}_{-2}$ with $n = 2^{16}$ on the right. In both cases, the RQMC estimate is closer to the true density, and on the left it oscillates less. If we repeat this experiment several times, the red curve would vary much more than the green one across the realizations.

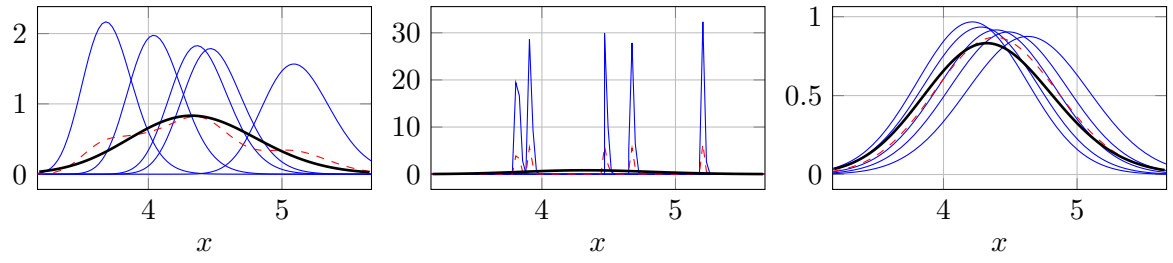

Figure 2 **Five realizations of the density conditional on $\mathcal{G}_{-k}$ (blue), their average (red), and the true density (thick black) for $k = 1$ (left), $k = 2$ (middle), and $k = 3$ (right), for the cantilever example.**

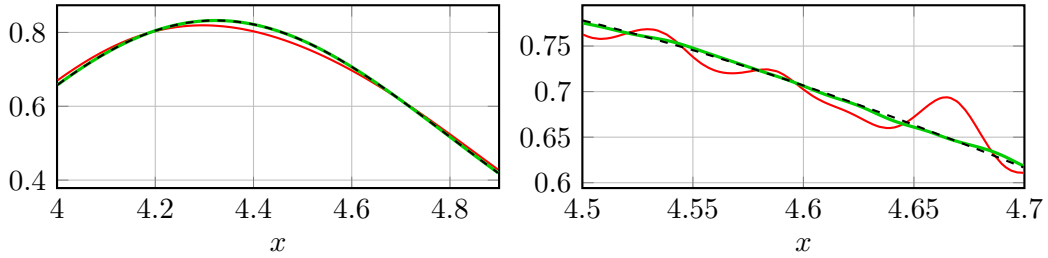

Figure 3 **The CDE under MC (red), under RQMC (green, very close to the dashed line) and the true density (black, dashed) for $\mathcal{G}_{-1}$ with $n = 2^{10}$ (left) and for $\mathcal{G}_{-2}$ with $n = 2^{16}$ (right), for the cantilever example.**

4.3. A stochastic activity network

In this example, the conditioning for the CDE must hide more than one random variable. We consider an acyclic directed graph $G = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is a finite set of nodes and $\mathcal{A} = \{a_j = (\alpha_j, \beta_j), j = 1, \ldots, d\}$ a finite set of arcs (directed links) where a_j goes from α_j to β_j. There is a source node having only outcoming arcs, a sink node having only incoming arcs, and each arc belongs to at least one path going from the source to the sink. There can be at most one arc for each pair (α_j, β_j) (no parallel arcs). Each arc j has random length Y_j. These Y_j are assumed independent with continuous cdf's F_j, density f_j, and can be generated by inversion: $Y_j = F_j^{-1}(U_j)$ where $U_j \sim U(0,1)$. The length of the longest path from the source to the sink is a random variable X and the goal is to estimate the density of X.

This general model has several applications. The arcs a_j may represent activities having random durations and the graph represents precedence relationships between all activities of a project. Activity a_j cannot start before all activities j' with $\beta_{j'} = \alpha_j$ are completed. Then X represents the duration of the project if all activities are started as soon as allowed. This type of *stochastic activity network* (SAN) is widely used in project management for all types of projects (e.g., construction, software, etc.), communication, transportation, etc. For example, the graph may represent a large railway network in which each activity corresponds to a train stopping at a station, or a train covering a given segment of its route, or a minimal spacing between trains, etc. Precedence relationships are needed because railways are shared, there are ordering and distancing rules between trains, passengers have connections between trains, trains are merged or split at certain points, etc. The travel time of one passenger in this network turns out to be the length X of the longest path in a subnetwork whose source and sink are the origin and destination of this passenger.

For our numerical experiments, we use a small example from Avramidis and Wilson (1996, 1998), who showed how to use CMC to estimate $\mathbb{E}[X]$ and some quantiles of the distribution of X. L'Ecuyer and Lemieux (2000) and L'Ecuyer and Munger (2012) used this same example to test the combination of CMC with RQMC to estimate $\mathbb{E}[X]$. The network is depicted in Fig. 4 and the cdf's F_j are given in Avramidis and Wilson (1996). Much larger networks can be handled in the same way. We will estimate the density of X over $[a, b] = [22, 106.24]$, which covers about 95% of the density.

Here, X is defined as the maximum length over several paths, and if we hide only a single random variable Y_j to implement the CDE, we run into the same problem as in Example 4: Assumption 1 does not hold, because $F(\cdot \mid \mathcal{G})$ has a jump. This means that we must hide more information (condition on less). Following Avramidis and Wilson (1996, 1998), we select a *uniformly directed cut* $\mathcal{L}$, which is a set of activities such that each path from the source to the sink contains exactly

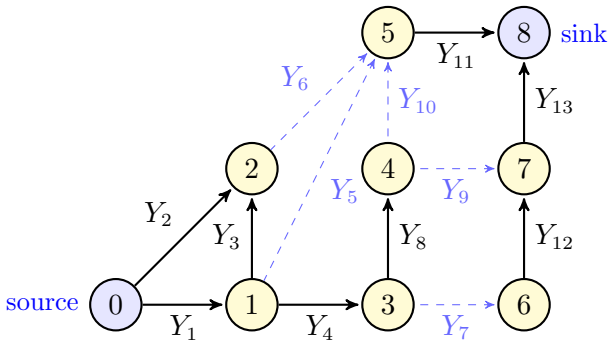

Figure 4 **A stochastic activity network, with the cut $\mathcal{L} = \{5, 6, 7, 9, 10\}$ shown in dashed light blue**

one activity from $\mathcal{L}$, and let $\mathcal{G}$ represent $\{Y_j, j \notin \mathcal{L}\}$. In Figure 4, $\{1, 2\}$, $\{11, 13\}$, $\{5, 6, 7, 9, 10\}$, and $\{2, 3, 5, 8, 9, 13\}$, are all valid choices of $\mathcal{L}$. The corresponding conditional cdf is

$$F(x \mid \mathcal{G}) = \mathbb{P}\left[X \le x \mid \{Y_j : j \notin \mathcal{L}\}\right] = \prod_{j \in \mathcal{L}} \mathbb{P}[Y_j \le x - P_j] = \prod_{j \in \mathcal{L}} F_j(x - P_j) \tag{16}$$

where P_j is the length of the longest path that goes through arc j when we exclude Y_j from that length. The conditional density is

$$f(x \mid \mathcal{G}) = \frac{\mathrm{d}}{\mathrm{d}x} F(x \mid \mathcal{G}) = \sum_{j \in \mathcal{L}} f_j(x - P_j) \prod_{l \in \mathcal{L}, l \neq j} F_l(x - P_j).$$

Under this conditioning, if the Y_j's are continuous variables with bounded variance, Assumption 1 holds, so $f(x \mid \mathcal{G})$ is an unbiased density estimator with uniformly bounded variance.

For our numerical experiments, we use the same cut $\mathcal{L} = \{5, 6, 7, 9, 10\}$ as Avramidis and Wilson (1996), indicated in light blue in Figure 4, even though there are other cuts with six links, which could possibly perform better because they hide more links. We could also compute the CDE with several choices of $\mathcal{L}$ and then take a convex combination. This approach scales nicely and works in exactly the same way for very large networks, with thousands of links. A simple adaptation also works for a stochastic max-flow problem, in which we want the density of the capacity of the minimal cut having the smallest capacity (L'Ecuyer et al. 2020).

The GLRDE method described in Section 2.5 does not work for this example. Indeed, with $X = h(\mathbf{Y})$ defined as the length of the longest path, for any j, the derivative $h_j(\mathbf{Y})$ is zero whenever arc j is not on the longest path, so we would need to select an arc j that is guaranteed to be on the longest path. But there is no such arc in general. We could perhaps apply a modified GLRDE that selects a cut instead of a single coordinate Y_j, but this is beyond the scope of this paper.

Table 2 and Figure 5 summarize our results. We see that for $n = 2^{19}$, the CDE outperforms the KDE by a factor of about 20 with MC, and by a factor of about $2^8 \approx 250$ with RQMC.

Table 2 **Values of $\hat{\nu}$ and e19 for the SAN example.**

		$\hat{\nu}$	e19
CDE	MC	0.96	25.6
	Lat+s	1.31	30.9
	Lat+s+b	1.17	29.6
	Sob+LMS	1.27	29.9
KDE	MC	0.78	20.9
	Lat+s	0.95	22.7
	Lat+s+b	0.93	22.0
	Sob+LMS	0.74	21.9

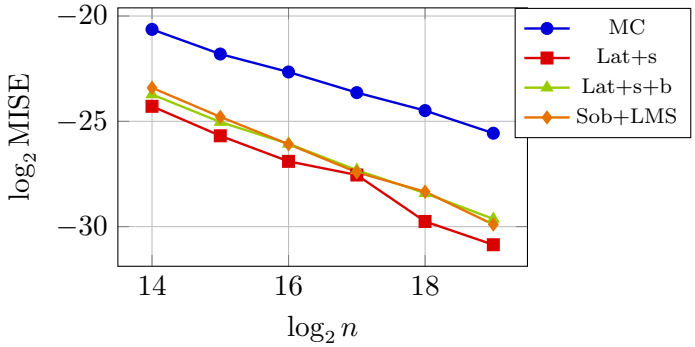

Figure 5 **MISE vs n in log-log scale, for the SAN example.**

4.4. Density of waiting times in a single queue

We adapted this example from Peng et al. (2020), mainly to compare our approach with the GLRDE proposed in their paper. What we use here is not ordinary CMC, but extended CMC, in which we condition on different information for each customer. This type of strategy would work for much larger queueing systems and many other types of systems that involve random delays. To estimate the density of the waiting times in a queueing system, the general idea is to hide sufficient information, for each customer in the system, so that its exact waiting time is unknown, but it has a density conditional on the known information, and we can compute this density easily. This is often easy to do even for large queueing systems. The hidden information can be the arrival time of the customer, the departure time of the previous customer, or something similar, selected so that we can compute the conditional density.

4.4.1. Model with independent days. We consider a single-server FIFO queue in which customers arrive from an arbitrary arrival process (not necessarily stationary Poisson) and the service times are independent, with continuous cdf G and density g. If W denotes the waiting time of a "random" customer, we want to estimate $p_0 = \mathbb{P}[W = 0]$ and the density f of W over $(0, \infty)$.

We first consider a system that starts empty and evolves over a fixed time horizon τ, which we call a *day*. Let T_j be the arrival time of the jth customer, $T_0 = 0$, $A_j = T_j - T_{j-1}$ the jth interarrival time, S_j the service time of customer j, and W_j the waiting time of customer j. Since the system starts

empty, we have $W_1 = 0$, and the Lindley recurrence gives us that $W_j = \max(0,\, W_{j-1} + S_{j-1} - A_j)$ for $j \geq 2$. At time τ, the arrival process stops, but service continues until all customers already arrived are served. The number of customers handled in a day is the random variable $N = \max\{j \geq 1 : T_j < \tau\}$. The cdf of W can be written as $F(0) = p_0$ and for $x > 0$, $F(x) = \mathbb{P}[W \leq x] = \mathbb{E}[\mathbb{I}(W \leq x)]$. The sequence of waiting times of all customers over an infinite number of independent successive days is a regenerative process that regenerates at the beginning of each day, so we can apply the renewal reward theorem, which gives

$$F(x) = \mathbb{E}[\mathbb{I}(W \leq x)] = \frac{\mathbb{E}\left[\mathbb{I}[W_1 \leq x] + \cdots + \mathbb{I}[W_N \leq x]\right]}{\mathbb{E}[N]}. \tag{17}$$

Since $\mathbb{E}[N]$ does not depend on x, we see that for $x > 0$, the density $f(x)$ is the derivative of the numerator with respect to x, divided by $\mathbb{E}[N]$.

To obtain a differentiable cdf estimator, we want to replace each indicator in the numerator by a conditional expectation. One simple way of doing this is to hide the service time S_{j-1} of the previous customer; that is, replace $\mathbb{I}[W_j \leq x]$ by

$$P_j(x) = \mathbb{P}[W_j \leq x \mid W_{j-1} - A_j] = \mathbb{P}[S_{j-1} \leq x + A_j - W_{j-1}] = G(x + A_j - W_{j-1}) \quad \text{for } x \geq 0.$$

This gives $P_j(0) = G(A_j - W_{j-1})$ (there is a probability mass at 0), whereas for $x > 0$, we have $P_j'(x) = \mathrm{d}P_j(x)/\mathrm{d}x = g(x + A_j - W_{j-1})$ and then, since N does not change when we change x,

$$f(x) = \frac{\mathbb{E}[D(x)]}{\mathbb{E}[N]} \qquad \text{where } D(x) = \sum_{j=1}^{N} g(x + A_j - W_{j-1}). \tag{18}$$

Note that we are not conditioning on the same information for all terms of the sum, so what we do is not exactly CMC, but *extended CMC* (Bratley et al. 1987). It nevertheless provides the required smoothing and an unbiased density estimator for the numerator of (17). In a multiserver queue, such as a call center with a large number of agents, one possibility would be to hide the arrival time A_j of the call, and compute the density of its waiting time conditional on the other information.

Often, for example if the arrival process is Poisson, $\mathbb{E}[N]$ can be computed exactly, in which case we only need to estimate $\mathbb{E}[D(x)]$ and we get an unbiased density estimator. Otherwise, the denominator $\mathbb{E}[N]$ can be estimated in the usual way, and we are then in the standard setting of estimating a ratio of expectations (Asmussen and Glynn 2007), for which we have unbiased estimators for the numerator and the denominator. We simulate n days, independently (with MC) or with n RQMC points, to obtain n realizations of $(N, D(x))$, say $(N_1, D_1(x)), \ldots, (N_n, D_n(x))$. The ratio estimator (CDE) of $f(x)$ is

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} D_i(x)}{\sum_{i=1}^{n} N_i}.$$

It can be computed at any $x \in [0, \infty)$. For independent realizations (with MC), the variance of $\hat{f}(x)$ can be estimated using the delta method for ratio estimators (Asmussen and Glynn 2007):

$$n\mathrm{Var}[\hat{f}(x)] \to \frac{\mathrm{Var}[D_i(x)] + \mathrm{Var}[N_i]f^2(x) - 2\mathrm{Cov}[D_i(x), N_i]f(x)}{(\mathbb{E}[N_i])^2}$$

asymptotically, when $n \to \infty$. This variance can be estimated by replacing the unknown quantities in this expression by their empirical values. This is consistent because the n pairs $(D_i(x), N_i)$, $i = 1, \ldots, n$, are independent. Alternatively, a confidence interval on $f(x)$ can also be computed with a bootstrap approach (Choquet et al. 1999).

In the RQMC case, the pairs $(D_i(x), N_i)$ are no longer independent. Then, to obtain an estimator of $f(x)$ for which we can estimate the variance, we make n_r independent replicates of the RQMC estimator of the pair $(\mathbb{E}[D(x)], \mathbb{E}[N])$, say $(\bar{D}_1(x), \bar{N}_1), \ldots, (\bar{D}_{n_r}(x), \bar{N}_{n_r})$, where each $(\bar{D}_j(x), \bar{N}_j)$ is the average of n pairs $(D_i(x), N_i)$ sampled by RQMC. We estimate the density $f(x)$ by the ratio of the two grand sums

$$\hat{f}_{\mathrm{rqmc},n_r}(x) = \frac{\sum_{j=1}^{n_r} \bar{D}_j(x)}{\sum_{j=1}^{n_r} \bar{N}_j}.$$

To estimate the variance, we use that

$$\mathrm{Var}[\hat{f}_{\mathrm{rqmc},n_r}(x)] \approx \frac{\mathrm{Var}[\bar{D}_j(x)] + \mathrm{Var}[\bar{N}_j]f^2(x) - 2\mathrm{Cov}[\bar{D}_j(x), \bar{N}_j]f(x)}{n_r(\mathbb{E}[N])^2}$$

and we replace all the unknown quantities in this expression by their empirical values.

Here, the required dimension of the RQMC points is the (random) total number of inter-arrival times A_j and service times S_j that we need to generate during the day. It is approximately twice the number of customers that arrive during the day. This number is unbounded, so the RQMC points must have unbounded (or infinite) dimension, and one must be able to generate the points without first selecting a maximal dimension. Recurrence-based RQMC point sets have this property; they can be provided for instance by ordinary or polynomial Korobov lattice rules (L'Ecuyer and Lemieux 2000, 2002), which are available in the hups package of SSJ (L'Ecuyer 2016).

4.4.2. Steady-state model. In a slightly different setting, we can assume that the single queue evolves in steady-state over an infinite time horizon, under the additional assumptions that the A_j's are i.i.d. and the S_j's are also i.i.d. Again, we want to estimate the density of the waiting time W of a random customer. In this case, the system regenerates whenever a new customer arrives in an empty system. The regenerative cycles can be much shorter on average than for the previous case, unless the day is very short or the utilization factor of the system is close to 1. The CDE has exactly the same form, apart from the different definition of regenerative cycle. In this case n represents the number of regenerative cycles, N_i is the number of customers in the ith cycle and $D_i(x)$ is the realization of $D(x)$ over the ith cycle.

In both settings, one could also hide A_j instead of S_{j-1}. The density estimator is similar and easy to derive. Intuition says that this should be a better choice if A_j has more variance than S_{j-1}.

4.4.3. The GLRDE estimator. Peng et al. (2020), Section 4.2.2., show how to construct a GLRDE for the density of the *sojourn time* of customer j in this single-queue model. The density of the *waiting time* can be estimated as follows. If the service times S_j are lognormal with parameters (μ, σ^2), we can write

$$X = W_j = \max(0,\, W_{j-1} + S_{j-1} - A_j) = \max(0,\, W_{j-1} + \exp[\sigma Z_{j-1} + \mu] - A_j) =: h(\mathbf{Y})$$

where Z_{j-1} has the standard normal density ϕ, and $\mathbf{Y} = (Y_1, Y_2, Y_3) = (Z_{j-1}, A_j, W_{j-1})$. When $W_j > 0$, taking the derivative of h with respect to $Y_1 = Z_{j-1}$ gives $h_1(\mathbf{Y}) = \exp[\sigma Z_{j-1} + \mu]\sigma = S_{j-1}\sigma$, $h_{11}(\mathbf{Y}) = S_{j-1}\sigma^2$, and these derivatives are 0 when $W_j = 0$. We also have $\partial \log \phi(x)/\partial x = -x$, and therefore for $x > 0$, $f(x) = \mathbb{E}[L(x)]/\mathbb{E}[N]$ where $L(x) = \sum_{j=1}^{N} \mathbb{I}[W_j \le x] \cdot \Psi_j$ and $\Psi_j = -(Z_{j-1} + \sigma)/(S_{j-1}\sigma)$. We can do n runs to estimate each of the two expectations in the ratio. This provides a very similar density estimator as with the CDE in (18), but here $L(x)$ is discontinuous in x, whereas $D(x)$ in (18) is continuous.

4.4.4. Numerical results. For a numerical illustration, suppose the time is in minutes, let the arrival process be Poisson with constant rate $\lambda = 1$, and the service times S_j lognormal with parameters $(\mu, \sigma^2) = (-0.7, 0.4)$. This gives $\mathbb{E}[S_j] = e^{-0.5} \approx 0.6065$ and $\mathrm{Var}[S_j] = e^{-1}(e^{0.4} - 1) \approx 0.18093$. For RQMC, we use infinite-dimensional RQMC points defined by Korobov lattice rules (L'Ecuyer and Lemieux 2000) selected with `Lattice Builder` (L'Ecuyer and Munger 2016) using order-dependent weights $\gamma_k = 0.005^k$ for projections of order k. We do not use Sobol' points because with the available software, there is an upper bound on the dimension.

Finite-horizon case. For the finite-horizon case, take $\tau = 60$, so $\mathbb{E}[N] = 60$, we only need to estimate the numerator, and we have an unbiased density estimator all over $[0, \infty)$. The results for $(a, b] = (0, 2.2]$ are in Table 3. Due to the large and random dimensionality of the required RQMC points, and more importantly the discontinuity of the derivative of the CDE with respect to the underlying uniforms (because of the max, the HK variation is infinite), it was unclear if RQMC could bring any significant gain for this example. The good surprise is that although RQMC does not improve $\tilde{\nu}$ significantly, it improves the IV itself by a factor of about $2^{7.5} \approx 180$ for $n = 2^{19}$, which is quite significant. We also see that CDE beats GLRDE by a factor of about 500 with MC and about 200 with RQMC.

Steady-state case. We performed a similar experiment using regenerative simulation for the steady-state model. The density is similar but not exactly the same as in the finite-horizon case. The results are in Table 4. They are similar to those of the finite-horizon case, with similar empirical convergence rates, and the IV for $n = 2^{19}$ is again about 180 times smaller with CDE+RQMC compared to CDE+MC. The IV for GLRDE with $n = 2^{19}$ is roughly 300 times larger than with

Table 3 Values of $\hat{\nu}$ and e19 for the single queue example, finite-horizon case.

		$\hat{\nu}$	e19
CDE	MC	1.00	24.8
	Lat+s	0.99	32.3
	Lat+s+b	1.02	32.3
GLRDE	MC	1.00	15.8
	Lat+s	1.03	24.6
	Lat+s+b	1.08	25.0

CDE with MC and 200 times larger than with CDE with RQMC. The only important difference is that here, the IV is about 30 times *larger* than in the finite-horizon case, for all the methods. The explanation is that in the finite-horizon case, we simulate n runs with about 60 customers per run, whereas in the steady-state case, we have about 2.5 customers per regenerative cycle on average, so we simulate about 25 times fewer customers. Interestingly, the fact that we use much more coordinates of the RQMC points in the finite-horizon case (on average) makes no significant difference. A similar observation was made by L'Ecuyer and Lemieux (2000), Section 10.3, who compared finite-horizon runs of 5000 customers each on average, with regenerative simulation, in the context of estimating the probability of a large waiting time using RQMC. The reason why RQMC performs well even for a very large time horizon is that the integrand has *low effective dimension in the successive-dimensions sense* (as defined by these authors). Appendix C of the Supplement provides additional plots for this example.

Table 4 Values of $\hat{\nu}$ and e19 for the single queue example, steady-state case.

		$\hat{\nu}$	e19
CDE	MC	0.99	19.9
	Lat+s	1.04	27.6
	Lat+s+b	1.08	27.8
GLRDE	MC	0.99	11.5
	Lat+s	1.20	20.1
	Lat+s+b	1.21	20.4

4.5. Making a change of variable

In many situations, $X = h(\mathbf{Y})$ for a random vector $\mathbf{Y}$ and hiding a single coordinate of $\mathbf{Y}$ does not provide a very effective CDE. But sometimes, after an appropriate change of variable $\mathbf{Y} = g(\mathbf{Z})$, hiding one coordinate of the random vector $\mathbf{Z}$ can provide a much more effective CDE. We will use this technique in Section 4.6. We describe it here in a separate subsection because it can be useful for a much wider range of applications.

Specifically, let $\mathbf{Z}_{-j}$ denote the vector $\mathbf{Z}$ with Z_j (the jth coordinate) removed, and let $\gamma(z) = \gamma(z; \mathbf{Z}_{-j}) = h(g(z; \mathbf{Z}_{-j}))$ denote the value of $h(\mathbf{Y})$ as a function of $Z_j = z$ when $\mathbf{Z}_{-j}$ is fixed. We assume in the following that for almost any realization of $\mathbf{Z}_{-j}$, $\gamma(z; \mathbf{Z}_{-j})$ is a monotone non-decreasing and differentiable function of z, so that $\gamma^{-1}(x) = \inf\{z \in \mathbb{R} : \gamma(z) \geq x\}$ is well defined for

any x. We also assume that Z_j has density φ and is independent of $\mathbf{Z}_{-j}$ (to simplify). Conditional on $\mathbf{Z}_{-j}$, we have

$$\mathbb{P}[x < h(\mathbf{Y}) \le x+\delta \mid \mathbf{Z}_{-j}] = \mathbb{P}[x < \gamma(Z_j) \le x+\delta \mid \mathbf{Z}_{-j}] = \mathbb{P}[z < Z_j \le z+\Delta \mid \mathbf{Z}_{-j}] \approx \varphi(z)\Delta$$

where $z = \gamma^{-1}(x)$ and $z+\Delta = \gamma^{-1}(x+\delta)$. Taking the limit gives

$$f(x \mid \mathbf{Z}_{-j}) = \lim_{\delta\to 0} \frac{\mathbb{P}[z < Z_j \le z+\Delta \mid \mathbf{Z}_{-j}]}{\delta} = \lim_{\delta\to 0} \frac{\varphi(z)\Delta}{\delta} = \frac{\varphi(z)}{\gamma'(z)} = \frac{\varphi(\gamma^{-1}(x))}{\gamma'(\gamma^{-1}(x))},$$

assuming that the latter is well defined. In case there are closed-form formulas for γ^{-1} and γ', this CDE can be evaluated directly. Otherwise, $z = \gamma^{-1}(x)$ can often be computed by a few iterations of a root-finding algorithm. Since γ and its inverse γ^{-1} depend on $\mathbf{Z}_{-j}$, this could mean inverting a different function for each sample realization. Our next example will show that the approach could nevertheless bring a huge benefit.

4.6. A function of a multivariate normal vector

We consider a multivariate normal vector $\mathbf{Y} = (Y_1, \ldots, Y_s)^{\mathsf{t}}$ (where $^{\mathsf{t}}$ means transposed) defined via $Y_j = Y_{j-1} + \mu_j + \sigma_j Z_j$ with $Y_0 = 0$, the μ_j and $\sigma_j > 0$ are constants, and the Z_j are independent $\mathcal{N}(0,1)$ random variables, with cdf Φ and density ϕ. Let $X = \bar{S} = (S_1 + \cdots + S_s)/s$ where $S_j = S_0 e^{Y_j}$ for some constant $S_0 > 0$. We want to estimate the density of X over some interval $(a,b) = (K, K+c)$ where $K \ge 0$ and $c > 0$. This is the same as estimating the density of $\max(0, \bar{S} - K)$, which may represent the payoff of a financial contract, for example (Glasserman 2004). A simple way to define the CDE here is to hide Z_s. The conditional cdf is $\mathbb{P}[X \le x \mid \mathbf{Z}_{-s}] = \mathbb{P}[Z_s \le W(x)] = \Phi(W(x))$ where

$$W(x) = (\ln[sx - (S_1 + \cdots + S_{s-1})/S_0] - \ln S_0 - Y_{s-1} - \mu_s)/\sigma_s.$$

Taking the derivative with respect to x gives the unbiased CDE

$$f(x \mid \mathbf{Z}_{-s}) = \frac{\partial}{\partial x}\mathbb{P}[\bar{S} \le x \mid \mathbf{Z}_{-s}] = \phi(W(x))W'(x) \;=\; \frac{\phi(W(x))s}{[sx - (S_1 + \cdots + S_{s-1})/S_0]\sigma_s}. \tag{19}$$

Unfortunately, this *sequential* CDE is usually rather spiky, because hiding only this Z_s does not remove much information, and then the conditional density has a large variance.

We now describe a less obvious but more effective conditioning approach. The goal is to hide a variable that contains more information. For this, we generate the vector $\mathbf{Y}$ using a Brownian bridge construction in which the Z_j's are used in a different way, as follows (Caflisch et al. 1997, Glasserman 2004). Let $\bar{\mu}_j = \mu_1 + \cdots + \mu_j$ and $\bar{\sigma}_j = \sigma_1 + \cdots + \sigma_j$, for $j = 1, \ldots, s$. With this construction, we first sample $Y_s = \bar{\mu}_s + \bar{\sigma}_s Z_s$. Then, given $Y_s = y_s$, we put $j_2 = \lfloor s/2 \rfloor$, and we sample Y_{j_2} from its normal distribution conditional on $Y_s = y_s$, which is normal with mean $y_s\bar{\mu}_{j_2}/\bar{\mu}_s$ and variance $(\bar{\sigma}_s - \bar{\sigma}_{j_2})\bar{\sigma}_{j_2}/\bar{\sigma}_s$. This uses the fact that if X_1 and X_2 are independent and normal, then conditional

on $X_1+X_2=\bar{x}$, X_1 is normal with mean $\bar{x}\mathbb{E}[X_1]/\mathbb{E}[X_1+X_2]$ and variance $\mathrm{Var}[X_1]\mathrm{Var}[X_2]/\mathrm{Var}[X_1+X_2]$. Then we put $j_3=\lfloor j_2/2\rfloor$ and we sample Y_{j_3} conditionally on Y_{j_2}, then we put $j_4=\lfloor (j_2+s)/2\rfloor$ and we sample Y_{j_4} conditionally on (Y_{j_2},Y_s), and so on, until all the Y_j's are known.

For the CDE, we hide again Z_s, but now Z_s has much more impact on the payoff, because all the Y_j's depend on Z_s. This makes the conditional density much less straightforward to compute, but we can proceed as follows. To avoid sampling Z_s, we sample $Y_1,\ldots,Y_{s-1}$ conditional on $Z_s=z_s=0$, which will give say $Y_1^0,\ldots,Y_{s-1}^0$, and then write X as a function of $z=z_s$ conditional on these values, that is, conditional on $\mathbf{Z}_{-s}=(Z_1,\ldots,Z_{s-1})$. We have $Y_s=Y_s^0+\bar{\sigma}_s Z_s$ and $Y_j=Y_j^0+(\bar{\mu}_j/\bar{\mu}_s)\bar{\sigma}_s Z_s$. Then,

$$X=\bar{S}=\frac{S_0}{s}\sum_{j=1}^{s}e^{Y_j}=\frac{S_0}{s}\sum_{j=1}^{s}\exp[Y_j^0+Z_s(\bar{\mu}_j/\bar{\mu}_s)\bar{\sigma}_s].$$

This fits the framework of Section 4.5, with $j=s$,

$$\gamma(z)=\frac{S_0}{s}\sum_{j=1}^{s}\exp[Y_j^0+z(\bar{\mu}_j/\bar{\mu}_s)\bar{\sigma}_s]\quad\text{and}\quad\gamma'(z)=\frac{S_0}{s}\sum_{j=1}^{s}\exp[Y_j^0+z(\bar{\mu}_j/\bar{\mu}_s)\bar{\sigma}_s](\bar{\mu}_j/\bar{\mu}_s)\bar{\sigma}_s.$$

The CDE at $x=\gamma(z)$ is then $f(x\mid\mathbf{Z}_{-s})=\phi(z)/\gamma'(z)$. We call it the *bridge* CDE.

To compute this density at a specified x we need $z=\gamma^{-1}(x)$, We have no explicit formula for γ^{-1} in this case, but we can compute a root of $\gamma(z)-x=0$ numerically. To evaluate the density at the n_e evaluation points $e_1,\ldots,e_{n_e}$ in (a,b), we first compute $x_*=\gamma(0)$ and let j_* be the smallest j for which $e_j\ge x_*$. We compute $z=w_{j_*}$ such that $\gamma(w_{j_*})=e_{j_*}$. This can be done via Newton iteration, $z_k=z_{k-1}-(\gamma(z_{k-1})-e_{j^*})/\gamma'(z_{k-1})$, starting with $z_0=0$. Then, for $j=j_*+1,\ldots,n_e$, we use again Newton iteration to find $z=w_j$ such that $\gamma(w_j)=e_j$, starting at $z_0=w_{j-1}$. We do the same to find $z=w_j$ such that $\gamma(w_j)=e_j$ for $j=j_*-1,\ldots,1$, starting at $z_0=w_{j+1}$. This provides the point w_j required to evaluate the conditional density at e_j, for each j. We must repeat this procedure for each realization of $\mathbf{Z}_{-j}$, because the function γ depends on $\mathbf{Z}_{-j}$. However, the gain in accuracy is more significant than the cost of additional computations. This conditioning differs from the simpler ones used by Boyle et al. (1997) and Heidergott et al. (2015) for barrier options.

For a numerical illustration, we take $S_0=100$, $s=12$, $\mu_j=0.00771966$ and $\sigma_j=0.035033$ for all j, and $K=101$. We estimate the density of the payoff over $[a,b]=[101,\,128.13]$. To approximate the root of $\gamma(z)-x=0$ for the bridge CDE, we use five Newton iterations; doing more makes no significant difference. The results are in Table 5, with additional plots in the Supplement. RQMC with the bridge CDE performs extremely well. For example, for Sob+LMS, the MISE with $n=2^{19}$ is approximately $2^{-46.9}$, which is about 2^{19} (half a million) times smaller than for the same CDE with MC, and it decreases as $\mathcal{O}(n^{-2})$. With a KDE, the MISE with $n=2^{19}$ is about $2^{21}\approx 2$ million times larger with the same Sobol' points and $2^{26}\approx 67$ million times larger with MC. With the sequential CDE, RQMC is ineffective and the IV of the MC estimator is also quite large, as

expected. To illustrate the behavior of the sequential and bridge CDEs, Figure 6 plots five single realizations of each, using the same horizontal scale. The sequential CDE has much more spiky realizations than the bridge CDE, and this explains why the latter performs much better.

Table 5 Values of $\hat{\nu}$ and e19 for the Asian option, with sequential and bridge CDE constructions.

		$\hat{\nu}$	e19
sequential KDE	MC	0.78	20.4
	Sob+LMS	0.76	20.6
sequential CDE	MC	1.00	19.9
	Lat+s	1.07	20.3
	Lat+s+b	1.01	20.1
	Sob+LMS	1.00	20.0
bridge CDE	MC	1.04	27.9
	Lat+s	1.60	40.0
	Lat+s+b	1.74	45.0
	Sob+LMS	2.01	46.9

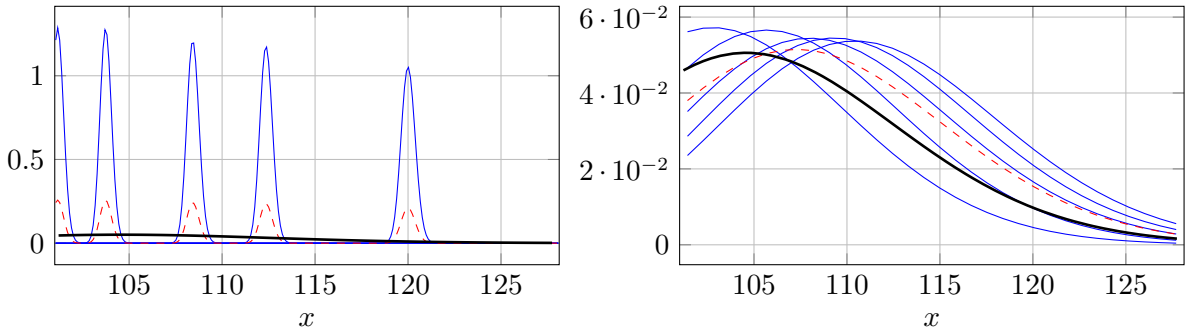

Figure 6 Five realizations of the density estimator (blue), their average (dashed red), and the true density (thick black) for the sequential CDE (left) and the bridge CDE (right), for the Asian option example.

The method introduced here works not only for a Brownian motion, but for more general Lévy processes as well. Some popular models in finance use a Lévy subordinator process to produce a random clock speed to model stochastic volatility, and a geometric Brownian process that evolves at that random speed. This includes the *variance-gamma* (VG) and the *normal inverse Gaussian* processes, for example. For these models, we can generate the subordinator as usual using a bridge method to obtain the random times at which the Brownian process is evaluated, and then apply the method that we just described to the resulting Brownian process. For the VG process, a more effective alternative could be to use the *double gamma bridge sampling* method described in Avramidis and L'Ecuyer (2006) and hide the first variable as we have done here. Yet another approach would be to do a monotone mapping between the Lévy process and a Brownian motion as explained in L'Ecuyer et al. (2008), estimate the density in the Brownian representation as we did here, and transform this density to the original Lévy process via the change of variable that corresponds to the mapping.

4.7. More examples

Additional examples are given in the Online Supplement. In the first one, X is a sum of independent normal random variables, with known density, and the purpose is to see how each estimator behaves as a function of the dimension (the number of summands) and of the relative variance of the one we hide. The second one is a (real-life) six-dimensional example in which X is the buckling strength of a steel plate. The third one is a multicomponent system in which each component fails at a certain random time, and we want to estimate the density of the failure time of the system. In the fourth one, we explain how accurate density estimation is useful to compute a confidence interval on a quantile or on the expected shortfall. This, alone, has many applications.

5. Summary and Guidelines

Here we provide a summary of the main conditions and some guidelines for applying the method. The primary task is to select the information $\mathcal{G}$ on which we condition, or equivalently, to decide what we hide. The main constraint is that $\mathcal{G}$ must be selected in a way that Assumption 1 is satisfied, at least in the region where we want to estimate the density. The key condition for this is that the conditional density must be well-defined in that region. In particular, the conditional CDF should never have jumps. A second requirement is that the conditional density can be calculated efficiently for all realizations of $\mathcal{G}$ (almost surely). There are often many possible choices of $\mathcal{G}$ for which these conditions are satisfied. The set of admissible choices is highly problem-dependent; it depends on the model and also on the variable X of interest.

In many cases, one must hide more than just a single input random variable. The hidden information can also be dynamically selected, in the sense that it may depend on the sample realization (for example, if X is the maximum of several independent random variables, one may generate all the variables and hide the maximum). Sometimes, it becomes much easier to apply the method after making an appropriate multivariate change of variable. This may require creativity, as we have shown in our examples.

When there are many choices for $\mathcal{G}$, finding the optimal one (e.g., to minimize the work-normalized MISE) can be difficult in general, because the MISE depends on many factors, but just finding a good one is usually much easier and sufficient. For comparable computing times, the best choices are often those for which the variance of the conditional density is the largest. As a rule of thumb, it is usually better to condition on lower-variance information and hide variables having a large variance contribution. The optimal choice also depends (in general) on the point x at which we want to estimate the density. In principle, one could optimize by using different conditioning for different intervals, but it is usually not worth the additional complications. When several good choices of $\mathcal{G}$ are available, selecting a few of them and taking a convex combination of the corresponding estimators can provide a more robust CDE than selecting only one $\mathcal{G}$.

When we want to combine the CDE with RQMC, additional properties come into play: we want to select $\mathcal{G}$ and also the formulation of the estimator as a function $\tilde{g}'$ of the vector $\mathbf{U}$ of underlying uniform random numbers in a way that the assumption of Proposition 3 is satisfied (if possible) and the variation of this function $\tilde{g}'$ is not too large. Proving these conditions in practice may be difficult, but one can always try RQMC empirically. Experience shows that it can often reduce the variance significantly, in particular when the effective dimension of $\tilde{g}'$ is small. This was illustrated in our examples.

6. Conclusion

We have examined a simple and very effective approach for estimating the density of a random variable generated by simulation from a stochastic model, by using a computable conditional density. The resulting CDE is unbiased and its MISE converges faster than for other popular density estimators such as the KDE. We have also shown how to further reduce the IV, and even improve its convergence rate, by combining the CDE with RQMC. Our numerical examples show that this combination can be very efficient. It sometimes reduces the MISE by factors over a million. The CDE approach also outperforms the recently proposed GLRDE method, and CDE+RQMC outperforms both GLRDE+RQMC and KDE+RQMC, in all our examples. RQMC tends to bring a larger improvement to the CDE than the KDE or GLRDE because the estimator usually has less variation as a function of the underlying uniforms.

The examples in the paper were selected to provide insight on key issues. We tried to avoid unnecessary complications in the models. But it would not be too difficult to derive CDEs for larger and more complicated versions of these models. We outlined some possibilities in the text. Suggested future work includes experimenting this methodology in more complicated applications, designing and exploring different types of conditioning, and perhaps adapting the Monte Carlo sampling strategies to make the method more effective for specific applications (e.g., by changing the way X is defined in terms of the basic input random variates). Its application to quantile and expected shortfall estimation also deserves further study.

Acknowledgments

This work has been supported by an IVADO Research Grant, an NSERC-Canada Discovery Grant, a Canada Research Chair, and an Inria International Chair, to P. L'Ecuyer. F. Puchhammer was also supported by Spanish and Basque governments fundings through BCAM (ERDF, ESF, SEV-2017-0718, PID2019-108111RB-I00, PID2019-104927GB-C22, BERC 2018e2021, EXP. 2019/00432, ELKARTEK KK-2020/00049), and the computing infrastructure of i2BASQUE academic network and IZO-SGI SGIker (UPV). The reviewers, the editors, and Julien Keutchayan gave several comments that improved the paper.

References

Asmussen S (2018) Conditional Monte Carlo for sums, with applications to insurance and finance. *Annals of Actuarial Science* 12(2):455–478.

Asmussen S, Glynn PW (2007) *Stochastic Simulation* (New York: Springer-Verlag).

Avramidis AN, L'Ecuyer P (2006) Efficient Monte Carlo and quasi-Monte Carlo option pricing under the variance-gamma model. *Management Science* 52(12):1930–1944.

Avramidis AN, Wilson JR (1996) Integrated variance reduction strategies for simulation. *Operations Research* 44:327–346.

Avramidis AN, Wilson JR (1998) Correlation-induction techniques for estimating quantiles in simulation experiments. *Operations Research* 46(4):574–591.

Ben Abdellah A, L'Ecuyer P, Owen A, Puchhammer F (2021) Density estimation by randomized quasi-Monte Carlo. *SIAM Journal on Uncertainty Quantification* 9(1):280–301.

Bingham D (2017) Virtual library of simulation experiments. `https://www.sfu.ca/~ssurjano/canti.html`.

Boyle P, Broadie M, Glasserman P (1997) Monte Carlo methods for security pricing. *Journal of Economic Dynamics and Control* 21(8-9):1267–1321.

Bratley P, Fox BL, Schrage LE (1987) *A Guide to Simulation* (New York, NY: Springer-Verlag), second edition.

Caflisch RE, Morokoff W, Owen A (1997) Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *J. of Computational Finance* 1(1):27–46.

Choquet D, L'Ecuyer P, Léger C (1999) Bootstrap confidence intervals for ratios of expectations. *ACM Transactions on Modeling and Computer Simulation* 9(4):326–348.

Cui Z, Fu MC, Hu JQ, Liu Y, Peng Y, Zhu L (2020) On the variance of single-run unbiased stochastic derivative estimators. *INFORMS Journal on Computing* Forthcoming.

Dick J, Pillichshammer F (2010) *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration* (Cambridge, U.K.: Cambridge University Press).

Dieudonné J (1969) *Foundations of Modern Analysis* (New York: Academic Press), second edition.

Efron B, Hastie T (2016) *Computer Age Statistical Inference* (New York: Cambridge University Press).

Fu M, Hu JQ (1997) *Conditional Monte Carlo* (Boston: Kluwer Academic).

Fu MC (2006) Sensitivity analysis in Monte Carlo simulation of stochastic activity networks. Alt FB, Fu MC, Golden BL, eds., *Perspectives in Operations Research*, 351–366, Operations Research/Computer Science Interfaces Series (Boston: Springer).

Fu MC, Hong LJ, Hu JQ (2009) Conditional Monte Carlo estimation of quantile sensitivities. *Management Science* 55(12):2019–2027.

Glasserman P (2004) *Monte Carlo Methods in Financial Engineering* (New York: Springer-Verlag).

Glynn PW (1987) Likelihood ratio gradient estimation: an overview. *Proceedings of the 1987 Winter Simulation Conference*, 366–375 (Piscataway, NJ: IEEE Press).

Gong WB, Ho YC (1987) Smoothed (conditional) perturbation analysis of discrete event dynamical systems. *IEEE Transactions on Automatic Control* AC-32(10):858–866.

Gu C, Qiu C (1993) Smoothing spline density estimation: Theory. *The Annals of Statistics* 21(1):217–234.

He Z, Wang X (2015) On the convergence rate of randomized quasi-monte carlo for discontinuous functions. *SIAM Journal on Numerical Analysis* 53(5):2488–2503.

Heidergott B, Leahu H, Volk-Makarewicz WM (2015) A smoothed perturbation analysis of parisian options. *IEEE Transactions on Automatic Control* 60(2):469–474.

Hickernell FJ (2002) Obtaining $O(N^{-2+\epsilon})$ convergence for lattice quadrature rules. Fang KT, Hickernell FJ, Niederreiter H, eds., *Monte Carlo and Quasi-Monte Carlo Methods 2000*, 274–289 (Berlin: Springer-Verlag).

Laub PJ, Salomone R, Botev ZI (2019) Monte Carlo estimation of the density of the sum of dependent random variables. *Mathematics and Computers in Simulation* 161:23–31.

Law AM (2014) *Simulation Modeling and Analysis* (New York: McGraw-Hill), fifth edition.

L'Ecuyer P (1990) A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science* 36(11):1364–1383.

L'Ecuyer P (2009) Quasi-Monte Carlo methods with applications in finance. *Finance and Stochastics* 13(3):307–349.

L'Ecuyer P (2016) SSJ: Stochastic simulation in Java, `http://simul.iro.umontreal.ca/ssj/`, accessed 9$^{\text{th}}$ August 2021.

L'Ecuyer P (2018) Randomized quasi-Monte Carlo: An introduction for practitioners. Glynn PW, Owen AB, eds., *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC 2016*, 29–52 (Berlin: Springer).

L'Ecuyer P, Buist E (2008) On the interaction between stratification and control variates, with illustrations in a call center simulation. *Journal of Simulation* 2(1):29–40.

L'Ecuyer P, Lemieux C (2000) Variance reduction via lattice rules. *Management Science* 46(9):1214–1235.

L'Ecuyer P, Lemieux C (2002) Recent advances in randomized quasi-Monte Carlo methods. Dror M, L'Ecuyer P, Szidarovszky F, eds., *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, 419–474 (Boston: Kluwer Academic).

L'Ecuyer P, Marion P, Godin M, Fuchhammer F (2020) A tool for custom construction of QMC and RQMC point sets. *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC 2020*, `https://arxiv.org/abs/2012.10263`.

L'Ecuyer P, Munger D (2012) On figures of merit for randomly-shifted lattice rules. Woźniakowski H, Plaskota L, eds., *Monte Carlo and Quasi-Monte Carlo Methods 2010*, 133–159 (Berlin: Springer-Verlag).

L'Ecuyer P, Munger D (2016) Algorithm 958: Lattice builder: A general software tool for constructing rank-1 lattice rules. *ACM Transactions on Mathematical Software* 42(2):Article 15.

L'Ecuyer P, Parent-Chartier JS, Dion M (2008) Simulation of a Lévy process by PCA sampling to reduce the effective dimention. *Proceedings of the 2008 Winter Simulation Conference*, 436–443 (Piscataway, NJ: IEEE Press).

L'Ecuyer P, Perron G (1994) On the convergence rates of IPA and FDC derivative estimators. *Operations Research* 42(4):643–656.

Lei L, Peng Y, Fu MC, Hu JQ (2018) Applications of generalized likelihood ratio method to distribution sensitivities and steady-state simulation. *Discrete Event Dynamic Systems* 28(1):109–125.

Lemieux C (2009) *Monte Carlo and Quasi-Monte Carlo Sampling* (Springer-Verlag).

Nakayama MK (2014a) Confidence intervals for quantiles using sectioning when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation* 24(4):Article 9.

Nakayama MK (2014b) Quantile estimation when applying conditional Monte Carlo. *2014 International Conference on Simulation and Modeling Methodologies, Technologies, and Applications (SIMULTECH)*, 280–285 (IEEE).

Nelson BL (2008) The MORE plot: Displaying measures of risk and error from simulation output. *Proceedings of the 2008 Winter Simulation Conference*, 413–416 (Piscataway, New Jersey: IEEE Press).

Niederreiter H (1992) *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63 of *SIAM CBMS-NSF Reg. Conf. Series in Applied Mathematics* (SIAM).

Owen AB (2003) Variance with alternative scramblings of digital nets. *ACM Transactions on Modeling and Computer Simulation* 13(4):363–378.

Parzen E (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33(3):1065–1076.

Peng Y, Fu MC, Heidergott B, Lam H (2020) Maximum likelihood estimation by Monte Carlo simulation: Towards data-driven stochastic modeling. *Operations Research* 68(6):1896–1912.

Peng Y, Fu MC, Hu JQ, Heidergott B (2018) A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Operations Research* 66(2):487–499.

Scott DW (2015) *Multivariate Density Estimation* (Wiley), second edition.

Serfling RJ (1980) *Approximation Theorems for Mathematical Statistics* (New York, NY: Wiley).

Silverman B (1986) *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).

Smith JS, Nelson BL (2015) Estimating and interpreting the waiting time for customers arriving to non-stationary queueing system. *Proceedings of the 2015 Winter Simulation Conference*, 2610–2621 (Piscataway, New Jersey: IEEE Press).

Sturrock DT, Pegden CD (2010) Recents innovations in Simio. *Proceedings of the 2010 Winter Simulation Conference*, 21–31 (Piscataway, New Jersey: IEEE Press).

Thiongane M, Chan W, L'Ecuyer P (2021) Learning-based prediction of conditional waiting time distributions in multiskill call centers. *Operations Research and Entreprise Systems, expanded selected papers from ICORES 2020* (Springer), URL `http://www.iro.umontreal.ca/~lecuyer/myftp/papers/delay-dist-forecast.pdf`, to appear.

Van der Vaart AW (2000) *Asymptotic Statistics* (Cambridge University Press).

Wand MP, Jones MC (1995) *Kernel Smoothing* (Chapman and Hall).

Yu J, Shi J, Liu A, Wang Y (2020) Smoothing spline semiparametric density models. *Journal of the American Statistical Association* 1–14, URL `http://dx.doi.org/10.1080/01621459.2020.1769636`.