Let's estimate all parameters as probabilities: Precise estimation using Chebyshev's inequality, Bernoulli distribution, and Monte Carlo simulations

Lubomír Štěpánek[†, ‡], Filip Habarta[†], Ivana Malá[†], Luboš Marek[†]
[†]Department of Statistics and Probability, [‡]Department of Mathematics
Faculty of Informatics and Statistics
Prague University of Economics and Business
W. Churchill's square 4, 130 67 Prague, Czech Republic
{lubomir.stepanek, filip.habarta, malai, marek}@vse.cz

Abstract—Regarding the parameter estimation task, besides the time effectiveness of the simulation, parameter estimates are required to be precise enough. Usually, the estimates are Monte Carlo-simulated using a prior estimated variability within a small sample. However, the problem with pre-estimated variability is that it can be estimated imprecisely or, even worse, underestimated, resulting in estimation bias. In this work, we address the abovementioned issue and suggest estimating all parameters as probabilities. Since the probability is not only finite but has its theoretical maximum as 1, using outcomes of Bernoulli and binomial distribution's upper-bounded variance and Chebyshev's inequality, the estimator's variability is theoretically upper-bounded within the Monte Carlo simulation and estimation process. It cannot be underestimated or estimated inaccurately; thus, its precision is ensured till a given decimal digit, with very high probability. If there is a known process that treats the parameter of interest in terms of probability, we can estimate how many iterations of the Monte Carlo simulation are needed to ensure parameter estimate on a given level of precision. Also, we analyze the asymptotic time complexity of the proposed estimation strategy and illustrate the approach on a short case study of π constant estimation.

I. Introduction

IN THIS work, we focus on the estimation of parameters that are of a non-probabilistic fashion, e.g., simulated estimates of claim amounts in actuaries [1] or simulated numbers of patients at risk of disease recurrence [2]. Typically, these parameters are hardly analytically derivable, thus estimated using Monte Carlo simulation and the following logic [3]. Firstly, within an initial Monte Carlo simulation, a number of iterations (100, 1000, and so) generating the parameter estimate is run, and the parameter estimates from individual iterations are averaged, and their standard deviation is calculated. Then, applying the central limit theorem, a confidence interval for the parameter is constructed, and the Monte Carlo simulation is repeated so many times that the interval is no wider than a given precision. A problem of the abovementioned approach is in the parameter's variability estimation within the initial Monte Carlo simulation. If the variability is underestimated, the confidence interval is falsely narrower than it should be, and the precision is, in fact, lower than expected. To overcome this issue, we rather refine the simulation logic – firstly, we find a function of the parameter equal to some probability, which is then simulated using Monte Carlo simulation. Since the probability has a theoretically-based upper bound, its variability is upper-bounded. Then, we use the properties of the Bernoulli distribution to estimate the largest possible variability of the parameter as a probability and Chebyshev's inequality to enumerate the number of iterations keeping the parameter estimate's precision. Due to Chebyshev's inequality, we do not need the assumption of the parameter estimate's normality, which makes the proposed approach more robust.

II. A traditional approach to Monte Carlo simulation and estimation of parameters of non-probabilistic fashion

Let us assume a parameter θ of a non-probabilistic fashion that can be estimated n times using point estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$. Then, calculation of the estimates' average and standard deviation, i.e.,

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i \qquad \text{and} \qquad \sigma_\theta = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\hat{\theta}_i - \bar{\theta})^2},$$

is feasible. To estimate parameter θ using Monte Carlo simulation on a given level of precision $1 - \varepsilon$, where $\varepsilon \gtrsim 0$, reached with probability $1 - \alpha$, one needs to know a number of iterations n of the simulation [3].

A. Principles of the traditional approach to Monte Carlo simulation and estimation of parameters of non-probabilistic fashion

Adopting the mathematical notation from the previous section, typical values of ε and α are, e.g., $\varepsilon = 0.001$ and $\alpha = 0.05$, respectively. A traditional approach to parameter θ estimation follows.

Thematic track: Computer Aspects of Numerical Algorithms

(i) Choose n_0 for initial Monte Carlo simulation that would pre-estimate parameter θ using individual estimates $\hat{\theta}_{0,1}, \hat{\theta}_{0,2}, \ldots, \hat{\theta}_{0,n_0}$. Typically, n_0 is chosen as $n_0 = 100$ or $n_0 = 1000$ or similar.

(ii) Calculate an average and a standard deviation of the pre-estimated parameter as

$$\bar{\theta}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} \hat{\theta}_{0,i} \quad \text{and} \quad \sigma_{0,\theta} = \sqrt{\frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (\hat{\theta}_{0,i} - \bar{\theta}_0)^2}.$$

(iii) Applying Ljapunov's central limit theorem [4], parameter θ should lie in an interval of

$$\left\langle \bar{\theta}_0 - u_{1-\alpha/2} \frac{\sigma_{0,\theta}}{\sqrt{n}}, \ \bar{\theta}_0 + u_{1-\alpha/2} \frac{\sigma_{0,\theta}}{\sqrt{n}} \right\rangle \quad (1)$$

in $(1 - \alpha)n$ cases of n total cases, thus, approximately with a probability $1 - \alpha$, where $u_{1-\alpha/2}$ is the $(1 - \alpha/2)$-th quantile of the standard normal distribution.

(iv) Number of iterations n of the main Monte Carlo simulation, outputting parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$, is chosen to keep precision $1 - \varepsilon$ with probability $1 - \alpha$, so the confidence interval's half length from formula (1) should be less than or equal to ε, thus

$$u_{1-\alpha/2} \frac{\sigma_{0,\theta}}{\sqrt{n}} \leq \varepsilon, \quad (2)$$

and, equivalently, the number of needed iterations is

$$n \geq \left(\frac{u_{1-\alpha/2} \cdot \sigma_{0,\theta}}{\varepsilon} \right)^2. \quad (3)$$

(v) Finally, parameter θ is estimated using $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i$, believed to keep precision $1 - \varepsilon$ with probability $1 - \alpha$.

B. Limitations of the traditional approach to Monte Carlo simulation and estimation of parameters of non-probabilistic fashion

Although the abovementioned approach works in general and is commonly applied, it can suffer from not meeting the asymptotic properties assumed by Ljapunov's central limit theorem when the confidence interval from formula (1) is constructed. This might happen particularly for low values of n or very high demands on precision, e.g., when $\varepsilon < 10^{-6}$. On a more practical note, inspecting formula (3), if the parameter's standard deviation σ_θ is underestimated by $\sigma_{0,\theta}$, i.e., when $\sigma_{0,\theta} < \sigma_\theta$, then also number n of iterations needed to keep imprecision $\leq \varepsilon$ is underestimated, which may result into imprecise, i.e., wrong (!) decimal digits staring the i-th digit behind (or before) the decimal point, where $i = \lfloor |\log_{10}(\varepsilon)| \rfloor$, if $\varepsilon < 1$ (or $\varepsilon > 1$, respectively).

C. The asymptotic time complexity of the traditional approach to Monte Carlo simulation and estimation of parameters of non-probabilistic fashion

Obviously, if one iteration of the Monte Carlo simulation takes τ units of time, then, since the simulation is repeated two times, firstly with n_0 iterations and secondly with $n \geq \left(\frac{u_{1-\alpha/2} \cdot \sigma_{0,\theta}}{\varepsilon} \right)^2$ iterations as comes from formula (4), the total asymptotic time complexity of the procedure, $\Theta(\dagger)$, is

$$\Theta(\dagger) = \Theta\left((n_0 + n)\tau \right) \geq$$
$$\geq \Theta\left(\left(n_0 + \left(\frac{u_{1-\alpha/2} \cdot \sigma_{0,\theta}}{\varepsilon} \right)^2 \right) \tau \right), \quad (4)$$

so, while $\Theta(\dagger)$ is linear in n_0 and n terms, it is quadratic in $\sigma_{0,\theta}$ and $\frac{1}{\varepsilon}$ terms.

III. A PROPOSED APPROACH TO MONTE CARLO SIMULATION AND ESTIMATION OF PARAMETERS OF (NON-)PROBABILISTIC FASHION

Let us suppose a parameter θ of non-probabilistic fashion. Besides the traditional approach for θ estimation as introduced above, we may assume a link function $f(\bullet)$ so that $f(\theta)$ has got a dimension of probability, so,

$$f(\theta) = P(\mathcal{T}), \quad (5)$$

where $P(\bullet)$ is a probability function as comes from σ-algebra, and $\mathcal{T}$ is a random event or a proposition consisting of random events. If occasionally θ would be apriori a probability, then the link function $f(\bullet)$ is an identity, i.e., $f(\theta) = \theta$, and the approach below still works. That is why we bound the prefix *non-* into brackets in the section title.

Thus, to estimate parameter θ of the (non-)probabilistic fashion, keeping precision $1 - \varepsilon$ with probability $1 - \alpha$, let us first assume a random variable X following Bernoulli distribution with an argument $P(\mathcal{T})$, i.e., $f(\theta)$ (probability of success). A sum of n independent Bernoulli trials follows the binomial distribution with arguments n (number of trials) and $f(\theta)$ (probability of success in each trial). After collecting n estimates $\hat{X}_i$ coming from the above mentioned Bernoulli distribution, we calculate $\frac{1}{n} \sum_{i=1}^{n} \hat{X}_i$ to estimate parameter $f(\theta)$. The number of trials n, i.e., a number of iterations of Monte Carlo simulation, is prior estimated using Chebyshev's inequality, also considering the terms of precision, $1 - \varepsilon$, and confidence probability, $1 - \alpha$.

A. Mathematical and statistical preliminaries of the proposed approach to Monte Carlo simulation and estimation

As we have seen, we need to revisit Bernoulli and binomial distribution and Chebyshev's inequality and their statistical properties. So let's start with Bernoulli and binomial distributions.

Definition 1 (Bernoulli distribution)**.** *A random variable X follows Bernoulli distribution with an argument $0 \leq p \leq 1$ (probability of success), if*

$$X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1 - p. \end{cases}$$

Formally, we write $X \sim Bernoulli(p)$.

Lemma 1 (Bernoulli distribution's expected value and variance). *Let random variable X follow Bernoulli distribution with an argument p. Then expected value of X is $\mathbb{E}(X) = p$ and variance of X is $\mathrm{var}(X) = p(1 - p)$.*

Proof. According to definition 1, it is $\mathbb{E}(X) = \sum_{i \in \{0,1\}} i \cdot P(X = i) = 0 \cdot (1-p) + 1 \cdot p = p$, and $\mathbb{E}\left(X^2\right) = \sum_{i \in \{0,1\}} i^2 \cdot P(X = i) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p$. Since routinely is $\mathrm{var}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2$, we get $\mathrm{var}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 = p - p^2 = p(1 - p)$. $\qquad\square$

Definition 2 (Binomial distribution). *A random variable X follows binomial distribution with arguments $n \in \mathbb{N}$ and $0 \leq p \leq 1$, if X is sum of n independent variables following Bernoulli distribution with an argument $0 \leq p \leq 1$ (i.i.d.). Formally, we write $X \sim binomial(n, p)$.*

Lemma 2 (Binomial distribution's expected value and variance). *Let random variable X follow binomial distribution with arguments n and p. Then expected value of X is $\mathbb{E}(X) = np$ and variance of X is $\mathrm{var}(X) = np(1 - p)$.*

Proof. According to definition 2, if $X \sim binomial(n, p)$, it is $X = Y_1 + Y_2 + \cdots + Y_n = \sum_{i=1}^{n} Y_i$, where $Y_i \sim \mathrm{Bernoulli}(p)$ for $\forall i \in \{1, 2, \ldots, n\}$. So, $\mathbb{E}(Y_i) = p$ (†) and $\mathrm{var}(Y_i) = p(1 - p)$ (‡). Thus,

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) \overset{\text{i.i.d.}}{=} \sum_{i=1}^{n} \mathbb{E}(Y_i) \overset{(\dagger)}{=} \sum_{i=1}^{n} p = np, \quad (6)$$

and also

$$\mathrm{var}(X) = \mathrm{var}\left(\sum_{i=1}^{n} Y_i\right) \overset{\text{i.i.d.}}{=} \sum_{i=1}^{n} \mathrm{var}(Y_i) \overset{(\ddagger)}{=} \sum_{i=1}^{n} p(1 - p) =$$
$$= np(1 - p). \quad (7)$$
$$\square$$

Lemma 3 (Binomial distribution's maximum variance). *Let random variable X follow binomial distribution with arguments n and p. Then maximum possible variance of X is $\mathrm{var}(X) = \frac{n}{4}$.*

Proof. According to lemma 2 and formula (7), if $X \sim binomial(n, p)$, it is $\mathrm{var}(X) = np(1-p)$. Let mark $p \equiv \frac{1}{2} + \delta$, where $\delta \in \left\langle -\frac{1}{2}, \frac{1}{2} \right\rangle$. Then, obviously,

$$\mathrm{var}(X) = np(1 - p) = n \left(\frac{1}{2} + \delta\right) \left(1 - \left(\frac{1}{2} + \delta\right)\right) =$$
$$= n \left(\frac{1}{2} + \delta\right) \left(\frac{1}{2} - \delta\right) = n \left(\frac{1}{4} - \delta^2\right) \leq \frac{n}{4}. \quad (8)$$

Thus, generally $\mathrm{var}(X) \leq \frac{n}{4}$, and $\mathrm{var}(X) = \frac{n}{4} = n\left(\frac{1}{4} - 0^2\right)$ for $\delta = 0$, so if and only if $p = \frac{1}{2} + \delta = \frac{1}{2} + 0 = \frac{1}{2}$. $\qquad\square$

Finally, let's revisit Markov's and Chebyshev's inequalities [5], that enables us to derive the number of needed iterations of Monte Carlo simulation.

Theorem 1 (Markov's inequality). *Let X be a non-negative random variable with expected value $\mathbb{E}(X)$. For $a > 0$ is*

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}. \quad (9)$$

Proof. Surely, since $a > 0$ and $X \geq 0$, it is $\mathbb{E}(X \mid X < a) \geq 0$ and $\mathbb{E}(X \mid X \geq a) \geq a$ (†). Because $\mathbb{E}(X \mid X < a) \geq 0$ and $P(X < a) \geq 0$, it is also $\mathbb{E}(X \mid X < a)P(X < a) \geq 0$. So, $\mathbb{E}(X) \geq \mathbb{E}(X) - \mathbb{E}(X \mid X < a)P(X < a)$. Also, due to the total expectations theorem, it is $\mathbb{E}(X) = \mathbb{E}(X \mid X < a)P(X < a) + \mathbb{E}(X \mid X \geq a)P(X \geq a)$, and $\mathbb{E}(X) - \mathbb{E}(X \mid X < a)P(X < a) = \mathbb{E}(X \mid X \geq a)P(X \geq a)$. Thus,

$$\mathbb{E}(X) \geq \mathbb{E}(X) - \mathbb{E}(X \mid X < a)P(X < a) =$$
$$= \mathbb{E}(X \mid X \geq a)P(X \geq a) \overset{(\dagger)}{\geq} a \cdot P(X \geq a),$$

and, finally,

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$
$$\square$$

Theorem 2 (Chebyshev's inequality). *Let X be a random variable with expected value $\mathbb{E}(X)$, and non-zero and finite variance $0 < \mathrm{var}(X) < \infty$. For $b > 0$ is*

$$P(|X - \mathbb{E}(X)| \geq b) \leq \frac{\mathrm{var}(X)}{b^2}. \quad (10)$$

Proof. If we realize that $\mathrm{var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$ (†) and formally put $X \equiv (X - \mathbb{E}(X))^2$ and $a \equiv b^2$ into formula (9) of Markov's inequality, we directly get Chebyshev's inequality,

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$
$$P\left((X - \mathbb{E}(X))^2 \geq b^2\right) \leq \frac{\mathbb{E}\left((X - \mathbb{E}(X))^2\right)}{b^2}$$
$$P\left(|X - \mathbb{E}(X)| \geq b\right) \overset{(\dagger)}{\leq} \frac{\mathrm{var}(X)}{b^2}.$$
$$\square$$

B. Number of needed iterations of Monte Carlo simulation for parameter estimation keeping the estimate's given precision

Let's assume a random variable X following Bernoulli distribution with an argument $P(\mathcal{T})$, i.e., $X \sim \mathrm{Bernoulli}(P(\mathcal{T}))$. Thus, probability of a success in each Bernoulli trial is $P(\mathcal{T}) = f(\theta)$. If we repeat Bernoulli trials n times, based on definition 1, we can get a random variable $\sum_{i=1}^{n} X_i$, where $X_i \sim \mathrm{Bernoulli}(P(\mathcal{T}))$ for $i \in \{1, 2, \ldots, n\}$. Since lemma 2, it is $\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = nP(\mathcal{T}) = nf(\theta)$ and $\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = nP(\mathcal{T})(1 - P(\mathcal{T})) = nf(\theta)(1 - f(\theta))$ (†). Using Chebyshev's inequality from formula (10), we get

$$P\left(\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)\right| \geq b\right) \leq \frac{\operatorname{var}\left(\sum_{i=1}^{n} X_i\right)}{b^2}$$

$$P\left(\left|\sum_{i=1}^{n} X_i - nf(\theta)\right| \geq b\right) \overset{(\dagger)}{\leq} \frac{nf(\theta)(1-f(\theta))}{b^2}$$

We can simplify the right-hand side using lemma 3, i.e., $\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) \leq \frac{n}{4}$, so

$$P\left(\left|\sum_{i=1}^{n} X_i - nf(\theta)\right| \geq b\right) \leq \frac{nf(\theta)(1-f(\theta))}{b^2} \overset{(8)}{\leq} \frac{n}{4b^2}$$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{nf(\theta)}{n}\right| \geq \frac{b}{n}\right) \leq \frac{n}{4b^2}$$

$$P\left(\left|\bar{X} - f(\theta)\right| \geq \frac{b}{n}\right) \leq \frac{n}{4b^2},$$

let's set $b \equiv n\varepsilon$, then we get

$$P\left(\left|\bar{X} - f(\theta)\right| \geq \frac{n\varepsilon}{n}\right) \leq \frac{n}{4(n\varepsilon)^2}$$

$$P\left(\left|\bar{X} - f(\theta)\right| \geq \varepsilon\right) \leq \frac{1}{4n\varepsilon^2},$$

and by setting the probability's uncertainty as $\frac{1}{4n\varepsilon^2} \leq \alpha$ is

$$P\left(\left|\bar{X} - f(\theta)\right| \geq \varepsilon\right) \leq \frac{1}{4n\varepsilon^2} \leq \alpha. \tag{11}$$

Formula (11) tells us that a probability of getting a distance between the parameter $f(\theta)$ and its estimate $\bar{X}$ greater than ε, is lower than α. Thus, to keep imprecision $\leq \epsilon$, i.e., to keep $\frac{1}{4n\varepsilon^2} \leq \alpha$, we need n iterations of the Monte Carlo simulation,

$$n \geq \frac{1}{4\alpha\varepsilon^2}, \tag{12}$$

and unlike (3), formula (12) does not include a stochastic term.

C. A scheme of the proposed approach to Monte Carlo simulation and estimation of parameters

The previous paragraphs and particularly formulas (11) and (12) suggest Monte Carlo simulation for not only probabilistic-like parameters, keeping non-underestimated precision, that consists of the following steps.

(i) Setting the tuning parameters of the simulation – precision $1 - \varepsilon$ and probability $1 - \alpha$.

(ii) Assuming formula (5), constructing a generative Bernoulli process $X \sim \text{Bernoulli}(P(\mathcal{T}))$. We want to estimate parameter θ's value since we don't know it using link function f and a different random process, known from theory, with outcome $P(\mathcal{T})$ where $P(\mathcal{T}) = f(\theta)$.

(iii) Repeating Bernoulli process n times, where $n \geq \frac{1}{4\alpha\varepsilon^2}$, and collecting the outcomes $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$.

(iv) Finally, averaging the outcomes, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} \hat{X}_i$, by getting estimate $f(\theta)$ on precision level $1 - \varepsilon$ with probability $1 - \alpha$.

(v) Parameter θ's estimate is $f^{-1}(f(\theta)) = f^{-1}(\bar{X})$. While parameter $f(\theta)$ is estimated with imprecision ε, parameter $\theta = f^{-1}(\bar{X})$ is estimated with imprecision $f^{-1}(\varepsilon)$.

An algorithm for the proposed Monte Carlo simulation is in Algorithm 1.

Algorithm 1: The proposed approach to Monte Carlo simulation and estimation of parameters of not only probabilistic fashion

---

Data: generative Bernoulli process with probability of success $P(\mathcal{T})$, link function f ensuring that $P(\mathcal{T}) = f(\theta)$, precision $1 - \varepsilon$, probability $1 - \alpha$

Result: parameter θ's estimate

1 $\boldsymbol{X} = \{\emptyset\}$ // a vector for;
2   // estimates saving;
3 $n_{\min} = \frac{1}{4\alpha\varepsilon^2}$ // # of iterations;
4 **for** $i = 1 : n_{\min}$ **do**
5   generate outcome $\hat{X}_i$ from $X \sim \text{Bernoulli}(P(\mathcal{T}))$;
6     $\boldsymbol{X} = \{\boldsymbol{X}, \hat{X}_i\}$ // update the vector;
7 **end**
8 estimate θ as $\theta = f^{-1}(\bar{\boldsymbol{X}}) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{X}_i\right)$;

D. The asymptotic time complexity of the proposed approach to Monte Carlo simulation and estimation of parameters

The simulation is repeated n times where $n \geq \frac{1}{4\alpha\varepsilon^2}$, as comes from formula (12). Assuming one iteration of the Monte Carlo simulation takes τ time units, the total asymptotic time complexity of the procedure, $\Theta(\ddagger)$, is

$$\Theta(\ddagger) = \Theta(n\tau) \geq \Theta\left(\frac{1}{4\alpha\varepsilon^2}\tau\right), \tag{13}$$

so, while $\Theta(\ddagger)$ is linear in n and $\frac{1}{\alpha}$ terms, it is quadratic in $\frac{1}{\varepsilon}$ term. To compare asymptotic time complexity $\Theta(\dagger)$ from formula (4) for the traditional estimation procedure and $\Theta(\ddagger)$ from formula (13) for the proposed one, assuming that $n_0 \ll n$, it is $\Theta(\ddagger) > \Theta(\dagger)$, since, in general, is $\frac{1}{\alpha} > u_{1-\alpha/2}$. Both functions $\frac{1}{\alpha}$ and $u_{1-\alpha/2}$ are monotonous and decreasing while α increases, but for $\alpha \leq 0.05$ is $u_{1-\alpha/2} \gtrsim 2$ while $\frac{1}{\alpha} \geq 20$. So, while the traditional approach is "faster" in terms of time complexity, it may suffer from false underestimating of the parameter estimate's variability.

E. Keeping the first k decimal digits precise in the proposed approach to Monte Carlo simulation and estimation

Due to avoiding the issue of variability coming from lemma 3 and Chebyshev's inequality (10), an appropriate setting of precision level $1 - \varepsilon$ can ensure the first k decimal digits are correctly estimated within the proposed simulation and estimation approach. Inspecting formula (11), we can realize that $\bar{X} - \varepsilon \leq f(\theta) \leq \bar{X} + \varepsilon$ with probability $1 - \alpha$. On

a given probability level $1 - \alpha$, to avoid rounding error up to k-th decimal digit, we should set ε as

$$\varepsilon < 0.5 \cdot 10^{-k}. \tag{14}$$

Moreover, if the inversion function to link function f is of an additive (linear) form, i.e., $\forall \xi, \eta \in \mathbb{R}$ is $f^{-1}(\xi + \eta) = f^{-1}(\xi) + f^{-1}(\eta)$, it is also $f^{-1}(\bar{X} - \varepsilon) \leq f^{-1}(f(\theta)) \leq f^{-1}(\bar{X} + \varepsilon)$, so, $f^{-1}(\bar{X}) - f^{-1}(\varepsilon) \leq \theta \leq f^{-1}(\bar{X}) + f^{-1}(\varepsilon)$, and we can estimate also the *real* imprecision level ε_θ for parameter θ, i.e., not only $f(\theta)$, as

$$f^{-1}(\varepsilon_\theta) < 0.5 \cdot 10^{-k}, \quad \text{or,} \quad \varepsilon_\theta < f(0.5 \cdot 10^{-k}). \tag{15}$$

IV. THE PROPOSED APPROACH TO MONTE CARLO SIMULATION AND ESTIMATION APPLIED: π ESTIMATION

Revisiting the well-known example of π constant estimation using Monte Carlo simulation, let us assume a quarter circle with a radius of 1 as in Fig. 1. For a random point $A = [x, y]$ in the unit square around the quarter circle, where $[x, y] \in \langle 0, 1 \rangle^2$, the generative Bernoulli process $X \sim$ Bernoulli($P(\mathcal{T})$) here returns number 1 if A lies in the quarter circle (in gray color in Fig. 1), otherwise it returns 0. Thus, the random event is $\mathcal{T} = \{A \in \text{quarter circle} \mid A \in \text{unit square}\}$ and $P(\mathcal{T}) = \frac{S_{\text{quarter circle}}}{S_{\text{unit square}}} = \frac{\pi \cdot 1^2}{4}/1 = \frac{\pi}{4}$. So, the Bernoulli process enables us to estimate $f(\theta) = \frac{\pi}{4}$, which implies the link function f as $f(\eta) = \frac{\eta}{4}$.
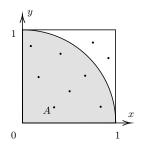

Fig. 1. A quarter circle in a unit square enabling estimation of $\frac{\pi}{4}$ parameter using Monte Carlo simulation of many points such as $A = [x, y] \in \langle 0, 1 \rangle^2$.

Both for traditional and the proposed approach, we repeated Monte Carlo simulation $m = 100$ times to evaluate how likely the k-th decimal digit is not correct, with $k \in \{1, 2, 3\}$. We set probability level $\alpha = 0.05$ and real imprecision $\varepsilon_\theta = f(0.4 \cdot 10^{-k}) = \frac{0.4 \cdot 10^{-k}}{4} = 0.1 \cdot 10^{-k}$ using formula (15). The number of simulation iterations was estimated using formulas (3) and (12). The initial number of iterations for the traditional approach needed for pre-estimating the estimate's standard deviation $\sigma_{0,\theta}$, was $n_0 = 100$. We used R programming language and environment [6] for the Monte-Carlo simulations. There are more numerical applications of R language to various fields in [7]–[9].

Results are in Table I. While the traditional approach did not always ensure the precise k-th digit, particularly (but rarely, in $\leq \alpha = 0.05 = 5\%$ of all cases) for $k = 2$ and $k = 3$, the proposed approach kept the k-th digit's precision every time.

Unlike the proposed method not considering a stochastic term (see formula (12)), the traditional one may suffer from a possible underestimate of initial estimate's variability $\sigma_{0,\theta}$ and needed number n of Monte Carlo iterations (see formula (3)).

TABLE I
PROPORTIONS OF CASES WHEN THE k-TH DIGIT WAS INCORRECT OUT OF $m = 100$ REPETITIONS (MARKED AS r) OF MONTE CARLO SIMULATION.

k	ε_θ	traditional approach		proposed approach	
		$n_{\min}$	r	$n_{\min}$	r
1	0.01	1,028	0.00	50,000	0.00
2	0.001	102,765	0.01	5,000,000	0.00
3	0.0001	10,276,423	0.01	500,000,000	0.00

V. CONCLUSIONS REMARKS

We introduced an alternative approach to Monte Carlo estimation, using refining all estimated parameters as probabilities. That enables us to apply Bernoulli trials with upper-bounded variability of the estimate and Chebyshev's inequality for a robust estimate of the number of iterations needed to ensure the estimate's precision on a given probability level.

VI. ACKNOWLEDGMENT

REFERENCES

[1] Søren Asmussen. "Conditional Monte Carlo for sums, with applications to insurance and finance". In: *Annals of Actuarial Science* 12.2 (Jan. 2018), pp. 455–478. DOI: 10.1017/s1748499517000252.

[2] Eleni G. Elia, Shirley Ge, Lisa Bergersen, et al. "A Monte Carlo Simulation Approach to Optimizing Capacity in a High-Volume Congenital Heart Pediatric Surgical Center". In: *Frontiers in Health Services* 1 (Feb. 2022). DOI: 10.3389/frhs.2021.787358.

[3] Christopher Mooney. *Monte Carlo Simulation*. SAGE Publications, Inc., 1997. DOI: 10.4135/9781412985116.

[4] Patrick Billingsley. *Probability and Measure*. en. 3rd ed. Wiley Series in Probability & Mathematical Statistics: Probability & Mathematical Statistics. Nashville, TN: John Wiley & Sons, May 1995.

[5] Gerold Alsmeyer. "Chebyshev's Inequality". In: *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, 2011, pp. 239–240. DOI: 10.1007/978-3-642-04898-2_167.

[6] R Core Team. *R: A language and environment for statistical computing*. manual. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[7] Patrícia Martinková, Lubomír Štěpánek, Adéla Drabinová, et al. "Semi-real-time analyses of item characteristics for medical school admission tests". In: *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2017. DOI: 10.15439/2017f380.

[8] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. "Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test". In: *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2020. DOI: 10.15439/2020f198.

[9] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. "Machine Learning at the Service of Survival Analysis: Predictions Using Time-to-Event Decomposition and Classification Applied to a Decrease of Blood Antibodies against COVID-19". In: *Mathematics* 11.4 (Feb. 2023), p. 819. DOI: 10.3390/math11040819.