The Robust Randomized Quasi Monte Carlo method, applications to integrating singular functions

Emmanuel Gobet, Matthieu Lerasle, David Métivier

The Robust Randomized Quasi Monte Carlo method, applications to integrating singular functions *

Emmanuel Gobet† Matthieu Lerasle‡ David Métivier§

September 15, 2023

ABSTRACT: We are given a simulation budget of B points to calculate an expectation $\mu = \mathbb{E}(F(U))$. The standard Monte Carlo method achieves a root mean squared risk of order $1/\sqrt{B}$, while a Randomized Quasi Monte Carlo (RQMC) method achieves an accuracy $\sigma_B \ll 1/\sqrt{B}$, as long as F is only assumed square integrable. The question we address in this work is, given a budget B and a confidence level $1-\delta$, what is the optimal size of error tolerance such that $\mathbb{P}(|\texttt{Est} - \mu| > \texttt{TOL}) \leq \delta$ for an estimator $\texttt{Est}$ to be determined? We show that a judicious choice of "robust" aggregation methods coupled with RQMC methods allows reaching the best $\texttt{TOL}$, with provable optimality as $B \to +\infty$. This study is supported by numerical experiments, ranging from bounded $F(U)$ to heavy-tailed $F(U)$, the latter being well suited to singular F.

KEYWORDS: Robust statistics; Quasi Monte Carlo methods; PAC bounds

MSC2020: 62G35; 62G15; 11K45

1 Introduction

Context. The Monte Carlo method for computing an integral is one of the most popular numerical schemes and has been recognized by the Society for Industrial and Applied Mathematics [10] among the 10 algorithms which have most influenced the development and practice of the engineering sciences during the 20th century. We are concerned with the evaluation of a d-dimensional integral of the form

$$\mu = \int_{[0,1]^d} F(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}(F(\boldsymbol{U})), \tag{1.1}$$

*This action benefited from the support of the Chair Stress Test, Risk Management and Financial Steering, led by the French ÃĽcole polytechnique and its foundation and sponsored by BNP Paribas and from the Chair Energies Durables, led by CEA, ÃĽcole polytechnique and EDF.

†CMAP, CNRS, ÃĽcole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau, France. emmanuel.gobet@polytechnique.edu

‡CREST, ENSAE, Institut Polytechnique de Paris, 5 Avenue Le Chatelier, 91120 Palaiseau, France. matthieu.lerasle@ensae.fr

§CMAP, CNRS, ÃĽcole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau, France. david.metivier@polytechnique.edu

with $F : [0,1]^d \mapsto \mathbb{R}$ and $\boldsymbol{U} \sim \mathcal{U}([0,1]^d)$. The above is written under a generic form of expectation with respect to (w.r.t.) the uniform distribution over the unit cube: it is not really a restriction since even in the case where the Quantity of Interest (QoI) is an expectation w.r.t. a more complex distribution, in practice we usually generate uniform samples and after some transformations specific to the sampling scheme, we are back to Eq. (1.1) with some F related to the QoI and to the sampling scheme.

The current work is devoted to the use of Randomized Quasi Monte Carlo Sequences (RQMCS) to evaluate μ and, specifically, how to derive some confidence intervals that are provably better than those obtained with usual independent random sequences (standard Monte Carlo method). Monte Carlo and Quasi Monte Carlo methods are discussed in several monographs [45, 60] and review articles [63, 5, 22]. For RQMCS, see [32], [15], [31] for reviews. From now on, we will use the term Monte Carlo alone when referring to the standard Monte Carlo method, i.e. i.i.d. uniform sampling. We fix a few notations:

- we make use of n independent mini-batches, each providing an estimate $(\overline{\mu}_N^{(i)}, i = 1, \ldots, n)$ of μ;
- the i-th mini-batch estimate is computed as an average of F along a sequence of N points $(\boldsymbol{R}_1^{(i)}, \ldots, \boldsymbol{R}_N^{(i)})$ so that

$$\overline{\mu}_N^{(i)} := \frac{1}{N} \sum_{j=1}^{N} F(\boldsymbol{R}_j^{(i)}). \tag{1.2}$$

  The RQMCS $(\boldsymbol{R}_1^{(i)}, \ldots, \boldsymbol{R}_N^{(i)})$ are usually generated using a low-discrepancy deterministic sequence including an appropriate randomization: for such techniques, see the Cranley-Patterson rotation [11], scrambled nets [47][16, Chapter 13]. Usually the size N being given (as a power of an integer in the case of (t, m, d)-nets), one has to choose a RQMCS of that size; for extensible sequence, see [23][16, Chapter 4]. The key properties of the RQMCS are that each $\boldsymbol{R}_j^{(i)}$ is distributed as $\mathcal{U}([0,1]^d)$, while having the N points fill more regularly the unit cube than usual independent random points so that

$$\sigma_N^2 := \mathrm{Var}\left(\overline{\mu}_N^{(i)}\right) = o(N^{-1}) \qquad \text{as } N \to +\infty, \tag{1.3}$$

  for appropriate functions F and RQMCS. See [49, Theorem 1] or [18, Corollary 1] showing the above estimate under the sole assumption that F is square integrable and when the RQMCS is based on a (t, m, d) scrambled net. For strong law of large numbers for $F(U) \in L^p$ $(p > 1)$, see the recent work [57]. In [54, Theorem 3], it is proved that $\sigma_N = \mathcal{O}\left(N^{-3/2}[\log(N)]^{(d-1)/2}\right)$ for smooth functions and scrambled nets. We refer to Section 4.3 for quantitative results. Unless specified explicitly all the convergence results will be meant for a given F in some functional space $\mathcal{F}$, as opposed to uniformly over all F in $\mathcal{F}$ as in [29]. However, we will compare our results (for a given F) with those in the latter reference.

All in all, as an estimator of μ, we can consider the Empirical Mean (EM) over the n mini-batches:

$$\overline{\mu}_{N,n} := \frac{1}{n} \sum_{i=1}^{n} \overline{\mu}_N^{(i)}. \tag{1.4}$$

Because $\boldsymbol{R}_j^{(i)}$ is uniformly distributed, $\overline{\mu}_{N,n}$ is advantageously an unbiased estimator of μ. The quadratic error equals its variance:

$$\mathbb{E}\left(|\overline{\mu}_{N,n}-\mu|^2\right)=\mathrm{Var}\left(\overline{\mu}_{N,n}\right)=\frac{\sigma_N^2}{n}=o\left(N^{-1}\right)n^{-1},$$

as $N\to\infty$, showing that the distribution of $\overline{\mu}_{N,n}$ concentrates much around its mean. We aim at designing non-asymptotic confidence intervals (CIs) of μ for estimators (possibly different from $\overline{\mu}_{N,n}$) which would efficiently leverage the fast convergence rate (1.3). One of our main results (see Theorem 5.1 for a precise statement) states that, for any uncertainty level $\delta\in(0,1)$, for some specific estimator $\widehat{\mu}_B:=\widehat{\mu}_B(\overline{\mu}_N^{(i)}:1\le i\le n)$ based on a budget of $B=n\times N$ points, we have

$$\mathbb{P}\left(|\widehat{\mu}_B-\mu|\le(\sigma_N\sqrt{N})L_{\delta,n}\sqrt{\frac{\log(2/\delta)}{B}}\right)\ge 1-\delta \tag{1.5}$$

where $L_{\delta,n}\approx\sqrt{2}$, $\sigma_N\sqrt{N}\to 0$ as $N\to+\infty$ under the sole assumption $\mathbb{E}\left(F^2(U)\right)<+\infty$, and with some fast rates to 0 when F satisfies some regularity/singularity estimates (see Theorem 4.4). Our results complement lower bound results of [29] by providing explicit estimators achieving their lower bounds (up to log terms) and by establishing convergence rates in new situations. See Section 5 for a detailed discussion.

To achieve this program, we will first analyze in Section 2 that the usual empirical mean (1.4) can not – in general – achieve the above bounds (1.5), thus we have to propose new estimators of μ based on different aggregations of $\overline{\mu}_N^{(i)}$. Combining robust estimator with established randomized algorithms for achieving better confidence properties has been done in the past with the median trick (or probability amplification) [27]. It is similar to what is called median-of-means (see Section 3.3). Here we investigate, in a more generic framework, robust estimators that have better theoretical guarantees and perform better numerically. In particular, these estimators should satisfy finite sample risk bounds like (1.5) that

- are proportional to the optimal rate $\sigma_N/\sqrt{n}$ with respect to n,N, thus fully benefiting from the LD property of RQMCS;
- are proportional to the optimal $\sqrt{\log 1/\delta}$ rate with respect to δ, to be relevant even for very small uncertaity levels δ.

Interestingly, this question is related to the work of Catoni [6]. In this paper, the author wondered if one can design an estimator of the expectation $\mu=\mathbb{E}(X)$ of a random variable X from an independent identically distributed (i.i.d.) sample $X_1,\dots,X_n$ with sub-Gaussian deviation tails, assuming only a finite second moment of each variable X_i. He proved that it is possible to build estimators $\hat{\mu}_{\delta,\sigma}$ depending on the uncertainty level δ and the variance σ such that

$$\mathbb{P}\left(|\hat{\mu}_{\delta,\sigma}-\mu|>C_n\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right)\le\delta\ ,$$

with leading constant C_n asymptotically as $n\to\infty$ not larger than the optimal value $\sqrt{2}$ and independent of the law of X, level $\delta\ge\exp(-cn)$ for some absolute constant c. Notice that this

bound exhibits precisely the behavior we are seeking for our estimators when applied to the random variables $X_i = \hat{\mu}_N^{(i)}$. Let us pause for a second to stress important points regarding this result: Compared to the empirical mean, there are indeed important restrictions to this result, that may be undesirable in our application.

- First, the Catoni estimator depends on the variance parameter σ^2, which is typically unknown in applications.
- Second, it also depends on the uncertainty level δ.
- Third, the uncertainty level δ has to be sufficiently large (at least of order $\exp(-cn)$).

The biggest problem to apply this result is the dependency of the estimator on σ^2 (look at the numerical experiments from Section 6 though where Catoni's strategy is applied, replacing σ^2 by an appropriate estimator). However, Catoni's work was followed by [12] that showed that the median-of-means principle can be used to build estimators achieving similar theoretical results (although with worse leading constants $C_n > \sqrt{2}$). Compared with Catoni's estimators, median-of-means estimators *do not depend on* σ and are robust to the presence of a few outliers. This proves that the first restriction we discussed can be overcome. The paper [12] also showed, however, that the dependency in the confidence level $1-\delta$ is mandatory without further information on the distribution of X and that the third restriction always holds unless X is sub-Gaussian with a link between its variance and its ψ_2-norm. Other robust estimators have then been proposed and analyzed, including Trimmed-mean estimators, that will be presented and discussed in detail in Section 3.

Asymptotic error assessment. Last but not least, to complete the state-of-the-art picture, let us mention some references about evaluating RQMC error asymptotically (although our focus is on non-asymptotic estimates). In numerical software packages, stopping criteria are sometimes proposed [2, 8] as approximate error assessment methods. Besides, in some cases, the distribution of $\bar{\mu}_N^{(i)}$ for N large can be close to a normal distribution: if so, $\overline{\mu}_{N,n}$ is also (approximately) normally distributed, so the Empirical Mean is exactly a (sub)-Gaussian estimators (and thus optimal); then, robust estimators are not expected to yield much improvement against EM. These cases of asymptotic normality are known for $(0, m, d)$-scrambled nets and for F with Lipschitz continuous first derivatives, see [36]; sufficient conditions are, furthermore, studied in [43]. However, the approximate normality of $\overline{\mu}_N^{(i)}$ may not hold, as exemplified by [33]. All in all, non-asymptotic approaches keep all its advantages compared to the asymptotic point of view. Last, we refer to the work [29] by Kunsch and Rudolf, which has been pointed out to us by the authors when our work was finalized and presented in conference: their work deals with theoretical asymptotic minimal error bounds for confidence intervals. In Section 5, we discuss how our approach fits in their framework by achieving the minimal error bounds.

Our contributions. This paper combines Robust (sub-Gaussian) Estimators to Quasi Monte Carlo method to exhibit Robust Randomized Quasi Monte Carlo (RRQMC) estimates that can be provably highly accurate with a large probability. To the best of our knowledge, this combination is

original. We show that as long as $F(\boldsymbol{U})$ is squared integrable, the RRQMC method converges still faster than the usual Monte Carlo method, and robust methods allow a high concentration of the statistical error around 0. We exhibit, in Theorem 5.1, the associated non-asymptotic confidence interval, and show that for square integrable integrands F or for smoother functional spaces, it reaches the predicted optimal bound (up to a logarithmic term) predicted in [29].

As a side contribution, we show in Theorem 4.4, how quick the Mean Square Error σ_N^2 decreases for integrands with corner singularities. This extends the previous result of [53] on the expected integration error (L_1-error) of these functions. Robust statistics are particularly relevant for this kind of singular function mimicking heavy-tailed distributions. Combining Theorem 5.1 and 4.4 together provides confidence intervals for this case.

Numerical experiments in Section 6 confirm how beneficial the combination of RQMCS and robust estimators is, even in the case of smooth and bounded F.

Outline. Section 2 investigates various theoretical concentration-of-measure inequalities for the empirical mean, in the perspective of arguing for variance reduction effect when deriving non-asymptotic confidence intervals. This overview is interesting on its own and serves as a warm-up before moving to robust mean estimators. In Section 3 we review the Robust Mean estimators existing in the literature and the best confidence bounds available. Section 4 is devoted to present some (Robust) Quasi Monte Carlo methods with their properties. In the last theoretical Section 5 we derive results combining the two previous approaches. In Section 6, we test numerically these various methods.

2 Are there theoretical guarantees that the empirical mean has improved confidence intervals when using a variance reduction method?

The discussion of this section goes beyond the framework of Randomized Quasi Monte Carlo method, and it actually encompasses any variance reduction method: it is aimed at advocating for the use of robust mean methods for benefiting from variance reduction, instead of simple empirical means (EM).

Namely, consider a variance reduction scheme from which we sample n i.i.d. $(\overline{\mu}_N^{(i)})_{i=1,\cdots,n}$ which are unbiased estimate of μ and let us study the accuracy of the empirical mean $\overline{\mu}_{N,n}$ defined in (1.4) in term of the supposedly reduced variance $\sigma_N^2 = \mathrm{Var}\left(\overline{\mu}_N^{(i)}\right)$. The parameter N here should be interpreted as a variance reduction cost equivalent to a sampling effort, that is we should have small $N\sigma_N^2$. We focus on the measure of accuracy through exhibiting CIs which size are proportional to σ_N, it takes the form

$$\mathbb{P}\left(|\overline{\mu}_{N,n} - \mu| \leq \sigma_N g(n,\delta)\right) \geq 1-\delta, \qquad \delta \in (0,1), \tag{2.1}$$

which reads as "Given a confidence level at least equal to $1-\delta$, what is the error bound obtained with that confidence level?". The parameter δ is the so-called *uncertainty*. We retrieve a full variance

reduction effect on the CI if g does depend only n and δ, but not on the distribution of $F(U)$. The inequality (2.1) is somewhat equivalent to the other representation

$$\mathbb{P}\left(|\overline{\mu}_{N,n} - \mu| \geq \sigma_N \varepsilon\right) \leq f(n, \varepsilon), \qquad \varepsilon \geq 0, \tag{2.2}$$

which reads as given an error tolerance $\sigma_N \varepsilon$, how likely does the EM achieve this error tolerance. Informally, the two viewpoints are related by $\varepsilon = g(n, \delta)$ and $\delta = f(n, \varepsilon)$.

Concentration-of-measure inequalities are the crux for deriving CIs such as (2.1). The aim of the following discussion is to show that none of the available concentration-of-measure inequalities is able to fit (2.1) or (2.2) with functions independent of the distribution $F(U)$.

Asymptotic bounds. The Central Limit Theorem gives, as $n \to +\infty$,

$$\mathbb{P}\left(|\overline{\mu}_{N,n} - \mu| \leq \phi^{-1}(1 - \delta/2)\frac{\sigma_N}{\sqrt{n}}\right) \to 1 - \delta \ , \tag{2.3}$$

where $\phi^{-1}(q)$ is the quantile function of the standard Normal distribution. For a 95% CI, we have $\phi(0.975) \simeq 1.96$. For later reference, note that $\phi^{-1}(1-\delta/2) \sim \sqrt{2\log(1/\delta)}$ as $\delta \to 0$. On the one hand, (2.3) is a nice bound which shows well the interest of variance reduction scheme for which σ_N is small compared to $1/\sqrt{N}$. Indeed, in terms of budget $B = n \cdot N$, the term in the bound can be written as $\sqrt{N}\sigma_N/\sqrt{B}$.

On the other hand, this bound is valid only as $n \to +\infty$, and it is not clear how the benefit of a small σ_N is transferred to a probable bound on $|\overline{\mu}_{N,n} - \mu|$ for a finite n. As will be discussed in the following, finite sample concentration inequalities for sums of independent random variables can be used to prove risk bounds for $\overline{\mu}_{N,n}$. However, these are never as good as the asymptotic ones, even under strong assumptions such as boundedness of $\overline{\mu}_N^{(i)}$, and they are deteriorating rapidly when these assumptions are relaxed to allow for unbounded $\overline{\mu}_N^{(i)}$. Before we move to other aggregation procedures of the $\overline{\mu}_N^{(i)}$, let us review classical and recent bounds that can be proved for the empirical mean.

Finite sample bounds for bounded $\overline{\mu}_N^{(i)}$. Assume that $\overline{\mu}_N^{(i)}$ is almost surely bounded by a constant c_N. This happens, for example, when F is bounded by c with $c_N \leq c$. Under this assumption, Hoeffding's inequality [4, Theorem 2.8] gives (2.2) with $f(n, \varepsilon) = 2\exp\left(-\frac{n\sigma_N^2\varepsilon^2}{2c_N^2}\right)$ and (2.1) with $g(n, \delta) = \frac{c_N}{\sigma_N}\sqrt{\log(2/\delta)/n}$. These bounds have a similar flavor as the asymptotic ones, with the important difference due to the ratio $r_N := \frac{c_N}{\sigma_N}$. This yields a huge downgrading of the bounds since when effective, the variance reduction gives $r_N \to +\infty$ as $N \to +\infty$.

In this situation where the standard deviation is much smaller than the sup-norm, it is well known that Hoeffding's inequality can be improved into Bennett's inequality [4, Theorem 2.9]:

$$f(n, \varepsilon) = 2\exp\left(-\frac{n\sigma_N^2}{c_N^2} h\left(\frac{c_N \varepsilon}{\sigma_N}\right)\right), \quad h(u) = (1+u)\log(1+u) - u \text{ for } u \geq 0.$$

The function h behaves like $u \mapsto u^2$ around 0 and like $u \mapsto u\log(u)$ when $u \to \infty$.

To understand the improvement brought by this bound, let us now state more precisely the asymptotic situation with respect to n and N that we will be interested in. Assume that c_N is a constant (which makes sense as it is typically asymptotically $|\mu| > 0$), we let δ be free to allow for very small uncertainty levels, and we will discuss the bounds in the case where the variance $n\sigma_N^2 \to 0$, as this is the case with the usual Monte Carlo sampling. Set the error tolerance $\varepsilon \asymp \frac{c_N}{\sigma_N} \log(1/\delta)/(n \log(c_N \log(1/\delta)/(n\sigma_N^2)))$. Fix first δ, then the quantity appearing in the function h in Benett's bound, $c_N\varepsilon/\sigma_N \asymp [n\sigma_N^2 \log(1/(n\sigma_N^2))]^{-1} \to \infty$. It follows that $h(c_N\varepsilon/\sigma_N) \asymp c_N\varepsilon/\sigma_N \log(c_N\varepsilon/\sigma_N)$ and thus $f(n,\varepsilon) = \delta$. For fixed uncertainty level $1-\delta$, this bound can be compared with the one we got from Hoeffding's inequality. First, we see that the dependency with respect to uncertainty level is slightly worse here, of order $\log(1/\delta)/\log\log(1/\delta)$ instead of $\sqrt{\log(1/\delta)}$ in Hoeffding's result. However, the bound derived from Bennett's inequality yields a substantial improvement with respect to n, N, the new bound being of order $\{n \log(1/n\sigma_N^2)\}^{-1}$ instead of $1/\sqrt{n}$. This risk bound benefits from small σ_N but remains larger than the asymptotic one $\sigma_N/\sqrt{n}$: Indeed,

$$\frac{\{n \log(1/n\sigma_N^2)\}^{-1}}{\sigma_N/\sqrt{n}} = \frac{1}{(n\sigma_N^2)^{1/2} \log(1/n\sigma_N^2)} \to \infty, \qquad \text{when } n\sigma_N^2 \to 0 \ .$$

Finally, both bounds, compared to the asymptotic result, are only valid under the assumption of bounded $\overline{\mu}_N^{(i)}$ (or F), which is still a strong restriction for the QoI (1.1).

Finite sample bounds under exponential moments for $\overline{\mu}_N^{(i)}$. In this paragraph and the following, we show that the boundedness assumption on F can be slightly relaxed without deteriorating Benett's risk bounds by more than a logarithmic factor. Consider first the case where $\overline{\mu}_N^{(i)}$ has finite exponential moments. More precisely, assume the following Bernstein's conditions

$$\mathbb{E}\left(|\overline{\mu}_N^{(i)} - \mu|^q\right) \leq \frac{q!}{2}\sigma_N^2 c_N^{q-2}, \qquad \text{for any integer} \quad q \geq 2,$$

where $c_N > 0$. It is easy to check that this assumption implies in particular that c_N is an upper bound on the Orlicz norm $\|\mu_N^{(i)}\|_{\psi_1} \lesssim c_N$ rather than on the sup-norm in Benett's case. In this case, Bernstein's inequality (see [4, Theorem 2.10]) ensures that (2.2) holds with

$$f(n,\varepsilon) = 2\exp\left(-\frac{n\sigma_N^2\varepsilon^2}{2(\sigma_N^2 + c_N\sigma_N\varepsilon/3)}\right). \tag{2.4}$$

Choosing $\varepsilon = \frac{c_N}{\sigma_N}\log(1/\delta)/n$, this bound becomes

$$\log(f(n,\varepsilon)) = \log 2 - \frac{c_N^2 \log(1/\delta)^2}{2(n\sigma_N^2 + c_N^2 \log(1/\delta)/3)} \ .$$

In the asymptotic $n\sigma_N^2 \to 0$, it yields $f(n,\varepsilon) \leq \delta$ (for δ small enough). Hence, the error bound is only slightly deteriorated compared to the previous paragraph, by extra logarithmic factors $\log(1/(n\sigma_N^2))$ in n, N and $1/\log\log(1/\delta)$ w.r.t. δ, while this result holds under the relaxed assumption. As for the sup-norm, the Orlicz norm c_N is typically asymptotically bounded but does not converge to 0. Therefore, the upper bound (2.4) does not benefit more (actually

slightly worse) from the better variance estimate from a variance reduction scheme. Finally, while this result holds for unbounded F, it still requires a strong sub-exponential behavior of $\mu_N^{(i)}$, which remains much stronger than the existence of a finite second moment.

Finite sample bounds for β-heavy-tailed $\overline{\mu}_N^{(i)}$. In [7], the case where the $\mu_N^{(i)}$ have β-heavy-tailed distributions (see formal definition below) was considered, to allow to cover in particular log-normal distributions (where all polynomial moments are finite, but all exponential moments are infinite). Suppose that there exists $\beta > 1$ such that the Orlicz norm $\|F(\boldsymbol{U})\|_\beta^{\mathtt{HT}}$ is finite, where

$$\|F(\boldsymbol{U})\|_\beta^{\mathtt{HT}} := \inf\left\{c > 0 : \ \mathbb{E}\left(\Psi_\beta\left(|F(\boldsymbol{U})|/c\right)\right) \leq 1\right\},$$
$$\text{with} \quad \Psi_\beta(x) := \exp((\ln(x+1))^\beta) - 1 \quad \text{for} \quad x \geq 0.$$

In this case, it follows from [7, Corollary 2.3] that there exists a universal constant $c > 0$ such that (2.2) holds with

$$f(n,\varepsilon) = 2\exp\left(-\left(\ln\left(1 + \frac{\varepsilon\sigma_N n}{c\left(\sigma_N\sqrt{n} + \|\overline{\mu}_N^{(1)}\|_\beta^{\mathtt{HT}}\Psi_{1/\beta}(n)\right)}\right)\right)^\beta\right).$$

Here, the sequence $\Psi_{1/\beta}(n) \to +\infty$ as $n \to +\infty$ but with $\Psi_{1/\beta}(n)/n^\gamma \to 0$ for any $\gamma > 0$. This bound yields, as in the previous examples, an error bound of order (for small ε)

$$\varepsilon \asymp \frac{\|\overline{\mu}_N^{(1)}\|_\beta^{\mathtt{HT}}\Psi_{1/\beta}(n)\log(1/\delta)^{1/\beta}}{\sigma_N n}, \qquad \text{when } n\sigma_N^2 \to 0 \ ,$$

for an uncertainty level $\log(f(\varepsilon)) = \delta$. The degradation $\log(1/\delta) \leftrightarrow \log(1/\delta)^{1/\beta}$ is the price to pay to assume only a finite β-heavy-tails for $\mu_N^{(i)}$. A strong feature of this result is that the dependency with respect to n, N remains $1/n$ (up to slow varying terms). Regarding the variance reduction problem, there is no reason implying $\|\overline{\mu}_N^{(1)}\|_\beta^{\mathtt{HT}}/\sigma_N$ remains bounded, so these bound still do not leverage fully the benefits of variance reduction.

Finite sample bounds under finite second moment for $\overline{\mu}_N^{(i)}$. In the extreme case where only the second moment of $\overline{\mu}_N^{(i)}$ is finite, then Chebyshev-Cantelli's inequality implies

$$\mathbb{P}\left(\overline{\mu}_{N,n} - \mu \geq \frac{\sigma_N}{\sqrt{\delta n}}\right) \vee \mathbb{P}\left(\overline{\mu}_{N,n} - \mu \leq -\frac{\sigma_N}{\sqrt{\delta n}}\right) \leq \delta \qquad \forall \delta \in (0,1) \ .$$

Here, the dependency with respect to n, N, $\frac{\sigma_N}{\sqrt{n}}$, is the targeted order of magnitude. On the other hand, the dependency on the uncertainty level $1/\sqrt{\delta}$ is significantly deteriorated compared with the previous results, of order $\log(1/\delta)^{1/\beta}$, $\log(1/\delta)$ or even $\sqrt{\log(1/\delta)}$ in the asymptotic case: for $\delta = 5$ % (resp. $\delta = 0.1$ %), we have $1/\sqrt{\delta} = 4.47$ (resp. 31.62), while $\log(1/\delta) = 3$ (resp. 6.91) and $\sqrt{\log(1/\delta)} = 1.73$ (resp. in 2.63) arising in (2.3).

To summarize, none of the finite sample concentration inequalities yield the desired bounds (2.1)-(2.2) for the empirical mean, showing that we can not guarantee the latter fully benefits from the variance reduction effect in terms of the deviation probability.

3 Theory of Robust Mean estimators

We present some standard robust estimators as well as some recent ones. In this whole section, we denote by $X_1, \dots, X_n$ a sequence of i.i.d. real random variables with expectation $\mu = \mathbb{E}(X_1)$ and finite variance $\sigma^2 = \mathrm{Var}(X_1)$. The quantity $\bar{\mu}_n$ is the empirical mean estimator of the $X_1, \dots, X_n$ sample. These results will be applied to our problem by considering $X_i = \overline{\mu}_N^{(i)}$, so $\sigma = \sigma_N$ and $\bar{\mu}_n = \overline{\mu}_{N,n}$. Let also $\widehat{\mu}_n$ denote a generic mean estimator built from $X_1, \dots, X_n$.

In this section, we will present the concentration inequalities with the convention Eq. (2.1), $(g(\delta), \delta)$ as it is customary in this literature.

3.1 Sub-Gaussian distributions

Using

$$\Phi^{-1}(1 - \delta/2) \leq \sqrt{2\log(2/\delta)}$$

and the Central Limit theorem, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{\mu}_n - \mu| \leq \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\right) \geq 1 - \delta$$

for $\delta \in (0, 1)$.

Definition 3.1. *A random variable X is called sub-Gaussian if it satisfies for all $\lambda \in \mathbb{R}$*

$$\mathbb{E}\left(\exp\left(\lambda\left(X - \mu\right)\right)\right) \leq \exp\left(\tilde{\sigma}^2\lambda^2/2\right)$$

where $\tilde{\sigma}$ is proportional to the ψ_2-norm of $X - \mu$.

It is easy to show using Chernoff's bound that sub-Gaussian random variables satisfy, for any $\delta \in (0, 1)$

$$\mathbb{P}\left(|\bar{\mu}_n - \mu| \leq \tilde{\sigma}\sqrt{\frac{2\log(2/\delta)}{n}}\right) \geq 1 - \delta.$$

Actually, it can be shown that the empirical mean only achieves this bound (w.r.t. n and δ, for all δ) if X is sub-Gaussian. In particular, thus, if the distribution of X_1 has a finite variance only, the empirical mean does not have sub-Gaussian tails. [6] shows that, under this assumption, Chebyshev's inequality is tight. In the following sections, we present several robust alternatives to the empirical mean. As explained in the introduction, these estimators typically depend on the uncertainty level δ that has therefore to be set by the statistician. This and our ultimate goal motivate the following definition.

Definition 3.2. *Given $\delta \in (0, 1)$, we say that an estimator of the mean μ is a δ-Sub-Gaussian mean estimators (SGME) $\widehat{\mu}_n \doteq \widehat{\mu}_n(X_1, \dots, X_n)$, if it satisfies*

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| \geq L\sigma\sqrt{\frac{\log(2/\delta)}{n}}\right) \leq \delta.$$

for some $0 < L < +\infty$.

Informally, δ-sub-Gaussian estimators are those we are interested in as they exhibit provable risk bounds of optimal order both with respect to n, σ and δ. It turns out that such estimators exist even under weak finite second moment assumptions on data, but they have to depend on the confidence level.

Remark 3.1 (Optimality of sub-Gaussian Mean Estimator)**.** *SGME are optimal in the sense that for any estimator of the mean $\widehat{\mu}_n$ and uncertainty level $\delta \in (0,1)$, there exists a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ such that*

$$\mathbb{P}\left(\widehat{\mu}_n - \mu \geq r\sigma\right) \leq \delta,$$

where $r \geq \phi^{-1}(1-\delta)$ [6, Proposition 6.1]. Since $\phi^{-1}(1-\delta) \sim \sqrt{2\log(1/\delta)}$ for small δ we must have $L \geq L_{\text{optimal}} = \sqrt{2}$ for any mean estimator working for all finite variance distributions. An estimator satisfying $L = \sqrt{2} + o(1)$ is referred to as "nearly optimal".

3.2 Trimmed Mean

Perhaps the most intuitive robust estimator is the trimmed mean, which removes the most extreme samples and takes the empirical mean on the remaining samples. The asymptotic behavior of this estimator is well known, see for example [64]. However, the non-asymptotic deviation bounds have only recently been investigated. We present here a result from [37, Theorem 6] attributed to [46] (work unpublished at the date of our work).

For $\alpha \leq \beta$, define the truncation function

$$\varphi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x > \beta, \\ x & \text{if } x \in [\alpha, \beta], \\ \alpha & \text{if } x < \alpha. \end{cases}$$

For $y_1, \ldots, y_m \in \mathbb{R}$, let $y_1^* \leq y_2^* \leq \cdots \leq y_m^*$ be a non-decreasing rearrangement.

Theorem 3.1 (Theorem 6 [37])**.** *Let n be an even integer. Let $X_1, \ldots, X_{n/2}, Y_1, \ldots, Y_{n/2}$ be i.i.d. random variables with mean μ and variance σ^2.*

1. *Let $\delta \in (0,1)$ be such that $n > (32/3)\log(8/\delta)$ and*
$$\varepsilon = \frac{32\log(8/\delta)}{3n}.$$
2. *Assume for simplicity that $\varepsilon n/2$ is an integer.*
3. *Let $\alpha = Y^*_{\varepsilon n/2}$ and $\beta = Y^*_{(1-\varepsilon)n/2}$ and set the trimmed mean estimator as*
$$\widehat{\mu}_n = \frac{2}{n}\sum_{i=1}^{n/2} \varphi_{\alpha,\beta}\left(X_i\right).$$

Then

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| \leq 19\sigma\sqrt{\frac{2\log(8/\delta)}{n}}\right) \geq 1 - \delta.$$

In words, the trimmed mean is a δ-sub-Gaussian estimator. It does not depend on σ but it is "structurally not" nearly optimal as a trimmed mean is based essentially on only half of the sample (the other half being used to compute the trimming levels α and β). The constant $L = 19$ appearing in the bound can probably be optimized, it differs from the constant $L = 9$ given in [37, Theorem 6] which is the result of a typo.

3.3 Median-of-means

An alternative to trimmed mean, that has theoretically the same kind of properties, is given by median-of-means. The idea is to partition the dataset into batches, take the mean on each batch, and then aggregate these estimators by taking their median. These estimators are structurally robust to the presence of a few outliers due to the median step, but they can also easily be shown to be δ-sub-Gaussian when the number of blocks is of order $\log(1/\delta)$. We present these estimators and this property formally in this section.

Definition 3.3. *The median-of-means estimator $\widehat{\mu}_n$ of a sample $X_1, \dots, X_n$ is defined as*

$$\widehat{\mu}_n = \text{median}\left(\bar{X}_{J_1}, \dots, \bar{X}_{J_k}\right) \tag{3.1}$$

where $\bar{X}_{J_l} = |J_l|^{-1} \sum_{i \in J_l} X_i$ and J_l is a partition of $\{1, \dots, n\}$, that is $\{1, \dots, n\} = \bigcup_{l=1}^{k} J_l$ and $J_l \bigcap J_{l'} = \emptyset$ for $l \neq l'$, each batch J_l has roughly the same size. In the following, we denote by $m = |J_l|$ assuming $n = km$.

We used the notation $|J| \doteq \text{Card}(J)$ to denote the cardinal of an ensemble. Note that the result depends on the arrangements of the sample.

Median-of-means estimators appeared independently in various communities, see for example [44, 1]. Their sub-Gaussian property was established in [12]. We present here a slightly tighter result.

Theorem 3.2 (Theorem 4.1 [12])**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let m, k be positive integers, assume that $n = mk$. Then the median-of-means estimator $\widehat{\mu}_n$ introduced in Definition 3.3, is a δ-sub-Gaussian estimator with $k = \lceil 8 \log(1/\delta) \rceil \leq n$, that is*

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| \leq \sigma \sqrt{\frac{32 \log(1/\delta)}{n}}\right) \geq 1 - \delta.$$

A permutation-invariant version of median-of-means estimators exists (also related to a higher order Hodges-Lehmann estimator), see the discussion in [42, Section 3.4]. There, the median is taken over all possible equi-partitions of the samples, $\text{median}\left(\bar{X}_J, J \in \mathcal{A}_n^{(k)}\right)$, $\mathcal{A}_n^{(k)} = \{J : J \subseteq \{1, \dots, n\}, |J| = k\}$. In this case, it is not clear how the deviation bounds are modified, but we expect a better estimate. Of course, $\left|\mathcal{A}_n^{(k)}\right| = \binom{n}{k}$ so this estimator is not tractable in practice.

3.4 Z–estimators

Parameter estimation often relies on minimization (or maximization) criteria, such as the Maximum Likelihood estimator. Another classical example is the problem of estimating a location parameter θ by minimizing the mean square error $\sum_i (X_i - \theta)^2$. The minimum is reached at the empirical mean $\theta^* = \bar{\mu}_n$. Generalizing that idea, M-estimators were introduced [26] to minimize the expression of the type $\sum_i \rho(X_i - \theta)$ for some well-chosen function ρ. Similarly, Z–estimators are defined as the zero of the expression $\sum_i \psi(X_i - \theta)$ where $\psi = \rho'$ is called the influence function. Here we consider Z–estimators, $\hat{\mu}_n$, defined as the zero of the following equation

$$\mathcal{R}_{n,\psi}(\theta) = \sum_{i=1}^{n} \psi\left(\alpha(X_i - \theta)\right), \tag{3.2}$$

where the influence function $\psi : \mathbb{R} \to \mathbb{R}$ is an antisymmetric non-decreasing function and $\alpha \in \mathbb{R}$ is a tuning parameter. Different choices of influence function yield different estimators. For example, $\psi(x) = x$ gives the empirical mean, while $\psi(x) = \text{sign}(x)$ gives a median estimator. The behavior at large $|x|$ determines the outliers' importance.

3.4.1 Huber's influence function

The Huber's influence function [26] is defined as

$$\psi(x) = \begin{cases} 1 & \text{if } x > 1, \\ x & \text{if } |x| \leq 1, \\ -1 & \text{if } x < -1. \end{cases} \tag{3.3}$$

Essentially all large values above some threshold dictated by α are treated with the same weight. In this case, the following result shows that the Huber M–estimator is sub-Gaussian.

Theorem 3.3 (See for example [39, Section 3]). *Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $\delta \in (0, 1)$ be such that $n > 2\log(2/\delta)$. Define the Huber's mean estimator $\widehat{\mu}_n$ as the zero of Eq. (3.2) with influence function Eq. (3.3) and parameter*

$$\alpha = \sqrt{\frac{2\log(4/\delta)}{n\sigma^2}}. \tag{3.4}$$

Then Huber's mean estimator is a δ-sub-Gaussian estimator, that is

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| < 8\sigma\sqrt{\frac{2\log(4/\delta)}{n}} \right) \geq 1 - \delta.$$

Notice that, contrary to the previously introduced estimators, Huber's estimator depends on the variance σ^2 through the tuning parameter α. In practice, this undesirable feature can be partially overcome by taking an estimator of σ^2.

3.4.2 Catoni's influence function

Catoni in [6] proposes the following influence function

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2) & \text{if } x \leq 0, \\ -\log(1 - x + x^2/2) & \text{if } x > 0. \end{cases} \tag{3.5}$$

This function is not bounded, and the resulting estimators are therefore still sensitive to very large outliers. However, the logarithmic growth at infinity allows reducing the importance of "reasonable outliers" arising in i.i.d. samples from finite two moments distributions, as shown by the following result.

Theorem 3.4 ([6]). *Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $\delta \in (0,1)$ be such that $n > 2\log(2/\delta)$. Define Catoni's mean estimator $\widehat{\mu}_n$ as the zero of Eq. (3.2) with the influence function Eq. (3.5) and the parameter*

$$\alpha = \sqrt{\frac{2\log(2/\delta)}{n\sigma^2\left(1 + \frac{2\log(2/\delta)}{n - 2\log(2/\delta)}\right)}}. \tag{3.6}$$

Then

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| < \sigma\sqrt{\frac{2\log(2/\delta)}{n - 2\log(2/\delta)}}\right) \geq 1 - \delta. \tag{3.7}$$

The sub-Gaussian bound satisfied by Catoni's estimator is therefore tight up to a $\frac{1}{1 - \frac{2\log(2/\delta)}{n}} = 1 + o(1)$ term when $\log(1/\delta)/n \to 0$. As Huber's estimators, the tuning parameter defining Catoni's estimators depends on the variance σ^2 which might be unknown. When an interval $[\sigma_1, \sigma_2]$ is known such that $\sigma \in [\sigma_1, \sigma_2]$, Catoni [6] showed that this dependency can be removed using Lepski's method. However, the bounds are only tight when the ratio σ_2/σ_1 is bounded and this extra step makes the leading constant in the risk bound bigger than its optimal value $\sqrt{2}$.

3.4.3 Lee Valiant estimator

In a recent work, Lee and Valiant [34] propose a tight SGME up to a term going to zero, that does not require any knowledge on the variance.

Definition 3.4. *For a given δ, define the median-of-means estimator Eq. (3.1) $\kappa \doteq \kappa(X_1, \dots, X_n)$ computed on $k = \log(\frac{1}{\delta}) \leq n$ groups with $\delta \geq e^{-n}$ and k an integer. The Lee Valiant estimator is then defined as*

$$\widehat{\mu}_n = \kappa + \frac{1}{n}\sum_{i=1}^{n}(X_i - \kappa)(1 - \min(\alpha(X_i - \kappa)^2, 1)) \tag{3.8}$$

where the parameter α is the solution of the monotonic, piecewise-linear equation

$$\sum_{i=1}^{n}\min(\alpha(X_i - \kappa)^2, 1) = \frac{1}{3}\log\left(\frac{1}{\delta}\right). \tag{3.9}$$

Theorem 3.5 ([34]). *Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $\delta \in (0,1)$ be such that $\delta \geq e^{-n}$ and assume that $k = \log(\frac{1}{\delta})$ is an integer. Then, Lee-Valiant's estimator $\widehat{\mu}_n$ introduced in Definition 3.4, satisfies*

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| < \sigma(1+o(1))\sqrt{\frac{2\log(1/\delta)}{n}}\right) \geq 1-\delta, \tag{3.10}$$

where the $o(1)$ term goes to zero when $(\delta, \log(1/\delta)/n) \to (0,0)$.

In this result, the $o(1)$ term is not explicit, this is also the case in the proof of this theorem in [34]. This means that the risk bound cannot be used directly to derive confidence intervals in applications. When $\kappa = 0$, Lee-Valiant's estimator can be reformulated as an M–estimator. The condition (3.9) guarantees that "$\frac{1}{3}\log\left(\frac{1}{\delta}\right)$ of the samples are discarded".

3.4.4 Minsker-Ndaoud's Estimator

In an even more recent work, Minsker and Ndaoud [42] propose another SGME that does not require any information on the variance and describe its properties in detail. The idea is, as for median-of-means, to divide the sample $X_1, \dots, X_n$ into k blocks of similar size and take the empirical mean $\bar{X}_l$ of each block J_l. However, instead of computing the median of the block, they propose a weighted mean where each weight is inversely proportional to the empirical variance $\bar{\sigma}_l$.

Definition 3.5. *Let $p_{\mathrm{MN}} \geq 1$. Minsker-Ndaoud's estimator $\widehat{\mu}_n$ of a sample $X_1, \dots, X_n$ is defined as*

$$\widehat{\mu}_n = \frac{\sum_{l=1}^{k} \bar{X}_l / \bar{\sigma}_l^{p_{\mathrm{MN}}}}{\sum_{l=1}^{k} 1/\bar{\sigma}_l^{p_{\mathrm{MN}}}}, \tag{3.11}$$

where $\bar{X}_l = |J_l|^{-1} \sum_{i \in J_l} X_i$ is the empirical mean over the block J_l, and $\bar{\sigma}_l$ denotes the empirical standard deviation. The sample $\{1, \dots, n\} = \bigcup_{l=1}^{k} J_l$ is divided into k blocks, $J_l \bigcap J_{l'} = \emptyset$ for $l \neq l'$, of size $m = |J_l|$ assuming $n = km$.

The weights give less importance to the block with large variance (and possibly outliers). Note that for $p_{\mathrm{MN}} = 0$, it corresponds to the empirical mean, while for $p_{\mathrm{MN}} \to \infty$ only the block with the smallest variance is considered. In [42] the authors consider the case $p_{\mathrm{MN}} = 1$ and $p_{\mathrm{MN}} = 2$. We state a simplified δ dependent version of their main theorem, showing that the estimator (3.11) is an SGME.

Theorem 3.6 ([42, Theorem 3.1 with Lemma 3.2]). *Let $X_1, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $\delta \in (0,1)$ be such that $\delta \gtrsim e^{-\sqrt{n/\log(n)}}$. Then, Minsker-Ndaoud's*

estimator $\hat{\mu}_n$ *introduced in Definition 3.5 with* $k = \log(3/\delta)$ *blocks of size* $m = n/k$ *satisfies for* $p_{\text{MN}} \geq 1$

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| < \sigma C_{p_{\text{MN}}} \sqrt{\frac{1 + \log(3/\delta)}{n}}\right) \geq 1 - \delta,$$

for some $C_{p_{\text{MN}}} > 0$ *constant depending only on* p_{MN}.

Notice that the range of δ for which this result applies is more restrictive than in the other results. The constant $C_{p_{\text{MN}}} > 0$ is sub-optimal in this result, but it is proved in the paper that the estimator is also asymptotically normal [42], with optimal variance (i.e., $\sqrt{n}(\widehat{\mu}_n - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2)$), which is an interesting feature that suggests a good practical behavior.

3.4.5 Discussion

We have defined five "robust" estimators, all satisfying sub-Gaussian bounds with different L. Up to recently, Catoni's estimators had the best theoretical guarantees (see Eq. (3.7)) given some information on the variance σ^2 because $L = \sqrt{2}(1 + o(1)) \approx L_{\text{optimal}}$. In practice, this yields important improvements over other SGME like median-of-means and Trimmed Mean that do not request knowledge of the variance but have larger L. However, recent results like Lee Valiant [34] show that the best of both worlds is possible: they obtain quasi optimal bound (see Eq. (3.10)) without any knowledge on the variance. In our practical experiments in Section 6, we confirm this impression: the estimators from Lee Valiant and Minsker Ndaoud look like they have the best practical performances. The theoretical non-asymptotic deviation bound obtained by Minsker and Ndaoud, see Theorem 3.6, does not express explicitly the leading constant. However, as suggested by their asymptotic result, it seems like this constant can be proved to be very close to optimal, even for reasonable p_{MN}.

4 Theory of Quasi Monte Carlo methods

Say that we wish to approximate the QoI Eq.(1.1) by a quadrature rule, i.e., an expression of the form

$$\mu \simeq \sum_{j=1}^{N} w_j F(\boldsymbol{x}_j) \doteq \bar{\mu}_N$$

for a given sequence of point $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ and weights $(w_1, \ldots, w_N)$. Common examples of deterministic quadrature rules are tensored one-dimensional rules (of Newton, Gauss, Clenshaw-Curtis, etc. types) or sparse grids (Smolyack's quadrature rules). Generally speaking, the error convergence rate depends on the dimension d and on the regularity of the function F, see [65] and references therein for an overview. The purpose of Monte Carlo, Quasi Monte Carlo and Randomized Quasi Monte Carlo methods is to obtain convergence rates that are completely or almost independent of the dimension d. The most classical approach is to use equal weights $w_j = 1/N$.

4.1 Monte Carlo

Let $\boldsymbol{U}_1, \dots, \boldsymbol{U}_N$ be N independent vectors uniformly distributed over $[0,1]^d$. For square integrable functions F, set $\sigma^2 = \mathrm{Var}\left(F(\boldsymbol{U})\right)$ where $\boldsymbol{U} \sim \mathcal{U}([0,1]^d)$: we have $\mathrm{Var}\left(\bar{\mu}_N\right) = \frac{\sigma^2}{N}$, i.e., randomly choosing the quadrature point $\boldsymbol{U}_1, \dots, \boldsymbol{U}_N$ leads to an RMSE of order $\mathcal{O}(N^{-1/2})$. This result is usually worse than the deterministic quadrature rules in small dimensions, but it is both independent of the dimension d and the integrand regularity.

4.2 Quasi Monte Carlo

Instead of using N independent uniformly distributed random variables, quasi Monte Carlo methods use deterministic sequences $(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N)$ that asymptotically fills well the unit hypercube (like random sequences), thus the name quasi Monte Carlo (QMC) methods. We review briefly the basic properties of these QMC methods. Let us define the multidimensional interval

$$[\boldsymbol{a}, \boldsymbol{b}) = \prod_{k=1}^{d} [a_k, b_k) \doteq \{\boldsymbol{x} \in \mathbb{R}^d \mid a_k \leq x_k < b_k, k = 1, \dots, d\},$$

assuming $a_k < b_k$ for any k. To measure how well a point sequence $(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N)$ fills the d-dimensional cube $[0,1]^d$, following [45, Section 2.1], we introduce the star discrepancy defined as

$$D_N^*\left(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N\right) \doteq \sup_{\boldsymbol{x} \in [0,1]^d} |\Delta(\boldsymbol{x}; \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N)|, \quad \text{where} \quad \Delta(\boldsymbol{x}; \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N) \doteq \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\boldsymbol{\xi}_j \in [0, \boldsymbol{x})} - \prod_{k=1}^{d} x_k.$$

If the d-dimensional cube was filled homogeneously, the proportion $N^{-1} \sum_{j=1}^{N} 1_{\boldsymbol{\xi}_j \in [0, \boldsymbol{x})}$ of points inside the multidimensional interval $[0, \boldsymbol{x})$ would exactly be equal to its Lebesgue measure, thus $\Delta(\boldsymbol{x}; \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N) = 0$. The perfect homogeneity is out of reach, but the error can be made quite small.

Definition 4.1. *A d-dimensional infinite sequence $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots$ is said to have Low Discrepancy (LD) if*

$$D_N^*(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_N) = \mathcal{O}\left(N^{-1} (\log N)^d\right)$$

as $N \to \infty$.

4.2.1 Koksma-Hlawka error bound

The Koksma-Hlawka inequality [28, 24] provides an error bound for the finite N estimate of any sequence $(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_N)$,

$$\left| \frac{1}{N} \sum_{j=1}^{N} F(\boldsymbol{u}_j) - \mu \right| \leq V_{\mathrm{HK}}(F) D_N^*(\boldsymbol{u}_1, \dots, \boldsymbol{u}_N), \tag{4.1}$$

where $V_{\mathrm{HK}}(F)$ is the Hardy-Krause variation of the function F. In words, $V_{\mathrm{HK}}(F)$ is defined as the supremum over all multidimensional intervals partitions of the variations of F between intervals'

vertices (see [45, p.19] for details). If $V_{\mathrm{HK}}(F) < +\infty$, we write $F \in \mathrm{BVHK}([0,1]^d)$ or simply $F \in \mathrm{BVHK}$. Observe that if F is unbounded then $F \notin \mathrm{BVHK}$; this means that for a heavy-tailed random variable $F(U)$, the above (4.1) is not informative since the right-hand side is infinite.

If F is smooth with continuous mixed partial derivatives, then $F \in \mathrm{BVHK}$ and we have

$$V_{\mathrm{HK}}(F) = \sum_{\emptyset \neq u \subseteq \{1,\dots,d\}} \int_{[0,1]^{|u|}} \left| \frac{\partial^{|u|} F}{\partial \boldsymbol{x}_u}(\boldsymbol{x}_u : \mathbf{1}_{-u}) \right| \mathrm{d}\boldsymbol{x}_u,$$

where for a non-empty set of coordinates u and for $\boldsymbol{x} = (x_1, \cdots, x_d) \in [0,1]^d$, $\boldsymbol{x}_u$ stands for the components $\boldsymbol{x}_j$ such that $j \in u$, $\boldsymbol{x}_u : \mathbf{1}_{-u}$ denotes the point $\boldsymbol{y} \in [0,1]^d$ with $y_j = x_j$ for $j \in u$ and $y_j = 1$ for $j \notin u$, see [15, p.176]. It is known that the inequality (4.1) is tight for smooth functions, see [45, Theorem 2.12]. Besides, extension of (4.1) to functions F satisfying fractional regularity conditions is achieved in [13].

All in all, for any LD sequence as in Definition 4.1 and $F \in \mathrm{BVHK}$ we have

$$\left| \frac{1}{N} \sum_{j=1}^{N} F(\boldsymbol{\xi}_j) - \mu \right| = \mathcal{O}\left(N^{-1} (\log N)^d \right),$$

which is asymptotically better than MC. However, this estimate is completely deterministic, necessarily biased, and error bounds are difficult to evaluate in practice since the bounding terms $V_{\mathrm{HK}}(F)$ and $D_N^*(\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_N)$ are not easily tractable.

4.2.2 Digital nets and sequence

The construction of LD sequences satisfying Definition 4.1 with good properties is an active research area. Ideally, one wants both the LD sequence to have good theoretical guarantees and efficient algorithms to generate large N sequences in large dimensions d. Moreover, as we will discuss in Section 4.3, we will be interested in a sequence that can be randomized while preserving their LD characteristics. For detailed reviews on LD sequence, see [45, 15]. Here, we only focus on digital net/sequence [16] because of their well-studied properties (see Section 4.3).

Let us define (t, m, d)-nets and (t, d)-sequences, also referred to as digital nets and sequences. Let $b \geq 2$ an integer base in which to represent real numbers. Consider the elementary intervals $E_{\boldsymbol{k}}(\boldsymbol{c})$ in base b which are subintervals of $[0,1)^d$ of the form

$$E_{\boldsymbol{k}}(\boldsymbol{c}) = \prod_{j=1}^{d} \left[\frac{c_j}{b^{k_j}}, \frac{c_j + 1}{b^{k_j}} \right),$$

for integers k_j and c_j, with $k_j \geq 0$ and $0 \leq c_j < b^{k_j}$.

Definition 4.2. *Let m and t be integers with $0 \leq t \leq m$. A sequence $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ of $N = b^m$ points in $[0,1)^d$ is called a (t, m, d)-net in base b if every elementary interval $E_{\boldsymbol{k}}(\boldsymbol{c})$ in base b of volume b^{t-m} contains precisely b^t points of the sequence.*

The parameter t defines the quality of the net, with smaller values implying better equidistribution. From [45, Theorem 4.10], it is known that a (t, m, d)-net in base b satisfies the LD property of Definition 4.1.

Definition 4.3. *The infinite sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \in [0,1)^d$ is a (t,d)-sequence in base b if for all $k \geq 0$ and $m \geq t$ the sequence $\boldsymbol{x}_{kb^m+1}, \ldots, \boldsymbol{x}_{(k+1)b^m}$ is a (t,m,d)-net in base b.*

Digital nets are "closed" sequences of points (with a fixed number of points $N = b^m$) whereas digital sequence are "open" infinite sequences, which has the advantage of providing as many points as desired by the user.

4.3 Randomized Quasi Monte Carlo

One drawback with QMC estimation is that it provides a (deterministic) error estimate (Koksma-Hlawka inequality (4.1)) which can be very loose and is as hard as the QoI to compute. Randomized low discrepancy sequences have been introduced to allow easier error estimation. The deterministic LD sequence $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N)$ is transformed into a random $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)$ that is still LD (with probability 1) with $\boldsymbol{R}_j \sim \mathcal{U}([0,1]^d)$ for all $j \in \{1, \ldots, N\}$. This new sequence is called a Randomized Quasi Monte Carlo Sequence (RQMCS). Hence, the RQMCS is composed of identically distributed uniform random variables which are *not* independent. The dependency is the key difference with Monte Carlo sampling. Given a RQMCS of N points $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N \in [0,1]^d$, the estimator of the QoI is

$$\bar{\mu}_N = \frac{1}{N} \sum_{j=1}^{N} F(\boldsymbol{R}_j). \tag{4.2}$$

Due to the uniform distribution of each point, the RQMC estimate is unbiased.

4.3.1 Nested Uniform Scrambling (NUS)

There are several ways to randomize a LD sequence $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N)$ which can affect the convergence rates, see [32, 15, 31] for reviews. Scrambling is a randomization method that is typically applied to (t, m, d)-net (sequence) by shuffling each digit of each coordinate of the $\boldsymbol{R}_j$. Among the different class of scrambling, there is the nested uniform scramble [47] (sometimes referred to as Owen scrambling). It preserves their (t, m, d)-net (sequence) character [47, Proposition 1] with probability 1 and each point is uniformly distributed [47, Proposition 2] (independently of the digital net/sequence nature of the sequence). The Nested Uniform Scrambling (NUS) has strong theoretical guarantees, we will expose some in the following. We will from now on refer to QMC sequences scrambled as in [47] simply as scrambled sequences. Recently, it has been proved in [57] that for any $F \in L^p$ with $p > 1$, the estimator (4.2) based on a scrambled (t, d)-sequence follows a strong law of large numbers.

4.3.2 Variance of scrambled nets

In view of the discussion in the introduction, the standard deviation σ_N given by

$$\sigma_N^2 := \mathrm{Var}\left(\frac{1}{N} \sum_{j=1}^{N} F(\boldsymbol{R}_j) \right)$$

plays an important role. We expect that for an RQMCS, it goes to 0 faster than N^{-1} (the case of MC method).

General bounds. In a series of works [47, 48], Owen has shown that for nested uniform scrambling, the above variance is bounded by N^{-1} up to a multiplicative factor, for any value of N, see Theorem 4.1. Moreover, in [49, Theorem 1], it is shown that $N\sigma_N^2 \to 0$ under the sole assumption of square integrable F. We summarize these results in the following statement, which definitely justifies to use RQMC as a variance reduction technique.

Theorem 4.1. *Let $\boldsymbol{R}_1, \cdots, \boldsymbol{R}_N$ be the points of nested uniform scrambling of a (t, m, d)-net in base b where $N = b^m$. Assume that F is such that $\sigma^2 := \mathrm{Var}\left(F(\boldsymbol{U})\right) < \infty$. Then*

$$\lim_{N \to +\infty} N\sigma_N^2 = 0 \tag{4.3}$$

and we have

$$\sigma_N^2 \leq b^t \left(\frac{b+1}{b-1}\right)^d \frac{\sigma^2}{N}. \tag{4.4}$$

The limit (4.3) is valid for each given F, whereas (4.4) ensures an uniform bound

$$\sup_{F:\mathrm{Var}(F(\boldsymbol{U}))\leq 1} N\sigma_N^2 \leq b^t \left(\frac{b+1}{b-1}\right)^d.$$

The proof relies on an ANOVA Haar wavelet decomposition of the variance using the structure of (t, m, d) scrambled net. The above non-asymptotic bound should be seen as a worst-case variance estimate, showing that the RQMC variance cannot be worse than the Monte Carlo variance up to a factor, given a budget of N points; this is cheering. The factor growths exponentially fast with the dimension d. This does not contradict Eq. (4.3), it just says that for a fixed N, there might exist a worst case function F reaching this bound.

The first asymptotic result is really appealing, and states that RQMC based on an NUS is asymptotically always better than MC. It will be of main importance when combined with robust mean estimators of Section 3.

Bounds for $F \in$ BVHK. When the integrand F satisfies some regularity properties (typically $F \in$ BVHK), the previous variance reduction can be improved. It directly stems from the Koksma-Hlawka bound (4.1), assuming that the convergence in Definition 4.1 is valid with a deterministic upper-bound:

$$\sigma_N^2 = \mathbb{E}\left((\bar{\mu}_N - \mu)^2\right) \leq \mathbb{E}\left(\left(D_N^*\left(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N\right) V_{\mathrm{HK}}(F)\right)^2\right) = \mathcal{O}\left(\frac{\log(N)^{2d}}{N^2}\right).$$

For the scrambling of a digital net, the above can be refined as follows. The result is due to [54, Theorem 3].

Theorem 4.2. *Assume that the function F is smooth*[1]*, in the sense that $\frac{\partial^{|u|}}{\partial \boldsymbol{x}_u} F(\boldsymbol{x})$ is continuous on $[0,1]^d$ for all possible $u \subseteq \{1, \cdots, d\}$. If the $\boldsymbol{R}_i$'s are a nested uniform scramble of (t, m, d)-net in base b, then*

$$\sigma_N^2 = \mathcal{O}\left(\frac{\log(N)^{d-1}}{N^3}\right).$$

as $N \to +\infty$.

The logarithmic factor can be avoided by using the randomized Frolov quadrature rule, see [66], which is however less easy to implement than the scrambled nets.

Bounds when $F \notin$ BVHK. Because $F \in$ BVHK implies that F is bounded (discarding heavy-tailed $F(\boldsymbol{U})$), the previous assumption does not cover many practical situations. However, the case of bounded and fractional smooth functions can be covered by the extension of the Koksma Hlawka inequality by [13]. The case of discontinuous F is studied in [21, 20]. Therefore, we now focus our subsequent discussion on unbounded F, to handle fat or heavy-tailed $F(\boldsymbol{U})$, which is to us a primary challenge.

To obtain more precise scaling of the variance σ_N^2, one needs assumptions on F singularities.

Definition 4.4. *The function F on $[0,1]^d$ has corner singularities no worse than $\prod_{k=1}^d x_k^{-A_k}$ if*

$$\left|\frac{\partial^{|u|} F(\boldsymbol{x})}{\partial \boldsymbol{x}_u}\right| \leq C \prod_{k=1}^d x_k^{-A_k - \mathbb{1}_{\{k \in u\}}}$$

holds for all $u \subseteq \{1, 2, \ldots, d\}$, some $A_k \in (0,1)$ and some $C < \infty$. Here x_k is the k-th component of the vector $\boldsymbol{x}$.

Setting $p^\star = (\max_k A_k)^{-1}$, observe that $F(\boldsymbol{U})$ is in L^p for any $p < p^\star$. Thus, the condition $0 < A_k < 1$ ensures just that F is at least integrable (but potentially singular).

In the case of such an F, the integration error can be estimated, see [52, Theorem 5.7].

Theorem 4.3. *Let $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N \sim \mathcal{U}([0,1]^d)$ with $\mathbb{E}\left(D_N^*\left(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N\right)\right) = \mathcal{O}\left(N^{-1+\varepsilon}\right)$ for all $\varepsilon > 0$. If F satisfies Definition 4.4, then for any $\varepsilon > 0$,*

$$\mathbb{E}\left(|\bar{\mu}_N - \mu|\right) = \mathcal{O}\left(N^{-1+\varepsilon+\max_k A_k}\right).$$

Appendix A includes a proof of this result. The result in [52] is in fact slightly more generic, as it considers singularities possibly both at the origin and $\mathbf{1} = (1, \ldots, 1)$. The result is extended to arbitrary singularity point [53, Theorem 1] and arbitrary RQMC sequence (not only scrambled nets).

We provide the analog theorem for the variance. This is one of our contributions. Before, we introduce the useful concept of joint density between random pairs of sequence points.

[1] mixed smooth in the terminology of [54].

Definition 4.5. *Let $(\boldsymbol{R}_1, \dots, \boldsymbol{R}_N)$ be a sequence of random points such that for each $\boldsymbol{R}_i \sim \mathcal{U}([0,1]^d)$ and $\mathbb{E}(F(\boldsymbol{U})) = \mu$ and $\mathrm{Var}(F(\boldsymbol{U})) = \sigma^2 < \infty$. Let I, J be two different integers valued in $\{1, \dots, N\}$ picked with uniform distribution (independent of $(\boldsymbol{R}_1, \dots, \boldsymbol{R}_N)$). Define by $\Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ the joint distribution of $(\boldsymbol{R}_I, \boldsymbol{R}_J)$: for any measurable function $\ell : [0,1]^d \times [0,1]^d \mapsto \mathbb{R}^+$,*

$$\mathbb{E}(\ell(\boldsymbol{R}_I, \boldsymbol{R}_J)) := \int_{[0,1]^d \times [0,1]^d} \ell(\boldsymbol{x}, \boldsymbol{y}) \Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}).$$

With this definition at hand, we have, for any square integrable function G,

$$\begin{aligned} \frac{1}{N(N-1)} \sum_{i \neq j} \mathbb{E}(G(\boldsymbol{R}_i) G(\boldsymbol{R}_j)) &= \mathbb{E}(G(\boldsymbol{R}_I) G(\boldsymbol{R}_J)) \\ &= \int_{[0,1]^d \times [0,1]^d} G(\boldsymbol{x}) G(\boldsymbol{y}) \Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}). \end{aligned} \tag{4.5}$$

We are now in a position to state a new variance bound, whose proof is postponed to Appendix A.

Theorem 4.4. *Let $\boldsymbol{R}_1, \dots, \boldsymbol{R}_N$ be a RQMCS: each $\boldsymbol{R}_i \sim \mathcal{U}([0,1]^d)$ and for any $\varepsilon > 0$, there exists (deterministic) $D_\varepsilon < \infty$ with $\mathbb{P}\left(D_N^*(\boldsymbol{R}_1, \dots, \boldsymbol{R}_N) \leq D_\varepsilon N^{-1+\varepsilon}\right) = 1$ for all N large enough.*

Assume that F satisfies Definition 4.4 (with $\max_k A_k < \frac{1}{2}$) and that the joint distribution $\Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ has a density (w.r.t. Lebesgue's measure on $[0,1]^d \times [0,1]^d$) uniformly bounded (in N), then for any $\varepsilon > 0$,

$$\sigma_N^2 = \mathcal{O}\left(N^{-2+\varepsilon+2\max_k A_k}\right).$$

As a consequence of Theorem 4.4, for a square integrable function (i.e., the condition $\max_k A_k < 1/2$), RQMC is asymptotically better than MC: this is consistent with Theorem 4.1. In addition, the smaller $\max_k A_k$, the faster the convergence.

We shall mention that without the condition on the joint distribution $\Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$, one can still get that $\sigma_N^2 = \mathcal{O}\left(N^{-1+\varepsilon+2\max_k A_k}\right)$ (see Remark (A.1) in Appendix A) which is better than MC when $\max_k A_k < 1/4$ ($F(\boldsymbol{U})$ has four finite polynomial moments).

By leveraging [69, Theorem 3.6], we additionally prove that the property of bounded density is satisfied for a $(0, m, d)$-net in base $b \geq 2$. See Appendix B for the proof.

Note that the hypothesis on the existence of deterministic D_ε is satisfied for the scrambled (t, m, d)-net, as it is a (t, m, d)-net to which we can apply the deterministic bound from [45, Theorem 4.6 and Theorem 4.10] on the discrepancy.

5 Error bounds for Robust Quasi Monte Carlo

Combining the previous results of robust mean estimation (Section 3) and RQMCS with scrambled nets (Section 4.3), one obtains the following "Robust Quasi Monte Carlo" finite sample result.

Theorem 5.1. *Let $F : [0,1]^d \mapsto \boldsymbol{R}$ such that $F(\boldsymbol{U})$ (with $\boldsymbol{U} \sim \mathcal{U}([0,1]^d)$) has mean μ and finite variance σ^2. Consider a (t, m, d)-net in base b where $N = b^m$, and n independent replicas of the related nested uniform scramble, which we denote $(\boldsymbol{R}_j^{(i)} : 1 \leq j \leq N)_{1 \leq i \leq n}$. Set the i-th RQMC-empirical mean $\overline{\mu}_N^{(i)}$ as in (1.2).*

Then, for a given $\delta > 0$, define a sub δ-Gaussian estimator

$$\widehat{\mu}_B := \widehat{\mu}_n(\bar{\mu}_N^{(1)}, \ldots, \bar{\mu}_N^{(n)})$$

as defined in Definition 3.2 (with some $L_{\delta,n} \in [\sqrt{2}, \infty)$), assuming that the restriction condition between δ and n is satisfied: the above estimator has an evaluation budget equal to $B = N \times n$. Then

$$\mathbb{P}\left(|\widehat{\mu}_B - \mu| \leq (\sigma_N \sqrt{N}) L_{\delta,n} \sqrt{\frac{\log(2/\delta)}{B}} \right) \geq 1 - \delta \tag{5.1}$$

where $\sigma_N \sqrt{N} \to 0$ as $N \to +\infty$.

The above Theorem requires several important comments. To alleviate the discussion, we assume that $L_{\delta,n}$ depends neither on δ nor on n, which is merely the case for all the estimators proposed in Section 3.

- Theorem 5.1 is to the best of our knowledge, the first result stating a non-asymptotic error bound for RQMC making explicit the advantage of variance reduction (since $\sigma_N \sqrt{N} \to 0$ as $N \to +\infty$) owing to RQMCS, under the sole assumption of square integrable $F(\boldsymbol{U})$.

- This estimate, with the upper bound $\sigma_N \sqrt{N} \leq \sigma b^{t/2} \left(\frac{b+1}{b-1}\right)^{d/2}$ of (4.4), is consistent with the lower bound result by [29, Theorem 2.3] in the case of non-smooth F (in their notation, it corresponds to $r = 0$, $p = q = 2$). The latter reference states that the B-th minimal probabilistic Monte Carlo error at uncertainty δ, uniformly over the square integrable functions F (with second moment at most equal to 1) is asymptotically (in B) equal to

  $$\sqrt{\frac{\log(1/\delta)}{B}},$$

  up to constant[2]. Observe that since the latter reference considers the rate uniformly in F, it can not catch the improvement $\sigma_N \sqrt{N}$ for each specific F as we do. Besides, our bound is valid for any given B and not only for asymptotic large B.
  As a consequence, scrambled (t, m, d)-nets coupled with robust mean techniques achieve the optimal lower bound of [29, Theorem 2.3], for functions F under the sole assumption $\mathbb{E}\left(F^2(U)\right) < +\infty$.

- Furthermore, the above bound (5.1) highlights well the effect of variance reduction due to $\sigma_N \sqrt{N} \to 0$ as $N \to +\infty$ in general, with some quantitative rate under some assumptions on F as those studied in Theorem 4.2 or Theorem 4.3.

  - For a mixed smooth function F, i.e. $\frac{\partial^{|u|}}{\partial \boldsymbol{x}_u} F(\cdot)$ is continuous on $[0,1]^d$ for all possible $u \subseteq \{1, \cdots, d\}$, combining the Theorems 4.2 and 5.1 and by taking n fixed, it readily follows that the probabilistic Monte Carlo error at uncertainty δ is equal to $c \frac{\log(B)^{(d-1)/2}}{B} \sqrt{\frac{\log(2/\delta)}{B}}$.

[2] we neglect the usual condition that $B \geq c \log(1/\delta)$ which is often required both in [29] and the definition of sub δ-Gaussian estimators in Section 3

Up to the logarithm $\log(B)$ factor, this rate achieves the lower bound of [29, Section 4 on Mixed Smoothness Space, Equation (28)], showing the quasi-optimality of scrambled (t, m, d)-net in this Robust RQMC approach.

- The generalization of this estimator for more regular functions in the mixed smoothness sense of order $r \in \mathbb{N}^*$ [29, Section 4] can be made using higher order scrambling methods: it includes smooth functions with continuous partial derivatives $\frac{\partial^{r_1+\cdots+r_d}}{\partial_{\boldsymbol{x}_1}^{r_1}\cdots\partial_{\boldsymbol{x}_d}^{r_d}}F(\cdot)$ for all $(r_1, \cdots, r_d) \in \{0, \cdots, r\}^d$ [14, Section 3.2.2]. According to [14, Theorem 10], the variance of quasi Monte Carlo integration with sufficiently high-order scrambled nets (generalization of Owen scrambling) yields $\sigma_N = O(N^{-r-1/2}(\log N)^{d(r+1)/2})$. This combined with a robust estimator at uncertainty δ produces an asymptotic error bound as $c\frac{\log(B)^{d(r+1)/2}}{B^r}\sqrt{\frac{\log(2/\delta)}{B}}$. This is once again an optimal bound up to a logarithmic term [29, Section 4 on Mixed Smoothness Space, Equation (28)]. Note that for $f \in \mathcal{C}^r([0,1]^d)$, the optimal convergence rate is $B^{-1/2-r/d}$, see [9] for a recent computationally tractable implementation. In this case, there are no log terms, and it can also be combined with a robust estimator at uncertainty δ.
- The function with corner singularities like in Theorem 4.3 is not covered by the optimality results in [29]. Our Theorem 5.1 shows that the probabilistic Monte Carlo error at uncertainty δ is equal to $\frac{c}{B^{\frac{1}{2}-\max_k A_k-\varepsilon}}\sqrt{\frac{\log(2/\delta)}{B}}$ for any $\varepsilon > 0$. Knowing whether this rate is optimal is an open question.

- The rate of the form $\sigma_N\sqrt{N}L\sqrt{\frac{\log(2/\delta)}{B}}$ shows that, to improve the probabilistic Monte Carlo error at uncertainty δ, it is theoretically better to increase N than to increase n, under the constraint of a fixed budget B. This heuristic will be confirmed in our next numerical experiments, see Figure 6.

- While working on the revision of this article, the authors have been aware of another approach combining median-of-means and RQMC [59, 58]. In their work, they focus on Linear Matrix Scrambling (LMS) [40] which is a weaker form of scrambling than the nested uniform scrambling considered in this paper, but LMS is numerically more memory efficient. They show that with high probability, LMS provides an estimation with a small variance; however, with low probability, it gives a very wrong estimation. Using the robust median-of-means mitigates these extreme cases, which can be seen as outliers or contaminated samples. It is interesting to note that the outlier does not come from an extreme distribution or singular integrand but from the randomization method. A strong smoothness assumption on F is required for their result.

6 Numerical Experiments

We compare the Monte Carlo (MC) approach and Randomized Quasi Monte Carlo (RQMC) for mean estimator. We will use the following abbreviations: EM: Empirical Mean, MoM: median-of-means, LV: Lee Valiant, MN: Minsker Ndaoud, CA: Catoni estimator and HU: Huber. Often we

report results for LV and not for MoM since LV is an improved version of MoM. We do not report results for the Trimmed Mean estimator because its theoretical and numerical performance is worse than others robust estimators, in addition the condition $n > (32/3)\log(8/\delta)$ in Theorem 3.1 is not satisfied for our simulations where $n = 56$. In all the following for the Minsker Ndaoud estimator (Definition 3.5) we use $p_{\text{MN}} = 1$. The number of realizations used to obtain the distributions of $\widehat{\mu}_n$ is denoted by M.

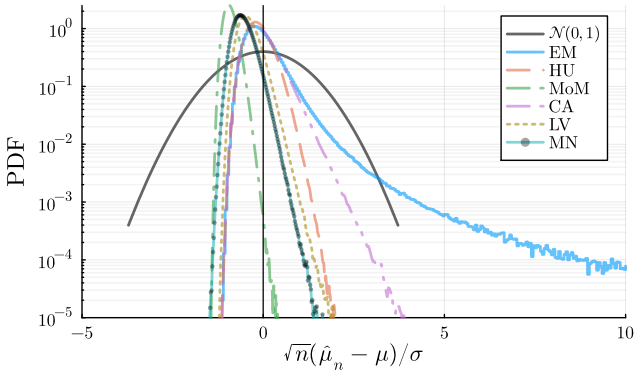
6.1 Illustration of Robust Mean Estimators

We illustrate the robustness of the estimator presented in Section 3 in the "most extreme case" where the sample $X_1, \dots, X_n$ comes from a Pareto distribution whose density function is $f_X(x) = \theta/x^{1+\theta}\mathbb{1}_{x\geq 1}$ with finite second moment, i.e., $\theta > 2$. For these examples, we use the α parameter of the Catoni (3.6) and Huber (3.4) estimators, the exact standard deviation σ of the Pareto distribution. Figure 1 shows the distribution of different (standardized) estimators μ_n for different θ compared with the Gaussian distribution. Because of the relatively small number of samples $n = 56$ and the heavy tail character of the distribution, the behavior of all estimators (including EM) is far from normality. However, if EM estimator distribution suffers from heavy tails, other robust estimators are much less affected. It is clear that the robust estimators HU, CA, LV, MN and especially MoM trade robustness against bias. CA and HU seem like the robust estimator with the smallest bias. However, these two estimators are both computed using the exact standard deviation σ, which is not known in general. The best estimators requiring no knowledge about the variance σ^2 seem to be LV and MN.

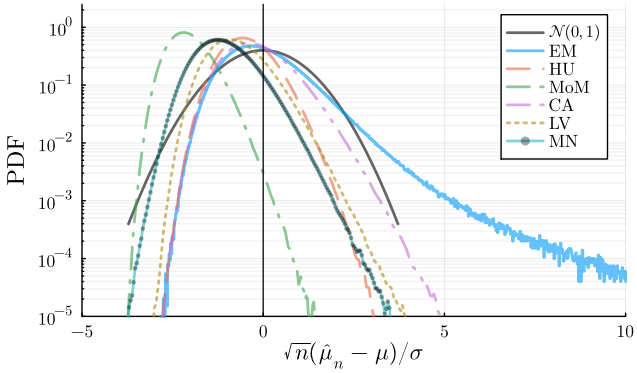
6.2 Tests for Robust Quasi Monte Carlo

6.2.1 QMC Numerical Methods and Software

Choosing between all the different available LD sequences is difficult and one is never consistently better than others in terms of variance reduction, for example [35] finds that Lattice can in some cases outperform digital nets and in some other not. See discussion [55, Chapter 16 End Notes]. While most theoretical results use NUS for scrambling [47], other methods like digital shift or Linear Matrix Scrambling [40] (or both combined) are commonly used and implemented in software like `scipy.stats.qmc` [68], `qmcpy` [8] because they are much faster and their variance reduction performances look as good as NUS. See [51] for a comparison of some scrambling methods. In our numerical example, we use the NUS [47] implemented in `Julia` [3]. The authors of this paper contributed to the collaborative Julia package QuasiMonteCarlo.jl [67] by adding the several randomization methods used in this work. Some of this code was inspired by the `R` code described on Owen's personal webpage [56]. The authors also developed a Julia package RobustMeans.jl [41] containing all the robust mean estimators used in this paper. The code (script, plot instruction, notebooks and package version) to reproduce the figures presented in this paper are available at the GitHub repository `https://github.com/dmetivie/Robust-Randomized-Quasi-Monte-Carlo-paper-code`. Note that for practitioners, the actual implementation time of QMC and RQMC methods is a major concern and one can prefer fast implementation to strong theoretical guarantees.



(a) $\theta = 2.1$



(b) $\theta = 3.1$

Figure 1: Distribution of $\sqrt{n}(\hat{\mu}_n - \mu)/\sigma$ for different estimators. The X are i.i.d. random variables of Pareto distribution with parameter θ. The number of samples is $n = 56$ and uncertainty level $\delta = 3e^{-8} \simeq 0.1\%$. $M = 10^7$.

6.2.2 Integrand F

Several papers have provided careful numerical tests, see for example [35, 62, 50, 17, 30]. Different types of "difficult" integrands F can be considered:

- The latent dimension d of $\boldsymbol{x}$ is typically large. This is a common case encountered in financial or engineering problems.
- F is not much smooth, for instance it is only $\mathcal{C}^1([0,1]^d)$, or it oscillates a lot.
- F is square integrable but not more (fat-tailed distributions). This is typical in risk management applications.

Such functions are found in the literature and also in applications such as finance where F represent some gain, and μ is the average gain or the risk evaluation.

We consider integrands of the form [55, Chapter 15]

$$F(\boldsymbol{x}) = F_{\boldsymbol{\beta},G}(\boldsymbol{x}) = \prod_{k=1}^{d}(1 + \beta_k G_k(x_k)) \tag{6.1}$$

where for all $k \in \{1, \cdots, d\}$, $G_k : [0,1] \mapsto \mathbb{R}$ satisfies

$$\int_0^1 G_k(x)\mathrm{d}x = 0, \qquad \int_0^1 G_k^2(x)\mathrm{d}x = 1 \tag{6.2}$$

for $\boldsymbol{\beta} \in \mathbb{R}^d$. It is easy to show that

$$\mu = 1 \quad \text{and} \quad \sigma^2 = \prod_{k=1}^{d}(1+\beta_k^2) - 1.$$

We now consider $G_k(\boldsymbol{x}) = G_\varepsilon^{(1)}(\boldsymbol{x})$ for all $k \in \{1, \cdots, d\}$ with

$$G_\varepsilon^{(1)}(x) = \frac{\sqrt{2\varepsilon}(2\varepsilon+1)}{|1-2\varepsilon|}\left(\frac{1}{x^{\frac{1}{2}-\varepsilon}} - \frac{2}{2\varepsilon+1}\right). \tag{6.3}$$

The function $G_\varepsilon^{(1)}$ satisfies the conditions (6.2) for $\varepsilon \in (0, 1/2) \cup (1/2, +\infty)$. Moreover, for such ε, $F_{\boldsymbol{\beta},\boldsymbol{G}} \in L^2([0,1))$. The $\boldsymbol{\beta}$ control how each dimension participates in the integrand. We choose

$$\beta_k = \frac{\beta}{\log(1+k)},$$

so that the larger dimensions participate less than the first (but the logarithmic decay remains very gentle compared to the exponential ones often considered in the literature). For $\varepsilon \in (0, 1/2)$, F has singularities and thus is unbounded.

6.2.3 Numerical Results

For all the tests, $M = 10^5$ realizations are used to obtain the distributions. We compute the sample variance $\hat{\sigma}_N^2 = \frac{1}{M\times n-1}\sum_{i=1}^{n\times M}(\overline{\mu}_N^{(i)} - \bar{\mu}_N)^2$ where $\bar{\mu}_N = \frac{1}{M\times n}\sum_{i=1}^{n\times M}\overline{\mu}_N^{(i)}$ to estimate the mean. We use this estimator to calibrate the α parameter for Catoni (3.6) and Huber (3.4) estimators.

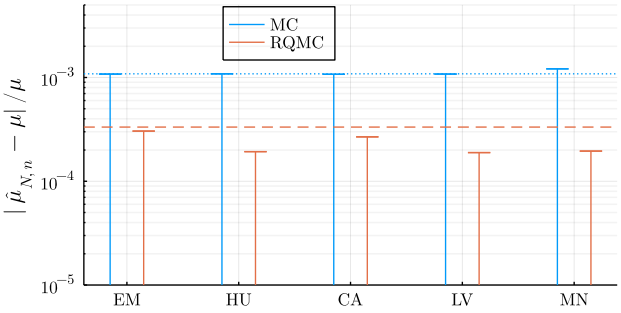
In Figures 2, 3 and 4, we show examples of Monte Carlo (MC) vs. Randomized Quasi Monte Carlo integration in dimension $d = 17$ with both non-singular and singular integrands. In Figure 2, we show only the empirical $(1-\delta)$-quantile for the absolute error $|\hat{\mu}_{N,n} - \mu|/\mu$ while in Figure 3, we show the distribution of the standardized error compared with a Normal distribution. This is an example where F is bounded and smooth, yet the EM even at $\delta = 5\%$ (Figure 3a) performs worse for RQMCS than robust estimators.

In Figure 4, we show an example with a singular integrand. We represent for the absolute error $|\hat{\mu}_{N,n} - \mu|/\mu$ the estimate quantile $\hat{Q}_{1-\delta}$ (top of the whisker) and estimated quantiles $\hat{Q}_{1/4}, \hat{Q}_{3/4}$ forming the box. The dots are the remaining outliers.

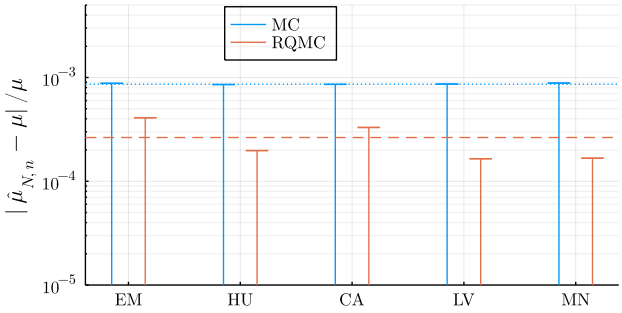
We now demonstrate numerically that for a given budget, $B = N \times n$, the best choice is to take n as small as possible. For example, in Figure 6 with the Lee Valiant estimator (3.8), we fix $B = n \times N = 458752$ and vary n and N. The box plots clearly show that with $N = 8192$ and $n = 56$ the estimator has the smallest error.

In Figure 5, we change the integrand F by changing the G_ε to

$$G_\varepsilon^{(2)}(x) = (x^{\varepsilon-1/2}|x-1/2|^{\varepsilon-1/2} - \eta_1^{(\varepsilon)})/\eta_2^{(\varepsilon)} \tag{6.4}$$

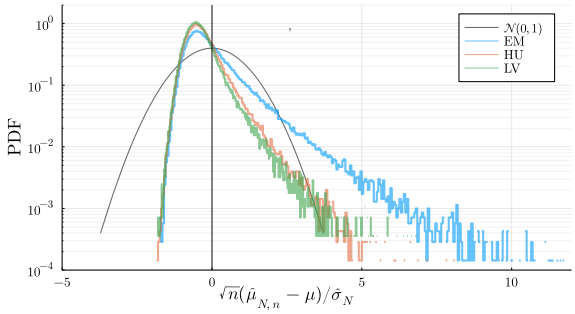


(a) $\delta = 3e^{-4} \simeq 0.055$, $n = 12$, $n(\hat{\sigma}_N^{(\mathrm{MC})})^2 \simeq 4 \times 10^{-5}$ and $n(\hat{\sigma}_N^{(\mathrm{RQMC})})^2 \simeq 4 \times 10^{-6}$.

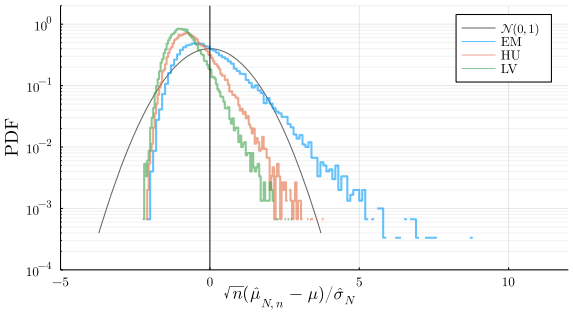


(b) $\delta = 3e^{-8} \simeq 0.001$, $n = 56$, $n(\hat{\sigma}_N^{(\mathrm{MC})})^2 \simeq 2 \times 10^{-4}$ and $n(\hat{\sigma}_N^{(\mathrm{RQMC})})^2 \simeq 2 \times 10^{-5}$

Figure 2: Estimated quantile $\hat{Q}_{1-\delta}$ of the distribution of $|\hat{\mu}_{N,n} - \mu|/|\mu|$ where $F = F_{\boldsymbol{\beta}, G_\varepsilon^{(1)}}$ is defined in Eqs. (6.1), (6.3) with $\varepsilon = 2 \times 10^5$ and β such that $\sigma = 1/2$. Here F is smooth and bounded of range $[-0.01, 200]$. We use $d = 17$, $N = 2^{16}$. The mean estimator acronyms are defined in the text. The RQMC is based on the first N points of a nested uniform scrambled of the Sobol' sequence. The horizontal lines show the $\hat{\sigma}_N \Phi^{-1}(1 - \delta/2)/\sqrt{n}$ values.



(a) $\delta = 3e^{-4} \simeq 0.055$, $n = 12$



(b) $\delta = 3e^{-8} \simeq 0.001$, $n = 56$

Figure 3: Same data as Figure 2 but only for the RQMCS. We depict the full standardized distribution against the normal distribution.

where $\eta_1^{(\varepsilon)}, \eta_2^{(\varepsilon)}$ ensure that Eqs. (6.2) are satisfied. These can be computed explicitly or numerically but are not detailed here. This function has multiple singularities at point which coordinate is either 0 or 1/2. The results are very similar to the case with corner singularity, indicating that robust estimators $\hat{\mu}_{N,n}$ still performs well.

Observations:

- Case where F is bounded and smooth ($\varepsilon = 2 \times 10^5$) – Figures 2 and 3. The range of F is approximately $[-0.01, 200]$ and most of the distribution is located around -0.01. As announced with the MC approach, each $\overline{\mu}_N^{(i)}$ is expected to approximately follow a Normal distribution because N is large. Thus, the estimate $\hat{\mu}_{N,n}$ is also approximately Normal, which is observed since all estimators are almost on the theoretical Normal quantile level $\sigma_N^{(\mathrm{MC})} \Phi^{-1}(1-\delta/2)/\sqrt{n}$. For RQMCS, the Central Limit theorem generally does not apply or if it applies, the asymptotic behavior might be long to reach even for very large N. Here despite the function being smooth

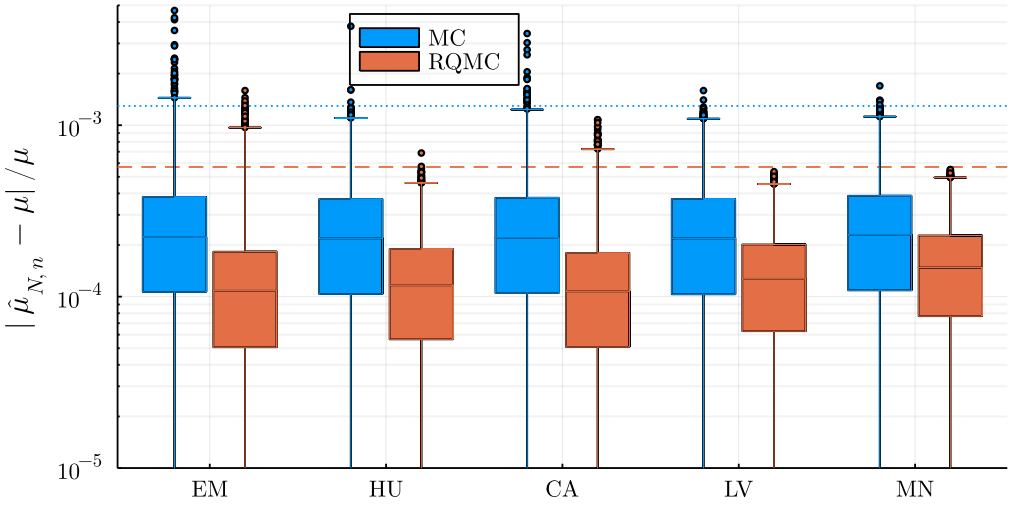


Figure 4: Box plot distribution of $|\hat{\mu}_{N,n} - \mu|/|\mu|$ where $F = F_{\boldsymbol{\beta},G_\varepsilon^{(1)}}$ is defined in Eqs. (6.1), (6.3) with $\varepsilon = 0.05$ and β such that $\sigma \simeq 0.5$. We use $d = 17$, $n = 56$, $N = 2^{15}$, $\delta = 3e^{-8} \simeq 0.001$. The mean estimator acronyms are defined in the text. The RQMC is based on the first N points of a nested uniform scrambled of the Sobol's sequence. The upper whisker is set to extend up to $\hat{Q}_{1-\delta}$. The horizontal lines show the $\hat{\sigma}_N \Phi^{-1}(1-\delta/2)/\sqrt{n}$ values.

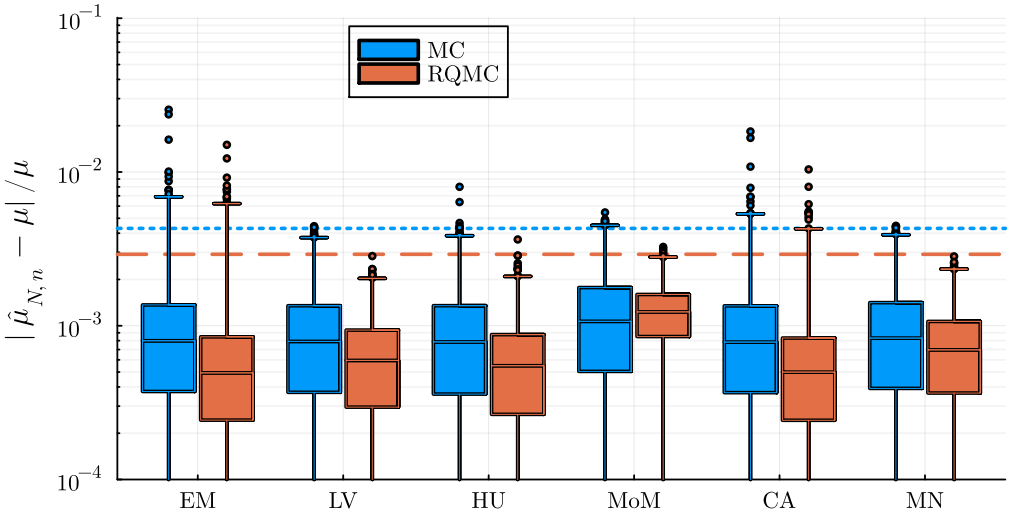


Figure 5: Box plot distribution of $|\hat{\mu}_{N,n} - \mu|/|\mu|$ where $F = F_{\boldsymbol{\beta},G_\varepsilon^{(2)}}$ is defined in Eqs. (6.1),(6.4) with $\varepsilon = 0.04$ and β such that $\sigma \simeq 0.5$. We use $d = 7$, $n = 56$, $N = 2^{11}$, $\delta = 3e^{-8} \simeq 0.001$. The mean estimator acronyms are defined in the text. The RQMC is based on the first N point of a nested uniform scrambled of the Sobol's sequence. The upper whisker is set to extend up to $\hat{Q}_{1-\delta}$. The horizontal lines show the $\hat{\sigma}_N \Phi^{-1}(1-\delta/2)/\sqrt{n}$ values.

and the scrambling method as in the Loh Central Limit theorem [36, Theorem 1] (the condition on the dimension d is not respected, so the theorem does not apply directly) the $\overline{\mu}_N^{(i)}$ are far from being Normal random variables. Moreover, when $n\sigma_N^2$ is smaller, robust estimators

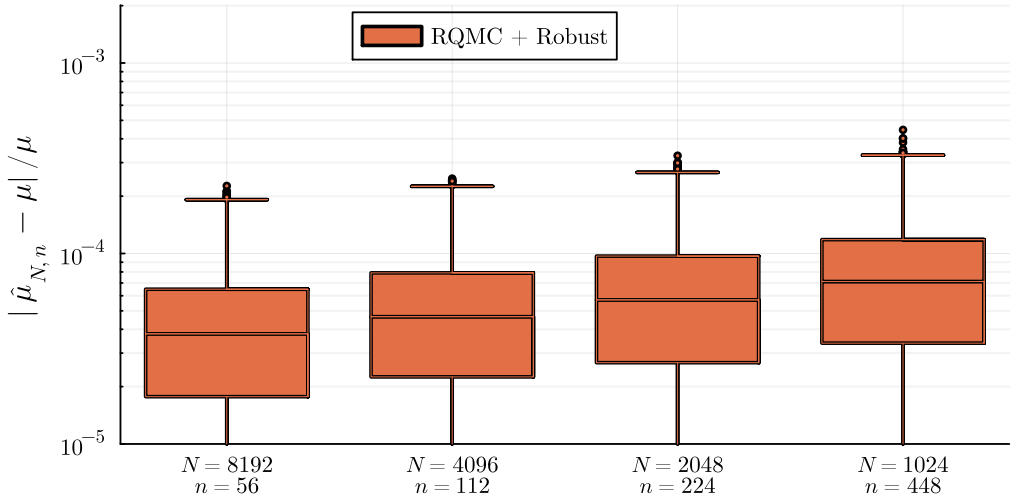


Figure 6: For a fixed budget, $B = n \times N = 458752$ this is the Box plot distribution of $|\hat{\mu}_{N,n} - \mu|/|\mu|$ where $F = F_{\boldsymbol{\beta}, G_\varepsilon^{(1)}}$ is defined in Eqs. (6.1), (6.3) with $\varepsilon = 0.35$ and β such that $\sigma \simeq 0.5$. We use $d = 10$, $\delta = 3e^{-8} \simeq 0.001$. The robust mean estimator used here is the Lee Valiant Eq. (3.8). The RQMC is based on the first N point of a nested uniform scrambled of the Sobol's sequence. The upper whisker is set to extend up to $\hat{Q}_{1-\delta}$.

are expected to be better than EM as discussed in the Introduction. This is observed for uncertainty levels of 5%, Figure 3a and even more so for smaller $\delta = 0.001$, Figure 3b.

- When F is singular at the origin as $x^{-1/2+\varepsilon}$, mimicking a heavy tail distribution, the advantages of combining Robust estimators with RQMCS are even more striking – Figures 4 and 5. Central Limit theorem regime is not completely reached even for MC simulations, despite $N = 2^{15}$.
- As expected – Figure 6 – the smaller n and larger N combination produce the best result by fully taking advantage of the Quasi Monte Carlo variance reduction.
- Even integrand not singular as in Definition 4.4 can benefit from robust RQMC.

7 Conclusion

In this paper, we are interested with robust and efficient estimation methods for very generic integrals. We want to provide accurate error bounds for the estimate. So far, the main tool is to use asymptotic normality, which might either be long to be satisfied or not valid when using RQMCS with generic integrands.

We review various sub Gaussian estimators from the most well known e.g., median-of-means, Huber Z-estimator, to the newest e.g., Lee Valiant and Minsker Ndaoud. These estimators are compared in terms of theoretical advantage and disadvantage, as well as in a numerical illustration. We then introduce the concept of (Randomized) Quasi Monte Carlo methods with a special focus on scrambled nets which are known to always produce asymptotically better estimate than MC as

long as $F \in L^2$. We prove a new asymptotic scaling for the variance when F is smooth but singular at the origin.

Our paper proposes a new methodology combining the two previous concepts to fully exploit the variance reduction $\sigma_N = o(1/N)$ provided by Randomized QMC thanks to sub Gaussian estimators. First it is used to build non-asymptotic quasi-optimal – up to logarithmic terms – confidence intervals which are opposed to classical concentration inequalities which are always not optimal with respect to their variance dependence in the regime $n\sigma_N^2$ small. Then, we illustrate numerically both for bounded smooth integrands and for singular ones, the practical benefits of our Robust Randomized Monte Carlo methodology.

This paper only deals with one-dimensional valued integrands $F : [0,1]^d \to \mathbb{R}$. One natural question is how does the proposed methodology generalize in $\mathbb{R}^q$ for $q > 1$? The sole concept of sub Gaussian estimator is different. Some recent efforts, e.g., [38] have been made to find robust multidimensional estimators. These are defined as estimators satisfying $\hat{\mu}_n(X_1, \dots, X_n)$ for a sample of random vector $X_1, \dots, X_n$ (up to multiplicative constants)

$$\mathbb{P}\left(\|\hat{\mu}_n - \mu\| \geq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max} \log(1/\delta)}{n}} \right) \leq \delta$$

where $\|\cdot\|$ is the Euclidean norm, Σ is the covariance matrix, and $\lambda_{\max}$ its maximum eigenvalue. How do variance reduction methods such as Quasi Monte Carlo influence sub Gaussian bounds? Do we still expect a strong gain over the Empirical Mean estimator and multidimensional concentration inequality? These issues are left for further investigation.

A Variance of unbounded integrands

Our purpose is to establish Theorems 4.3 and 4.4. We use the standard notations from Section 4.

Framework. We consider functions satisfying Definition 4.4 (with $A_k \in (0,1)$) where there is only one singularity at the origin. Essentially, we treat here the case of smooth functions with a singularity at the origin.

Low-Variation extensions of F. Following ideas by Sobol' [61] and Owen [52, Section 2.3], we extend F into $\tilde{F} \in \mathrm{BVHK}$ as follows. Given $\eta \in (0,1)$ whose value is set later, define the hyperbolic region avoiding 0 (the location of the singularity of F) by

$$K_\eta := \left\{ \boldsymbol{x} \in [0,1]^d \mid \prod_{1 \leq k \leq d} x_k \geq \eta \right\},$$

The next construction ensures that $F = \tilde{F}$ on K_η and that

$$|\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})| \leq \tilde{B} \prod_{k=1}^{d} x_k^{-A_k}, \qquad \forall \boldsymbol{x} \notin K_\eta, \tag{A.1}$$

for some constant $\tilde{B}$. Note that K_η contains the anchor point $\mathbf{1}$, in the sense that $\boldsymbol{x} \in K_\eta$ implies $[\boldsymbol{x}, \mathbf{1}] \subseteq K_\eta$. Under this condition, we can write

$$F(\boldsymbol{x}) = F(\mathbf{1}) + \sum_{u \neq \emptyset} \int_{[\mathbf{1}_u, \boldsymbol{x}_u]} \frac{\partial^{|u|} F}{\partial \boldsymbol{x}_u}(\boldsymbol{z}_u : \mathbf{1}_{-u}) \mathrm{d}\boldsymbol{z}_u,$$

where we recall that for $\boldsymbol{z} = (z_1, \cdots, z_d) \in [0,1]^d$, $\boldsymbol{z}_u$ stands for the components z_j such that $j \in u$ and $\boldsymbol{z}_u : \mathbf{1}_{-u}$ denotes the point $\boldsymbol{y} \in [0,1]^d$ with $y_j = z_j$ for $j \in u$ and $y_j = 1$ for $j \notin u$. The low variation extension is

$$\tilde{F}(\boldsymbol{x}) = F(\mathbf{1}) + \sum_{u \neq \emptyset} \int_{[\mathbf{1}_u, \boldsymbol{x}_u]} \mathbb{1}_{\boldsymbol{z}_u : \mathbf{1}_{-u} \in K_\eta} \frac{\partial^{|u|} F}{\partial \boldsymbol{x}_u}(\boldsymbol{z}_u : \mathbf{1}_{-u}) \mathrm{d}\boldsymbol{z}_u.$$

Indeed, by the fundamental theorem of calculus $\tilde{F} = F$ on $\boldsymbol{x} \in K_\eta$. In addition, from [52, Proof of Theorem 5.5] and taking advantage of Definition 4.4, we get $\tilde{F} \in \mathrm{BVHK}$, with $V_{\mathrm{HK}}(\tilde{F}) = \mathcal{O}(\eta^{-\max_k A_k})$ assuming the largest A_k is unique and $\mathcal{O}(\eta^{-\max_k A_k - \varepsilon})$ for any $\varepsilon > 0$ otherwise. Last, this extension $\tilde{F}$ satisfies the error bound (A.1), see [52, Lemma 5.1].

If we consider the RQMC sequence $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)$ not necessarily inside K_η, the estimator (4.2) satisfies

$$\begin{aligned}
&|\bar{\mu}_N - \mu| \\
&\leq \left| \frac{1}{N} \sum_{j=1}^N \tilde{F}(\boldsymbol{R}_j) - \int_{[0,1]^d} \tilde{F}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| + \int_{[0,1]^d} |\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} + \frac{1}{N} \sum_{j=1}^N \left| F(\boldsymbol{R}_j) - \tilde{F}(\boldsymbol{R}_j) \right| \\
&\underset{(4.1)}{\leq} D_N^*(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) V_{\mathrm{HK}}(\tilde{F}) + \int_{K_\eta^c} |\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} + \frac{1}{N} \sum_{j=1}^N \left| F(\boldsymbol{R}_j) - \tilde{F}(\boldsymbol{R}_j) \right|. \qquad \text{(A.2)}
\end{aligned}$$

Bound on L^1-Error. Let us prove Theorem 4.3. Take the expectation in (A.2):

$$\mathbb{E}(|\bar{\mu}_N - \mu|) \leq \mathbb{E}(D_N^*(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)) V_{\mathrm{HK}}(\tilde{F}) + 2 \int_{K_\eta^c} |\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}.$$

From [52, Lemma 5.4], we get

$$\int_{\prod_k x^k < \eta} |\widetilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq \mathcal{O}\left(\eta^{1-\varepsilon-\max_k A_k}\right). \qquad \text{(A.3)}$$

Hence, we obtain

$$\mathbb{E}(|\bar{\mu}_N - \mu|) = \mathcal{O}\left(N^{-1+\varepsilon/2}\right) \mathcal{O}(\eta^{-\max_k A_k - \varepsilon/2}) + \mathcal{O}\left(\eta^{1-\varepsilon-\max_k A_k}\right),$$

whence the result when we take $\eta = N^{-1}$. Theorem 4.3 is proved. □

Bound on the variance. To get the analog result for the variance, take the square of (A.2) and the expectation: it gives

$$\mathrm{Var}\left(\bar{\mu}_N\right) \leq 3\,\mathbb{E}\left(D_N^*\left(\boldsymbol{R}_1,\ldots,\boldsymbol{R}_N\right)^2\right) V_{\mathrm{HK}}(\tilde{F})^2 + 3\left(\int_{K_\eta^c} |\tilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}\right)^2 + 3\,\mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^N \left|F\left(\boldsymbol{R}_i\right) - \tilde{F}\left(\boldsymbol{R}_i\right)\right|\right)^2\right).$$

Setting $\eta = N^{-1}$, the two first terms on the above right-hand side are $\mathcal{O}\left(N^{-2+\varepsilon+2\max_k A_k}\right)$ using the same estimates as before (i.e., (A.3) and the bound on $V_{\mathrm{HK}}(\tilde{F})$). It remains to upper-bound the last term. Expanding the square of the average, using the decomposition (4.5) and that the joint distribution $\Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ has a uniformly bounded density (by a constant C_Ψ), it readily follows

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^N \left|F\left(\boldsymbol{R}_i\right) - \tilde{F}\left(\boldsymbol{R}_i\right)\right|\right)^2\right) \\
&\quad = \frac{1}{N}\int_{[0,1]^d} |F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + \frac{N-1}{N}\int_{[0,1]^d\times[0,1]^d} |F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x})|\,|F(\boldsymbol{y}) - \tilde{F}(\boldsymbol{y})| \Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \\
&\leq \frac{1}{N}\int_{K_\eta^c} |F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + C_\Psi\left(\int_{K_\eta^c} |F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}\right)^2. \qquad \text{(A.4)}
\end{aligned}$$

Similarly to (A.3) and using (A.1), we derive

$$\int_{\prod_k x^k < \eta} |\widetilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \leq \mathcal{O}\left(\eta^{1-\varepsilon-2\max_k A_k}\right), \qquad \text{(A.5)}$$

for any $\varepsilon > 0$. Taking $\eta = N^{-1}$, plugging (A.5) and (A.3) into (A.4) gives

$$\begin{aligned}
\mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^N \left|F\left(\boldsymbol{R}_i\right) - \tilde{F}\left(\boldsymbol{R}_i\right)\right|\right)^2\right) &= \mathcal{O}\left(\frac{1}{N} N^{-1+\varepsilon+2\max_k A_k} + (N^{-1+\varepsilon+\max_k A_k})^2\right) \\
&= \mathcal{O}\left(N^{-2+2\varepsilon+2\max_k A_k}\right).
\end{aligned}$$

This gives (the arbitrary ε values are adjusted to correspond)

$$\mathrm{Var}\left(\bar{\mu}_N\right) = \mathcal{O}\left(N^{-2+\varepsilon+2\max_k A_k}\right).$$

This asymptotic regime is better than Monte Carlo $\mathcal{O}(N^{-1})$ when $\max_k A_k < 1/2$. □

Remark A.1. *In the case where the assumption of bounded density for the joint density distribution Ψ is not available, we can still use the rough bound*

$$\mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^N |F(\boldsymbol{R}_i) - \tilde{F}(\boldsymbol{R}_i)|\right)^2\right) \leq \mathbb{E}\left(|F(\boldsymbol{R}_1) - \tilde{F}(\boldsymbol{R}_1)|^2\right) = \int_{\prod_k x^k < \eta} |\widetilde{F}(\boldsymbol{x}) - F(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}.$$

With the use $\eta = N^{-1}$ and (A.5), we obtain

$$\mathrm{Var}\left(\bar{\mu}_N\right) = \mathcal{O}\left(N^{-1+\varepsilon+2\max_k A_k}\right),$$

which for $\max_k A_k < 1/4$ is better than Monte Carlo.

B Bounded density for the joint distribution $\Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$

This joint distribution function Ψ is studied in detail for a base b digital scrambled (t, m, d)-nets in [69]. Their definition of base b digital scramble definition includes NUS and other types of scrambles. This boundedness hypothesis of Ψ is satisfied all over $[0,1]^{2d}$ for digital nets $(0, m, d)$ in base b, where we can obtain using [69, Lemma 3.4 and Theorem 3.6],

$$\frac{\Psi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})}{\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}} \leq \frac{2^{d+1}b^{m+d}}{(b^m-1)(b-1)^d} \leq 2^{d+1} \sup_{N\geq 2} \left(\frac{N}{N-1}\right)^{d+1} = 4^{d+1}$$

uniformly in $\boldsymbol{x}$, $\boldsymbol{y}$. □

References

[1] N. Alon, Y. Matias, and M. Szegedy. The Space Complexity of Approximating the Frequency Moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[2] C. Bayer, H. Hoel, E. von Schwerin, and R. Tempone. On non-asymptotic optimal stopping criteria in Monte Carlo simulations. *SIAM Journal on Scientific Computing*, 36(2):A869–A885, 2014.

[3] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence.* Clarendon Press, Oxford, 2013.

[5] R. E. Caflisch. Monte-Carlo and Quasi-Monte-Carlo methods. *Acta numerica*, 7:1–49, 1998.

[6] O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'I.H.P. Probabilités et statistiques*, 48(4):1148–1185, 2012.

[7] L. Chamakh, E. Gobet, and W. Liu. Orlicz norms and concentration inequalities for β-heavy tailed random variables. *In minor revision for Bernoulli*, 2021.

[8] S.-C. T. Choi, F. J. Hickernell, R. Jagadeeswaran, M. J. McCourt, and A. G. Sorokin. Quasi-Monte Carlo software. *arXiv:2102.07833*, Oct. 2021.

[9] N. Chopin and M. Gerber. Higher-order stochastic integration through cubic stratification, June 2023.

[10] B. Cipra. The Best of the 20th Century: Editors Name Top 10 Algorithms. *SIAM News* `https://archive.siam.org/news/news.php?id=637`, 33(4), 2000.

[11] R. Cranley and T. N. L. Patterson. Randomization of number theoretic methods for multiple integration. *SIAM Journal on Numerical Analysis*, 13(6):904–914, 1976.

[12] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[13] J. Dick. Koksma–Hlawka type inequalities of fractional order. *Annali di Matematica Pura ed Applicata*, 187(3):385–403, 2008.

[14] J. Dick. Higher order scrambled digital nets achieve the optimal rate of the root mean square error for smooth integrands. *The Annals of Statistics*, 39(3):1372–1398, June 2011.

[15] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the Quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[16] J. Dick and F. Pillichshammer. *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press, 2010.

[17] H. Faure and C. Lemieux. Generalized Halton sequences in 2008: A comparative study. *ACM Transactions on Modeling and Computer Simulation*, 19(4):15:1–15:31, 2009.

[18] M. Gerber. On integration methods based on scrambled nets of arbitrary size. *Journal of Complexity*, 31(6):798–816, 2015.

[19] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. Wiley, New York, 1986.

[20] Z. He. Quasi-Monte Carlo for discontinuous integrands with singularities along the boundary of the unit cube. *Mathematics of Computation*, 87(314):2857–2870, 2018.

[21] Z. He and X. Wang. On the convergence rate of randomized Quasi–Monte Carlo for discontinuous functions. *SIAM Journal on Numerical Analysis*, 53(5):2488–2503, 2015.

[22] P. Hellekalek. On the assessment of random and quasi-random point sets. In *Random and quasi-random point sets*, pages 49–108. Springer, 1998.

[23] F. J. Hickernell, H. S. Hong, P. L'Écuyer, and C. Lemieux. Extensible lattice sequences for quasi-Monte Carlo quadrature. *SIAM Journal on Scientific Computing*, 22(3):1117–1138, 2000.

[24] E. Hlawka. Funktionen von beschränkter Variatiou in der Theorie der Gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54(1):325–333, 1961.

[25] P. Huber and E. Ronchetti. *Robust Statistics*. Wiley, New York, second edition, 2009.

[26] P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[27] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, Jan. 1986.

[28] J. Koksma. Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1. *Mathematica B (Zutphen)*, 11(7-11):43, 1942.

[29] R. J. Kunsch and D. Rudolf. Optimal confidence for Monte Carlo integration of smooth functions. *Advances in Computational Mathematics*, 45(5):3095–3122, Dec. 2019.

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

[31] P. L'Ecuyer. Randomized Quasi-Monte Carlo: An Introduction for Practitioners. In A. B. Owen and P. W. Glynn, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, Springer Proceedings in Mathematics & Statistics, pages 29–52, Cham, 2018. Springer International Publishing.

[32] P. L'Ecuyer and C. Lemieux. Recent advances in randomized Quasi-Monte Carlo methods. In M. Dror, P. L'Ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, International Series in Operations Research & Management Science, pages 419–474. Springer US, New York, NY, 2002.

[33] P. L'Ecuyer, D. Munger, and B. Tuffin. On the distribution of integration error by randomly-shifted lattice rules. *Electronic Journal of Statistics*, 4:950–993, 2010.

[34] J. C. Lee and P. Valiant. Optimal Sub-Gaussian Mean Estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683, Feb. 2022.

[35] C. Lemieux and P. L'Ecuyer. Efficiency improvement by lattice rules for pricing Asian options. In *1998 Winter Simulation Conference. Proceedings (Cat. No.98CH36274)*, volume 1, pages 579–585, 1998.

[36] W.-L. Loh. On the asymptotic distribution of scrambled net quadrature. *The Annals of Statistics*, 31(4):1282–1324, 2003.

[37] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[38] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, Dec. 2019.

[39] T. Mathieu. Concentration study of M-estimators using the influence function. *Electronic Journal of Statistics*, 16(1):3695–3750, Jan. 2022.

[40] J. Matoušek. On the l2-discrepancy for anchored boxes. *Journal of Complexity*, 14(4):527–556, 1998.

[41] D. Métivier. `RobustMeans.jl`. https://github.com/dmetivie/RobustMeans.jl, 2022.

[42] S. Minsker and M. Ndaoud. Robust and efficient mean estimation: An approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036–6070, 2021.

[43] M. Nakayama and B. Tuffin. Sufficient conditions for a central limit theorem to assess the error of randomized Quasi-Monte Carlo methods. In *2021 - Winter Simulation Conference*, page 1, Dec. 2021.

[44] A. S. Nemirovskij and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[45] H. Niederreiter. *Random number generation and quasi-Monte-Carlo methods*, volume 63 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[46] R. Oliveira and P. Orenstein. The sub-Gaussian property of trimmed means estimators. *Unpublished*, IMPA, 2019.

[47] A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Lecture Notes in Statistics, pages 299–317, New York, NY, 1995. Springer.

[48] A. B. Owen. Monte Carlo Variance of Scrambled Net Quadrature. *SIAM Journal on Numerical Analysis*, 34(5):1884–1910, 1997.

[49] A. B. Owen. Scrambling Sobol' and Niederreiter–Xing Points. *Journal of Complexity*, 14(4):466–489, 1998.

[50] A. B. Owen. The dimension distribution and quadrature test functions. *Statistica Sinica*, 13(1):1–17, 2003.

[51] A. B. Owen. Variance with alternative scramblings of digital nets. *ACM Transactions on Modeling and Computer Simulation*, 13(4):363–378, 2003.

[52] A. B. Owen. Halton sequences avoid the origin. *SIAM Review*, 48(3):487–503, 2006.

[53] A. B. Owen. Quasi-Monte Carlo for integrands with point singularities at unknown locations. In H. Niederreiter and D. Talay, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 403–417. Springer, 2006.

[54] A. B. Owen. Local antithetic sampling with scrambled nets. *The Annals of Statistics*, 36(5):2319–2343, 2008.

[55] A. B. Owen. *Monte Carlo Theory, Methods and Examples*. Stanford University, 2013.

[56] A. B. Owen. Nested uniform scrambled sobol' points. https://artowen.su.domains/code/rsobol.R, 2017.

[57] A. B. Owen and D. Rudolf. A strong law of large numbers for scrambled net integration. *SIAM Review*, 63(2):360–372, 2021.

[58] Z. Pan and A. Owen. Super-polynomial accuracy of one dimensional randomized nets using the median of means. *Mathematics of Computation*, 92(340):805–837, Mar. 2023.

[59] Z. Pan and A. B. Owen. Super-polynomial accuracy of multidimensional randomized nets using the median-of-means, Aug. 2022.

[60] I. H. Sloan. *Lattice methods for multiple integration*. Oxford University Press, 1994.

[61] I. M. Sobol'. Computation of improper integrals by means of equidistributed sequences. *Doklady Akademii Nauk SSSR*, 210:278–281, 1973.

[62] I. M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001.

[63] J. Spanier and E. H. Maize. Quasi-random methods for estimating integrals using relatively small samples. *SIAM review*, 36(1):18–44, 1994.

[64] S. M. Stigler. The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3):472–477, 1973.

[65] V. Temlyakov. *Multivariate approximation*, volume 32. Cambridge University Press, 2018.

[66] M. Ullrich. A Monte Carlo method for integration of multivariate smooth functions. *SIAM Journal on Numerical Analysis*, 55(3):1188–1200, 2017.

[67] Various contributors. `QuasiMonteCarlo.jl`. https://github.com/SciML/QuasiMonteCarlo.jl, 2019.

[68] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[69] J. Wiart, C. Lemieux, and G. Y. Dong. On the dependence structure and quality of scrambled (t, m, s)-nets. *Monte Carlo Methods and Applications*, 27(1):1–26, 2021.