

Monte Carlo and Quasi-Monte Carlo Density Estimation

Pierre L'Ecuyer

Joint work with **Amal Ben Abdellah, Art B. Owen, and Florian Puchhammer**



CORS/INFORMS 2022

Vancouver, June 2022

What this talk is about

Monte Carlo (MC) simulation is widely used to estimate the expectation $\mathbb{E}[X]$ of a random variable X and compute a confidence interval on $\mathbb{E}[X]$. $\text{MSE} = \text{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

But simulation usually provides information to do much more! The output data can be used to estimate the entire distribution of X , e.g., the cumulative distribution function (cdf) F of X , defined by $F(x) = \mathbb{P}[X \leq x]$, or its density f defined by $f(x) = F'(x)$.

What this talk is about

Monte Carlo (MC) simulation is widely used to estimate the expectation $\mathbb{E}[X]$ of a random variable X and compute a confidence interval on $\mathbb{E}[X]$. $\text{MSE} = \text{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

But simulation usually provides information to do much more! The output data can be used to estimate the entire distribution of X , e.g., the cumulative distribution function (cdf) F of X , defined by $F(x) = \mathbb{P}[X \leq x]$, or its density f defined by $f(x) = F'(x)$.

If X_1, \dots, X_n are n indep. realizations of X , the empirical cdf

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

is unbiased for $F(x)$ at all x , and $\text{Var}[\hat{F}_n(x)] = \mathcal{O}(n^{-1})$.

What this talk is about

Monte Carlo (MC) simulation is widely used to estimate the expectation $\mathbb{E}[X]$ of a random variable X and compute a confidence interval on $\mathbb{E}[X]$. $\text{MSE} = \text{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

But simulation usually provides information to do much more! The output data can be used to estimate the entire distribution of X , e.g., the cumulative distribution function (cdf) F of X , defined by $F(x) = \mathbb{P}[X \leq x]$, or its density f defined by $f(x) = F'(x)$.

If X_1, \dots, X_n are n indep. realizations of X , the empirical cdf

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

is unbiased for $F(x)$ at all x , and $\text{Var}[\hat{F}_n(x)] = \mathcal{O}(n^{-1})$.

For a continuous r.v. X , the density f provides a better visual idea of the distribution.

Here we focus on estimating the density f of X over $[a, b] \subset \mathbb{R}$.

(Density is with respect to Lebesgue measure.)

Setting

Classical **density estimation** in statistics was developed in the context where X_1, \dots, X_n are **given independent observations of X** and one estimates the **density f of X** from that.

Leading method: **kernel density estimator (KDE)**; $\text{MSE}[\hat{f}_n(x)] = \mathcal{O}(n^{-4/5})$.

Setting

Classical **density estimation** in statistics was developed in the context where X_1, \dots, X_n are **given independent observations of X** and one estimates the **density f of X** from that.

Leading method: **kernel density estimator** (KDE); $\text{MSE}[\hat{f}_n(x)] = \mathcal{O}(n^{-4/5})$.

In this talk, we assume that X_1, \dots, X_n are **generated by simulation** from a model.

We can choose n and we have some freedom on how the simulation is performed.

Setting

Classical **density estimation** in statistics was developed in the context where X_1, \dots, X_n are **given independent observations of X** and one estimates the **density f of X** from that.

Leading method: **kernel density estimator (KDE)**; $\text{MSE}[\hat{f}_n(x)] = \mathcal{O}(n^{-4/5})$.

In this talk, we assume that X_1, \dots, X_n are **generated by simulation** from a model.

We can choose n and we have some freedom on how the simulation is performed.

Questions:

1. Is it possible to obtain **unbiased density estimators** whose variance converges as $\mathcal{O}(n^{-1})$ or better, using clever sampling strategies? How?
2. **How can we benefit from RQMC to estimate density?** Can we improve the convergence rate of the error?

Small example: a stochastic activity network (SAN)

Precedence relations between activities. Activity k has random duration Y_k (length of arc k) with known cdf $F_k(y) := \mathbb{P}[Y_k \leq y]$.

Project duration X = (random) length of longest path from source to sink.

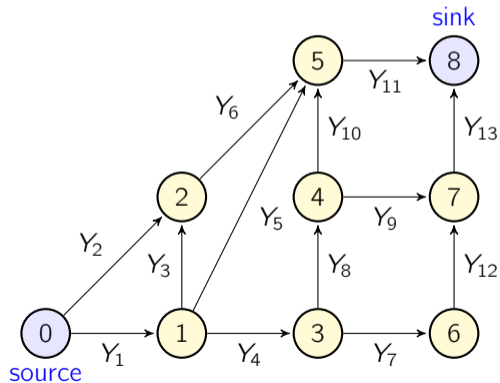
Specific case (Avramidis and Wilson 1998):

$Y_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, 2, 4, 11, 12$;

$Y_k \sim \text{Expon}(1/\mu_k)$ otherwise;

μ_1, \dots, μ_{13} : 13.0, 5.5, 7.0, 5.2, 16.5, 14.7,
10.3, 6.0, 4.0, 20.0, 3.2, 3.2, 16.5;

$\sigma_k = \mu_k/4$.



Small example: a stochastic activity network (SAN)

Precedence relations between activities. Activity k has random duration Y_k (length of arc k) with known cdf $F_k(y) := \mathbb{P}[Y_k \leq y]$.

Project duration $X =$ (random) length of longest path from source to sink.

Specific case (Avramidis and Wilson 1998):

$Y_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, 2, 4, 11, 12$;

$Y_k \sim \text{Expon}(1/\mu_k)$ otherwise;

μ_1, \dots, μ_{13} : 13.0, 5.5, 7.0, 5.2, 16.5, 14.7,

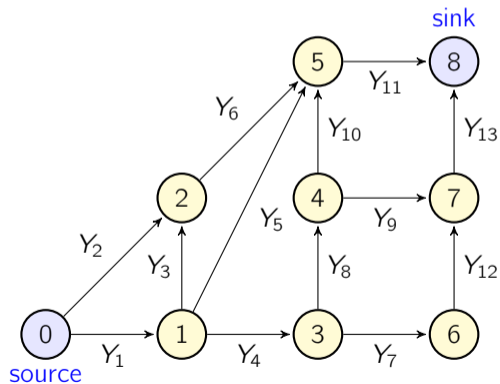
10.3, 6.0, 4.0, 20.0, 3.2, 3.2, 16.5;

$\sigma_k = \mu_k/4$.

Mean $\mathbb{E}[X] \approx 64.2$

C.I. for $\mathbb{E}[X]$ could be, e.g., [64.05, 64.32].

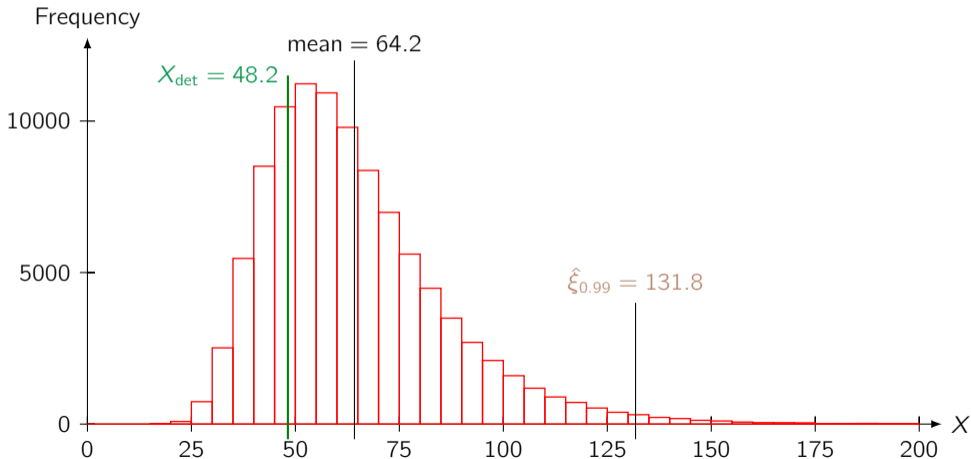
That's all?



Results of an experiment with $n = 100\,000$ independent runs.

The histogram gives an idea of the density of X .

Much more information than a C.I. on $\mathbb{E}[X]$. Can we do better than a histogram?



Density estimation

Want to **estimate the density** of $X = h(\mathbf{Y}) = h(Y_1, \dots, Y_s)$, assuming we know how to get Monte Carlo samples of \mathbf{Y} from its multivariate distribution.

Suppose we **estimate the density f over a finite interval $[a, b]$** .

Let $\hat{f}_n(x)$ denote the density estimator at x , with sample size n .

Density estimation

Want to **estimate the density** of $X = h(\mathbf{Y}) = h(Y_1, \dots, Y_s)$, assuming we know how to get Monte Carlo samples of \mathbf{Y} from its multivariate distribution.

Suppose we **estimate the density f over a finite interval $[a, b]$** .

Let $\hat{f}_n(x)$ denote the density estimator at x , with sample size n .

We use simple error measures:

$$\text{MISE} = \text{mean integrated squared error} = \int_a^b \mathbb{E}[(\hat{f}_n(x) - f(x))^2] dx = \text{IV} + \text{ISB}$$

$$\text{IV} = \text{integrated variance} = \int_a^b \text{Var}[\hat{f}_n(x)] dx$$

$$\text{ISB} = \text{integrated squared bias} = \int_a^b (\mathbb{E}[\hat{f}_n(x)] - f(x))^2 dx$$

To minimize the MISE, we may need to **balance** the IV and ISB.

Density estimation

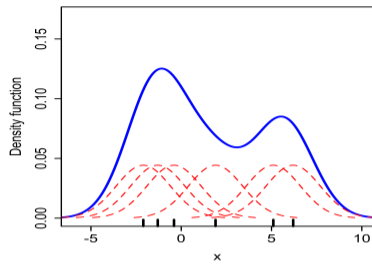
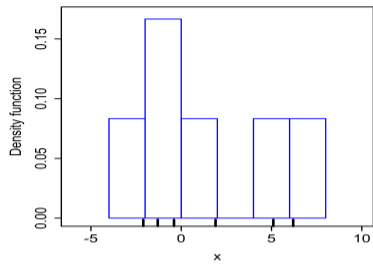
Histogram: Partition $[a, b]$ in m intervals of size $h = (b - a)/m$ and define

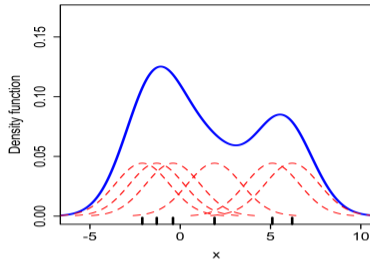
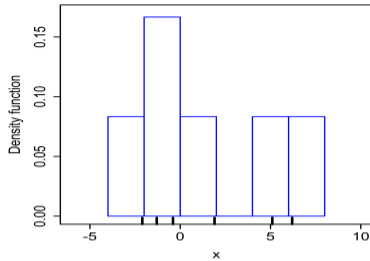
$$\hat{f}_n(x) = \frac{n_j}{nh} \text{ for } x \in I_j = [a + (j - 1)h, a + jh), \quad j = 1, \dots, m$$

where n_j is the number of observations X_i that fall in interval j .

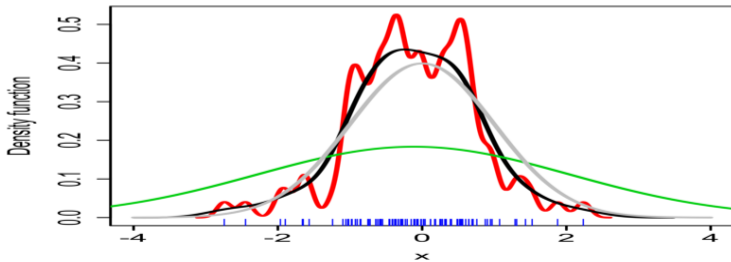
Kernel Density Estimator (KDE) : Select kernel k (unimodal symmetric density centered at 0) and **bandwidth** $h > 0$ (horizontal stretching factor for the kernel). The KDE is

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

$n = 6$ 



$n = 6$



$n = 100$

Credit: Drleft at English Wikipedia / CC BY-SA (<https://creativecommons.org/licenses/by-sa/3.0>)

<https://commons.wikimedia.org/w/index.php?curid=73892711>

https://commons.wikimedia.org/wiki/File:Comparison_of_1D_histogram_and_KDE.png

Asymptotic convergence with Monte Carlo for smooth f

Here we assume independent random samples (Monte Carlo or given data).

For **histograms and KDEs**, when $n \rightarrow \infty$ and $h \rightarrow 0$:

$$\text{AMISE} = \text{AIV} + \text{AISB} \sim \frac{C}{nh} + Bh^\alpha.$$

The asymptotically optimal h is

$$h^* = \left(\frac{C}{B\alpha n} \right)^{1/(\alpha+1)}$$

and it gives $\text{AMISE} = Kn^{-\alpha/(1+\alpha)}$.

For any $g : \mathbb{R} \rightarrow \mathbb{R}$, define

$$R(g) = \int_a^b (g(x))^2 dx, \quad \mu_r(g) = \int_{-\infty}^{\infty} x^r g(x) dx, \quad \text{for } r = 0, 1, 2, \dots$$

	C	B	α	h^*	AMISE
Histogram	1	$\frac{R(f')}{12}$	2	$(nR(f')/6)^{-1/3}$	$\mathcal{O}(n^{-2/3})$
KDE	$\mu_0(k^2)$	$\frac{(\mu_2(k))^2 R(f'')}{4}$	4	$\left(\frac{\mu_0(k^2)}{(\mu_2(k))^2 R(f'') n} \right)^{1/5}$	$\mathcal{O}(n^{-4/5})$

To estimate h^* , one can estimate $R(f')$ and $R(f'')$ via KDE (plugin).

This is true under the simplifying [assumption that \$h\$ must be the same all over \$\[a, b\]\$](#) .

One may also [vary the bandwidth](#) over $[a, b]$.

Randomized quasi-Monte Carlo (RQMC)

Suppose $X = h(\mathbf{Y}) = g(\mathbf{U}) \in \mathbb{R}$ where $\mathbf{U} = (U_1, \dots, U_s) \sim U(0, 1)^s$.

Monte Carlo: $X_i = g(\mathbf{U}_i)$ for $\mathbf{U}_1, \dots, \mathbf{U}_n$ independent $U(0, 1)^s$. Estimate the mean $\mathbb{E}[X]$ by

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n g(\mathbf{U}_i).$$

Randomized quasi-Monte Carlo (RQMC)

Suppose $X = h(\mathbf{Y}) = g(\mathbf{U}) \in \mathbb{R}$ where $\mathbf{U} = (U_1, \dots, U_s) \sim U(0, 1)^s$.

Monte Carlo: $X_i = g(\mathbf{U}_i)$ for $\mathbf{U}_1, \dots, \mathbf{U}_n$ independent $U(0, 1)^s$. Estimate the mean $\mathbb{E}[X]$ by

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n g(\mathbf{U}_i).$$

RQMC: Take $\mathbf{U}_1, \dots, \mathbf{U}_n$ as RQMC points and compute again

$$\hat{\mu}_{n,\text{rqmc}} = \frac{1}{n} \sum_{i=1}^n g(\mathbf{U}_i).$$

Both unbiased. $\text{Var}[\hat{\mu}_n] = \mathcal{O}(1/n)$. $\text{Var}[\hat{\mu}_{n,\text{rqmc}}]$ is often $\mathcal{O}(n^{-2+\epsilon})$ or even $\mathcal{O}(n^{-3+\epsilon})$.

QMC point sets: integration lattices, polynomial lattices, digital nets, ...

Randomizations: random shift mod 1, random digital shift, scrambles, ...

RQMC variance bounds

There are various Cauchy-Schwarz-type inequalities of the form

$$\text{Var}[\hat{\mu}_{n,\text{rqmc}}] \leq V^2(g) \cdot D^2(P_n)$$

for all g in some Hilbert space or Banach space \mathcal{H} , where $V(g) = \|g - \mu\|_{\mathcal{H}}$ is the variation of g , and $D(P_n)$ is the discrepancy of $P_n = \{\mathbf{U}_1, \dots, \mathbf{U}_n\}$ (defined by an expectation in the RQMC case).

RQMC variance bounds

There are various Cauchy-Schwarz-type inequalities of the form

$$\text{Var}[\hat{\mu}_{n,\text{rqmc}}] \leq V^2(g) \cdot D^2(P_n)$$

for all g in some Hilbert space or Banach space \mathcal{H} , where $V(g) = \|g - \mu\|_{\mathcal{H}}$ is the **variation** of g , and $D(P_n)$ is the **discrepancy** of $P_n = \{\mathbf{U}_1, \dots, \mathbf{U}_n\}$ (defined by an expectation in the RQMC case).

Classical **Koksma-Hlawka**: $D(P_n) = D^*(P_n)$ is the star discrepancy and

$$V(g) = V_{\text{HK}}(g) = \sum_{\emptyset \neq \mathbf{v} \subseteq \mathcal{S}} \int_{[0,1]^{|\mathbf{v}|}} \left| \frac{\partial^{|\mathbf{v}|}}{\partial \mathbf{u}_{\mathbf{v}}} g(\mathbf{u}_{\mathbf{v}}, \mathbf{1}) \right| d\mathbf{u}_{\mathbf{v}}, \quad (\text{Hardy-Krause (HK) variation})$$

Variance bounds are conservative; RQMC often works well empirically, sometimes even when $V_{\text{HK}}(f) = \infty$.

Combining RQMC with the KDE

KDE density estimator at a single point x :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{x - g(\mathbf{U}_i)}{h}\right) = \frac{1}{n} \sum_{i=1}^n \tilde{g}(\mathbf{U}_i).$$

With RQMC points \mathbf{U}_i , this is an RQMC estimator of $\mathbb{E}[\tilde{g}(\mathbf{U})] = \mathbb{E}[\hat{f}_n(x)]$.

RQMC **does not change the bias**, but **may reduce** $\text{Var}[\hat{f}_n(x)]$, and then the IV and MISE.

Combining RQMC with the KDE

KDE density estimator at a single point x :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{x - g(\mathbf{U}_i)}{h}\right) = \frac{1}{n} \sum_{i=1}^n \tilde{g}(\mathbf{U}_i).$$

With RQMC points \mathbf{U}_i , this is an RQMC estimator of $\mathbb{E}[\tilde{g}(\mathbf{U})] = \mathbb{E}[\hat{f}_n(x)]$.

RQMC **does not change the bias**, but **may reduce** $\text{Var}[\hat{f}_n(x)]$, and then the IV and MISE.

To **prove** RQMC variance bounds, we need **bounds on the variation of \tilde{g}** .

Partial derivatives:

$$\frac{\partial^{|\mathbf{v}|}}{\partial \mathbf{u}_{\mathbf{v}}} \tilde{g}(\mathbf{u}) = \frac{1}{h} \frac{\partial^{|\mathbf{v}|}}{\partial \mathbf{u}_{\mathbf{v}}} k\left(\frac{x - g(\mathbf{u})}{h}\right).$$

We assume they exist and are uniformly bounded. E.g., Gaussian kernel k .

Combining RQMC with the KDE

KDE density estimator at a single point x :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{x - g(\mathbf{U}_i)}{h}\right) = \frac{1}{n} \sum_{i=1}^n \tilde{g}(\mathbf{U}_i).$$

With RQMC points \mathbf{U}_i , this is an RQMC estimator of $\mathbb{E}[\tilde{g}(\mathbf{U})] = \mathbb{E}[\hat{f}_n(x)]$.

RQMC **does not change the bias**, but **may reduce** $\text{Var}[\hat{f}_n(x)]$, and then the IV and MISE.

To **prove** RQMC variance bounds, we need **bounds on the variation of \tilde{g}** .

Partial derivatives:

$$\frac{\partial^{|\mathbf{v}|}}{\partial \mathbf{u}_{\mathbf{v}}} \tilde{g}(\mathbf{u}) = \frac{1}{h} \frac{\partial^{|\mathbf{v}|}}{\partial \mathbf{u}_{\mathbf{v}}} k\left(\frac{x - g(\mathbf{u})}{h}\right).$$

We assume they exist and are uniformly bounded. E.g., Gaussian kernel k .

But when expanding via the chain rule, we obtain terms in h^{-j} for $j = 2, \dots, |\mathbf{v}| + 1$.

The term for $|\mathbf{v}| = s$ grows as $h^{-s-1} k^{(s)}((g(\mathbf{u}) - x)/h) \prod_{j=1}^s g_j(\mathbf{u}) = \mathcal{O}(h^{-s-1})$ when $h \rightarrow 0$.

Can make it $\mathcal{O}(h^{-s})$ via a change of variables.

An AIV upper bound that we were able to prove

Assumption. Let $g : [0, 1]^s \rightarrow \mathbb{R}$ be **piecewise monotone** in each coordinate u_j when the other coordinates are fixed. Assume that all first-order partial derivatives of g are continuous and that $\|g_{\mathfrak{w}_1} g_{\mathfrak{w}_2} \cdots g_{\mathfrak{w}_\ell}\|_1 < \infty$ for all selections of non-empty, mutually disjoint index sets $\mathfrak{w}_1, \dots, \mathfrak{w}_\ell \subseteq \mathcal{S} = \{1, \dots, s\}$.

Proposition Then the Hardy-Krause variation of \tilde{g} satisfies

$$V_{\text{HK}}(\tilde{g}) \leq c_j h^{-s} + \mathcal{O}(h^{-s+1}) \quad \text{for each } j.$$

Corollary. With RQMC point sets having $D^*(P_n) = \mathcal{O}(n^{-1+\epsilon})$ for all $\epsilon > 0$ when $n \rightarrow \infty$, we obtain

$$\text{AIV} = \mathcal{O}(n^{-2+\epsilon} h^{-2s}) \quad \text{for all } \epsilon > 0.$$

By picking h to minimize the AMISE bound, we get $\text{AMISE} = \mathcal{O}(n^{-4/(2+s)+\epsilon})$.

Worse than MC when $s \geq 4$. The factor h^{-2s} hurts! **But this is only an upper bound.**

Why not take the sample derivative of an estimator of F ?

We want to estimate the density $f(x) = F'(x)$.

A simple unbiased estimator of F is the empirical cdf

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

However $d\hat{F}_n(x)/dx = 0$ almost everywhere, so this cannot be a useful density estimator!

We need a smoother estimator of F .

Conditional Monte Carlo (CMC) for density estimation

Idea: Replace indicator $\mathbb{I}[X_i \leq x]$ by its conditional cdf given filtered information:

$$F(x | \mathcal{G}) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x | \mathcal{G}]$$

where the sigma-field \mathcal{G} contains not enough information to reveal X but enough to compute $F(x | \mathcal{G})$, and is chosen so that the following holds:

Conditional Monte Carlo (CMC) for density estimation

Idea: Replace indicator $\mathbb{I}[X_i \leq x]$ by its conditional cdf given filtered information:

$$F(x | \mathcal{G}) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x | \mathcal{G}]$$

where the sigma-field \mathcal{G} contains not enough information to reveal X but enough to compute $F(x | \mathcal{G})$, and is chosen so that the following holds:

Assumption 1. For all realizations of \mathcal{G} , $F(x | \mathcal{G})$ is a continuous function of x over $[a, b]$, differentiable except perhaps over a denumerable set of points $D(\mathcal{G}) \subset [a, b]$, and for which $f(x | \mathcal{G}) = F'(x | \mathcal{G}) = dF(x | \mathcal{G})/dx$ (when it exists) is bounded uniformly in x by a random variable Γ such that $\mathbb{E}[\Gamma^2] \leq K_\gamma < \infty$.

Proposition CDE: Under Ass. 1, for $x \in [a, b]$, $\mathbb{E}[f(x | \mathcal{G})] = f(x)$ and $\text{Var}[f(x | \mathcal{G})] < K_\gamma$.

Proposition: If $\mathcal{G} \subset \tilde{\mathcal{G}}$ both satisfy Assumption 1, then $\text{Var}[f(x | \mathcal{G})] \leq \text{Var}[f(x | \tilde{\mathcal{G}})]$.

Conditional density estimator (CDE) with sample size n : $\hat{f}_{\text{cde},n}(x) = \frac{1}{n} \sum_{i=1}^n f(x | \mathcal{G}^{(i)})$

where $\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(n)}$ are n independent “realizations” of \mathcal{G} . $\text{Var}[\hat{f}_{\text{cde},n}(x)] = \mathcal{O}(n^{-1})$.

Example: displacement of cantilever beam (Bingham 2017)

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

where $\kappa = 5 \times 10^5$, $w = 4$, $t = 2$, Y_1, Y_2, Y_3 independent normal, $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$,

Description	Symbol	μ_j	σ_j
Young's modulus	Y_1	2.9×10^7	1.45×10^6
Horizontal load	Y_2	500	100
Vertical load	Y_3	1000	100

We estimate the density of X over $[3.1707, 5.6675]$, which covers about 99% of the density (it clips 0.5% on each side).

CDE estimator

Conditioning on $\mathcal{G}_{-1} = \{Y_2, Y_3\}$ means hiding Y_1 . We have

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

For $x > 0$,

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

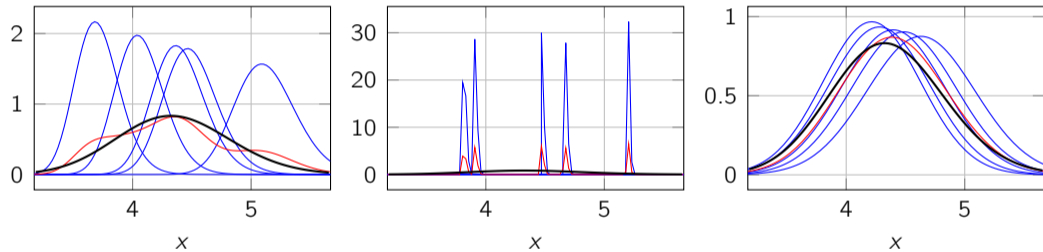
and

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

For $n = 2^{19}$, the MISE is about $2^{-14.7}$ for the usual KDE+MC and $2^{-22.8}$ for the CDE with \mathcal{G}_{-3} ; the MISE is divided by about $2^8 = 256$.

The MISE rate goes from $\mathcal{O}(n^{-4/5})$ to around $\mathcal{O}(n^{-1})$ with the CDE.

We can also hide Y_2 or Y_3 instead of Y_1 .



Five realizations of conditional density $f(\cdot | \mathcal{G}_{-k})$ (blue), their average (red), and true density (thick black), for $k = 1$ (left), $k = 2$ (middle), and $k = 3$ (right).

We see that hiding Y_3 seems to perform the best.

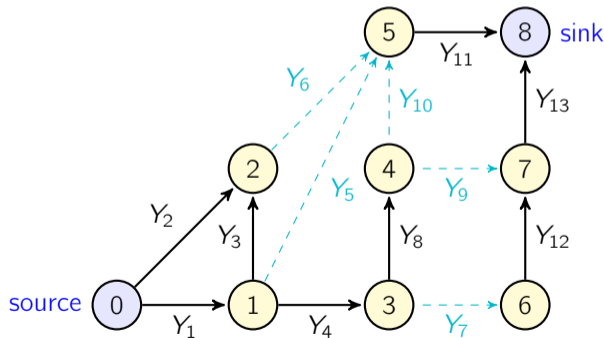
CMC for the SAN Example

X = length of longest path. **CMC estimator of $\mathbb{P}[X \leq x]$:**

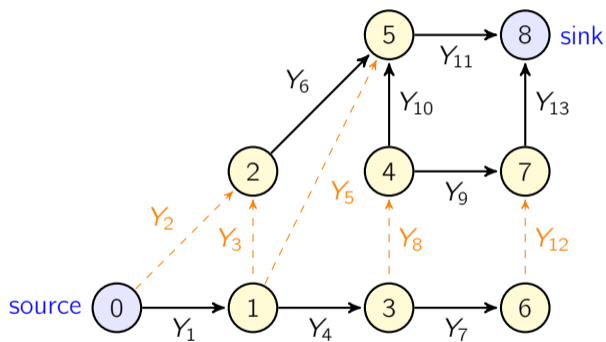
Pick a **minimal cut \mathcal{L}** between source and sink, and let $F(x | \mathcal{G}) = \mathbb{P}[X \leq x | \{Y_j, j \notin \mathcal{L}\}]$.

Ex.: $\mathcal{L} = \{5, 6, 7, 9, 10\}$ and $Y_j = F_j^{-1}(U_j)$. This estimator **continuous in the U_j 's and in x** .

(Erasing a single Y_j does not work: it does not make the conditional cdf continuous.)



Another minimal cut: $\mathcal{L} = \{2, 3, 5, 8, 12\}$.



For each $j \in \mathcal{L}$, let P_j be the length of the longest path that goes through arc j when we exclude Y_j from that length. Then

$$F(x | \mathcal{G}) = \mathbb{P} [X < x | \{Y_j : j \notin \mathcal{L}\}] = \prod_{j \in \mathcal{L}} F_j(x - P_j)$$

and

$$f(x | \mathcal{G}) = \sum_{j \in \mathcal{L}} f_j(x - P_j) \prod_{l \in \mathcal{L}, l \neq j} F_l(x - P_l),$$

if f_j exists for all $j \in \mathcal{L}$.

Under this conditioning, the cdf of every path length is continuous in x , and so is $F(\cdot | \mathcal{G})$, and Assumption 1 holds, so $f(x | \mathcal{G})$ is an **unbiased density estimator**.

For each $j \in \mathcal{L}$, let P_j be the length of the longest path that goes through arc j when we exclude Y_j from that length. Then

$$F(x | \mathcal{G}) = \mathbb{P} [X < x | \{Y_j : j \notin \mathcal{L}\}] = \prod_{j \in \mathcal{L}} F_j(x - P_j)$$

and

$$f(x | \mathcal{G}) = \sum_{j \in \mathcal{L}} f_j(x - P_j) \prod_{l \in \mathcal{L}, l \neq j} F_l(x - P_l),$$

if f_j exists for all $j \in \mathcal{L}$.

Under this conditioning, the cdf of every path length is continuous in x , and so is $F(\cdot | \mathcal{G})$, and Assumption 1 holds, so $f(x | \mathcal{G})$ is an unbiased density estimator.

When we replace a KDE by the CDE in our example, empirically, the MISE rate goes from $\mathcal{O}(n^{-4/5})$ to $\mathcal{O}(n^{-1})$ and the MISE for $n = 2^{19}$ is divided by about 25 to 30.

Applying RQMC to the CDE

Now we want to sample the CDE using RQMC points.

For this, we must rewrite the CDE as a function of $\mathbf{u} \in [0, 1)^s$:

$$\begin{aligned}F(x \mid \mathcal{G}) &= \tilde{g}(x, \mathbf{u}), \\f(x \mid \mathcal{G}) &= \tilde{g}'(x, \mathbf{u}) = d\tilde{g}(x, \mathbf{u})/dx\end{aligned}$$

for some $\tilde{g} : [a, b] \times [0, 1)^s$ for which $\tilde{g}'(x, \cdot)$ has bounded variation for each x .

Applying RQMC to the CDE

Now we want to sample the CDE using RQMC points.

For this, we must rewrite the CDE as a function of $\mathbf{u} \in [0, 1)^s$:

$$\begin{aligned}F(x | \mathcal{G}) &= \tilde{g}(x, \mathbf{u}), \\f(x | \mathcal{G}) &= \tilde{g}'(x, \mathbf{u}) = d\tilde{g}(x, \mathbf{u})/dx\end{aligned}$$

for some $\tilde{g} : [a, b] \times [0, 1)^s$ for which $\tilde{g}'(x, \cdot)$ has bounded variation for each x .

CDE sample: $\tilde{g}'(x, \mathbf{U}_1), \dots, \tilde{g}'(x, \mathbf{U}_n)$ where $\{\mathbf{U}_1, \dots, \mathbf{U}_n\}$ is an RQMC point set.

Applying RQMC to the CDE

Now we want to sample the CDE using RQMC points.

For this, we must rewrite the CDE as a function of $\mathbf{u} \in [0, 1)^s$:

$$\begin{aligned} F(x | \mathcal{G}) &= \tilde{g}(x, \mathbf{u}), \\ f(x | \mathcal{G}) &= \tilde{g}'(x, \mathbf{u}) = d\tilde{g}(x, \mathbf{u})/dx \end{aligned}$$

for some $\tilde{g} : [a, b] \times [0, 1)^s$ for which $\tilde{g}'(x, \cdot)$ has bounded variation for each x .

CDE sample: $\tilde{g}'(x, \mathbf{U}_1), \dots, \tilde{g}'(x, \mathbf{U}_n)$ where $\{\mathbf{U}_1, \dots, \mathbf{U}_n\}$ is an RQMC point set.

If $\tilde{g}'(x, \cdot)$ has **bounded variation**, then we can get an $\mathcal{O}(n^{-2+\epsilon})$ rate for the MISE, and sometimes better. This holds in several examples that we tried.

If $\tilde{g}'(x, \cdot)$ has **unbounded variation**, RQMC may still reduce the IV, but no guarantee.

A likelihood ratio density estimator (LRDE)

Back to $X = h(\mathbf{Y})$ where \mathbf{Y} has known density f_Y over $R \subseteq \mathbb{R}^d$. We have

$$F(x) = \mathbb{P}[h(\mathbf{Y}) \leq x] = \int_{\mathbb{R}^d} \mathbb{I}[h(\mathbf{y}) \leq x] f_Y(\mathbf{y}) d\mathbf{y}.$$

Want to change the integrand into a continuous function of x .

A likelihood ratio density estimator (LRDE)

Back to $X = h(\mathbf{Y})$ where \mathbf{Y} has known density f_Y over $R \subseteq \mathbb{R}^d$. We have

$$F(x) = \mathbb{P}[h(\mathbf{Y}) \leq x] = \int_{\mathbb{R}^d} \mathbb{I}[h(\mathbf{y}) \leq x] f_Y(\mathbf{y}) d\mathbf{y}.$$

Want to change the integrand into a continuous function of x .

Main idea: Make a change of variable $\mathbf{y} \mapsto \mathbf{z} = \mathbf{z}(x)$ of the form $\mathbf{y} = \varphi(\mathbf{z}; x)$, for a family of one-to-one functions $\{\varphi(\cdot; x), x \in [a, b]\}$ such that $\{h(\varphi(\mathbf{z}; x)) \leq x\} \equiv \{\tilde{h}(\mathbf{z}) \leq 1\}$ for a function \tilde{h} independent of x when \mathbf{z} is given.

In case $\tilde{R} = \varphi^{-1}(R) \stackrel{\text{def}}{=} \varphi^{-1}(R; x)$ is independent of x , the integration domain remains independent of x after the change of variable, and we can rewrite

$$F(x) = \int_{\tilde{R}} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] f_Y(\varphi(\mathbf{z}; x)) |J_\varphi(\mathbf{z}; x)| d\mathbf{z},$$

where $|J_\varphi(\mathbf{z}; x)|$ is the Jacobian of the transformation $\mathbf{y} = \varphi(\mathbf{z}; x)$.

Take a given $x = x_0$. In a small open neighborhood of x_0 , we can write

$$F(x) = \int_{\tilde{R}} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] L(\mathbf{z}; x, x_0) f_Y(\varphi(\mathbf{z}; x_0)) |J_\varphi(\mathbf{z}; x_0)| d\mathbf{z}$$

where

$$L(\mathbf{z}; x, x_0) = \frac{f_Y(\varphi(\mathbf{z}; x)) |J_\varphi(\mathbf{z}; x)|}{f_Y(\varphi(\mathbf{z}; x_0)) |J_\varphi(\mathbf{z}; x_0)|}.$$

is the **likelihood ratio** between the density of \mathbf{z} at x and at x_0 . Under appropriate conditions:

$$\begin{aligned} f(x) = F'(x) &= \frac{d}{dx} \int_{\tilde{R}} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] L(\mathbf{z}; x, x_0) f_Y(\varphi(\mathbf{z}; x_0)) |J_\varphi(\mathbf{z}; x_0)| d\mathbf{z} \\ &= \int_{\tilde{R}} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \left(\frac{d}{dx} L(\mathbf{z}; x, x_0) \right) f_Y(\varphi(\mathbf{z}; x_0)) |J_\varphi(\mathbf{z}; x_0)| d\mathbf{z} \\ &= \int_{\tilde{R}} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \left(\frac{d}{dx} L(\mathbf{z}; x, x_0) \right) \frac{f_Y(\varphi(\mathbf{z}; x)) |J_\varphi(\mathbf{z}; x)|}{L(\mathbf{z}; x, x_0)} d\mathbf{z} \\ &= \int_{\tilde{R}} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \left(\frac{d}{dx} \ln L(\mathbf{z}; x, x_0) \right) f_Y(\varphi(\mathbf{z}; x)) |J_\varphi(\mathbf{z}; x)| d\mathbf{z} \\ &= \int_R \mathbb{I}[h(\mathbf{y}) \leq x] S(\mathbf{y}, x) f_Y(\mathbf{y}) d\mathbf{y} = \mathbb{E}[\mathbb{I}[h(\mathbf{Y}) \leq x] S(\mathbf{Y}, x)] \end{aligned}$$

where

$$S(\mathbf{y}, x) = \left(\nabla(\ln f_Y)(\mathbf{y})^t \nabla_x \varphi(\mathbf{z}; x) + \frac{d \ln |J_\varphi(\mathbf{z}; x)|}{dx} \right) \Big|_{\mathbf{z}=\varphi^{-1}(\mathbf{y}; x)}$$

is the **score function** associated with L .

This gives the unbiased **likelihood ratio density estimator** (LRDE)

$$\hat{f}(x) = \mathbb{I}[h(\mathbf{Y}) \leq x] S(\mathbf{Y}, x)$$

where $\mathbf{Y} \sim f_Y$. Here, \mathbf{Y} can have a multivariate distribution for which conditioning is hard.

where

$$S(\mathbf{y}, x) = \left(\nabla(\ln f_Y)(\mathbf{y})^t \nabla_x \varphi(\mathbf{z}; x) + \frac{d \ln |J_\varphi(\mathbf{z}; x)|}{dx} \right) \Big|_{\mathbf{z}=\varphi^{-1}(\mathbf{y}; x)}$$

is the [score function](#) associated with L .

This gives the unbiased [likelihood ratio density estimator](#) (LRDE)

$$\hat{f}(x) = \mathbb{I}[h(\mathbf{Y}) \leq x] S(\mathbf{Y}, x)$$

where $\mathbf{Y} \sim f_Y$. Here, \mathbf{Y} can have a multivariate distribution for which conditioning is hard.

This LR approach has been widely used to estimate the derivative of $\mathbb{E}[h(\mathbf{Y})]$ with respect to a parameter of the distribution of \mathbf{Y} (e.g., Glynn 1987, L'Ecuyer (1990), Asmussen and Glynn 2007).

Assumption LR. For all $x \in [a, b]$, the mapping $\varphi(\cdot; x)$ has continuous partial derivatives and $|J_\varphi(\cdot; x)| \neq 0$. With probability 1 over the realizations of $\mathbf{Y} = \varphi(\mathbf{Z}; x)$, $f_Y(\varphi(\mathbf{Z}; x))|J_\varphi(\mathbf{Z}; x)|$ is a continuous function of x over $[a, b]$, and is differentiable except perhaps at a countable set of points $D(\mathbf{Y})$. There is also a random variable Γ defined over the same probability space as \mathbf{Y} with $\mathbb{E}[\Gamma^2] < \infty$ and

$$\sup_{x \in [a, b]} |\mathbb{I}[h(\mathbf{Y}) \leq x] S(\mathbf{Y}, x)| \leq \Gamma.$$

Assumption LR. For all $x \in [a, b]$, the mapping $\varphi(\cdot; x)$ has continuous partial derivatives and $|J_\varphi(\cdot; x)| \neq 0$. With probability 1 over the realizations of $\mathbf{Y} = \varphi(\mathbf{Z}; x)$, $f_Y(\varphi(\mathbf{Z}; x))|J_\varphi(\mathbf{Z}; x)|$ is a continuous function of x over $[a, b]$, and is differentiable except perhaps at a countable set of points $D(\mathbf{Y})$. There is also a random variable Γ defined over the same probability space as \mathbf{Y} with $\mathbb{E}[\Gamma^2] < \infty$ and

$$\sup_{x \in [a, b]} |\mathbb{I}[h(\mathbf{Y}) \leq x]S(\mathbf{Y}, x)| \leq \Gamma.$$

Proposition LR. Suppose $\tilde{R} = \varphi^{-1}(R; x)$ is independent of x . Under Assumption LR, the LRDE is unbiased for the density $f(x)$ at x for all $x \in [a, b]$, and its MISE is bounded by $(b - a)\mathbb{E}[\Gamma^2]$.

What if $\tilde{R}(x) = \varphi^{-1}(R; x)$ depends on x ?

In that case, when taking the derivative w.r.t. x , we must account for the displacement of the boundary. According to Leibniz's integral rule, we must add to $F'(x)$ the term

$$B(\mathbf{y}, x) = \int_{\partial\tilde{R}(x)} [\dots] \mathbf{b}(\mathbf{z}(x), x) d\mathbf{z} = \int_{\partial R} \mathbb{I}[h(\mathbf{y}) \leq x] \mathbf{b}(\varphi^{-1}(\mathbf{y}, x), x) f_Y(\mathbf{y}) d\mathbf{y}$$

where $\mathbf{b}(\mathbf{z}(x), x)$ is the **rate of displacement of the boundary $\partial\tilde{R}(x)$** as a function of x in the outward normal direction.

This term can be easy to estimate, e.g., when R is a rectangular box.

GLR density estimators

Peng et al. (2020) proposed an adaptation of a [generalized likelihood ratio](#) (GLR) method of Peng et al. (2018) to density estimation. An improved version can be found in Peng et al. (2021).

The estimators are somewhat similar to our LRDEs, although the assumptions are different and somewhat stronger.

Cantilever beam example

Estimated $\text{MISE} = Kn^{-\hat{\nu}}$. MISE with $n = 2^{19}$ is 2^{-e19} .

	$\hat{\nu}$				e19			
	KDE	\mathcal{G}_{-3}	CDE-c	LRDE	KDE	\mathcal{G}_{-3}	CDE-c	LRDE
MC	0.76	0.99	0.98	1.03	15.8	22.8	22.5	16.8
Lat+s	1.03	2.06	2.04	1.55	21.9	41.6	41.9	26.4
Lat+s+b	0.93	2.27	2.25	1.25	21.0	46.8	47.0	24.7
Sob+LMS	0.97	2.21	2.21	1.31	21.5	45.7	46.1	25.6

The MISE decreases roughly as $\mathcal{O}(n^{-2})$ or better for CDE+RQMC.

For $n = 2^{19}$, the MISE is about $2^{-15.8}$ for the usual KDE+MC and 2^{-47} for the new CDE+RQMC; i.e., MISE is divided by more than $2^{31} \approx 2 \text{ billions}$.

SAN Example

		\hat{D}	e19
KDE	MC	0.78	20.9
	Lat+s	0.95	22.7
	Sob+LMS	0.74	21.9
CDE	MC	0.96	25.6
	Lat+s	1.31	30.9
	Sob+LMS	1.27	29.9
LRDE	MC	1.00	20.5
	Lat+s	1.22	23.5
	Sob+LMS	1.16	24.6

With CDE+RQMC, we observe a convergence rate near $\mathcal{O}(n^{-1.3})$ for the MISE.

For $n = 2^{19}$, by using the new CDE+RQMC rather than the usual KDE+MC, the MISE is divided by about **500** to **1000**.

LRDE does not perform as well as CDE.

Conclusion

- ▶ Combining a KDE with RQMC can reduce the MISE and sometimes improve its convergence rate, even though our MISE bounds converge faster only when the dimension is very small.
- ▶ The CDE is an unbiased density estimator with better convergence rate. Combining it with RQMC can provide an even better rate, and sometimes huge MISE reductions.
- ▶ When we cannot find \mathcal{G} for which Assumption 1 holds and $f(x | \mathcal{G})$ is easy to compute, the LRDE and the GLR can be good unbiased alternatives.
The LRDE is often discontinuous in \mathbf{U} , in which case it does not combine well with RQMC. Perhaps in some cases, one can add CMC to them before applying RQMC.
- ▶ **Extensions:** Density estimation for a function of the state of a Markov chain, using Array-RQMC. Generalization to multivariate output.

Some references

- ▶ S. Asmussen. Conditional Monte Carlo for sums, with applications to insurance and finance, *Annals of Actuarial Science*, 12, 2: 455-478, 2018.
- ▶ A. Ben Abdellah, P. L'Ecuyer, A. B. Owen, and F. Puchhammer. Density estimation by Randomized Quasi-Monte Carlo. *SIAM Journal on Uncertainty Quantification*, **9**, 1, 280–301, 2021.
- ▶ J. Dick and F. Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, Cambridge, U.K., 2010.
- ▶ M. Fu and J.-Q. Hu. *Conditional Monte Carlo*. Kluwer Academic, 1997.
- ▶ P. W. Glynn. Likelihood ratio gradient estimation: an overview. *Proceedings of the 1987 Winter Simulation Conference*, 366–375.
- ▶ P. J. Laub, R. Salomone, Z. I. Botev. Monte Carlo estimation of the density of the sum of dependent random variables. *Mathematics and Computers in Simulation* 161: 23–31, 2019.
- ▶ P. L'Ecuyer. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science* 36: 1364–1383, 1990.

- ▶ P. L'Ecuyer. Randomized quasi-Monte Carlo: An introduction for practitioners. In P. W. Glynn and A. B. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2016*, 29–52, 2018.
- ▶ P. L'Ecuyer and G. Perron. On the Convergence Rates of IPA and FDC Derivative Estimators for Finite-Horizon Stochastic Simulations. *Operations Research*, 42 (4):643–656, 1994.
- ▶ P. L'Ecuyer and F. Puchhammer, “Density Estimation by Monte Carlo and Quasi-Monte Carlo,” in *Monte Carlo and Quasi Monte Carlo Methods: MCQMC 2020*, A. Keller, Ed., Springer-Verlag, 2022.
- ▶ P. L'Ecuyer, F. Puchhammer, and A. Ben Abdellah. Monte Carlo and Quasi-Monte Carlo Density Estimation via Conditioning. *INFORMS Journal on Computing*, 2022.
- ▶ Y. Peng, M. C. Fu, B. Heidergott, H. Lam. Maximum likelihood estimation by Monte Carlo simulation: Towards data-driven stochastic modeling. *Operations Research*, 68, 1896–1912, 2020.
- ▶ Y. Peng, M. C. Fu, J. Q. Hu, B. Heidergott. A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Operations Res.* 66, 487–499, 2018.
- ▶ Y. Peng, M. C. Fu, J.-Q. Hu, P. L'Ecuyer, and B. Tuffin, “A Generalized Likelihood Ratio Method for Stochastic Models with Uniform Random Numbers As Inputs,” revised December 2021.
- ▶ F. Puchhammer and P. L'Ecuyer, “Likelihood Ratio Density Estimation for Simulation Models,” submitted to *Proceedings of the 2022 Winter Simulation Conference*.
- ▶ D. W. Scott. *Multivariate Density Estimation*. Wiley, 2015.