# An Evaluation of Estimation Techniques
# for Probabilistic Reachability

Mariia Vasileva　　　　Paolo Zuliani

School of Computing, Newcastle University, UK

`vasileva.m2@newcastle.ac.uk`　　　`paolo.zuliani@newcastle.ac.uk`

We evaluate numerically-precise Monte Carlo (MC), Quasi-Monte Carlo (QMC) and Randomised Quasi-Monte Carlo (RQMC) methods for computing probabilistic reachability in hybrid systems with random parameters. Computing reachability probability amounts to computing (multidimensional) integrals. In particular, we pay attention to QMC methods due to their theoretical benefits in convergence speed with respect to the MC method. The Koksma-Hlawka inequality is a standard result that bounds the approximation of an integral by QMC techniques. However, it is not useful in practice because it depends on the variation of the integrand function, which is in general difficult to compute. The question arises whether it is possible to apply statistical or empirical methods for estimating the approximation error. In this paper we compare a number of interval estimation techniques based on the Central Limit Theorem (CLT), and we also introduce a new approach based on the CLT for computing confidence intervals for probability near the borders of the [0,1] interval. Based on our analysis, we provide justification for the use of the developed approach and suggest usage guidelines for probability estimation techniques.

## 1　Introduction

Reachability is one of the fundamental problems in verification and model checking. Given a system model and a set of "goal" states (indicating (un)wanted behaviour), does the system eventually reach these states? The generalisation of this problem for stochastic systems is called probabilistic reachability, and it amounts to compute the probability that the system reaches a goal state.

Checking reachability in hybrid systems is an undecidable problem for all but the simplest systems (timed automata) [2]. Formal verification of hybrid systems can include verifying satisfiability of formulas involving real variables, which is known to be an undecidable problem when, *e.g.*, trigonometric functions are involved [20]. In order to combat the undecidability of general sentences over the reals, Gao, Avigad and Clarke defined the notion of $\delta$-complete decision procedure [11]. This approach has been extended to a bounded probabilistic reachability method with statistically valid enclosures for the probability that a hybrid system can reach a goal state within a given time bound and number of steps [18]. In particular, we consider the $k$-step reachability probability for parametric hybrid systems with random parameters. The ProbReach tool [17] implements such a method and computes under- and over-approximation of the reachability probability, which amounts to computing multi-dimensional integrals. There are three possible ways to compute such integrals - formal, Monte-Carlo (MC) and Quasi-Monte Carlo (QMC). The number of system evolutions to explore in order to accurately compute integrals grows exponentially with respect to the number of dimensions [20]. This motivates a combination of MC and QMC methods and numerical decision procedures in order to define efficient, numerically accurate estimation techniques.

It is well-known that MC methods are based on the Law of Large Numbers and random sampling. Instead, QMC methods are based on *deterministic* sampling from so-called quasi-random sequences. The

error estimation of the QMC method can be computed theoretically by the Koksma-Hlawka inequality. Unfortunately, in practice its use is connected with a number of calculation difficulties [14]. The terms of quasi-random sequences are statistically dependent, so the Central Limit Theorem (CLT) can not be used for estimating the integration error. At the same time we can successfully use the CLT for estimating the error of the *Randomised* Quasi-Monte Carlo (RQMC) methods.

The aim of this paper is to compare different interval estimation techniques, in particular in the extreme cases of probability close to 0 or 1. A problem of many confidence interval (CI) techniques is that the actual coverage probability of the interval near the boundaries (0 and 1) can be poor [16] [6].

The paper is structured as follows: In Section 2, we briefly introduce probabilistic reachability for stochastic parametric hybrid systems. In Section 3, we present integral estimation methods including MC, QMC and RQMC. Additionally, we consider recent approach to QMC variance calculation for statistical error estimation. In Section 4, we examine CI error estimation with approaches based on the standard CLT interval and on the *Beta* function. In Section 5, we empirically compare those CI estimation techniques on five benchmarks, and derive usage guidelines. In Section 6, we provide conclusions and suggest future work in the area.

## 2   Probabilistic Reachability

Hybrid systems provide a framework for modelling real-world systems that combine continuous and discrete dynamics [2]. In particular parametric hybrid systems (PHSs) [18] represent continuous and discrete dynamic behavior dependent on initial parameters, which remain unchanged during the system evolution. Such systems can both flow, described by a differential equation and jump, described by difference equations or control graphs.

In order to combat the undecidability of reasoning over real sentences Gao *et al.* [11] defined $\delta$-complete decision procedures, which correctly decide whether a slightly relaxed sentence is satisfiable or not. Let $\delta \in \mathbb{Q}^+ \cup \{0\}$ be a constant and $\phi$ a bounded $\Sigma_1$-sentence in the standard form: $\phi = \exists^{I_1} x_1, ..., \exists^{I_n} x_n : \bigwedge_{i=1}^m (\bigvee_{j=1}^{k_i} f_{ij}(x_1, ..., x_n) = 0)$, where the $f_{ij}(x_1, ..., x_n)$ are compositions of Type 2-computable functions (these are essentially "numerically computable" real functions, including transcendental functions and solutions of differential equations). The $\delta$-weakening of $\phi$ is the formula: $\phi^\delta = \exists^{I_1} x_1, ..., \exists^{I_n} x_n : \bigwedge_{i=1}^m (\bigvee_{j=1}^{k_i} |f_{ij}(x_1, ..., x_n)| \leq \delta)$. The bounded $\delta$-SMT problem asks for the following: given a sentence $\phi$ of the above form and $\delta \in \mathbb{Q}^+$, correctly decide one of the following:

- **unsat**: $\phi$ is false,

- $\delta$-**true**: $\phi^\delta$ is true.

If the two cases overlap either decision can be returned. Standard bounded reachability questions over PHSs can be coded as $\Sigma_1$ sentences and "$\delta$-decided" by $\delta$-complete decision procedures [13].

In this paper we are concerned with *stochastic* PHS, which introduce random parameters to an otherwise deterministic PHS. Bounded $k$-step reachability in PHSs aims at finding the **probability** that for the given initial conditions, the system reaches a goal state in $k$ discrete transitions within a given finite time. It can be shown that this probability can be computed as an integral of the form $\int_G d\mathbb{P}$, where $G$ denotes the set of all random parameter values for which the system reaches a goal state in $k$ steps, and $\mathbb{P}$ is the probability measure associated with the random parameters [18].

# 3 Integral Estimation Methods

## 3.1 Monte Carlo Method

Consider the integral $I = \int_a^b f(y)dy$, and a random variable $U$ on $[a,b]$. The expectation of $f(U)$ is:

$$\mathbb{E}[f(U)] = \int_a^b f(y)\varphi(y)dy$$

where $\varphi$ is the density of $U$. If $U$ is uniformly distributed on $[a,b]$, then the integral becomes:

$$I = \int_a^b f(y)dy = (b-a)\mathbb{E}[f(U)].$$

If we take $N$ points, uniformly distributed on $[a,b]$, and compute the sample mean $\frac{1}{N}\sum_{i=1}^N f(u_i)$, we obtain the MC integral estimation:

$$\int_a^b f(y)dy \approx (b-a)\frac{1}{N}\sum_{i=1}^N f(u_i) \tag{1}$$

According to the Strong Law of Large Numbers, this approximation is convergent (for $N \to \infty$) to $I$ with probability one. The variance of the MC estimator (1) is:

$$Var(MC) = \int_a^b \dots \int_a^b \left(\frac{1}{N}\sum_{i=1}^N f(u_i) - I\right)^2 du_1 \dots du_N = \frac{\sigma_f^2}{N} \tag{2}$$

MC integration error mean equals to $\frac{\sigma_f^2}{N}$, where $\sigma_f^2$ is the integrand variance, which is assumed to exist. In practice, the integrand variance is often unknown. That is why the next estimation for the CI is instead used:

$$\widehat{\sigma}_f^2 = \frac{1}{N-1}\sum_{i=1}^N \left(f(u_i) - \frac{1}{N}\sum_{i=1}^N f(u_i)\right)^2$$

This estimator possesses the unbiasedness property: $\mathbb{E}[\widehat{\sigma}_f^2] = \sigma_f^2$.

## 3.2 Quasi-Monte Carlo Method

QMC methods can be regarded as a deterministic counterpart to classical MC method. Unlike MC integration, which uses estimates (1) with randomly selected points, QMC methods select the points $u_i$ *deterministically*. Specifically, QMC techniques produce deterministic sequences of points that provide the best-possible spread over the integration domain. These deterministic sequences are often referred to as low-discrepancy sequences. The Sobol sequence [19] is a well-known example of low-discrepancy sequence. In Figure 1, we present a simple example of the comparison between Sobol and pseudorandom points distribution. An effective way to use the QMC method is by performing a change of variables to reduce integration to the $[0,1]$ domain. When we need to integrate over a large domain $[a,b]$, that avoids multiplying the QMC estimate by a large factor $(b-a)$ as required by (1).

A QMC advantage with respect to MC is that its error is $O\left(\frac{1}{N}\right)$, while the MC error is $O\left(\frac{1}{\sqrt{N}}\right)$, where $N$ is the sample size. The Koksma-Hlawka inequality bounds the error of QMC estimates, but in practical applications it is very hard to estimate [14], thereby hampering the use of QMC methods. As such, other methods for estimating the QMC error need to be developed.
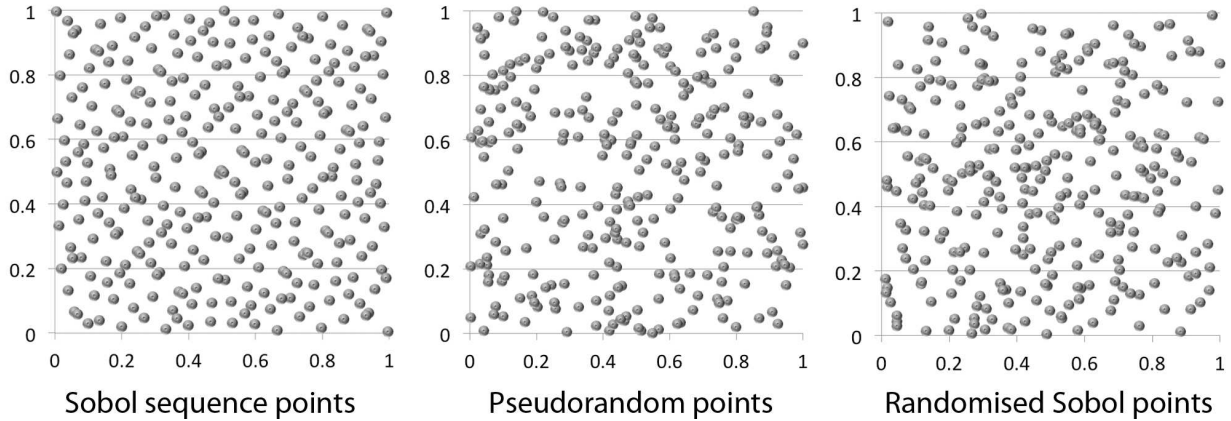
Figure 1: Sobol sequence, uniform pseudorandom and randomised Sobol sequence points (obtained by transformation $\Gamma = (\mathfrak{X} + \varepsilon) mod\, 1$, where $\varepsilon$ is a random sample from MC sequence and $\mathfrak{X}$ is low-discrepancy sample from Sobol sequence) distribution in the 2-dimensional unit space. The comparison is based on the first 300 points of sequences.

**Qint Cubature Method.**    Ermakov and Antonov [5] have recently introduced a new method for QMC variance estimation. To construct an estimate of the integral $I$ they use the set of random quadrature formulas, which were introduced by the Ermakov-Granovsky theorem [10]. This theorem allows us to construct $N$-point formulas with two important properties: the unbiasedness property for integral $I$ and the accuracy property for the considered Haar system. The nodes of the formula are random variables with distribution density:

$$\phi(u_1, u_2, ..., u_N) = \begin{cases} \frac{N^N}{N!} & \text{if } (u_1, u_2, ..., u_N) \in Lat(i_1, i_2, ..., i_N) \\ 0 & \text{otherwise} \end{cases}$$

where $Lat(i_1, i_2, ..., i_N)$ is a Latin set that relates to the permutation $(i_1, i_2, ..., i_N)$ and can be defined by the next condition:

$$(u_1, u_2, ..., u_N) \in Lat(i_1, i_2, ..., i_N) \Leftrightarrow \forall j \in \{1, 2, ..., N\} u_j \in \mathcal{U}_{i_j}$$

where $\mathcal{U}_{i_j}$ is a set of permuted orthonormal Haar functions [5].

The variance of the constructed cubature formula $Cub[f] = \frac{1}{N} \sum_{i=1}^{N} f(u_i)$ can be calculated as:

$$\mathbb{D}Cub[f] = \int_{\mathcal{U}^N} Cub[f]^2 d\phi - \left( \int_{\mathcal{U}^N} Cub[f] d\phi \right)^2 =$$

$$= \mathbb{D}MC[f] + \frac{1}{N}(a_1 + a_2 + ... + a_N)^2 - a_1^2 - a_2^2 - ... - a_N^2 = \mathbb{D}_{MC}[f] - \frac{1}{N} \sum_{i<j}(a_i - a_j)^2,$$

where $\mathbb{D}MC$ is the variance of MC method (2) and $a_i = \int_{\mathcal{U}_i} f(u)\mu(du)$ for $i = 1, 2, ..., N$.

In other words, we can redefine the integral estimation variance as:

$$Var(QMC) = Var(MC) - \frac{1}{N} \sum_{i<j}(a_i - a_j)^2 \ . \tag{3}$$

### 3.3 Randomised Quasi-Monte Carlo

As discussed earlier, the practical application of QMC is limited by the difficulty of computing an estimate of the integration error. However, allowing randomization into the deterministic QMC procedure enables constructing CIs. A *Randomised* QMC (RQMC) procedure can be described as follows. Suppose that $\mathfrak{X} = \{x_1, ..., x_n\}$ is a deterministic low-discrepancy set. By means of a transformation $\tilde{\mathfrak{X}} = \Gamma(\mathfrak{X}, \varepsilon)$ a finite set $\tilde{\mathfrak{X}}$ is generated by the random variable $\varepsilon$ and has the same quasi-random properties as set $\mathfrak{X}$ (see Figure 1). For a randomised set $\tilde{\mathfrak{X}}_i$ we construct a RQMC estimate similar to (1):

$$RQMC_{j,n} = \frac{1}{n} \sum_{i=1}^{n} f(\tilde{\mathfrak{X}}_{i,j}) \tag{4}$$

for $0 < j \leqslant r$, where $r$ is the total number of different pseudo-random sequences. Then, we take their average for overall RQMC estimation (4):

$$RQMC_n = \frac{1}{r} \sum_{j=1}^{r} RQMC_{j,n}. \tag{5}$$

If we choose the $\Gamma$ transformation in such a way that each of the estimates $RQMC_{j,n}$ has the unbiasedness property, *i.e.*, $\forall j \, \mathbb{E}\left[RQMC_{j,n}\right] = I$, (*e.g.* $\Gamma = (\mathfrak{X} + \varepsilon) \bmod 1$), then the estimator (5) will also be unbiased, *i.e.*, $\mathbb{E}[RQMC_n] = I$. By independence of the samples used in (4) and (5), we have that:

$$Var(RQMC_n) = \frac{\sum_{j=1}^{r} Var(RQMC_{j,n})}{r}.$$

Thus, we have the following variance estimation:

$$\widehat{Var}(RQMC_n) = \frac{1}{r(r-1)} \sum_{j=1}^{r} \left( RQMC_{j,n} - RQMC_n \right)^2.$$

### 3.4 Validated MC and QMC

Let us consider a hybrid system $H$ with random parameters only. For any parameter value $p$ from the initial parameters distribution we introduce the Bernoulli random variable $X$, which takes 1 if system $H$ reaches the goal in $k$ steps for $p$ and 0 otherwise. Since in general we can not sample $X$ because of undecidability, we instead consider two Bernoulli random variables: $X_{sat}$, which takes 1 if we can *correctly decide* that system $H$ reaches the goal in $k$ steps for $p$ and 0 otherwise; $X_{usat}$, which takes 0 if we can *correctly* decide that system $H$ does not reach the goal and 1 otherwise [18]. Therefore:

$$X_{sat} \leqslant X \leqslant X_{usat}$$

and thus:

$$\mathbb{E}[X_{sat}] \leqslant \mathbb{E}[X] \leqslant \mathbb{E}[X_{usat}].$$

By the definition of expectation, and denoting $P_R$ as the domain of the random parameters of $H$, we get:

$$\int_{P_R} X_{sat}(p) dp \leqslant \int_{P_R} X(p) dp \leqslant \int_{P_R} X_{usat}(p) dp. \tag{6}$$

We take the sample approximation of (6) and obtain

$$\frac{1}{N} \sum_{i=1}^{N} X_{sat}(p_i) \leqslant \frac{1}{N} \sum_{i=1}^{N} X(p_i) \leqslant \frac{1}{N} \sum_{i=1}^{N} X_{usat}(p_i)$$

where the $p_i$'s can be sampled by using low-discrepancy sequences for QMC methods or pseudo-random sequences for MC methods.

## 4    Confidence Interval Estimation and Error Analysis

In the following we shall use the notation blow:
- $\tilde{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - the sample mean.
- $C_a = Quant\left(1 - \frac{a}{2}\right)$ - the inverse cumulative distribution function of a normal random variable with mean 0 and standard deviation 1; parameter $a$ defines the confidence level at $1 - a$.
- $\hat{p} = n_s/n$ - the binomially-distributed proportion, where: $n_s$ - the number of "successes" and $n_f$ - the number of "failures" in Bernoulli trial process; $n$ - the total number of Bernoulli "trials".
- $\hat{q} = 1 - \hat{p}$.

### 4.1    Intervals Based on the Standard CLT Interval

**Modified Central Limit Theorem interval.**    First, we consider the case when the sample $x_i$ is extracted from the normal distribution $N(\mu, \sigma^2)$ with unknown parameter $\mu$ and known $\sigma^2$, where $\mu$ is the mean or expectation of the distribution and $\sigma^2$ is the variance. Here, $\mu$ can be approximated by the sample mean: $\mu \approx \tilde{X}$. To clarify this approximation, we construct a CI covering the parameter $\mu$ with a given confidence probability:

$$CI_{CLT} = \left(\tilde{X} - C_a\frac{\sigma}{\sqrt{n}}; \tilde{X} + C_a\frac{\sigma}{\sqrt{n}}\right) \tag{7}$$

If the variance $\sigma^2$ is unknown, we can use the same CI by replacing $\sigma$ with the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \tilde{X})^2}$. This method is widely used for estimating the distribution of the error regarding the binomially-distributed proportions. Many related works [7, 6, 8] note that the $CI_{CLT}$ approximation can be poor when applied to Bernoulli trials with $\hat{p}$ close to 0 or 1.

In order to resolve this problem, we introduce a new method for variance estimation, which uses a sequential estimation of the sample standard deviation and calculates $CI_{CLT}$ (7) at every new sample. Our solution simply approximates the sample standard deviation with $\frac{1}{n^2}$ at the initial stages of the computation if $\hat{p}$ is equal to 0 (or 1) and propagates it through the computation until the necessary number of samples to construct the interval are obtained. We show the advantages of this approach in Section 5.

**Wilson interval.**    It was introduced by Wilson in 1927 in his fundamental work [9] and uses the inversion of the CLT interval. Additionally, it involves a modified center by quantile formula mean value. The interval has the following form:

$$CI_W = \left(\frac{n_s + \frac{C_a^2}{2}}{n + C_a} - \frac{C_a\sqrt{n}}{n + C_a^2}\sqrt{\hat{p}\hat{q} + \frac{C_a^2}{4n}}; \frac{n_s + \frac{C_a^2}{2}}{n + C_a} + \frac{C_a\sqrt{n}}{n + C_a^2}\sqrt{\hat{p}\hat{q} + \frac{C_a^2}{4n}}\right) \tag{8}$$

This interval has some obvious advantages - it can not exceed probability boundaries, and it can be easily calculated even if $\hat{p}$ is 0 or 1. At the same time, $CI_W$ has downward spikes when $\hat{p}$ is close to 0 and 1, because it is formed by an inverted CLT approximation.

**Agresti-Coull interval.**    This method was introduced by Agresti and Coull in 1998 [1]. One of the most interesting features of this CI is that it makes a crucial assumption about $n_s$ and $n_f$. This interval formally adds two successes and two failures to the obtained values in case of 95% confidence level and then uses the CLT method. The interval can be constructed as follows:

$$CI_{AC} = \left(\tilde{X} - \frac{1}{n + C_a^2}\left(n_s + \frac{1}{2}C_a^2\right); \tilde{X} + \frac{1}{n + C_a^2}\left(n_s + \frac{1}{2}C_a^2\right)\right) \tag{9}$$

Additionally, this interval can be modified by using the center of the Wilson interval (8) in place of $\hat{p}$:

$$CI_{AC_W} = \left( \frac{n_s + \frac{C_a^2}{2}}{n + C_a} - C_a\sqrt{\hat{p}\hat{q}(n + C_a^2)}; \left(\frac{n_s + \frac{C_a^2}{2}}{n + C_a} - C_a\sqrt{\hat{p}\hat{q}(n + C_a^2)}\right) \right). \tag{10}$$

**Logit interval.** The Logit interval is based on a transformation of the standard interval [8]. It uses the empirical logit transformation: $\lambda = ln(\frac{\hat{p}}{1-\hat{p}}) = ln(\frac{n_s}{n-n_s})$. The variance of $\lambda$ is: $\widehat{Var}(\lambda) = \frac{n}{n_s(n-n_s)}$ and the Logit interval can be estimated as:

$$CI_L = \left( \frac{e^{\lambda_L}}{1 + e^{\lambda_L}}, \frac{e^{\lambda_U}}{1 + e^{\lambda_U}} \right) \tag{11}$$

where the lower bound transformation is $\lambda_L = \lambda - C_a\sqrt{\widehat{Var}(\lambda)}$ and the upper bound transformation is $\lambda_U = \lambda + C_a\sqrt{\widehat{Var}(\lambda)}$.

**Anscombe interval.** This interval was proposed by Anscombe in 1956 [4] and is based on the Logit interval (11). The key difference is in $\lambda$ and $\widehat{Var}(\lambda)$ estimation, where $\lambda$ is defined as $\lambda = ln(\frac{n_s + \frac{1}{2}}{n - n_s + \frac{1}{2}})$ and the variance is $\widehat{Var}(\lambda) = \frac{(n+1)(n+2)}{n(n_s+1)(n-n_s+1)}$. On this basis, the Anscombe interval $CI_{Anc}$ is estimated in the same way as Logit interval (11).

**Arcsine interval.** It uses a variance-stabilising transformation of $\hat{p}$. In 1948, Anscombe introduced an improvement [3] for achieving better variance stabilisation by replacing $\hat{p}$ to $p^\dagger = \frac{n_s + 3/8}{n + 3/4}$, obtaining

$$CI_{Arc} = \left( \sin(\arcsin(\sqrt{p^\dagger}) - \frac{C_a}{2\sqrt{n}})^2, \sin(\arcsin(\sqrt{p^\dagger}) + \frac{C_a}{2\sqrt{n}})^2 \right). \tag{12}$$

### 4.2 Alternative Intervals Based on the Beta-Function

**Bayesian interval.** This method is based on the assumption that the (unknown) probability $p$ to estimate is a random quantity [21]. The Bayesian interval is also called "credible", because it computes the posterior distribution of the unknown quantity by using its prior distribution and Bayes theorem. The prior distribution can be constructed by means of the *Beta* distribution, which is widely used for computing inferences on $p$. If $p$ has a prior distribution $Beta(\alpha, \beta)$ then $p$ has posterior distribution $Beta(n_s + \alpha, n - n_s + \beta)$. We can construct a Bayesian equal-tailed interval by the formula:

$$CI_B = \left( Beta^{-1}(\frac{a}{2}, n_s + \alpha, n - n_s + \beta), Beta^{-1}(1 - \frac{a}{2}, n_s + \alpha, n - n_s + \beta) \right) \tag{13}$$

where, $Beta^{-1}(a, \alpha, \beta)$ is the inverse of the cumulative distribution function of $Beta(\alpha, \beta)$.

**Jeffreys interval.** The Jeffreys interval is a Bayesian interval and uses the Jeffreys prior [15], which involves a non-informative prior invariant transformation given by the *Beta* distribution with parameters $(\frac{1}{2}, \frac{1}{2})$. The probability density function of the *Beta* distribution is $f(x; \alpha, \beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$, where $0 \leq x \leq 1$, $\alpha, \beta > 0$ and $B$ is beta function. We can form Jeffreys equal-tailed interval by (13) with parameters $(\frac{1}{2}, \frac{1}{2})$.

**Clopper-Pearson interval.**    This method was introduced by Clopper and Pearson in 1934 [7] and is based on the inversion of binomial test, rather than on approximations. The Clopper-Pearson interval is:

$$CI_{CP} = \left( Beta^{-1}(\frac{a}{2}, n_s, n - n_s + 1), Beta^{-1}(1 - \frac{a}{2}, n_s + 1, n - n_s) \right) \ . \tag{14}$$

The $CI_{CP}$ interval states that the computed coverage probability is always above or equal to the $1 - a$ confidence level. In practice, it can be achieved in cases when $n$ is large enough, while in general, the actual coverage can exceed $1 - a$. We can conclude from the equation (14) that due to the absence of the $\alpha$ and $\beta$ parameters, the appropriate result can be achieved only by increasing number of "trials".

## 5   Results

We apply CI estimation methods, based on the standard interval with the RQMC technique and Bayesian CI estimation method with the MC technique. All our results are the average of 10 runs: in the RQMC case 10 sequences were obtained by changing the pseudo-random points $\varepsilon$ of the equation $\Gamma = (\mathfrak{X} + \varepsilon) mod\, 1$, while the Sobol sequence points $\mathfrak{X}$ remain the same; in the MC case we used the same 10 pseudo-random points sequences, which were used for RQMC calculation.

### 5.1   Border Probability Cases

The true probability values, which are shown in Section 5.1 and Section 5.2 were obtained via pseudo-random number generation that produces boolean values according to a Bernoulli distribution.

**Intervals based on CLT and Bayesian interval.**    The comparison of the different CIs estimation techniques for extreme probability cases (near 0 bound) with accuracy $\varepsilon = 5 \times 10^{-3}$, which is presented in Figure 2, shows that all intervals except the Arcsin interval (12) (see plot $c = 0.99$ of Figure 2 for probability=0.001) contain the true probability value within their bounds. The Bayesian method tends to overestimate the true probability values according to their increase while $CI_{CLT}$ tends to underestimate them. Also, it is interesting to note that the most accurate center value is returned by the Agresti-Coull interval. The reason why $CI_{CLT}$ tends to include the true probability value near the upper bound of the interval is directly related to the number of samples. As it is shown in Figure 2 for true probability values 0.007 - 0.01, the $CI_{CLT}$ center is moving up evenly to the true probability value with the increase of the confidence value. It echoes the number of samples grow for obtaining the necessary confidence level. For the other true probability values (0.001-0.006) although this drift is saved, it can not be seen from the Figure, because of the small difference in the number of samples for all confidence levels, which causes the CI center to move wave-like.

The results in Figure 2 also demonstrate that the CIs based on the standard interval can have interval size smaller than its nominal value even for "large" sample sizes. It can be seen that every confidence level from 0.99 to 0.99999 displays further instances of the inadequacy of the CIs size. Also, Figure 2 shows that in spite of the "large" sample size (according to the confidence level), the size of the $CI_B$, $CI_{CLT}$ and $CI_{AC_w}$ intervals decreases significantly as $p$ moves toward 0. Also, CIs based on the standard interval have interval size changes because of two reasons: absence of a posteriori estimate and skewness of the underlying binomial distribution.

In Figure 3 we plot the number of samples that different CI estimation techniques used to return intervals with accuracy $\varepsilon = 5 \times 10^{-3}$ for different confidence levels. It can be clearly seen from the plots that with increasing of the confidence all CIs based on the standard interval outperform the Bayesian

CI. The plot with $c = 0.99999$ in Figure 3 displays that the best techniques in the number of samples from the best to the worst are: $CI_{CLT}$, $Qint$, $CI_{Arc}$, $CI_W$, $CI_L$, $CI_{Ans}$, $CI_{AC_W}$ and $CI_B$. The $CI_L$ and $CI_{Anc}$ techniques always show almost the same results near the bounds, because of the modification of the $CI_L$. Initially, $CI_L$ is not able to deal with probability values near the bounds according to its $\lambda$ formula (see Section 4.1). It has been modified to use the Anscombe estimation formula in cases when $\hat{p} = 0$ or $\hat{p} = 1$. It is also important to note that the difference in number of samples between $CI_{CLT}$, $CI_{Arc}$ and $CI_B$ for extreme probability cases is significant. For example in the plot with $c = 0.9999$ of Figure 3 the number of samples used to obtain interval for $p = 0.005$ equals to 1,078 for $CI_{CLT}$, 2,662 for the $CI_{Arc}$ and 4,440 for $CI_B$.

This trend is not preserved with the increase of the probability value from 0 to 0.5 and with the decrease from 1 to 0.5, respectively. Figure 4 shows that the difference in number of samples between all CIs (except $CI_{Arc}$) is almost undetectable. At the same time, $CI_{Arc}$ shows very "bad" results in comparison with the others, as opposed to its results for probability values at the extremes.

Summarizing, for probability values near the bounds (0 or 1) the modified CLT method achieves better results in number of samples in comparison with the others (see Figure 3). For probability values away from the bounds, CLT, Wilson, Agresti–Coull, Logit and Anscombe methods are all very similar (see Figure 4), and so for such probabilities we come to the conclusion that the CLT interval should be recommended, due to its simplest form. Meanwhile for smaller sample sizes, the $CI_{CLT}$ is strongly preferable to the others and so might be the choice where sampling cost is paramount.

**Qint method results.** In Figure 2 and Figure 3 we also plotted the results of the recently developed Qint algorithm. In our research we used Qint with $n = k \times 2^s$, where $k = 2$. These parameters were used to form $n$ points of the Sobol sequence $x_i$ with numbers $i \in I_{k,s} = \{1, 2, ..., k \times 2^s\}$. These parameters were chosen on the basis of the original study of the Qint method [5] as the most universal and reliable. As it was described earlier, Qint uses a cubature randomization method and provides the integral estimation variance (3). This formula is used to obtain a CI by calculating the standard interval (7) with our modification.

In Figure 2 we display the Qint intervals distribution for border probability values. We can see from the plots that Qint CI always contains the true probability value. At the same time for all confidence levels from 0.99 to 0.99999 and for true probability values 0.006-0.01 Qint shows better centration than $CI_B$ and $CI_{CLT}$. The greatest differences between the Qint CLT center result and the true probability values are: 0.00245 for $c = 0.99$ (p=0.004), 0.00191 for $c = 0.999$ (p=0.004), 0.00168 for $c = 0.9999$ (p=0.003), 0.00141 for $c = 0.99999$ (p=0.004), while for example this difference for $CI_B$ reaches 0.00518 for $c = 0.99$ (p=0.007), 0.00235 for $c = 0.999$ (p=0.008), 0.00181 for $c = 0.9999$ (p=0.008), 0.00143 for $c = 0.99999$ (p=0.008).

We can see in Figure 3 that, as it was expected, Qint uses fewer samples than other CIs but $CI_{CLT}$. Our modification allows the Qint algorithm to return intervals even if $n_s = 0$, which significantly decreases the final sample size for 10 runs. With the increase in $n$, which leads to further sampling and better reflects the behavior of the underlying random process, the effectiveness of the method decreases, and the benefit no longer seem so significant.

The fact that with the chosen parameters Qint can not outperform our modified $CI_{CLT}$ leads us to the conclusion that our use of the standard deviation formula with $\frac{1}{n^2}$ lower bound is a rather effective and simple solution. However, the deep range of the possible parameters variation as well as novelty of the Qint algorithm lead us to believe that further research towards their comparison is needed.
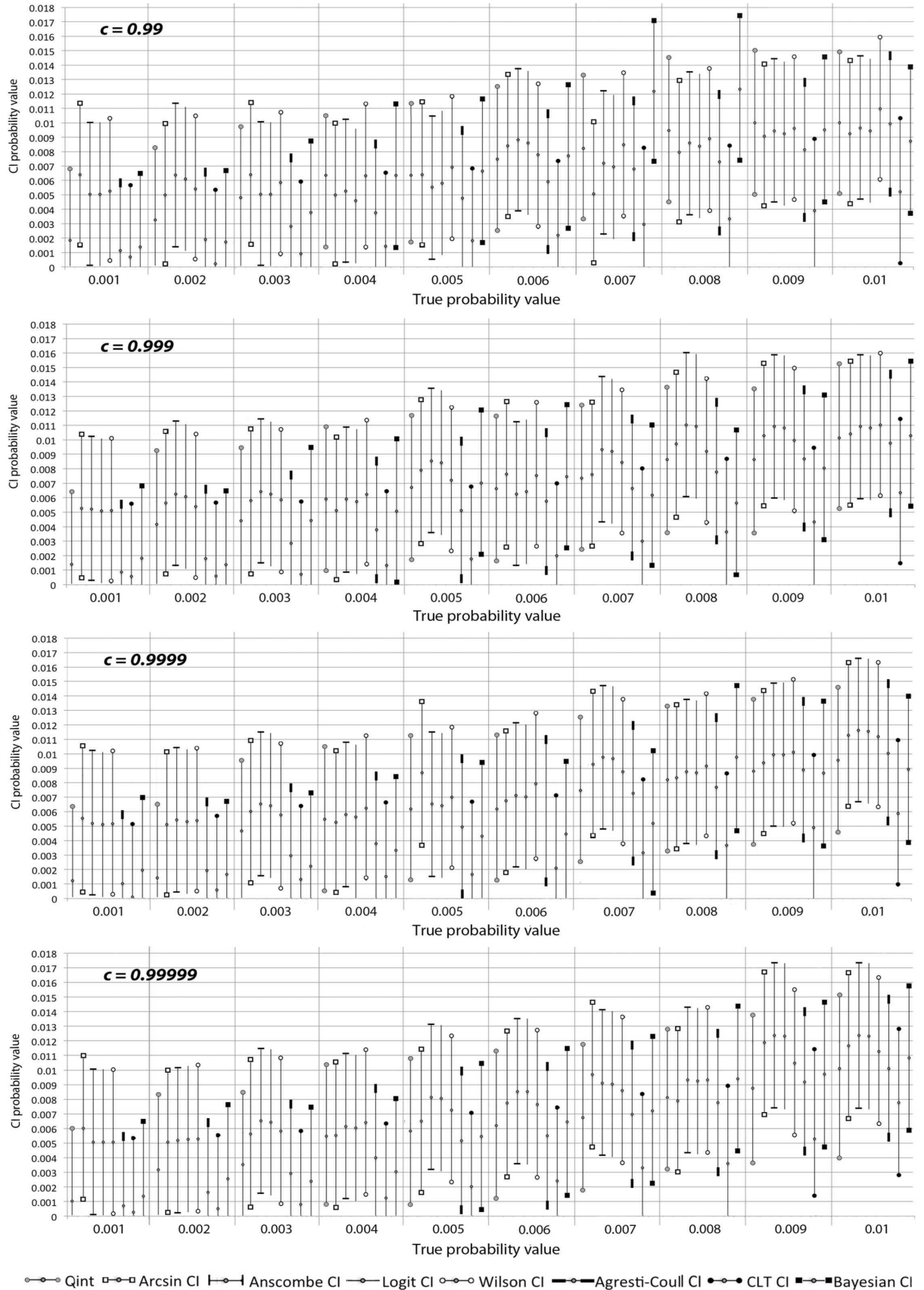
Figure 2: Comparison of confidence interval distribution for probability values near 0, interval size equal to $10^{-2}$ and $c$ - confidence level.
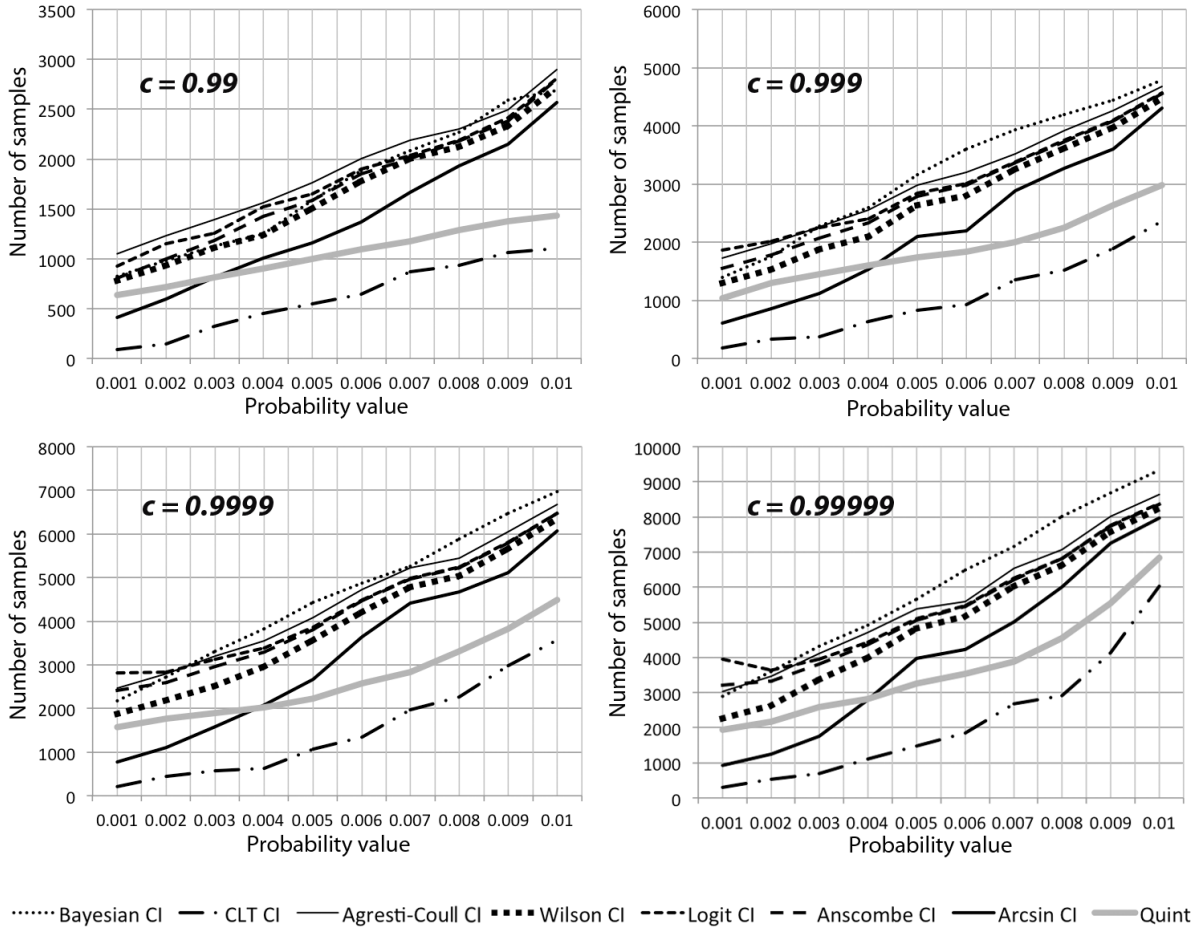
Figure 3: Comparison of sample size for probability values near 0, interval size equal to $10^{-2}$ and $c$ - confidence level.
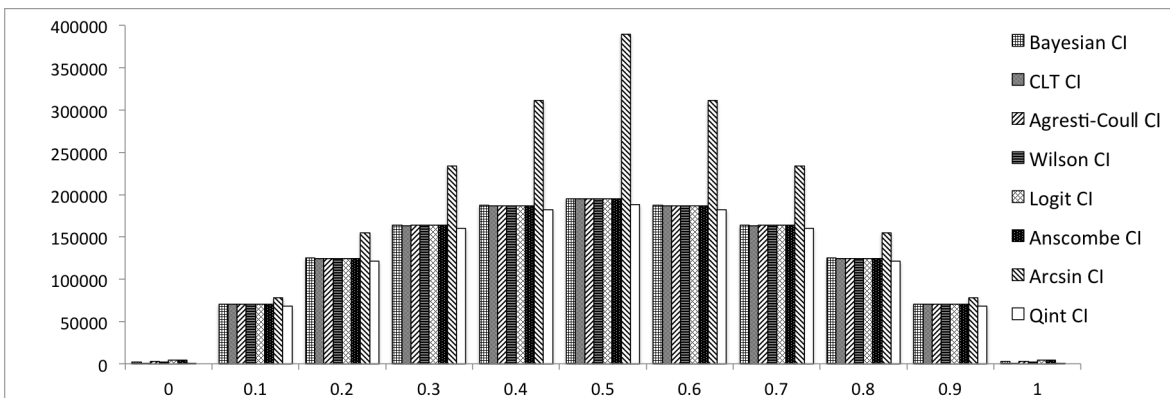


Figure 4: Comparison of sample size for probability values from 0 to 1, interval size equal to $10^{-2}$ and confidence level equal to 0.99999.

## 5.2   MC and QMC Error Comparison

Another key difference between the Bayesian CI and the CIs based on CLT is the use of MC and QMC techniques for interval calculation. As it was described in Section 3.2, the QMC advantage in the error size holds for all tested models. In the cases where the true error rate could not be detected due to the probability value extremely close to 0 ("Bad" model type min and Collision (Basic) model type min), we have that the MC absolute error line equals the true probability value, because $n_s = 0$ was obtained. The chaotic coverage properties of the MC method are far more persistent than they are appreciated. The chaotic behavior does not disappear even when $n$ is quite large and the true probability $p$ is not near the boundaries. For instance, in Figure 5 it is visible that even when $n$ is quite large (*i.e.*, tends to 10,000 samples) the actual absolute error value of the MC method reaches $5 \times 10^{-3}$. Hence we can conclude that CIs estimation techniques based on MC are misleading and defective in several respects and should not be trusted [6].

A notable phenomenon, which was noticed for the MC and QMC probability calculation is that the actual coverage probability contains non-negligible oscillations as both $p$ and $n$ vary. There exist some "unlucky" pairs $(p,n)$ such that the corresponding absolute error is much greater than the results for smaller $n$. The phenomenon of oscillation is both in $n$, for fixed $p$, and in $p$, for fixed $n$. Furthermore, drastic changes in coverage occur in nearby $p$ for fixed $n$ and in nearby $n$ for fixed $p$ [6]. We can see it on the simple example in Figure 6.

## 5.3   Tested Models Results

The results in Section 5.3 were obtained via the ProbReach tool [17] for computing bounded reachability in stochastic parametric hybrid systems and the dReal 3 solver [12] for analyzing (standard) bounded reachability question. Five different hybrid models were chosen for our experiments can be found at: https://github.com/dreal/probreach/tree/master/model.

**Intervals based on CLT and Bayesian interval.**   Based on our model set, we provide in Table 1 a comparison of the CIs described in the Section 4, obtained via ProbReach with precision $\delta = 10^{-3}$, interval size $10^{-2}$ and true probability value $P$ is either analytically computed single probability values or formally computed absolute (non-statistical) intervals. These parameters were chosen according to previous work [17] as a fine trade-off between the precision of the results and the CPU time. Each model was verified separately with different confidence level from 0.99 to 0.99999. The lowest confidence level is often used in similar work, while the highest can provide reasonable results for real-world complex models.

As it can be seen in Table 1, all the intervals for the various techniques overlap. The modified $CI_{CLT}$ approach shows very similar results to the $CI_B$, which can be regarded as a successful implementation. The key difference in the interval sizes can be found in the results of the "Bad" model Type min and the Collision (Basic) model Type min. From the results we can conclude that the true probability value is very close to 0. This allows the Bayesian, CLT and Agresti-Coull methods to form intervals, which in reality are half of the proposed interval size $10^{-2}$, while the other techniques return "fully" sized intervals. That happens because $CI_B$ is using posterior distribution to form the interval. At the same time, the $CI_{CLT}$ and $CI_{AC_W}$ calculations of the mean value return the result, which is quite close to zero. Thus, the next step of the interval bounds computation cuts the negative part of the interval. This trend is holding for all probability values within $[0, 0.001]$.

Table 1 also shows that with the increase of the confidence level the interval's precision is growing,

**Confidence level c=0.99**

| Model | Type | P | $CI_B$ | $CI_{CLT}$ | $CI_{AC_W}$ | $CI_W$ | $CI_L$ | $CI_{Ans}$ | $CI_{Arc}$ | $Qint$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Good | max | 0.1 | [0.09671, 0.10671] | [0.09564, 0.10564] | [0.09632, 0.10632] | [0.09574, 0.10574] | [0.09575, 0.10575] | [0.09577, 0.10577] | [0.09559, 0.10559] | [0.09147, 0.10147] |
| | min | 0.1 | [0.09529, 0.10529] | [0.0956, 0.1056] | [0.09666, 0.10666] | [0.09679, 0.10679] | [0.09678, 0.10678] | [0.0968, 0.1068] | [0.09639, 0.10639] | [0.09164, 0.10164] |
| Bad | max2 | 0.95001 | [0.94416, 0.95416] | [0.94495, 0.95493] | [0.94422, 0.95422] | [0.94397, 0.95397] | [0.94396, 0.95396] | [0.94392, 0.95392] | [0.94735, 0.95735] | [0.94459, 0.95459] |
| | max2 | 0.88747 | [0.8825, 0.8925] | [0.88028, 0.89028] | [0.88031, 0.88031] | [0.88019, 0.89019] | [0.8803, 0.8902] | [0.88019, 0.89019] | [0.88325, 0.89325] | [0.88136, 0.89136] |
| | min | $4\times10^{-7}$ | [0, 0.00525] | [0, 0.005] | [0, 0.00483] | [0, 0.00955] | [0.00005, 0.00959] | [0.00005, 0.00959] | [0.00131, 0.00959] | [0,0.005] |
| Deceleration | max | [0.08404, 0.08881] | [0.08471, 0.09471] | [0.08802, 0.09802] | [0.08817, 0.09817] | [0.08685, 0.09685] | [0.08614, 0.09614] | [0.0863, 0.0963] | [0.08963, 0.09932] | [0.08852, 0.09852] |
| | min | [0.04085, 0.04275] | [0.03835, 0.04861] | [0.03861, 0.04861] | [0.03854, 0.04854] | [0.03884, 0.04884] | [0.03886, 0.04886] | [0.0389, 0.0489] | [0.03873, 0.04873] | [0.03337, 0.04337] |
| Collision (Basic) | max | [0.96567, 0.97254] | [0.96371, 0.97381] | [0.96873, 0.97873] | [0.9684, 0.9784] | [0.96851, 0.97851] | [0.96875, 0.97875] | [0.96853, 0.97853] | [0.96851, 0.97851] | [0.96301, 0.97301] |
| | min | [0, 0.00201] | [0, 0.00525] | [0, 0.005] | [0, 0.00483] | [0, 0.00955] | [0.00005, 0.00959] | [0.00005, 0.00959] | [0.00131, 0.00959] | [0,0.005] |
| Collision (Extended) | max | [0.35751, 0.49961] | [0.42267, 0.43675] | [0.42418, 0.4342] | [0.42187, 0.43187] | [0.42345, 0.43345] | [0.42463, 0.43463] | [0.42457, 0.43457] | [0.42385, 0.43385] | [0.42342, 0.43342] |
| | min | [0.04296, 0.06311] | [0.0482, 0.0582] | [0.04772, 0.05772] | [0.04785, 0.05785] | [0.04823, 0.05823] | [0.04812, 0.05812] | [0.0481, 0.0581] | [0.04757, 0.05772] | [0.04618, 0.05618] |
| Collision (Advanced) | max | [0.14807, 0.31121] | [0.2072, 0.2172] | [0.20873, 0.21872] | [0.21872, 0.2185] | [0.20854, 0.21854] | [0.20854, 0.21854] | [0.20855, 0.21855] | [0.20111, 0.21111] | [0.20167, 0.21166] |
| | min | [0.02471, 0.05191] | [0.02631, 0.03631] | [0.03045, 0.04045] | [0.03016, 0.04016] | [0.03001, 0.04] | [0.03001, 0.04] | [0.03016, 0.04016] | [0.03164, 0.04164] | [0.0304, 0.0404] |
| Anesthesia | n/a | [0.00916, 0.04222] | [0.01361, 0.02361] | [0.01339, 0.02332] | [0.01374, 0.02374] | [0.01373, 0.02373] | [0.01318, 0.02318] | [0.01311, 0.02311] | [0.01592, 0.02592] | [0.01815, 0.02815] |

**Confidence level c=0.99999**

| Model | Type | P | $CI_B$ | $CI_{CLT}$ | $CI_{AC_W}$ | $CI_W$ | $CI_L$ | $CI_{Ans}$ | $CI_{Arc}$ | $Qint$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Good | max | 0.1 | [0.09499, 0.10499] | [0.09378, 0.10378] | [0.09386, 0.10386] | [0.09389, 0.10389] | [0.09391, 0.10391] | [0.09392, 0.10392] | [0.09405, 0.10405] | [0.09512, 0.10512] |
| | min | 0.1 | [0.09419, 0.10419] | [0.09667, 0.10667] | [0.09668, 0.10668] | [0.09677, 0.10677] | [0.09671, 0.10671] | [0.09679, 0.10679] | [0.09675, 0.10675] | [0.09525, 0.10525] |
| Bad | max2 | 0.95001 | [0.94525, 0.95525] | [0.94579, 0.95579] | [0.94564, 0.95564] | [0.94548, 0.95548] | [0.94545, 0.95545] | [0.94543, 0.95543] | [0.94735, 0.95735] | [0.94543, 0.95543] |
| | max2 | 0.88747 | [0.88215, 0.89215] | [0.88055, 0.89055] | [0.88057, 0.89057] | [0.88046, 0.89046] | [0.88046, 0.89046] | [0.88046, 0.89046] | [0.88325, 0.89325] | [0.88052, 0.89052] |
| | min | $4\times10^{-7}$ | [0, 0.00517] | [0, 0.00319] | [0, 0.00494] | [0, 0.00984] | [0, 0.00992] | [0, 0.00992] | [0.00445, 0.0139] | [0,0.005] |
| Deceleration | max | [0.08404, 0.08881] | [0.08613, 0.09613] | [0.08624, 0.09624] | [0.08312, 0.09312] | [0.08725, 0.09725] | [0.08725, 0.09725] | [0.08726, 0.09726] | [0.08746, 0.09746] | [0.08737, 0.09735] |
| | min | [0.04085, 0.04275] | [0.03514, 0.04514] | [0.03919, 0.04919] | [0.03918, 0.04918] | [0.03942, 0.04942] | [0.03943, 0.04943] | [0.03944, 0.04944] | [0.039, 0.049] | [0.03377, 0.04377] |
| Collision (Basic) | max | [0.96567, 0.97254] | [0.96359, 0.97359] | [0.96241, 0.97241] | [0.96767, 0.9767] | [0.96892, 0.96892] | [0.96689, 0.97589] | [0.96683, 0.97583] | [0.96863, 0.97863] | [0.96462, 0.97462] |
| | min | [0, 0.00201] | [0, 0.00517] | [0, 0.00319] | [0, 0.00494] | [0, 0.00984] | [0, 0.00992] | [0, 0.00992] | [0.00445, 0.0139] | [0,0.005] |
| Collision (Extended) | max | [0.35751, 0.49961] | [0.42651, 0.43652] | [0.42719, 0.43724] | [0.42757, 0.43757] | [0.42656, 0.43656] | [0.41774, 0.42774] | [0.41779, 0.42779] | [0.42745, 0.43745] | [0.42875, 0.43875] |
| | min | [0.04296, 0.06311] | [0.04979, 0.05979] | [0.04766, 0.05766] | [0.04764, 0.05764] | [0.04748, 0.05748] | [0.04745, 0.05745] | [0.04776, 0.05776] | [0.05776, 0.05673] | [0.04576, 0.05576] |
| Collision (Advanced) | max | [0.14807, 0.31121] | [0.20515, 0.21519] | [0.20558, 0.21563] | [0.20533, 0.21533] | [0.20531, 0.21531] | [0.20547, 0.21547] | [0.20547, 0.21547] | [0.20385, 0.21385] | [0.20453, 0.21453] |
| | min | [0.02471, 0.05191] | [0.03011, 0.04015] | [0.02902, 0.03902] | [0.02954, 0.03945] | [0.03956, 0.04956] | [0.03861, 0.04861] | [0.03887, 0.04887] | [0.0363, 0.04363] | [0.03031, 0.04031] |
| Anesthesia | n/a | [0.00916, 0.04222] | [0.01284, 0.02284] | [0.01513, 0.02511] | [0.01623, 0.02623] | [0.01545, 0.02545] | [0.01557, 0.02557] | [0.01562, 0.02562] | [0.01385, 0.02385] | [0.01852, 0.02852] |

Table 1: Results for CI computation obtained via ProbReach, with solver precision $\delta=10^{-3}$ and interval size equal to $10^{-2}$, **Type** - extremum type and **P** - true probability value, where single point values were analytically computed and interval values are numerically guaranteed.

| Model | Type | $c$ | $CI_B$ | $CI_{CLT}$ | $CI_{AC_W}$ | $CI_W$ | $CI_L$ | $CI_{Ans}$ | $CI_{Arc}$ | $Qint$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Good | max | 0.99 | 24252 | 24025 | 24038 | 24027 | 24035 | 24034 | 26681 | 23136 |
| | min | 0,99 | 23451 | 23248 | 24256 | 24250 | 24253 | 24252 | 26894 | 23245 |
| Bad | max | 0.99 | 13118 | 12670 | 12841 | 12817 | 12833 | 12832 | 23006 | 11726 |
| | max2 | 0.99 | 27498 | 26954 | 26960 | 26955 | 26958 | 26958 | 40442 | 25734 |
| | min | 0.99 | 2590 | 96 | 961 | 688 | 680 | 680 | 347 | 94 |
| Deceleration | max | 0.99 | 22842 | 22393 | 22673 | 22517 | 22628 | 22623 | 24365 | 20318 |
| | min | 0.99 | 11224 | 11073 | 11114 | 11086 | 11104 | 11104 | 11570 | 9798 |
| Collision (Basic) | max | 0.99 | 9581 | 9318 | 9653 | 9463 | 9386 | 9381 | 10643 | 8222 |
| | min | 0.99 | 2590 | 96 | 961 | 688 | 680 | 680 | 347 | 94 |
| Collision (Extended) | max | 0.99 | 65109 | 64804 | 64854 | 64841 | 64932 | 64930 | 104637 | 62485 |
| | min | 0.99 | 13624 | 13257 | 13486 | 13375 | 13326 | 13320 | 14737 | 12869 |
| Collision (Advanced) | max | 0.99 | 44370 | 43602 | 43645 | 43640 | 43644 | 43643 | 51734 | 43524 |
| | min | 0.99 | 9500 | 9081 | 9094 | 9085 | 9090 | 9089 | 9282 | 9080 |
| Anesthesia | n/a | 0.99 | 5801 | 4847 | 5024 | 4952 | 4928 | 4919 | 5522 | 4804 |
| Good | max | 0.99999 | 70422 | 69484 | 69582 | 69496 | 69530 | 69529 | 77262 | 68456 |
| | min | 0,99999 | 71898 | 71286 | 71339 | 71293 | 71321 | 71321 | 79369 | 68994 |
| Bad | max | 0.99999 | 37388 | 36518 | 36771 | 36629 | 36687 | 36868 | 60006 | 36164 |
| | max2 | 0.99999 | 79306 | 79097 | 79125 | 79101 | 79118 | 79118 | 96442 | 77892 |
| | min | 0.99999 | 5797 | 124 | 2766 | 1963 | 4136 | 4136 | 572 | 94 |
| Deceleration | max | 0.99999 | 65248 | 65233 | 65330 | 65299 | 65320 | 65319 | 72114 | 59882 |
| | min | 0.99999 | 33147 | 32969 | 33133 | 33018 | 33060 | 33060 | 34231 | 29096 |
| Collision (Basic) | max | 0.99999 | 25279 | 24711 | 24834 | 24789 | 24934 | 24933 | 26045 | 23016 |
| | min | 0.99999 | 5797 | 124 | 2766 | 1963 | 4136 | 4136 | 572 | 94 |
| Collision (Extended) | max | 0.99999 | 191466 | 190776 | 191253 | 190894 | 191485 | 191472 | 376294 | 185456 |
| | min | 0.99999 | 41153 | 38942 | 39745 | 39473 | 39537 | 39541 | 47923 | 37608 |
| Collision (Advanced) | max | 0.99999 | 131517 | 129746 | 131185 | 129845 | 129934 | 129933 | 183405 | 127486 |
| | min | 0.99999 | 27305 | 25657 | 25835 | 25736 | 25792 | 25791 | 29362 | 24569 |
| Anesthesia | n/a | 0.99999 | 16197 | 15453 | 15834 | 15634 | 15734 | 15733 | 17845 | 15314 |

Table 2: Samples size comparison for confidence interval computation obtained via ProbReach, with solver $\delta$ precision equal to $10^{-3}$ and interval size equal to $10^{-2}$, **Type** - extremum type and $c$ - confidence level.
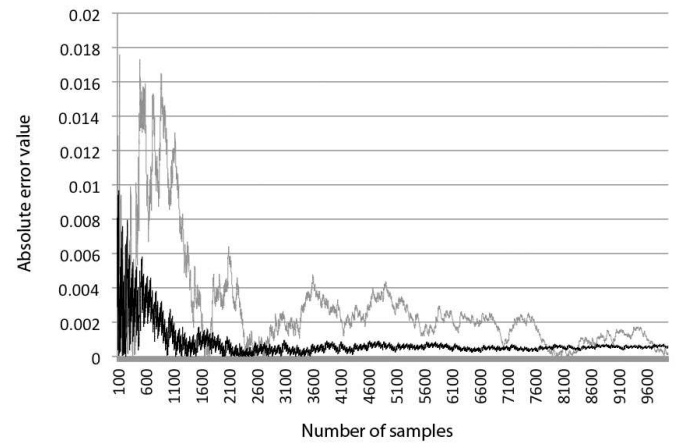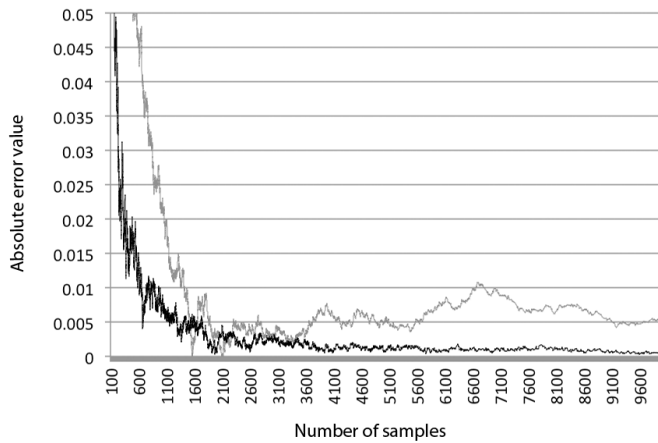


Figure 5: MC (grey line) and QMC (black line) absolute error with respect to the number of samples. Model: Collision advanced, type - max.



Figure 6: MC (grey line) and QMC (black line) absolute error with respect to the number of samples. Model: Bad, type - max2.

which in turn is directly related to the usage of the inverse cumulative distribution function for normal random variable with given confidence level in formulas for $CI_{CLT}$ (7), $CI_W$ (8), $CI_{AC_W}$ (10) and $CI_{Arc}$ (12). It also results in the increase the sample size $n$ for $CI_L$ and $CI_{Anc}$.

The comparison of the obtained intervals (see Table 1) with the true probability value or interval $\boldsymbol{P}$ shows that all CIs contain the single probability values, but $CI_{Acr}$ (see "Bad" type min model of Table 1), and all CIs overlap with the true probability intervals. We can also note that the true probability intervals of the Collision Extended, Collision Advanced, and Anesthesia models contain all confidence intervals for all confidence levels. The reasons why Collision Basic and Deceleration models' true probability intervals do not contain CIs are their size, which is very small ($< 0.01$).

Table 2 provides very interesting results with respect to the number of samples, which were used to find CIs obtained via ProbReach with solver precision $\delta = 10^{-3}$ and interval size $10^{-2}$. The number of samples varies for different models and types. As it was noted earlier for Figure 4, the number of samples needed for the computation grows from the bounds to the center of the [0,1] interval. The presented models show different behaviour and probability results. The most important outcome is that all CIs (except $CI_{Arc}$) show better result in number of points with respect to $CI_B$. The best result was shown by $CI_{CLT}$. It shows that the proposed CLT modification can provide reasonable results for RQMC calculation in comparison with the well-established Bayesian MC integral calculation.

**Qint method results.** A comparison of the Qint method's confidence intervals is presented in Table 1. All Qint intervals also contain single probability values and overlap with true probability intervals. The original Qint algorithm is not able to provide results for "Bad" type min and Collision Basic type min models, because for very small probability values like $4 \times 10^{-7}$ and [0, 0.00201] it could not detect $n_s > 0$ for the chosen confidence levels and interval size. Due to this reason the original Qint algorithm was changed by modifying the CLT method described in Section 4.1. From the results we see that that the Qint algorithm shows great potential, which is connected with the very fast convergence rate of the QMC method and with finding an appropriate partition (in terms of the parameters $k, s$).

Table 2 allows us to compare Qint's sample sizes with those of other CIs. However, $CI_{CLT}$ had an advantage in the number of samples for small probability values near the border (see Figure 3), where we can clearly see that for bigger true probability values, which is presented in the tested models except "Bad" type min and Collision Basic type min, this trend is not preserved. On the contrary, Qint uses fewer number of samples than other CIs and $CI_{CLT}$ in particular (see also Figure 4). For the tested models set with confidence $c$=0.99999, Qint used on average between 1,850 and 24,802 fewer samples than other CIs techniques.

# 6    Conclusion and Future Work

In this paper we provide comprehensive evaluation of CIs calculation techniques based on the Monte Carlo (MC) and Quasi-Monte Carlo (QMC) techniques. The experiments show that our modified CLT technique is usable in practice even for complex dynamics and for probabilities close to the bounds. The QMC-based calculation techniques we consider have excellent convergence and efficiency especially when the number of samples is small. Based on our analysis of CIs, we suggest that our results can be used as guidelines for probability estimation techniques. In the future we plan to provide specification for the CI calculation by the Qint technique for the chosen models and also extend the test model range. Possible variations of the Qint technique are also of interest. In particular, joining the CLT and Qint methods is the next challenge for further research.

# References

[1] Alan Agresti & Brent A. Coull (1998): *Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions*. The American Statistician 52(2), pp. 119–126.

[2] Rajeev Alur, Costas Courcoubetis, Thomas A. Henzinger & Pei-Hsin Ho (1992): *Hybrid Automata: An Algorithmic Approach to the Specification and Verification of Hybrid Systems*. In: Hybrid Systems, LNCS 736, pp. 209–229.

[3] Frank J. Anscombe (1948): *The Transformation of Poisson, Binomial and Negative-Binomial Data*. Biometrika 35(3/4), pp. 246–254.

[4] Frank J. Anscombe (1956): *On Estimating Binomial Response Relations*. Biometrika 43(3/4), pp. 461–464.

[5] Anton A. Antonov & Sergei M. Ermakov (2015): *Empirically estimating error of integration by quasi-Monte Carlo method*. Vestnik St. Petersburg University: Mathematics 47(1), pp. 1–8.

[6] Lawrence D. Brown, Tony T. Cai & Anirban Dasgupta (2001): *Interval Estimation for a Binomial Proportion*. Statistical Science 16(2), pp. 128–133.

[7] Cynthia J. Clopper & Egon S. Pearson (1934): *The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial*. Biometrika 26(4), pp. 404–413.

[8] Natalie Dean & Marcello Pagano (2015): *Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys*. Journal of Survey Statistics and Methodology 3(4), pp. 484–503.

[9] Wilson B. Edwin (1927): *Probable Inference, the Law of Succession, and Statistical Inference*. Journal of the American Statistical Association 22(158), pp. 209–212.

[10] Sergei M. Ermakov (1975): *The Monte Carlo Method and Related Questions*. Nauka: Moscow. In Russian.

[11] Sicun Gao, Jeremy Avigad & Edmund M. Clarke (2012): *Delta-Decidability over the Reals*. In: LICS, pp. 305–314.

[12] Sicun Gao, Soonho Kong & Edmund M. Clarke (2013): *dReal: An SMT Solver for Nonlinear Theories over the Reals*. In: CADE-24, LNCS 7898, pp. 208–214.

[13] Sicun Gao, Soonho Kong & Edmund M. Clarke (2013): *Satisfiability modulo ODEs*. In: FMCAD, pp. 105–112.

[14] Michael Gnewuch, Anand Srivastav & Carola Winzen (2009): *Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems*. J. Complexity 25(2), pp. 115–127.

[15] Kalpana K. Mahajan, Sangeeta Arora & Kamaljit Kaur (2015): *Bayesian estimation for Gini index and a poverty measure in case of pareto distribution using Jeffreys' prior*. MASA 10(1), pp. 63–72.

[16] Vivek Pradhan & Tathagata Banerjee (2008): *Confidence Interval of the Difference of Two Independent Binomial Proportions Using Weighted Profile Likelihood*. Communications in Statistics - Simulation and Computation 37(4), pp. 645–659.

[17] Fedor Shmarov & Paolo Zuliani (2015): *ProbReach: Verified Probabilistic $\delta$-Reachability for Stochastic Hybrid Systems*. In: HSCC, ACM, pp. 134–139.

[18] Fedor Shmarov & Paolo Zuliani (2016): *Probabilistic Hybrid Systems Verification via SMT and Monte Carlo Techniques*. In: HVC, LNCS 10028, pp. 152–168.

[19] Ilya M. Sobol' (1967): *On the distribution of points in a cube and the approximate evaluation of integrals*. USSR Computational Mathematics and Mathematical Physics 7(4), pp. 86 – 112.

[20] Joseph F. Traub, Grzegorz W. Wasilkowski & Henryk Woźniakowski (1988): *Information-based Complexity*. Academic Press.

[21] Paolo Zuliani, André Platzer & Edmund M. Clarke (2013): *Bayesian statistical model checking with application to Stateflow/Simulink verification*. Formal Methods in System Design 43(2), pp. 338–367.