Monte Carlo and Quasi-Monte Carlo for Density Estimation

Pierre L'Ecuyer

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

MCQMC 2020 **Virtual-Oxford, August 2020**

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 $\mathbb{E}[X]$. $\quad \text{MSE} = \text{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

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 $\mathbb{E}[X]$. $\quad \mathrm{MSE} = \mathrm{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

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 $\mathbb{E}[X]$. $\quad \text{MSE} = \text{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

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 $X_1, \ldots, X_n$ 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 $\mathbb{E}[X]$. $\quad \text{MSE} = \text{Var}[\bar{X}_n] = \mathcal{O}(n^{-1})$.

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 $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ are **given independent observations of** X and one estimates the density f of X from that.
Leading method: kernel density estimator (KDE); $\mathrm{MSE}[\hat{f}_n(x)] = \mathcal{O}(n^{-4/5})$.

Setting

Classical density estimation in statistics was developed in the context where $X_1, \ldots, X_n$ are **given independent observations of** X and one estimates the density f of X from that.
Leading method: kernel density estimator (KDE); $\mathrm{MSE}[\hat{f}_n(x)] = \mathcal{O}(n^{-4/5})$.

In this talk, we assume that $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$ are **given independent observations of** X and one estimates the density f of X from that. Leading method: kernel density estimator (KDE); $\mathrm{MSE}[\hat{f}_n(x)] = \mathcal{O}(n^{-4/5})$.

In this talk, we assume that $X_1, \ldots, X_n$ 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?

Six-course menu

1. What happens if we combine a KDE with RQMC?
   [Lunchtime discussion with Owen and Hickernell at workshop in Banff in 2015.]

2. Using conditional Monte Carlo (CMC) can provide an unbiased conditional density estimator (CDE) $\hat{f}_n$ for which $\mathbb{E}[\hat{f}_n(x) - f(x)]^2 = \mathcal{O}(n^{-1})$.

3. Under appropriate conditions, CDE + RQMC can give $\mathbb{E}[\hat{f}_n(x) - f(x)]^2 = \mathcal{O}(n^{-2+\epsilon})$ or even better.

4. An unbiased density estimator can also be obtained via the likelihood ratio (LR) method for derivative estimation. Is this LRDE RQMC-friendly?

5. Another density estimator was proposed very recently based on a generalized likelihood ratio (GLR) method. Is it RQMC-friendly?

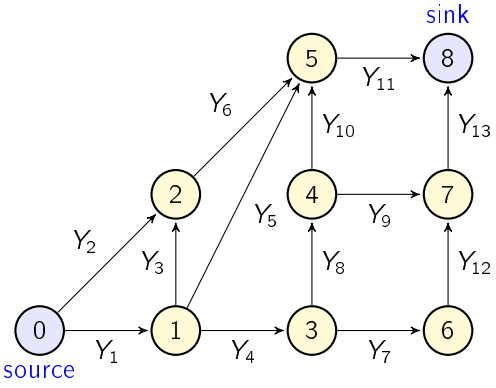
6. Numerical examples and comparisons.

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;
$\mu_1, \ldots, \mu_{13}$: 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)

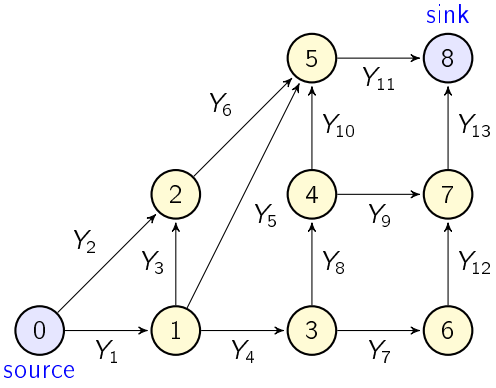
Precedence relations between activities. Activity k has random duration Y_k (length of arc k) with known cdf $F_k(y) := \mathbb{P}[Y_k \leq y]$.

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

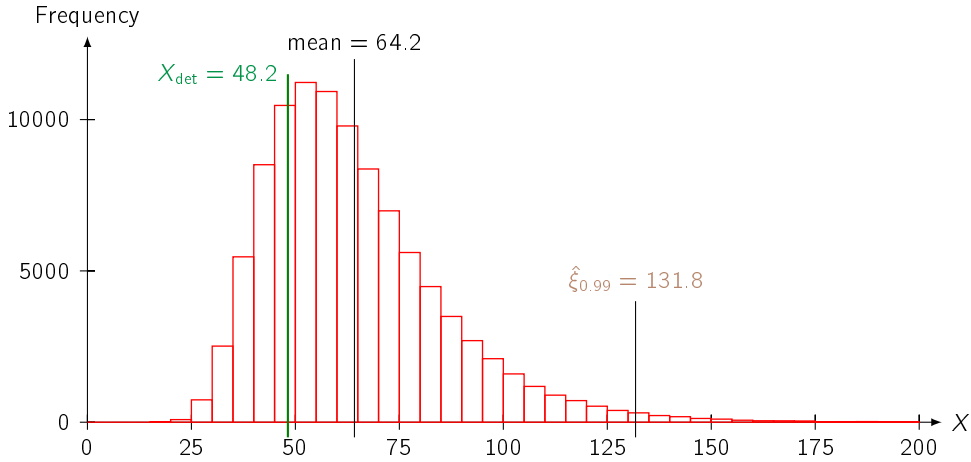
Specific case (Avramidis and Wilson 1998):
$Y_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, 2, 4, 11, 12$;
$Y_k \sim \text{Expon}(1/\mu_k)$ otherwise;
$\mu_1, \ldots, \mu_{13}$: 13.0, 5.5, 7.0, 5.2, 16.5, 14.7, 10.3, 6.0, 4.0, 20.0, 3.2, 3.2, 16.5;
$\sigma_k = \mu_k/4$.

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 $\mathbb{E}[X]$. How can we do better?



Density estimation

Want to estimate the density of $X = h(\mathbf{Y}) = h(Y_1, \ldots, Y_s)$, 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 $X = h(\mathbf{Y}) = h(Y_1, \ldots, Y_s)$, 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:

$$
\begin{aligned}
\text{MISE} &= \text{mean integrated squared error} = \int_a^b \mathbb{E}[(\hat{f}_n(x) - f(x))^2]\mathrm{d}x \\
&= \text{IV} + \text{ISB} \\
\text{IV} &= \text{integrated variance} = \int_a^b \mathrm{Var}[\hat{f}_n(x)]\mathrm{d}x \\
\text{ISB} &= \text{integrated squared bias} = \int_a^b (\mathbb{E}[\hat{f}_n(x)] - f(x))^2\mathrm{d}x
\end{aligned}
$$

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

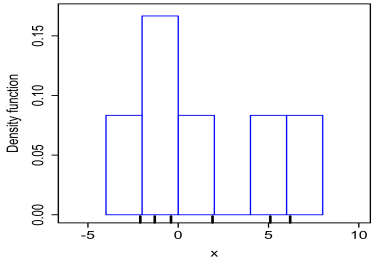
$$\hat{f}_n(x) = \frac{n_j}{nh} \quad \text{for } x \in I_j = [a + (j-1)h, a + jh), \quad j = 1, ..., m$$

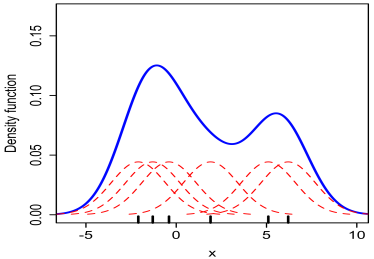
where n_j is the number of observations X_i that fall in interval j.

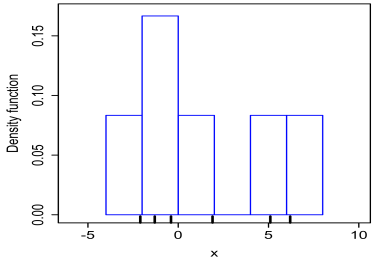
Kernel Density Estimator (KDE) : Select kernel k (unimodal symmetric density centered at 0) and bandwidth $h > 0$ (horizontal stretching factor for the kernel). The KDE is

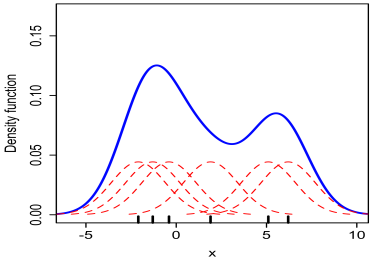
$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right).$$

$n = 6$

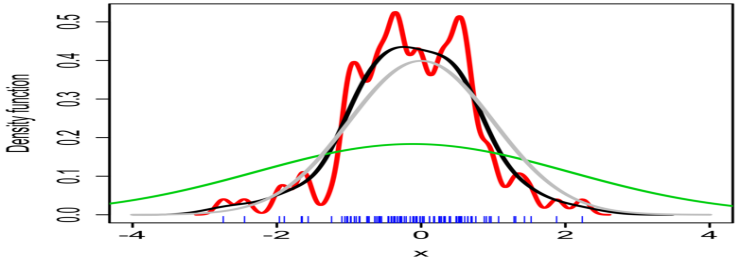

Density function
0.00
0.05
0.10
0.15
-5
0
5
10
x


Density function
0.00
0.05
0.10
0.15
-5
0
5
10
x





$n = 6$



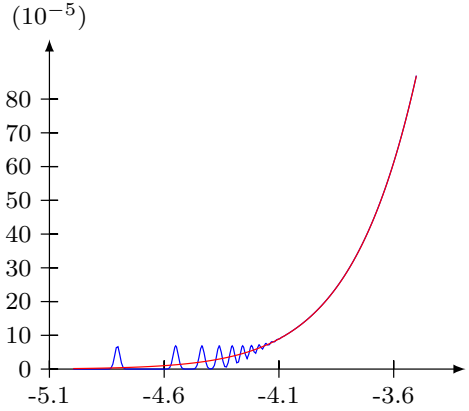
$n = 100$

Example of a KDE in $s = 1$ dimension

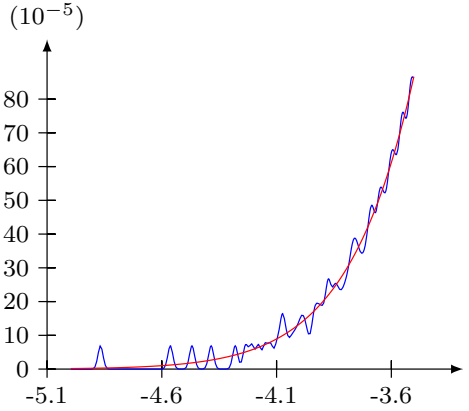
KDE (blue) vs true density (red) with $n = 2^{19}$:

Here we take $U_1, \ldots, U_n$ in $(0, 1)$ and put $X_i = F^{-1}(U_i)$.

midpoint rule for the U_i's



stratified sample for the U_i's



Asymptotic convergence with Monte Carlo for smooth f

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

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

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

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 $g : \mathbb{R} \to \mathbb{R}$, define

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

	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 $\mathbf{U} = (U_1, \ldots, U_s) \sim U(0,1)^s$.

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

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=0}^{n-1} g(\mathbf{U}_i).$$

Randomized quasi-Monte Carlo (RQMC)

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

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

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=0}^{n-1} g(\mathbf{U}_i).$$

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

$$\hat{\mu}_{n,\text{rqmc}} = \frac{1}{n} \sum_{i=0}^{n-1} g(\mathbf{U}_i).$$

Both unbiased. $\mathrm{Var}[\hat{\mu}_n] = \mathcal{O}(1/n)$. $\mathrm{Var}[\hat{\mu}_{n,\text{rqmc}}]$ 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-Shwartz-type inequalities of the form

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

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 $P_n = \{\mathbf{U}_0, \ldots, \mathbf{U}_{n-1}\}$ (defined by an expectation in the RQMC case).

RQMC variance bounds

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

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

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 $P_n = \{\mathbf{U}_0, \ldots, \mathbf{U}_{n-1}\}$ (defined by an expectation in the RQMC case).

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

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

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

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 $\mathrm{Var}[\hat{f}_n(x)]$, 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 $\mathrm{Var}[\hat{f}_n(x)]$, and then the IV and MISE.

To prove RQMC variance bounds, we need bounds on the variation of $\tilde{g}$.
Partial derivatives:

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

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 $\mathrm{Var}[\hat{f}_n(x)]$, and then the IV and MISE.

To prove RQMC variance bounds, we need bounds on the variation of $\tilde{g}$.
Partial derivatives:

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

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 $j = 2, \ldots, |\mathfrak{v}| + 1$.
The term for $|\mathfrak{v}| = s$ grows as $h^{-s-1} k^{(s)}\left((g(\mathbf{u}) - x)/h\right) \prod_{j=1}^{s} g_j(\mathbf{u}) = \mathcal{O}(h^{-s-1})$ when $h \to 0$.
Can make it $\mathcal{O}(h^{-s})$ via a change of variables.

An AIV upper bound that we were able to prove

Assumption. Let $g : [0,1]^s \to \mathbb{R}$ 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 $\mathfrak{w}_1, \ldots, \mathfrak{w}_\ell \subseteq \mathcal{S} = \{1, \ldots, s\}$.

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

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

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

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

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

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

An AIV upper bound that we were able to prove

Assumption. Let $g : [0,1]^s \to \mathbb{R}$ 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 $\mathfrak{w}_1, \ldots, \mathfrak{w}_\ell \subseteq \mathcal{S} = \{1, \ldots, s\}$.

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

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

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

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

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

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

Questions?

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 $\mathrm{d}\hat{F}_n(x)/\mathrm{d}x = 0$ 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:

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

where the sigma-field $\mathcal{G}$ contains not enough information to reveal X but enough to compute $F(x \mid \mathcal{G})$, 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:

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

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

Assumption 1. For all realizations of $\mathcal{G}$, $F(x \mid \mathcal{G})$ 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 $f(x \mid \mathcal{G}) = F'(x \mid \mathcal{G}) = \mathrm{d}F(x \mid \mathcal{G})/\mathrm{d}x$ (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 $x \in [a, b]$, $\mathbb{E}[f(x \mid \mathcal{G})] = f(x)$ and $\mathrm{Var}[f(x \mid \mathcal{G})] < K_\gamma$.

Conditional Monte Carlo (CMC) for density estimation

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

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

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

Assumption 1. For all realizations of $\mathcal{G}$, $F(x \mid \mathcal{G})$ 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 $f(x \mid \mathcal{G}) = F'(x \mid \mathcal{G}) = \mathrm{d}F(x \mid \mathcal{G})/\mathrm{d}x$ (when it exists) is bounded uniformly in x by a random variable Γ such that $\mathbb{E}[\Gamma^2] \le K_\gamma < \infty$.

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

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

Conditional Monte Carlo (CMC) for density estimation

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

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

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

Assumption 1. For all realizations of $\mathcal{G}$, $F(x \mid \mathcal{G})$ 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 $f(x \mid \mathcal{G}) = F'(x \mid \mathcal{G}) = \mathrm{d}F(x \mid \mathcal{G})/\mathrm{d}x$ (when it exists) is bounded uniformly in x by a random variable Γ such that $\mathbb{E}[\Gamma^2] \le K_\gamma < \infty$.

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

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

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

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

Baby example: a sum of independent random variables

$X = Y_1 + \cdots + Y_d$, where the Y_j are independent and continuous with cdf F_j and density f_j, and $\mathcal{G}$ is defined by hiding Y_k for an arbitrary k:

$$X \stackrel{\text{def}}{=} Y_1 + \cdots + \ Y_k \ + \cdots + Y_d.$$

Baby example: a sum of independent random variables

$X = Y_1 + \cdots + Y_d$, where the Y_j are independent and continuous with cdf F_j and density f_j, and $\mathcal{G}$ is defined by hiding Y_k for an arbitrary k:

$$\mathcal{G} = \mathcal{G}_{-k} = S_{-k} \quad \stackrel{\text{def}}{=} Y_1 + \cdots + \cancel{Y_k} + \cdots + Y_d.$$

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

The CDE for X becomes $f(x \mid \mathcal{G}_{-k}) = F'(x \mid \mathcal{G}_{-k}) = f_k(x - S_{-k})$. Shifted density of Y_k.

Baby example: a sum of independent random variables

$X = Y_1 + \cdots + Y_d$, where the Y_j are independent and continuous with cdf F_j and density f_j, and $\mathcal{G}$ is defined by hiding Y_k for an arbitrary k:

$$\mathcal{G} = \mathcal{G}_{-k} = S_{-k} \quad \stackrel{\text{def}}{=} Y_1 + \cdots + \not{Y}_k + \cdots + Y_d.$$

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

The CDE for X becomes $f(x \mid \mathcal{G}_{-k}) = F'(x \mid \mathcal{G}_{-k}) = f_k(x - S_{-k})$. Shifted density of Y_k.

Asmussen (2018) proposed and studied the CDE for this special case, with $k = d$ and same F_j for all j.

Baby example: a sum of independent random variables

$X = Y_1 + \cdots + Y_d$, where the Y_j are independent and continuous with cdf F_j and density f_j, and $\mathcal{G}$ is defined by hiding Y_k for an arbitrary k:

$$\mathcal{G} = \mathcal{G}_{-k} = S_{-k} \quad \stackrel{\text{def}}{=} Y_1 + \cdots + \not{Y}_k + \cdots + Y_d.$$

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

The CDE for X becomes $f(x \mid \mathcal{G}_{-k}) = F'(x \mid \mathcal{G}_{-k}) = f_k(x - S_{-k})$. Shifted density of Y_k.

Asmussen (2018) proposed and studied the CDE for this special case, with $k = d$ and same F_j for all j.

When the Y_j have different distributions, we usually want to hide the one with largest variance, but not always. We have examples where the optimal choice of k depends on x. Even better: hide more than one if possible.

Baby example: a sum of independent random variables

$X = Y_1 + \cdots + Y_d$, where the Y_j are independent and continuous with cdf F_j and density f_j, and $\mathcal{G}$ is defined by hiding Y_k for an arbitrary k:

$$\mathcal{G} = \mathcal{G}_{-k} = S_{-k} \quad \stackrel{\text{def}}{=} Y_1 + \cdots + \not{Y_k} + \cdots + Y_d.$$

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

The CDE for X becomes $f(x \mid \mathcal{G}_{-k}) = F'(x \mid \mathcal{G}_{-k}) = f_k(x - S_{-k})$. Shifted density of Y_k.

Asmussen (2018) proposed and studied the CDE for this special case, with $k = d$ and same F_j for all j.

When the Y_j have different distributions, we usually want to hide the one with largest variance, but not always. We have examples where the optimal choice of k depends on x. Even better: hide more than one if possible.

Interpretation: The (random) conditional density replaces the kernel in the KDE. No bias and no need to choose a kernel and a bandwidth.

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

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

For $x > 0$,

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

and

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

Suppose we condition on $\mathcal{G}_{-2} = \{Y_1, Y_3\}$ instead, i.e., hide Y_2. We have

$$X \leq x \quad \text{if and only if} \quad Y_2^2 \leq w^4 \left((xY_1/\kappa)^2 - Y_3^2/t^4 \right) \stackrel{\text{def}}{=} W_2.$$

Suppose we condition on $\mathcal{G}_{-2} = \{Y_1, Y_3\}$ instead, i.e., hide Y_2. We have

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

If $W_2 \le 0$, then $F'(x \mid \mathcal{G}_{-2}) = 0$. If $W_2 > 0$,

$$F(x \mid \mathcal{G}_{-2}) = \mathbb{P}[-\sqrt{W_2} \le Y_2 \le \sqrt{W_2} \mid W_2] = \Phi((\sqrt{W_2} - \mu_2)/\sigma_2) - \Phi(-(\sqrt{W_2} + \mu_2)/\sigma_2)$$

and

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

Suppose we condition on $\mathcal{G}_{-2} = \{Y_1, Y_3\}$ instead, i.e., hide Y_2. We have

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

If $W_2 \le 0$, then $F'(x \mid \mathcal{G}_{-2}) = 0$. If $W_2 > 0$,

$$F(x \mid \mathcal{G}_{-2}) = \mathbb{P}[-\sqrt{W_2} \le Y_2 \le \sqrt{W_2} \mid W_2] = \Phi((\sqrt{W_2} - \mu_2)/\sigma_2) - \Phi(-(\sqrt{W_2} + \mu_2)/\sigma_2)$$

and

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

For conditioning on $\mathcal{G}_{-3}$, same analysis as for $\mathcal{G}_{-2}$, by symmetry, and we get

$$f(x \mid \mathcal{G}_{-3}) = F'(x \mid \mathcal{G}_{-3}) = \frac{\phi((\sqrt{W_3} - \mu_3)/\sigma_3) + \phi(-(\sqrt{W_3} + \mu_3)/\sigma_3)}{t^4 x (Y_1/\kappa)^2/(\sigma_3 \sqrt{W_3})} > 0.$$

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

Instead of choosing a single conditioning k, we can take a convex combination:

$$\hat{f}(x) = \alpha_1 f(x \mid \mathcal{G}_{-1}) + \alpha_2 f(x \mid \mathcal{G}_{-2}) + \alpha_3 f(x \mid \mathcal{G}_{-3}),$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. This is equivalent to taking $f(x \mid \mathcal{G}_{-1})$ as the main estimator and the differences $f(x \mid \mathcal{G}_{-2}) - f(x \mid \mathcal{G}_{-1})$ and $f(x \mid \mathcal{G}_{-3}) - f(x \mid \mathcal{G}_{-1})$ as control variates (CV). We can use CV theory (least-squares regression) to optimize the α_j's.

Instead of choosing a single conditioning k, we can take a convex combination:

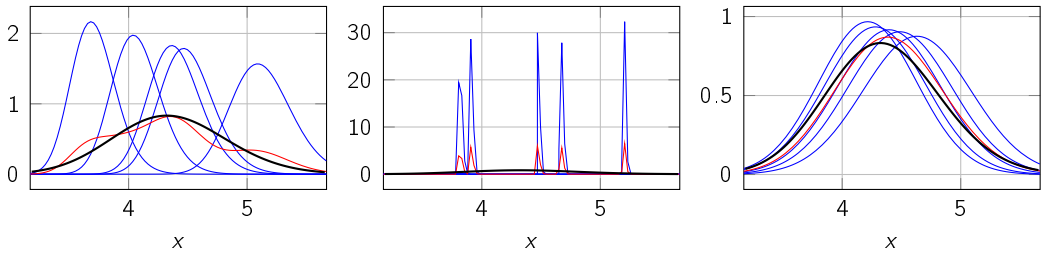
$$\hat{f}(x) = \alpha_1 f(x \mid \mathcal{G}_{-1}) + \alpha_2 f(x \mid \mathcal{G}_{-2}) + \alpha_3 f(x \mid \mathcal{G}_{-3}),$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. This is equivalent to taking $f(x \mid \mathcal{G}_{-1})$ as the main estimator and the differences $f(x \mid \mathcal{G}_{-2}) - f(x \mid \mathcal{G}_{-1})$ and $f(x \mid \mathcal{G}_{-3}) - f(x \mid \mathcal{G}_{-1})$ as control variates (CV). We can use CV theory (least-squares regression) to optimize the α_j's.

$\hat{\nu}$ (MISE $\approx K n^{-\hat{\nu}}$)					e19 (MISE $= 2^{-\mathrm{e19}}$ for $n = 2^{19}$)				
KDE	$\mathcal{G}_{-1}$	$\mathcal{G}_{-2}$	$\mathcal{G}_{-3}$	comb.	KDE	$\mathcal{G}_{-1}$	$\mathcal{G}_{-2}$	$\mathcal{G}_{-3}$	comb.
0.80	0.97	0.98	0.99	0.98	**14.7**	19.3	14.5	**22.8**	22.5

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 $2^8 = \mathbf{256}$.

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



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

Why is that? Hint: $w^4 = 256$ whereas $t^4 = 16$.

Example: discontinuity issues

Let $X = \max(Y_1, Y_2)$ where Y_1 and Y_2 are independent and continuous.
With $\mathcal{G} = \mathcal{G}_{-2}$ (we hide Y_2):

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

If $F_2(y) > 0$, this function is discontinuous at $x = y$, so Assumption 1 does not hold.
The method does not work in this case.

Example: discontinuity issues

Let $X = \max(Y_1, Y_2)$ where Y_1 and Y_2 are independent and continuous.
With $\mathcal{G} = \mathcal{G}_{-2}$ (we hide Y_2):

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

If $F_2(y) > 0$, this function is discontinuous at $x = y$, so Assumption 1 does not hold.
The method does not work in this case.

One possible trick: generate both Y_1 and Y_2, hide the maximum, and take the density of the max, conditional on the min.
If $Y_1 = y_1$ is the min, the CDE will be $f(x \mid \mathcal{G}) = f_2(y_2 \mid Y_2 > y_1)$.

Example: discontinuity issues

Let $X = \max(Y_1, Y_2)$ where Y_1 and Y_2 are independent and continuous.
With $\mathcal{G} = \mathcal{G}_{-2}$ (we hide Y_2):

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

If $F_2(y) > 0$, this function is discontinuous at $x = y$, so Assumption 1 does not hold.
The method does not work in this case.

One possible trick: generate both Y_1 and Y_2, hide the maximum, and take the density of the max, conditional on the min.
If $Y_1 = y_1$ is the min, the CDE will be $f(x \mid \mathcal{G}) = f_2(y_2 \mid Y_2 > y_1)$.

Same problem and same trick for $X = \min(Y_1, Y_2)$.

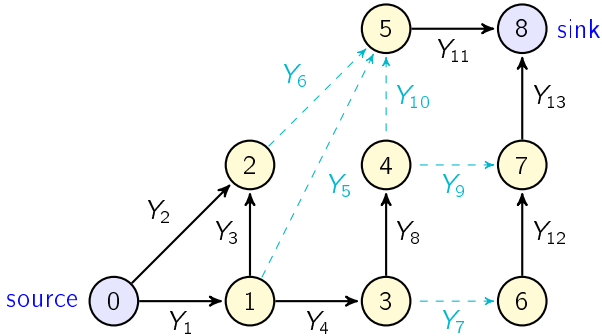
CMC for the SAN Example

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

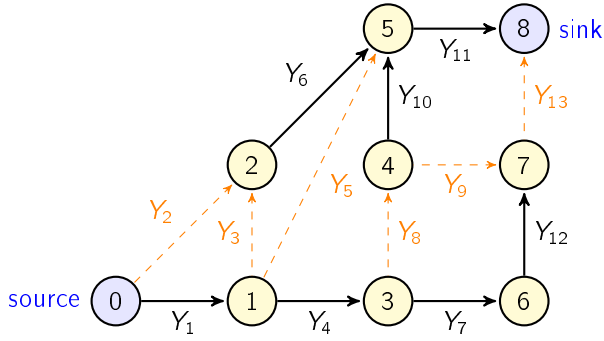
Pick a minimal cut $\mathcal{L}$ between source and sink, and let $F(x \mid \mathcal{G}) = \mathbb{P}[X \le x \mid \{Y_j,\, j \notin \mathcal{L}\}]$.

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 singe Y_j does not work: it does not make the conditional cdf continuous.)



Another minimal cut: $\mathcal{L} = \{2, 3, 5, 6, 9, 13\}$.



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

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

and

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

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 $F(\cdot \mid \mathcal{G})$, and Assumption 1 holds, so $f(x \mid \mathcal{G})$ 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

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

and

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

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 $F(\cdot \mid \mathcal{G})$, and Assumption 1 holds, so $f(x \mid \mathcal{G})$ 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.

Waiting-time distribution in a single-server queue

FIFO queue, arbitrary arrival process, independent service times with cdf G and density g.

The system starts empty and evolves over a day of length τ.
T_j = arrival time of customer j, $T_0 = 0$,
$A_j = T_j - T_{j-1}$ = jth interarrival time,
S_j = service time of customer j,
W_j = waiting time of customer j.

Waiting-time distribution in a single-server queue

FIFO queue, arbitrary arrival process, independent service times with cdf G and density g.

The system starts empty and evolves over a day of length τ.
T_j = arrival time of customer j, $T_0 = 0$,
$A_j = T_j - T_{j-1}$ = jth interarrival time,
S_j = service time of customer j,
W_j = waiting time of customer j.

Lindley recurrence: $W_1 = 0$ and $W_j = \max(0, W_{j-1} + S_{j-1} - A_j)$ for $j \geq 2$.

Random number of customers in the day: $N = \max\{j \geq 1 : T_j < \tau\}$.

Let W be the waiting time of a "random" customer.
Want to estimate $p_0 = \mathbb{P}[W = 0]$ (easy) and density f of W over $(0, \infty)$.

For a random customer over an infinite number of days, we have (renewal reward theorem):[29]

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

The density $f(x)$ is the derivative of the numerator with respect to x, divided by $\mathbb{E}[N]$.

For a random customer over an infinite number of days, we have (renewal reward theorem):[29]

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

The density $f(x)$ is the derivative of the numerator with respect to x, divided by $\mathbb{E}[N]$.

Cannot take the derivative inside the expectation.
CMC: hide the service time S_{j-1} of the previous customer. Replace $\mathbb{I}[W_j \le x]$ by

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

For a random customer over an infinite number of days, we have (renewal reward theorem):[29]

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

The density $f(x)$ is the derivative of the numerator with respect to x, divided by $\mathbb{E}[N]$.

Cannot take the derivative inside the expectation.
CMC: hide the service time S_{j-1} of the previous customer. Replace $\mathbb{I}[W_j \le x]$ by

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

For $x > 0$, we have $P_j'(x) = \mathrm{d}P_j(x)/\mathrm{d}x = g(x + A_j - W_{j-1})$ and

$$\hat{f}(x) = \frac{1}{\mathbb{E}[N]} \sum_{j=1}^{N} P_j'(x).$$

This is extended CMC: we condition on different information for different customers.
We replicate this for n days and take the average. Often, we already know $\mathbb{E}[N]$.

Other possibilities: Can also hide A_j for customer j, etc. Steady-state case.

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 $\mathbf{u} \in [0,1)^s$:

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

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 $\mathbf{u} \in [0,1)^s$:

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

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

CDE sample: $\tilde{g}'(x, \mathbf{U}_1), \ldots, \tilde{g}'(x, \mathbf{U}_n)$ where $\{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ 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 $\mathbf{u} \in [0,1)^s$:

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

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

CDE sample: $\tilde{g}'(x, \mathbf{U}_1), \ldots, \tilde{g}'(x, \mathbf{U}_n)$ where $\{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ 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.

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 $\mathbf{u} \in [0,1)^s$:

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

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

CDE sample: $\tilde{g}'(x, \mathbf{U}_1), \ldots, \tilde{g}'(x, \mathbf{U}_n)$ where $\{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ 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.

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 $\mathbf{u} \in [0,1)^s$:

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

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

CDE sample: $\tilde{g}'(x, \mathbf{U}_1), \ldots, \tilde{g}'(x, \mathbf{U}_n)$ where $\{\mathbf{U}_1, \ldots, \mathbf{U}_n\}$ 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.

Questions?

A likelihood ratio density estimator (LRDE)

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

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

Want to change the integrand into a continuous function of x, so we can take the derivative w.r.t. x inside the integral.

A likelihood ratio density estimator (LRDE)

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

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

Want to change the integrand into a continuous function of x, so we can take the derivative w.r.t. x inside the integral.

Main idea: Make a change of variable $\mathbf{y} \mapsto \mathbf{z} = \mathbf{z}(x)$ of the form $\mathbf{y} = \varphi_x(\mathbf{z})$, with Jacobian $|J_x(\mathbf{z})|$, so that $\tilde{h}(\mathbf{z}) = h(\varphi_x(\mathbf{z}))/x$ no longer depends on x for given $\mathbf{z}$. Then rewrite

$$F(x) = \int_{\mathbb{R}^d} \mathbb{I}[\tilde{h}(\mathbf{z}) \le 1]\, f_Y(\varphi_x(\mathbf{z}))|J_x(\mathbf{z})|\, \mathrm{d}\mathbf{z}.$$

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

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

where

$$L(\mathbf{z}; x, x_0) = \frac{f_Y(\varphi_x(\mathbf{z}))|J_x(\mathbf{z})|}{f_Y(\varphi_{x_0}(\mathbf{z}))|J_{x_0}(\mathbf{z})|}$$

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

$$\begin{aligned}
F'(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \int_{\mathbb{R}^d} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \; L(\mathbf{z}; x, x_0) \; f_Y(\varphi_{x_0}(\mathbf{z}))|J_{x_0}(\mathbf{z})|\mathrm{d}\mathbf{z} \\
&= \int_{\mathbb{R}^d} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \left(\frac{\mathrm{d}}{\mathrm{d}x} L(\mathbf{z}; x, x_0) \right) f_Y(\varphi_{x_0}(\mathbf{z}))|J_{x_0}(\mathbf{z})|\mathrm{d}\mathbf{z} \\
&= \int_{\mathbb{R}^d} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \left(\frac{\mathrm{d}}{\mathrm{d}x} L(\mathbf{z}; x, x_0) \right) \frac{f_Y(\varphi_x(\mathbf{z}))|J_x(\mathbf{z})|}{L(\mathbf{z}; x, x_0)} \mathrm{d}\mathbf{z} \\
&= \int_{\mathbb{R}^d} \mathbb{I}[\tilde{h}(\mathbf{z}) \leq 1] \left(\frac{\mathrm{d}}{\mathrm{d}x} \ln L(\mathbf{z}; x, x_0) \right) f_Y(\varphi_x(\mathbf{z}))|J_x(\mathbf{z})|\mathrm{d}\mathbf{z} \\
&= \int_{\mathbb{R}^d} \mathbb{I}[h(\mathbf{y}) \leq x] S(\mathbf{y}, x) f_Y(\mathbf{y}) \mathrm{d}\mathbf{y}
\end{aligned}$$

where

$$\begin{aligned} S(\mathbf{y}, x) &= S(\varphi_x(\mathbf{z}), x) = \frac{\mathrm{d} \ln L(\mathbf{z}; x, x_0)}{\mathrm{d}x} = \frac{\mathrm{d} \ln(f_Y(\varphi_x(\mathbf{z}))|J_x(\mathbf{z})|)}{\mathrm{d}x} \\ &= (\nabla(\ln f_Y)(\mathbf{y})) \cdot (\nabla_x \varphi_x(\mathbf{z})) + \frac{\mathrm{d} \ln |J_x(\mathbf{z})|}{\mathrm{d}x} \end{aligned}$$

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

$$\begin{aligned} S(\mathbf{y}, x) &= S(\varphi_x(\mathbf{z}), x) = \frac{\mathrm{d} \ln L(\mathbf{z}; x, x_0)}{\mathrm{d}x} = \frac{\mathrm{d} \ln(f_Y(\varphi_x(\mathbf{z}))|J_x(\mathbf{z})|)}{\mathrm{d}x} \\ &= (\nabla(\ln f_Y)(\mathbf{y})) \cdot (\nabla_x \varphi_x(\mathbf{z})) + \frac{\mathrm{d} \ln |J_x(\mathbf{z})|}{\mathrm{d}x} \end{aligned}$$

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 $\mathbf{Y}$ (Glynn 1987, Asmussen and Glynn 2007).

Assumption LR. With probability 1 over realizations of $\mathbf{Y} = \varphi_x(\mathbf{Z})$, $f_Y(\varphi_x(\mathbf{Z}))|J_x(\mathbf{Z})|$ is continuous in x over $[a, b]$ and is differentiable in x except perhaps at a countable set of points $D(\mathbf{Y}) \subset [a, b]$. There is also a random variable Γ defined over the same probability space as $\mathbf{Y}$, such that $\mathbb{E}[\Gamma^2] \leq K_\gamma$ for some constant $K_\gamma < \infty$, and for which

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

Assumption LR. With probability 1 over realizations of $\mathbf{Y} = \varphi_x(\mathbf{Z})$, $f_Y(\varphi_x(\mathbf{Z}))|J_x(\mathbf{Z})|$ is continuous in x over $[a, b]$ and is differentiable in x except perhaps at a countable set of points $D(\mathbf{Y}) \subset [a, b]$. There is also a random variable Γ defined over the same probability space as $\mathbf{Y}$, such that $\mathbb{E}[\Gamma^2] \leq K_\gamma$ for some constant $K_\gamma < \infty$, and for which

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

Proposition LR. Under Assumption LR, $\mathbb{I}[h(\mathbf{Y}) \leq x]\, S(\mathbf{Y}, x)$ is an unbiased estimator of $f(x)$ at almost all $x \in [a, b]$, with variance bounded uniformly by K_γ.

Example. Let $X = h(\mathbf{Y}) = Y_1 + \cdots + Y_d = \mathbf{1} \cdot \mathbf{Y}$ where $\mathbf{Y}$ has multivariate density f_Y. Take $\mathbf{y} = \varphi_x(\mathbf{z}) = x\mathbf{z}$, which gives $\tilde{h}(\mathbf{z}) = h(\mathbf{y})/x = \mathbf{1} \cdot \mathbf{y}/x = \mathbf{1} \cdot \mathbf{z}$, $|J_x(\mathbf{z})| = x^d$, and

$$S(\mathbf{y}, x) = \left(d + (\nabla(\ln f_Y)(\mathbf{y})) \cdot \mathbf{y}\right)/x.$$

Laub, Salomone, Botev (2019) proved that $\mathbb{I}[\mathbf{1} \cdot \mathbf{Y} \leq x]\, S(\mathbf{Y}, x)$ is an unbiased estimator of the density of X for this special case. Their paper motivated our more general LRDE.

Example. Let $X = h(\mathbf{Y}) = Y_1 + \cdots + Y_d = \mathbf{1} \cdot \mathbf{Y}$ where $\mathbf{Y}$ has multivariate density f_Y. Take $\mathbf{y} = \varphi_x(\mathbf{z}) = x\mathbf{z}$, which gives $\tilde{h}(\mathbf{z}) = h(\mathbf{y})/x = \mathbf{1} \cdot \mathbf{y}/x = \mathbf{1} \cdot \mathbf{z}$, $|J_x(\mathbf{z})| = x^d$, and

$$S(\mathbf{y}, x) = \left(d + (\nabla(\ln f_Y)(\mathbf{y})) \cdot \mathbf{y}\right)/x.$$

Laub, Salomone, Botev (2019) proved that $\mathbb{I}[\mathbf{1} \cdot \mathbf{Y} \leq x]\, S(\mathbf{Y}, x)$ is an unbiased estimator of the density of X for this special case. Their paper motivated our more general LRDE.

If $Y_1, \ldots, Y_d$ are independent with $Y_j \sim f_j$, then $\ln f_Y(\mathbf{y}) = \sum_{j=1}^d \ln f_j(y_j)$ and

$$S(\mathbf{y}, x) = \frac{d + \nabla(\ln f_Y)(\mathbf{y}) \cdot \mathbf{y}}{x} = \frac{1}{x}\left(d + \sum_{j=1}^d \frac{f_j'(y_j)}{f_j(y_j)}\right).$$

Example. Let $X = h(\mathbf{Y}) = Y_1 + \cdots + Y_d = \mathbf{1} \cdot \mathbf{Y}$ where $\mathbf{Y}$ has multivariate density f_Y. Take $\mathbf{y} = \varphi_x(\mathbf{z}) = x\mathbf{z}$, which gives $\tilde{h}(\mathbf{z}) = h(\mathbf{y})/x = \mathbf{1} \cdot \mathbf{y}/x = \mathbf{1} \cdot \mathbf{z}$, $|J_x(\mathbf{z})| = x^d$, and

$$S(\mathbf{y}, x) = \left(d + (\nabla(\ln f_Y)(\mathbf{y})) \cdot \mathbf{y}\right)/x.$$

Laub, Salomone, Botev (2019) proved that $\mathbb{I}[\mathbf{1} \cdot \mathbf{Y} \leq x]\, S(\mathbf{Y}, x)$ is an unbiased estimator of the density of X for this special case. Their paper motivated our more general LRDE.

If $Y_1, \ldots, Y_d$ are independent with $Y_j \sim f_j$, then $\ln f_Y(\mathbf{y}) = \sum_{j=1}^d \ln f_j(y_j)$ and

$$S(\mathbf{y}, x) = \frac{d + \nabla(\ln f_Y)(\mathbf{y}) \cdot \mathbf{y}}{x} = \frac{1}{x}\left(d + \sum_{j=1}^{d} \frac{f_j'(y_j)}{f_j(y_j)}\right).$$

In this independent case, an alternative (simpler) transformation is $\mathbf{y} = \varphi_x(\mathbf{z}) = (z_1 + x, z_2, \ldots, z_d)$. Then, $\nabla_x \varphi_x(\mathbf{z}) = (1, 0, \ldots, 0)^{\mathsf{t}}$, $|J_x(\mathbf{z})| = 1$, and $S(\mathbf{y}, x) = f_1'(y_1)/f_1(y_1)$, giving the estimator $\mathbb{I}[\mathbf{1} \cdot \mathbf{Y} \leq x] \cdot S(\mathbf{Y}, x)$.
This matches Example VII.5.7 of Asmussen and Glynn (2007).

A GLR density estimator

Peng et al. (2020) proposed an adaptation of a generalized likelihood ratio (GLR) method of Peng et al. (2018) to density estimation.

Suppose $X = h(\mathbf{Y}) = h(Y_1, \ldots, Y_d)$ where $Y_1, \ldots, Y_d$ are independent continuous random variables, and Y_j has cdf F_j and density f_j. For $j = 1, \ldots, d$, let $h_j(\mathbf{y}) := \partial h(\mathbf{y})/\partial y_j$, $h_{jj}(\mathbf{y}) := \partial^2 h(\mathbf{y})/\partial y_j^2$, and

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

Assumption GLR. (a) The Lebesgue measure of $h^{-1}((x-\epsilon, x+\epsilon))$ in $\mathbb{R}^d$ goes to 0 when $\epsilon \to 0$ (this means essentially that the density is bounded around x).
(b) The set $P(x) = \{\mathbf{y} \in \mathbb{R}^d : h(\mathbf{y}) \le x\}$ is measurable, the functions h_j, h_{jj}, and Ψ_j are well defined over it, and $\mathbb{E}[\mathbb{I}[X \le x] \cdot \Psi_j^2(\mathbf{Y})] < \infty$.

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

Proof and more details: see Peng et al (2020). Take a linear combination of the $\Psi_j(\mathbf{Y})$'s.

Experimental setting for numerical experiments

We tested the methods on some examples.
For each n considered, we compute each estimator with n samples,
evaluate it at a set of $n_e = 128$ evaluation points over $[a, b]$,
repeat this $n_r = 100$ times, compute the variance at each evaluation point to estimate the IV. For the KDE, we also estimated the ISB.

Experimental setting for numerical experiments

We tested the methods on some examples.
For each n considered, we compute each estimator with n samples,
evaluate it at a set of $n_e = 128$ evaluation points over $[a, b]$,
repeat this $n_r = 100$ times, compute the variance at each evaluation point to estimate the IV. For the KDE, we also estimated the ISB.

We repeat this for $n = 2^{14}, \ldots, 2^{19}$ and fit the model $\text{MISE} = \kappa n^{-\nu}$ by linear regression in log-log scale. We report $\hat{\nu}$ and also the MISE for $n = 2^{19}$ which is $2^{-\text{e19}}$.

Experimental setting for numerical experiments

We tested the methods on some examples.
For each n considered, we compute each estimator with n samples,
evaluate it at a set of $n_e = 128$ evaluation points over $[a, b]$,
repeat this $n_r = 100$ times, compute the variance at each evaluation point to estimate the IV. For the KDE, we also estimated the ISB.

We repeat this for $n = 2^{14}, \ldots, 2^{19}$ and fit the model $\text{MISE} = \kappa n^{-\nu}$ by linear regression in log-log scale. We report $\hat{\nu}$ and also the MISE for $n = 2^{19}$ which is $2^{-\text{e19}}$.

MC and RQMC Point sets:

- MC: Independent points,
- Lat+s: lattice rule with a random shift modulo 1,
- Lat+s+b: lattice rule with a random shift modulo 1 + baker's transformation,
- LMS: Sobol' points with left matrix scrambling (LMS) + digital random shift.

Cantilever beam example

Estimated MISE $= Kn^{-\hat{\nu}}$.

	$\hat{\nu}$					e19				
	KDE	$\mathcal{G}_{-3}$	CDE-c	LRDE	GLR	KDE	$\mathcal{G}_{-3}$	CDE-c	LRDE	GLR
MC	0.76	0.99	0.98	1.03	1.02	15.8	22.8	22.5	16.8	14.1
Lat+s	1.03	2.06	2.04	1.55	1.38	21.9	41.6	41.9	26.4	23.4
Lat+s+b	0.93	2.27	2.25	1.25	1.37	21.0	46.8	47.0	24.7	23.3
Sob+LMS	0.97	2.21	2.21	1.31	1.32	21.5	45.7	46.1	25.6	23.4

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 billions**.

SAN Example

Estimated MISE $= Kn^{-\hat{\nu}}$.

		$\hat{\nu}$	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. GLR does not apply to this example.

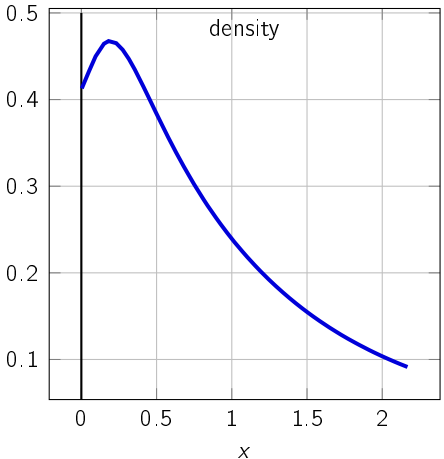
Waiting-time distribution in a single-server queue

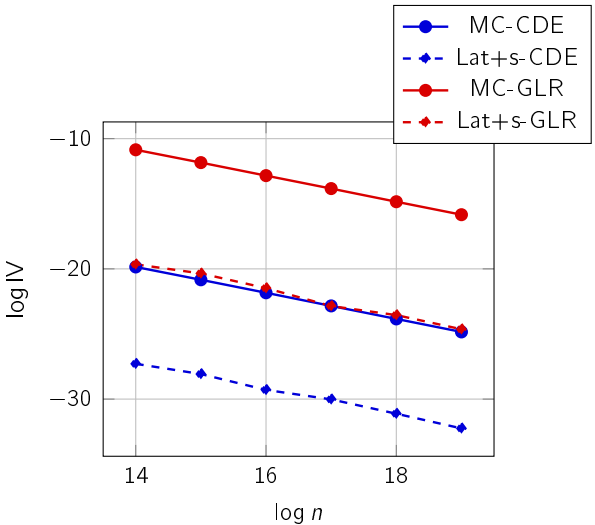
Take Poisson arrivals at rate $\lambda = 1$ over 60 time units. This gives $\mathbb{E}[N] = 60$.
Service times S_j lognormal$(\mu, \sigma^2) = (-0.7, 0.4)$.
This gives $\mathbb{E}[S_j] = e^{-0.5} \approx 0.6065$ and $\mathrm{Var}[S_j] = e^{-1}(e^{0.4} - 1) \approx 0.18093$.

Here the dimension s is random and unbounded, so we used only Korobov lattices, since they naturally provide unbounded dimension.

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

In terms of MISE for $n = 2^{19}$, CDE beats GLR by a factor of about $\mathbf{2^9 = 512}$ with MC and by a factor of about $\mathbf{2^7 = 128}$ with the randomly-shifted Korobov lattice.


density
0.5
0.4
0.3
0.2
0.1
0
0.5
1
1.5
2
x


MC-CDE
Lat+s-CDE
MC-GLR
Lat+s-GLR
−10
−20
−30
log IV
14
16
18
log n

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 $f(x \mid \mathcal{G})$ is easy to compute, the LRDE and the GLR can be good unbiased alternatives.
  Drawback: they do not get along so well with RQMC because they are often discontinuous in $\mathbf{U}$. Maybe 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.
- S. Asmussen and P. W. Glynn. Stochastic Simulation. Springer-Verlag, 2007.
- A. Ben Abdellah, P. L'Ecuyer, A. B. Owen, and F. Puchhammer. Density estimation by Randomized Quasi-Monte Carlo. https://arxiv.org/abs/1807.06133, 2018.
- 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*, 2017.
- 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, F. Puchhammer, and A. Ben Abdellah. Monte Carlo and Quasi-Monte Carlo Density Estimation via Conditioning. `https://arxiv.org/abs/1906.04607`. 2019.
- Y. Peng, M. C. Fu, B. Heidergott, H. Lam. Maximum likelihood estimation by Monte Carlo simulation: Towards data-driven stochastic modeling. Operations Research, 2020, to appear.
- Y. Peng, M. C. Fu, J. Q. Hu, B. Heidergott. A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. Operations Research 66(2): 487–499, 2018.
- D. W. Scott. Multivariate Density Estimation. Wiley, 2015.