Les Cahiers du GERAD ISSN: 0711–2440

Monte Carlo and quasi-Monte Carlo density estimation via conditioning

P. L'Ecuyer, F. Puchhammer,
A. Ben Abdellah

G–2019–41

June 2019

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Monte Carlo and quasi-Monte Carlo density estimation via conditioning

Pierre L'Ecuyer
Florian Puchhammer
Amal Ben Abdellah

GERAD & Département d'Informatique et de Recherche Opérationnelle, Université de Montréal (Québec) Canada, H3C 3J7

lecuyer@iro.umontreal.ca
florian.puchhammer@umontreal.ca
amal.ben.abdellah@umontreal.ca

June 2019
Les Cahiers du GERAD
G–2019–41

Abstract: 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 density estimators, the mean integrated square error converges slower 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 study an approach in which conditional Monte Carlo permits one to obtain a smooth estimator of the cumulative distribution function, whose sample derivative is an unbiased estimator of the density at any point, and therefore converges at a faster rate than the usual density estimators, under mild conditions. By combining this with randomized quasi-Monte Carlo to generate the sample, we can achieve an even faster rate.

Keywords: Density estimation, conditional Monte Carlo, quasi-Monte Carlo

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. We thank Julien Keutchayan for pointing out mistakes in a previous version.

1 Introduction

Simulation is commonly used to generate independent 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 (Asmussen and Glynn 2007, Law 2014). However, one can extract from the sample much more information than just a confidence interval on the mean. When n is reasonably large, one may in fact estimate the *entire distribution of* X. Simple ways of showing this information is to plot the empirical *cumulative distribution function* (cdf) or (more commonly) a histogram of the observations. The histogram provides a better visual insight on the distribution. When X is a continuous random variable, it can be interpreted as a rough sketch (or estimate) of the density of X.

There are more refined and efficient density estimation methods, the leading one being the *kernel density estimator* (KDE). Given n independent realizations of X, the mean integrated square error (MISE) of the KDE converges as $\mathcal{O}(n^{-4/5})$ in the best case, while a histogram can only achieve $\mathcal{O}(n^{-2/3})$. Both 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 rate for the KDE, histogram, and other similar density estimators stems from the presence of bias. For a histogram, taking wider rectangles reduces the variance but increases the bias by flattening out the short-range density variations, and 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.

Can we construct density estimators that avoid bias? When the n observations of X are given and nothing else is known, as traditionally assumed in classical non-parametric statistics, nothing known beats the rate of the KDE (Scott 2015). But if the observations of X are generated by simulation and we have control of how this is done, there are opportunities to improve the density estimators. This is the subject of the present paper. We study two main ideas and their combination.

The first general idea is to build a smooth estimator of the cdf via conditional Monte Carlo (CMC), and take the sample derivative of this estimator to estimate the density. We call it a *conditional density estimator* (CDE). Under appropriate conditions, the CDE is unbiased and its variance is bounded uniformly by a constant divided by n, so its MISE is $\mathcal{O}(n^{-1})$. This idea of using CMC was mentioned by Asmussen and Glynn (2007), page 146, Example 4.3, and further studied in Asmussen (2018), for the special case where the goal is to estimate the density of a sum of continuous random variables having a known distribution from which we can sample exactly. Asmussen (2018) simply "hides" the last term of the sum, meaning that this term is not generated or its actual value is erased, and he takes the conditional distribution of the sum given the other terms, to estimate the cdf, the density, the value at risk (VaR), and the conditional value at risk (CVaR) of the sum. CMC has also been studied for estimating the derivative of an expectation (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. Laub et al. (2019) propose a different (but related) approach that combines a clever change of variable with the likelihood ratio method for derivative estimation, to estimate the density of a sum of dependent random variables.

In this paper, we study the CMC method to estimate the cdf and density in a more general setting. We provide several examples in which X is not defined as a sum of random variables and where we hide (or erase) more than just one random variable to do the conditioning. We give conditions under which we can prove that the density estimator is unbiased. A key condition is that the conditional cdf must be a continuous function of the point x at which we estimate the density. Sometimes, this can be achieved only by hiding several variables and not only one. The variance of the density estimator may depend strongly on which variable(s) we hide, i.e., on what we are conditioning. We will illustrate this with several examples.

Once we have a smooth density estimator, the second strategy to further improve the convergence rate is to replace the independent uniform random numbers that drive the simulation by *randomized quasi-Monte Carlo* (RQMC) points. We prove in this paper that by combining these two strategies, under some conditions, we can obtain a density estimator whose MISE converges at a faster rate

than $\mathcal{O}(n^{-1})$, namely $\mathcal{O}(n^{-2+\epsilon})$ for any $\epsilon > 0$. We also observe this faster rate empirically on numerical examples. Ben Abdellah et al. (2019) studied the combination of RQMC with a KDE for density estimation. They 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, and they observed this faster rate empirically on examples. They also observed that the MISE reduction provided by RQMC degrades rapidly when the bandwidth is reduced (to reduce the bias) and that this degradation effect gets stronger when the dimension increases. The approach proposed in the present paper avoids this problem and is generally more effective than the KDE-RQMC combination.

The remainder is organized as follows. In Section 2, we define our general setting, recall some key facts about density estimators, introduce the general CMC method to build the CDEs considered in this paper, and prove some of their properties. In Section 3, we explain how to combine the CDE with RQMC and prove convergence properties for this combination. Section 5 reports experimental results with various examples. A conclusion is given in Section 6.

2 Model and CMC density estimator

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

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

The MISE can be decomposed as the sum of the *integrated variance* (IV) and the *integrated square bias* (ISB):

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

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 $\hat{f}_n$.

Historically, this type of density estimator was developed for the setting in which $X_1, \ldots, X_n$ are given a priori, and it works in exactly the same way if $X_1, \ldots, X_n$ are independent observations generated by simulation. However, when the observations are generated from a simulation model, there is often an opportunity to do better.

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 by taking its derivative we obtain $\mathrm{d}\hat{F}_n(x)/\mathrm{d}x = 0$ 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 continuous estimator of F can be constructed by CMC, as follows. 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. Under the following assumption, we prove that $F'(x \mid \mathcal{G}) = \mathrm{d}F(x \mid \mathcal{G})/\mathrm{d}x$ is an unbiased estimator of $f(x)$ whose variance is bounded uniformly in x. Note that 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 $D(\mathcal{G}) \subset [a, b]$. 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 $\sup_{x \in [a,b] \setminus D(\mathcal{G})} F'(x \mid \mathcal{G}) \le \Gamma$.*

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

Proof. We adapt the proof of Theorem 1 of L'Ecuyer (1990). By the mean value inequality theorem of Dieudonné (1969), Theorem 8.5.3, which is a form of mean value 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, $\mathrm{Var}[F'(x \mid \mathcal{G})] \le \mathbb{E}[\Gamma^2] \le K_\gamma$. □

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

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

It follows from Theorem 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

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

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

It is well known that in general, to estimate $\mathbb{E}[X]$, 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

$$\mathrm{Var}[\mathbb{E}[X \mid \mathcal{G}]] \le \mathrm{Var}[\mathbb{E}[X \mid \tilde{\mathcal{G}}]] \le \mathrm{Var}[X]. \tag{3}$$

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, (3) applies as well to the (conditional) cdf estimator.

However, applying it to the CDE, which is the derivative of the conditional cdf, 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.

Theorem 2 *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 (3) because F' is not an expectation; this is why the 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 Theorem 1, and similarly for $\tilde{\mathcal{G}}$. Using (3) with I in place of X gives

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

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 (4), 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 Theorem 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 Theorem 1, and also to its square, which is also valid because the square is bounded uniformly by Γ^2. This completes the proof. □

When none of $\mathcal{G}$ or $\tilde{\mathcal{G}}$ is a subset of the other, the previous theorem does not help, but the variances of the corresponding conditional density estimators may differ significantly. Generally speaking, we want to choose $\mathcal{G}$ so that the density of X conditional on $\mathcal{G}$ is spread out rather than being concentrated in a narrow peak. We give simple examples of this, to provide insight.

Example 1 Asmussen (2018) studied the special case where X is a sum of d independent continuous random variables, $X = Y_1 + \cdots + Y_d$, where the Y_j have common cdf F_0 and density f_0, and $\mathcal{G}$ is obtained by hiding Y_d. That is, we take

$$\mathcal{G} = S_{-d} \stackrel{\text{def}}{=} \sum_{j=1}^{d-1} Y_j,$$

so that

$$F(x \mid \mathcal{G}) = \mathbb{P}[X \le x \mid S_{-d}] = \mathbb{P}[Y_d \le x - S_{-d}] = F_0(x - S_{-d})$$

and the density estimator becomes $F'(x \mid \mathcal{G}) = f_0(x - S_{-d})$.

This easily extends to the case where the Y_j's have different continuous distributions and $\mathcal{G}$ is defined by hiding Y_k for an arbitrary k. We take

$$\mathcal{G} = \mathcal{G}_k = S_{-k} \stackrel{\text{def}}{=} \sum_{j=1,\, j\neq k}^{d} Y_j.$$

If F_k and f_k are the cdf and density of Y_k, we have

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

and the density estimator becomes $F'(x \mid \mathcal{G}_k) = f_k(x - S_{-k})$. This idea also works if the Y_j are not independent; it suffices to replace f_k by the density of Y_k conditional on the other Y_j's that we have generated. Of course, it is practical only if the conditional density is easy to compute.

Example 2 *The previous example can be generalized as follows. Let $X = h(Y_1, \ldots, Y_d)$ where the Y_j are as in Example 1, with cdf F_j and density f_j, and we define $\mathcal{G}_k$ again by erasing a continuous Y_k, except that in general $\mathcal{G}_k$ is not a sum. 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.*

As a simple illustration, if $X = (Y_1 + Y_2^2)/Y_3$ where Y_3 is always positive, has a continuous distribution, and $k = 3$, then $F(x \mid \mathcal{G}_3) = \mathbb{P}(X \le x \mid Y_1, Y_2) = \mathbb{P}(Y_3 \ge (Y_1 + Y_2^2)/x) = 1 - F_3((Y_1 + Y_2^2)/x)$ and the density estimator at x is

$$F'(x \mid \mathcal{G}_3) = f_3((Y_1 + Y_2^2)/x)(Y_1 + Y_2^2)/x^2.$$

If $k = 2$ instead, then $F(x \mid \mathcal{G}_2) = \mathbb{P}(X \le x \mid Y_1, Y_3) = \mathbb{P}(|Y_2| \le (Y_3 x - Y_1)^{1/2}) = F_2(Z(x)) - F_2(-Z(x))$ where $Z(x) = (Y_3 x - Y_1)^{1/2}$, and the density estimator at x is

$$F'(x \mid \mathcal{G}_2) = (f_2(Z(x)) + f_2(-Z(x)))\mathrm{d}Z(x)/\mathrm{d}x = (f_2(Z(x)) - f_2(-Z(x)))Y_3/(2Z(x)).$$

This second estimator could be problematic if $Z(x)$ can take values near 0; this shows that a good choice of k can be crucial in general.

Example 3 *The following tiny 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 $Y_2 \sim \mathcal{U}(0, \epsilon)$ for some small $\epsilon > 0$. 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.*

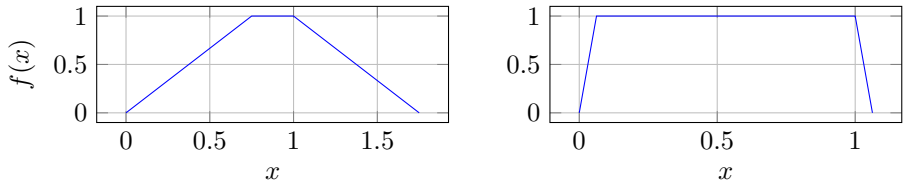

Figure 1: Exact density of X for the model in Example 3 with $\epsilon = 3/4$ (left) and $\epsilon = 1/16$ (right).

Using the notation of Example 1, 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, the density estimator with one sample is a uniform density, but the second one is over a narrow interval. Assumption 1 is easily

verified in both cases, since $F(\cdot \mid \mathcal{G}_j)$ is continuous and differentiable for $j = 1$ and 2. For $\mathcal{G}_1$ we can take $\Gamma = K_\gamma = 1$, and for $\mathcal{G}_2$ we can take $\Gamma = \epsilon^{-1}$ and $K_\gamma = \epsilon^{-2}$. When ϵ is small, $\mathcal{G} = \mathcal{G}_2$ gives a density estimator $\hat{f}_{\text{cde},n}$ which is a sum of high narrow peaks; it has much larger variance than with $\mathcal{G} = \mathcal{G}_1$.

The constants K_γ give upper bounds on the variance at any given x. 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 $\text{IV} = \epsilon/3$ for one sample. For a sample of size n, this gives $\text{IV} = \epsilon/(3n)$. 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 $\epsilon^{-1} f(x)(1-\epsilon f(x))$. Integrating over $[0, 1+\epsilon]$ gives $\text{IV} = 1/\epsilon - 1 + \epsilon/3$ for one sample.

Example 4 *This time X is the sum of two independent normal random variables $X = Y_1 + Y_2$ where $Y_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(0, \sigma_2^2)$, where $\sigma_1^2 + \sigma_2^2 = 1$, so $X \sim \mathcal{N}(0,1)$. 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 is easily verified with $\Gamma = \phi(0)/\sigma_2$ and $K_\gamma = \Gamma^2$, so this estimator is unbiased for $f(x) = \phi(x)$. Its variance is*

$$\begin{aligned}
\text{Var}[\phi((x-Y_1)/\sigma_2)/\sigma_2] &= \mathbb{E}[\exp[-(x-Y_1)^2/\sigma_2^2]/(2\pi\sigma_2^2)] - \phi^2(x) \\
&= \frac{1}{\sigma_2^2\sqrt{2\pi}} \mathbb{E}[\phi(\sqrt{2}(x-Y_1)/\sigma_2)] - \phi^2(x) \\
&= \frac{1}{\sigma_2\sqrt{2\pi(1+\sigma_1^2)}} \phi\left(\sqrt{2}x/\sqrt{1+\sigma_1^2}\right) - \phi^2(x). \qquad (5)
\end{aligned}$$

Example 5 *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 Assumption 1 does not hold. To illustrate this, let $X = \max(Y_1, Y_2)$ where Y_1 and Y_2 are two independent continuous random variables. 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 $x = y$.

Likewise, if $X = \min(Y_1, Y_2)$ instead, with $\mathcal{G} = \mathcal{G}_1$, we have

$$\mathbb{P}[X \le x \mid Y_1 = y] = \begin{cases} F_2(x) & \text{if } x < y; \\ 1 & \text{if } x \ge y, \end{cases}$$

Again, if $F_2(y) < 1$, this function is discontinuous at $x = y$, so Assumption 1 is not satisfied.

3 Combining RQMC with the CMC density estimator

In this section, we examine how we can use RQMC with the CDE, and under what conditions we can provably obtain a faster convergence rate for the IV of the resulting unbiased estimator. We first recall some basic facts about QMC and RQMC. Extensive coverages and tutorials can be found in Niederreiter (1992), Dick and Pillichshammer (2010), L'Ecuyer (2009), and L'Ecuyer (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

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

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

$$V_{\mathrm{HK}}(g) = \sum_{\emptyset \neq \mathfrak{v} \subseteq \mathcal{S}} \int_{[0,1]^{|\mathfrak{v}|}} |g_\mathfrak{v}(\mathbf{u}_\mathfrak{v}, \mathbf{1})| \, \mathrm{d}\mathbf{u}_\mathfrak{v}. \tag{6}$$

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:

$$|E_n| \leq V_{\mathrm{HK}}(g) \cdot D^*(P_n). \tag{7}$$

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

$$\mathrm{Var}[E_n] = \mathbb{E}[E_n^2] = \mathcal{O}(n^{-2+\epsilon}). \tag{8}$$

The definitions of variation and discrepancy given above are only one pair among a whole range of possibilities. In particular, there are versions of (7) with a discrepancy that converges as $\mathcal{O}(n^{-\alpha+\epsilon})$ for $\alpha > 1$, but with more restrictive definitions of the variation (stronger smoothness requirements on g). See Dick and Pillichshammer (2010).

To apply this theory and the variance bound (8) to the CDE, we must be able to write $F(x \mid \mathcal{G}) = \tilde{g}(x, \mathbf{u})$ and $F'(x \mid \mathcal{G}) = \tilde{g}'(x, \mathbf{u}) = \mathrm{d}g(x, \mathbf{u})/\mathrm{d}x$ for some function $\tilde{g} : [a, b] \times [0,1)^s$ for which $\tilde{g}'(x, \cdot)$ has bounded HK variation for each x. To prove the bounded variation, we need to take the partial derivative of $\tilde{g}'(x, \mathbf{u})$ with respect to each subset of coordinates of $\mathbf{u}$ and show that the integral of each such partial derivative is finite. To give an idea of how this can be done, we will do it for some examples. In many applications, this verification can be too complicated, and in some cases we may even know in advance that the HK variation is not bounded, but RQMC can still bring an important variance reduction, so it is often a good idea to just try it and see empirically.

Example 6 Consider a sum of random variables as in Example 1, with $\mathcal{G} = \mathcal{G}_k = S_{-k}$. Here, $\mathcal{G}_k$ is summarized in the single real number S_{-k}. We have $F(x \mid \mathcal{G}) = F_k(x - S_{-k})$ and $F'(x \mid \mathcal{G}) = f_k(x - S_{-k})$. Without loss of generality, let $k = d$. Suppose that each Y_j is generated by inversion from $U_j \sim \mathcal{U}(0,1)$, so $Y_j = F_j^{-1}(U_j)$ and $S_{-d} = F_1^{-1}(U_1) + \cdots + F_s^{-1}(U_s)$ with $s = d - 1$. This gives $\tilde{g}(x, \mathbf{u}) = F_d(x - S_{-d}) = F_d(x - F_1^{-1}(U_1) - \cdots - F_s^{-1}(U_s))$ and $\tilde{g}'(x, \mathbf{u}) = f_d(x - S_{-d}) = f_d(x - F_1^{-1}(U_1) - \cdots - F_s^{-1}(U_s))$. The partial derivatives of this last function are

$$\tilde{g}'_\mathfrak{v}(x, \mathbf{u}_\mathfrak{v}, \mathbf{1}) = f_d^{(|\mathfrak{v}|)}(x - S_{-d}) \prod_{j \in \mathfrak{v}} \frac{\partial (F_j^{-1}(U_j))}{\partial U_j}.$$

So the functions F_j^{-1} must be differentiable over $(0,1)$ for $j = 1, \ldots, d-1$, the density f_d must be s times differentiable, and the integral of $|\tilde{g}'_\mathfrak{v}(x, \mathbf{u}_\mathfrak{v}, \mathbf{1})|$ with respect to $\mathbf{u}_\mathfrak{v}$ must be finite.

For the special case of Example 3, with $\mathcal{G} = \mathcal{G}_2$ and $Y_1 = U_1$, we have $\tilde{g}'(x, \mathbf{u}) = \tilde{g}'(x, U_1) = \mathbb{I}[U_1 \leq x \leq \epsilon + U_1]/\epsilon = \mathbb{I}[x - \epsilon \leq U_1 \leq x]/\epsilon$. This function is not continuous, but its HK variation (not given by (6) in this case) is $2/\epsilon < \infty$. The behavior with $\mathcal{G} = \mathcal{G}_1$ is similar and the HK variation is 2.

For Example 4, if $\mathcal{G} = \mathcal{G}_2$, we have $Y_1 = \sigma_1\Phi^{-1}(U_1)$ where $U_1 \sim \mathcal{U}(0,1)$. Then, $F(x \mid \mathcal{G}_2) = F_2(x - Y_1) = \Phi((x - Y_1)/\sigma_2)$ and $F'(x \mid \mathcal{G}_2) = \phi((x - \sigma_1\Phi^{-1}(U_1))/\sigma_2)/\sigma_2 = \tilde{g}'(x, U_1)$. Taking the derivative with respect to u and noting that $\mathrm{d}\Phi^{-1}(u)/\mathrm{d}u = 1/(\phi(\Phi^{-1}(u)))$ yields

$$\tilde{g}'_{\mathfrak{v}}(x,u) = \frac{\phi'((x - \sigma_1\Phi^{-1}(u))/\sigma_2)\,\sigma_1}{\sigma_2^2\phi(\Phi^{-1}(u))}$$

for $\mathfrak{v} = \{1\} = \mathcal{S}$ (the only subset in this case). Integrating this with respect to u by making the change of variable $z = \Phi^{-1}(u)$ gives

$$\int_0^1 \tilde{g}'_{\mathfrak{v}}(x,u)\mathrm{d}u = \frac{\sigma_1}{\sigma_2^2}\int_{-\infty}^{\infty} |\phi'((x - \sigma_1 z)/\sigma_2)|\,\mathrm{d}z < \infty,$$

because $|\phi'(\cdot)|$ is bounded by $\phi(\cdot)$ multiplied by the absolute value of a polynomial of degree 1.

In the examples seen so far, $\mathcal{G}$ can be summarized by a single real number, but this is not always true in general and also not necessary.

4 Convex combination of conditional density estimators

When there are many possible choices of $\mathcal{G}$ for a given problem, one may try to pick the best one, but sometimes a better approach is to select more than one and take a convex linear combination of the corresponding CDEs as the final density estimator.

More specifically, suppose $\hat{f}_{0,n}, \ldots, \hat{f}_{q,n}$ are $q+1$ distinct unbiased density estimators. Typically, these estimators will be dependent and will be 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

$$\hat{f}_n(x) = \beta_0\hat{f}_{0,n}(x) + \cdots + \beta_q\hat{f}_{q,n}(x) = \hat{f}_{0,n}(x) - \sum_{\ell=1}^{q}\beta_\ell(\hat{f}_{0,n}(x) - \hat{f}_{\ell,n}(x)) \tag{9}$$

for all $x \in \mathbb{R}$, where $\beta_0 + \cdots + \beta_q = 1$. This combination 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. With this interpretation, the optimal coefficients β_ℓ can be estimated via standard control variate theory (Bratley et al. 1987, 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 easily, exactly as in standard least-squares linear regression, to obtain estimates of the optimal coefficients β_j. We did this for some of the examples in Section 5.

A more refined approach is to allow the coefficients β_j to depend on x:

$$\hat{f}_n(x) = \beta_0(x)\hat{f}_{0,n}(x) + \cdots + \beta_q(x)\hat{f}_{q,n}(x) = \hat{f}_{0,n}(x) - \sum_{\ell=1}^{q}\beta_\ell(x)(\hat{f}_{0,n}(x) - \hat{f}_{\ell,n}(x)), \tag{10}$$

where $\beta_0(x) + \cdots + \beta_q(x) = 1$ for all $x \in \mathbb{R}$. We can estimate the optimal coefficients by standard control variate theory at selected values of x, then for each $\ell \ge 1$, we 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 in a different setting. We did not implement it for this paper, but the additional flexibility can provide further variance reduction in general.

5 Numerical experiments

We now report on our numerical experiments with the CDE and CDE+RQMC on examples.

5.1 Experimental setting

Since the CDE is unbiased, we will measure its performance by the IV, which equals the MISE in this case. To approximate the IV estimator (2) 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 each evaluation point e_j. 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.

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 a power of 2 in our experiments, the logarithms are always in base 2. In our experiments, we estimated the IV for $n = 2^{14}, \ldots, 2^{19}$ (6 values) to fit the regression model. When reporting the results, 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. For the KDE, these values are for the MISE instead of the IV.

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); and (4) a Sobol' point set with a left random matrix scrambling and random digital shift (Sobol'+LMS). The short names in parentheses are used in the plots and tables. For the definitions are properties of these RQMC point sets, see Hickernell (1998), 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 $\gamma_{\mathfrak{u}} = \rho^{|\mathfrak{u}|}$, with ρ ranging from 0.05 to 0.8, depending on the example (a smaller ρ was used when the dimension s was smaller).

For each example, we tried CDEs based on different choices of $\mathcal{G}$. For some examples, we also tried a convex combination as explained in Section 4.

5.2 A Sum of normals

We start with a very simple example in which the density f is known beforehand, so there is no real need to estimate it, but this type of example is very convenient for testing the performance of our density estimators. Let $Z_1, \ldots, Z_d$ be independent standard normal random variables, i.e., with mean 0 and variance 1, and define

$$X = (a_1 Z_1 + \cdots + a_d Z_d)/\sigma, \quad \text{where} \quad \sigma^2 = a_1^2 + \cdots + a_d^2.$$

Then X is also standard normal, with density $f(x) = \phi(x) \stackrel{\text{def}}{=} \exp(-x^2/2)/\sqrt{2\pi}$ and cdf $\mathbb{P}[X \le x] = \Phi(x)$ for $x \in \mathbb{R}$. The term $a_j Z_j$ in the sum has variance a_j^2. We pretend that we do not know this and want to estimate $f(x)$ over the interval $[-2, 2]$, which contains slightly more than 95% of the density. We also tried larger intervals, such as $[-5, 5]$, and IVs for the CDE were almost the same.

To construct the CDE, we can define $\mathcal{G}_k$ as in Example 1, for any $k = 1, \ldots, d$. That is, we hide Z_k and estimate the cdf by

$$F(x \mid \mathcal{G}_k) = \mathbb{P}\left[a_k Z_k \le x\sigma - \sum_{j=1,\, j\ne k}^{d} a_j Z_j \,\middle|\, \mathcal{G}_k \right] = \Phi\left(\frac{x\sigma}{a_k} - \frac{1}{a_k} \sum_{j=1,\, j\ne k}^{d} a_j Z_j \right).$$

The CDE becomes

$$F'(x \mid \mathcal{G}_k) = \phi\left(\frac{x\sigma}{a_k} - \frac{1}{a_k} \sum_{j=1,\, j\ne k}^{d} a_j Z_j \right) \frac{\sigma}{a_k} = \phi\left(\frac{x\sigma}{a_k} - \frac{1}{a_k} \sum_{j=1,\, j\ne k}^{d} a_j \Phi^{-1}(U_j) \right) \frac{\sigma}{a_k} = \tilde{g}'(x, \mathbf{U})$$

for $x \in \mathbb{R}$, where $\mathbf{U} = (U_1, \ldots, U_{k-1}, U_{k+1}, \ldots, U_d)$, $Z_j = \Phi^{-1}(U_j)$, and the U_j are independent $\mathcal{U}(0,1)$ random variables. With this, Assumption 1 is easily verified.

For CMC+MC (independent sampling), we can obtain an exact formula for the variance of the CDE directly from Example 4, by taking in that example $Y_2 = a_k Z_k/\sigma$ and $Y_1 = X - Y_2$, whose variances are $\sigma_2^2 = (a_k/\sigma)^2$ and $\sigma_1^2 = 1 - \sigma_2^2$, and by plugging these values into (5). To prove that we have a better convergence rate with RQMC than with MC, we need to show that $V_{\text{HK}}(\tilde{g}'(x, \cdot) < \infty$ for any x. This can be shown by the same argument as in the last part of Example 6. Then we expect to observe a convergence rate near $\mathcal{O}(n^{-2})$, at least when s is small.

In our first experiment, we take $a_j = 1$ for all j, and $k = d$, as in Example 1. By symmetry, the true IV is the same for any other k. Table 1 reports the estimated rate $\hat{\nu}$ and the estimated value of $\text{e19} = -\log_2(\text{IV})$ for $n = 2^{19}$, for various values of d and selected sampling methods, for the CDE. The results for $k = d$ are in the rows marked CDE-1. We also tried computing the CDE for each choice of k from the same simulation data, and taking the average, which is the convex combination (9) with $\beta_\ell = 1/(q+1)$ for all ℓ. The results for this are in the rows labeled CDE-Avg in the table. Finally, for comparison, we also provide results for a KDE with a normal kernel and the bandwidth h selected using the methodology described in Ben Abdellah et al. (2019), with standard MC. For the KDE, the results are for the MISE instead of the IV.

Table 1: Values of $\hat{\nu}$ and e19 for a CDE, and a convex combination of CDEs, and a KDE, for a sum of $d = k$ normals with $a_j = 1$, over $[-2, 2]$.

		$\hat{\nu}$					e19				
		$d=2$	$d=3$	$d=5$	$d=10$	$d=20$	$d=2$	$d=3$	$d=5$	$d=10$	$d=20$
CDE-1	MC	0.99	0.98	1.02	1.00	1.02	22.1	21.4	20.8	19.8	19.2
	Lat+s	2.83	2.00	1.85	1.40	1.04	52.3	39.8	32.1	23.6	19.7
	Lat+s+b	2.69	2.11	1.69	1.14	1.05	50.5	41.5	31.1	21.8	20.0
	Sob+LMS	2.62	2.10	1.81	1.04	1.04	49.3	40.7	31.1	21.3	19.7
CDE-avg	MC	1.06	0.92	1.03	1.01	1.01	23.4	22.1	21.6	20.6	19.8
	Lat+s	2.79	1.84	1.33	1.19	1.05	53.3	39.8	32.2	23.0	20.6
	Lat+s+b	2.65	1.90	1.71	1.05	1.08	51.6	41.4	32.3	23.4	21.3
	Sob+LMS	2.60	2.10	1.92	1.02	1.03	49.8	42.0	33.0	22.7	20.5
KDE	MC	0.79	0.80	0.76	0.75	0.77	17.0	17.0	16.9	16.9	17.0

For MC, the estimated rates $\hat{\nu}$ agree with the (known) exact asymptotic ones, which are $\nu = 1$ for the CDE and $\nu = 0.8$ for the KDE. By looking at e19, we see that the MISE is smaller for the CDE than for the KDE, for example for $d = 2$ by a factor of about 32 for CDE-1 and about 70 for CDE-avg. For $d = 20$, the gains are more modest, with a factor between 4 and 8. The RQMC methods provide huge improvements for small d. We observe rates $\hat{\nu}$ larger than 2 for $d = 2$ and 3, and by looking at the

exponents e19, we see that for $d = 3$, for example, the MISE goes from 2^{-17} for the KDE to about 2^{-42} for CDE-1 with Sobol' points with LMS. This is a factor of about $2^{25} \approx 32$ millions. When d is large, such as $d = 20$, RQMC brings only a small gain.

Table 2: Values of $\hat{\nu}$ and e19 with a CDE for selected choices of $\mathcal{G}_k$, for a linear combination of $d = 11$ normals with $a_j^2 = 2^{1-j}$.

	$\hat{\nu}$				e19			
	$k=1$	$k=2$	$k=5$	$k=11$	$k=1$	$k=2$	$k=5$	$k=11$
MC	1.00	1.02	1.01	1.00	22.2	21.0	18.8	15.5
Lat+s	1.43	1.48	1.34	1.04	30.3	28.5	22.8	15.6
Lat+s+b	1.57	1.65	1.28	1.02	33.5	30.8	22.1	15.6
Sob+LMS	1.78	1.56	1.21	1.02	34.1	30.4	21.7	15.7

In our second experiment, we take $a_j^2 = 2^{1-j}$ for $j = 1, \ldots, d$. Now, the choice of k for the CDE makes a difference, and the best choice will obviously be $k = 1$, i.e., hide the term that has the largest variance. Note that with MC, $\mathrm{Var}[X] = 2 - 2^{-d}$, and when we apply CMC by hiding $a_k Z_k$ from the sum, we hide a term of variance $a_k^2 = 2^{1-k}$ and generate a partial sum S_{-k} of variance $2 - 2^{1-k} - 2^{-d}$. Both terms have a normal distribution with mean 0. The results of Example 4 hold with these variances. Table 2 reports the numerical results for $d = 11$ and $k = 1, 2, 5, 11$.

The MC rates $\hat{\nu}$ agree again with the theory, but here the IV depends very much on the choice of k, and this effect is more significant when k is smaller. For example, for Sobol' points, the IV with $k = 1$ is about 300,000 times smaller than with $k = 11$. The reason is that with $k = 11$, we hide only a variable having a very small variance, so the CDE for one sample is a high narrow peak, and the HK variation of $\tilde{g}'(x, \mathbf{u})$ is very large. For $k = 1$ or 2, we have the opposite and the integrand is much more RQMC-friendly.

5.3 Displacement of a cantilever beam

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

$$X = h(Y_1, Y_2, Y_3) = \frac{4\ell^3}{Y_1 w t}\sqrt{\frac{Y_2^2}{w^4} + \frac{Y_3^2}{t^4}} \tag{11}$$

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. This example fits the framework of Example 2, 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

$$F'(x \mid \mathcal{G}_1) = -\phi((W_1(x) - \mu_1)/\sigma_1)W_1'(x)/\sigma_1 = \phi((W_1(x) - \mu_1)/\sigma_1)W_1(x)/(x\sigma_1).$$

Suppose we condition on $\mathcal{G}_2$ instead, i.e., 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

$$F'(x \mid \mathcal{G}_2) = \frac{\phi((\sqrt{W_2(x)} - \mu_2)/\sigma_2) + \phi(-(\sqrt{W_2(x)} + \mu_2)/\sigma_2)}{(\sigma_2\sqrt{W_2(x)})/(w^4 x (Y_1/\kappa)^2)} > 0.$$

For conditioning 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 $W_2(x)$.

Table 3: Values of $\hat{\nu}$ and e19 with a CDE for each choice of $\mathcal{G}_k$ and for the best convex combination, for the cantilever beam model.

	$\hat{\nu}$				e19			
	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	comb.	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	comb.
MC	0.97	0.98	0.99	0.98	19.3	14.5	22.8	22.5
Lat+s	1.99	1.95	2.06	2.04	39.8	25.2	41.6	41.9
Lat+s+b	2.24	2.08	2.27	2.25	44.5	23.7	46.8	47.0
Sob+LMS	2.21	2.03	2.21	2.21	44.0	23.6	45.7	46.1

In addition to testing these three ways of conditioning, we also tested a convex combination of the three, as explained in Section 4, with coefficients β_ℓ that do not depend on x. Table 3 gives a summary of the results. For comparison, a KDE with MC gives $\hat{\nu} = 0.8$ and e19 $= 14.7$. Thus, the MISE is about $2^{-14.7}$ for KDE+MC, while it is around 2^{-46} for the CDE+RQMC; i.e., our new approach divides the MISE by more than $2^{31} \approx 2$ millions. The empirical convergence rate $\hat{\nu}$ is around 2 in all cases with the RQMC methods. We also observe that conditioning on $\mathcal{G}_2$ yields a much poorer IV reduction than for the other choices. The reason is that the conditional density in this case is a high narrow peak, similar to what we saw for $k = 11$ at the end of Example 5.2. To provide visual insight on what goes on, Figure 2 shows plots of five realizations of the conditional density for $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$. One can see the high narrow peaks for the realizations of $F'(\cdot \mid \mathcal{G}_2)$. Te average of the five realizations is shown in red and the true density in black. In Figure 3, we zoom 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), using RQMC (in green), as well as the "true density" (black, dashed), which was estimated with RQMC using a very large number of samples. On the left, we have $\mathcal{G}_1$ with $n = 2^{10}$, while on the right, we have $\mathcal{G}_2$ with $n = 2^{16}$. In both cases, the RQMC estimate is closer to the true density, and on the right it oscillates less. If we repeat this experiment several times, the red curve would vary much more than the green one across the realizations.

5.4 Buckling strength of a steel plate

This is a higher-dimensional example, with $d = 6$, taken from Schields and Zhang (2016). It models the buckling strength of a steel plate by

$$X = \left(\frac{2.1}{\Lambda} - \frac{0.9}{\Lambda^2}\right)\left(1 - \frac{0.75Y_5}{\Lambda}\right)\left(1 - \frac{2Y_6Y_2}{Y_1}\right), \tag{12}$$

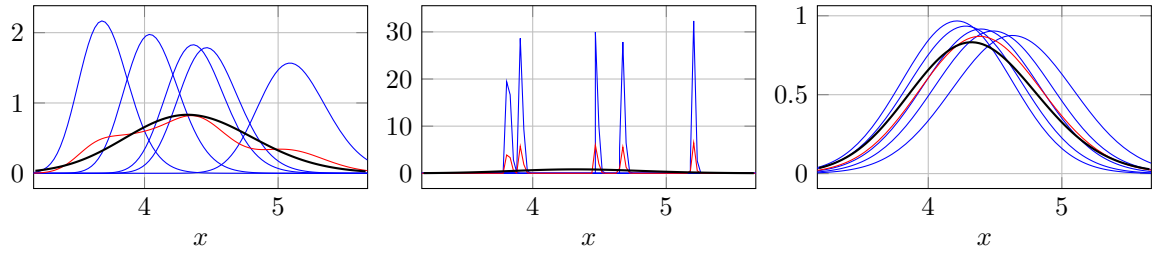

Figure 2: Five realizations of the density conditional on $\mathcal{G}_k$ (blue), their average (red), and the true density (thick black) for $k = 1$ (left), $k = 2$ (middle), and $k = 3$ (right).

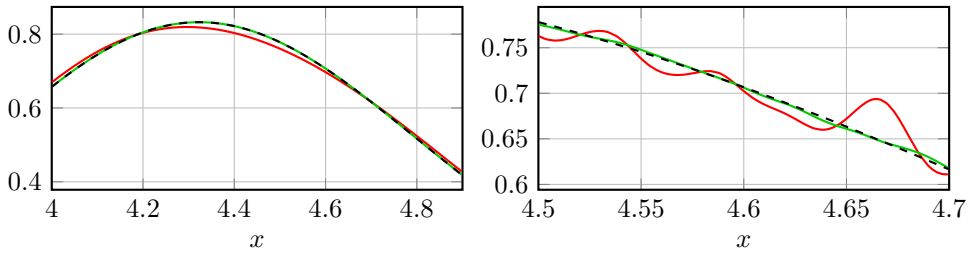

Figure 3: The CDE under MC (red), under RQMC (green) and the true density (black, dashed) for $\mathcal{G}_1$ with $n = 2^{10}$ (left) and for $\mathcal{G}_2$ with $n = 2^{16}$ (right).

where $\Lambda = (Y_1/Y_2)\sqrt{Y_3/Y_4}$, and $Y_1, \ldots, Y_6$ are independent random variables whose distributions are given in Table 4. Each distribution is either normal or lognormal, and the table gives the mean and the coefficient of variation (cv), which is the standard deviation divided by the mean. The goal is to estimate the density of X over $[a, b] = [0.5169, 0.6511]$, which contains about 99% of the density (leaving out 0.5% on each side). There is a nonzero probability of having $Y_4 \le 0$, in which case X is undefined, but this probability is extremely small and this has a negligible impact on the density estimator over $[a, b]$, so we just ignore it (alternatively we could truncate the density of Y_4). There are also tiny (negligible) probabilities that the density estimates below are negative and we ignore this in the same way, for simplicity.

Table 4: Distribution of each parameter for the buckling strength model.

parameter	distribution	mean	cv
Y_1	normal	23.808	0.028
Y_2	lognormal	0.525	0.044
Y_3	lognormal	44.2	0.1235
Y_4	normal	28623	0.076
Y_5	normal	0.35	0.05
Y_6	normal	5.25	0.07

For this example, computing the density of X conditional on $\mathcal{G}_5$ or $\mathcal{G}_6$ (i.e., when hiding Y_5 or Y_6) is relatively easy, so we will try and compare these two choices. If we hide one of the variables that appear in Λ, the CDE would be harder to compute (it would require to solve a polynomial equation of degree 4 for each sample), and we do not do it. Let us define

$$V_1 = \frac{2.1}{\Lambda} - \frac{0.9}{\Lambda^2}, \qquad V_2 = 1 - \frac{2Y_6Y_2}{Y_1}, \quad \text{and} \quad V_3 = 1 - \frac{3Y_5}{4\Lambda}.$$

Then we have

$$X \le x \quad \Leftrightarrow \quad Y_5 \ge \left(1 - \frac{x}{V_1V_2}\right)\frac{4\Lambda}{3}$$

and

$$F'(x \mid \mathcal{G}_5) = f_5\left(\left(1 - \frac{x}{V_1V_2}\right)\frac{4\Lambda}{3}\right)\frac{4\Lambda}{3V_1V_2} = \phi\left(\frac{(1 - x/(V_1V_2))\,4\Lambda/3 - 0.35}{0.0175}\right)\frac{4\Lambda}{0.0525 \cdot V_1V_2}.$$

Similarly,

$$F'(x \mid \mathcal{G}_6) = f_6\left(\left(1 - \frac{x}{V_1V_3}\right)\frac{Y_1}{2Y_2}\right)\frac{Y_1}{2Y_2V_1V_3} = \phi\left(\frac{(1 - x/(V_1V_3))\,Y_1/(2Y_2) - 5.25}{0.3675}\right)\frac{Y_1}{0.735 \cdot Y_2V_1V_3}.$$

Table 5 summarizes the results. We see again that the CDE with RQMC performs very well, and that it is much better to condition on $\mathcal{G}_6$ than on $\mathcal{G}_5$, and that combining the two provides no significant improvement. Figure 4 permits one to visualize the IV as a function of n in a log-log-scale for the CDE with $\mathcal{G}_5$ and $\mathcal{G}_6$. It unveils a slightly more erratic behavior of the MISE for the shifted lattice rule (Lat+s) than for the other methods; the performance depends on the choice of parameters of the lattice rule and their interaction with the particular integrand.

Table 5: Values of $\hat{\nu}$ and e19 with a CDE for $\mathcal{G}_5$, $\mathcal{G}_6$, and their combination, for the buckling strength model.

	$\hat{\nu}$			e19		
	$\mathcal{G}_5$	$\mathcal{G}_6$	comb.	$\mathcal{G}_5$	$\mathcal{G}_6$	comb.
MC	1.00	1.00	1.00	13.5	15.4	15.4
Lat+s	1.89	1.58	1.58	20.0	24.9	24.9
Lat+s+b	1.46	1.60	1.60	17.5	25.1	25.1
Sob+LMS	1.40	1.75	1.75	17.7	25.5	25.5

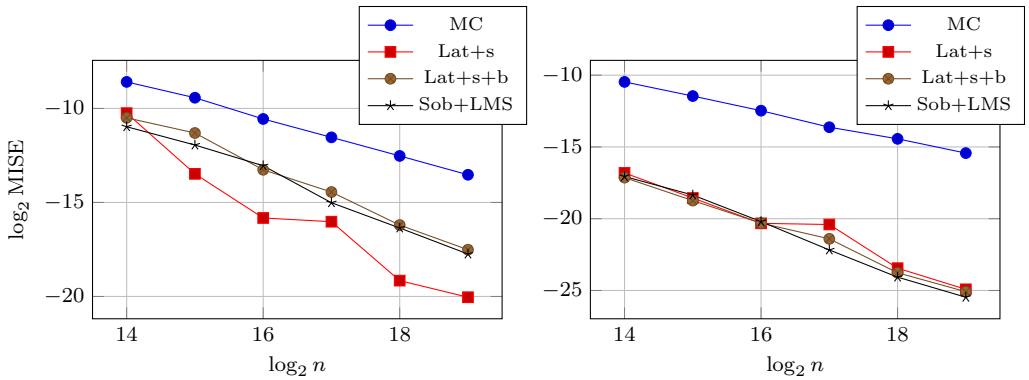

Figure 4: MISE vs n in log-log scale for the $\mathcal{G} = \mathcal{G}_5$ (left) and $\mathcal{G} = \mathcal{G}_6$ (right) for the buckling strength model.

5.5 A Stochastic activity network

In this example, the conditioning must hide more than one random variable. The example is taken from Avramidis and Wilson (1998) and was also used in L'Ecuyer and Lemieux (2000) and L'Ecuyer and Munger (2012). We consider the stochastic activity network of Figure 5, where arc j has random length Y_j for $j = 1, \ldots, 13$. The Y_j are independent with continuous cdf's F_j as given in L'Ecuyer and Munger (2012), and are generated by inversion: $Y_j = F_j^{-1}(U_j)$ where $U_j \sim U(0,1)$. The goal is to estimate the density of X which represents the length of the longest path from the source to the sink, over $[a, b] = [22, 106.24]$, which covers about 95% of the density.

Here, X is defined by a maximum (over all paths from 0 to 8), and if we hide only a single random variable Y_j, we run into the same problem as in Example 5: Assumption 1 does not hold, because $F(\cdot \mid \mathcal{G})$ has a jump. We must hide more (condition on less). We apply the conditioning proposed by Avramidis and Wilson (1998), for which we hide $\{Y_j,\, j \in \mathcal{L} \stackrel{\text{def}}{=} \{5, 6, 7, 9, 10\}\}$, and let $\mathcal{G}$ represent $\{Y_j,\, j \notin \mathcal{L}\}$. The corresponding conditional cdf is

$$F(x \mid \mathcal{G}) = \mathbb{P}\left[X < x \mid \{Y_j : j \notin \mathcal{L}\}\right] = \prod_{j \in \mathcal{L}} \mathbb{P}[Y_j \le x - P_j] = \prod_{j \in \mathcal{L}} F_j[x - P_j] \tag{13}$$

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, since the Y_j's are continuous variables with bounded variance, one can easily verify that Assumption 1 holds, so $F'(x \mid \mathcal{G})$ is an unbiased density estimator with uniformly bounded variance.

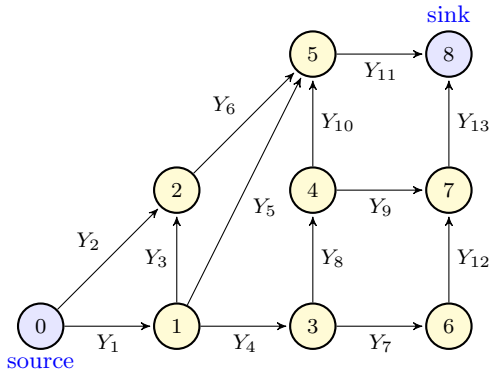

Figure 5: A stochastic activity network.

Table 6 summarizes 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 $2^9 \approx 500$ with RQMC. Figure 6 complements the table by showing the MISE as a function of n in a log-log-scale, for the same methods.

Table 6: Values of $\hat{\nu}$ and e19 with a CDE, 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

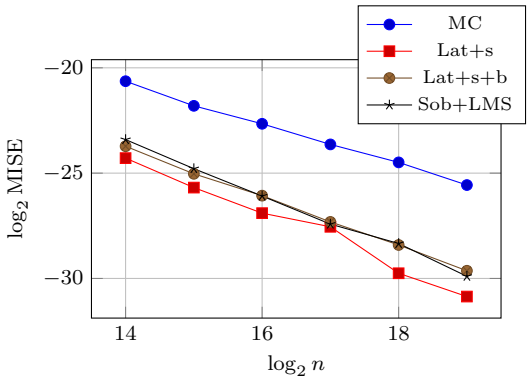

Figure 6: MISE vs n in log-log scale, for the SAN example.

6 Conclusion

We have examined a novel approach for estimating the density of a random variable generated by simulation from a stochastic model, by conditioning. The resulting CDE is unbiased and consequently its MISE converges at a faster rate 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 sampling. Our numerical examples show that this combination can be very efficient. It sometimes reduces the MISE by factors over a million (cf. the cantilever example).

Suggested future work includes experimenting this methodology on larger and more complicated stochastic models, designing and exploring different types of conditioning, and perhaps adapting the Monte Carlo sampling strategies (e.g., changing the way X is defined in terms the basic input random variates) to make the method more effective.

References

Asmussen, S., P. W. Glynn. 2007. Stochastic Simulation. Springer-Verlag, New York.

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

Avramidis, A. N., J. R. Wilson. 1998. Correlation-induction techniques for estimating quantiles in simulation experiments. Operations Research 46(4) 574–591.

Ben Abdellah, A., P. L'Ecuyer, A. Owen, F. Puchhammer. 2019. Density estimation by randomized quasi-Monte Carlo. Submitted manuscript.

Bingham, Derek. 2017. Virtual library of simulation experiments. https://www.sfu.ca/~ssurjano/canti.html.

Bratley, P., B. L. Fox, L. E. Schrage. 1987. A Guide to Simulation. 2nd ed. Springer-Verlag, New York, NY.

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

Dieudonné, J. 1969. Foundations of Modern Analysis. 2nd ed. Academic Press, New York.

Fu, M., J.-Q. Hu. 1997. Conditional Monte Carlo. Kluwer Academic, Boston.

Fu, Michael C., L. Jeff Hong, Jian-Qiang Hu. 2009. Conditional Monte Carlo estimation of quantile sensitivities. Management Science 55(12) 2019–2027.

Hickernell, F. J. 1998. Lattice rules: How well do they measure up? P. Hellekalek, G. Larcher, eds., Random and Quasi-Random Point Sets, Lecture Notes in Statistics, vol. 138. Springer-Verlag, New York, 109–166.

Laub, Patrick J., Robert Salomone, Zdravko I. Botev. 2019. Monte Carlo estimation of the density of the sum of dependent random variables. Mathematics and Computers in Simulation 161 23–31.

Law, A. M. 2014. Simulation Modeling and Analysis. 5th ed. McGraw-Hill, New York.

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/.

L'Ecuyer, P. 2018. Randomized quasi-Monte Carlo: An introduction for practitioners. P. W. Glynn, A. B. Owen, eds., Monte Carlo and Quasi-Monte Carlo Methods: MCQMC 2016. Springer, Berlin, 29–52.

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

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

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

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

Owen, A. B. 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.

Schields, M. D., J. Zhang. 2016. The generalization of latin hypercube sampling. Reliability Engineering and System Safety 148 96–108.

Scott, David W. 2015. Multivariate Density Estimation. 2nd ed. Wiley.