Sound Statistical Model Checking for Probabilities and Expected Rewards*

Carlos E. Budde[1,2], Arnd Hartmanns[3], Tobias Meggendorfer[4], Maximilian Weininger[5], and Patrick Wienhöft[6,7] ✉

[1] Technical University of Denmark, Lyngby, Denmark
[2] University of Trento, Trento, Italy
[3] University of Twente, Enschede, The Netherlands
[4] Lancaster University Leipzig, Germany
[5] Institute of Science and Technology Austria, Klosterneuburg, Austria
[6] Technical University Dresden, Germany · patrick.wienhoeft@tu-dresden.de
[7] Centre for Tactile Internet with Human-in-the-Loop (CeTI), Dresden, Germany

Abstract. Statistical model checking estimates probabilities and expectations of interest in probabilistic system models by using random simulations. Its results come with statistical guarantees. However, many tools use *unsound* statistical methods that produce incorrect results more often than they claim. In this paper, we provide a comprehensive overview of tools and their correctness, as well as of sound methods available for estimating probabilities from the literature. For expected rewards, we investigate how to bound the path reward distribution to apply sound statistical methods for bounded distributions, of which we recommend the Dvoretzky-Kiefer-Wolfowitz inequality that has not been used in SMC so far. We prove that even reachability rewards can be bounded in theory, and formalise the concept of limit-PAC procedures for a practical solution. The MODES SMC tool implements our methods and recommendations, which we use to experimentally confirm our results.

1 Introduction

Statistical model checking (SMC) [83] estimates quantities of interest by sampling a large number k of random runs from a compact executable model of a probabilistic system. Typical quantities of interest are reachability probabilities and expected rewards, to query for e.g. reliability or performance measures [12].

* This work was supported by the DFG through the Cluster of Excellence EXC 2050/1 (CeTI, project ID 390696704, as part of Germany's Excellence Strategy) and the TRR 248 (see perspicuous-computing.science, project ID 389792660), by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreements 101008233 (MISSION), 101034413 (IST-BRIDGE), and 101067199 (ProSVED), by the EU under NextGenerationEU projects D53D23008400006 (Smartitude) under MUR PRIN 2022 and PE00000014 (SERICS) under MUR PNRR, by the Interreg North Sea project STORM_SAFE, and by NWO VIDI grant VI.Vidi.223.110 (TruSTy).

A. Gurfinkel and M. Heule (Eds.): TACAS 2025, LNCS 15696, pp. 167–190, 2025.
https://doi.org/10.1007/978-3-031-90643-5_9

A *sound* statistical model checker delivers results guaranteed to be *probably approximately correct* (PAC), i.e. it returns a confidence interval I with $|I| \leq 2\varepsilon$ ("ε-approximately correct") such that the probability for I to contain the (unknown) true value x is higher than a given confidence level γ ("probably correct").

When applied judiciously, SMC can perform extremely well [82] and easily beat probabilistic model checking (PMC) [10,11] tools that rely on exhaustive state space exploration as well as partial exploration tools [5,53,64] in competitions when PAC results are allowed [26]. It is widely implemented in tools such as C-SMC [31], COSMOS [15], FIG [22], HYPEG [73], MODES [24] of the MODEST TOOLSET [46], MULTIVESTA [43,79], PLASMA LAB [60], PRISM [55], SBIP [69], or UPPAAL SMC [37]. It has been applied to case studies ranging from hardware [63, 77] over biology [84] to cybersecurity [23, 57]. New SMC tools such as SMC STORM [56] are now being developed in industrial contexts.

However, as we detail in Table 1, many existing and new SMC implementations either are unsound (i.e. they do not deliver PAC guarantees), or inefficient (i.e. they use statistical methods that need unnecessarily many samples). The unsoundness is often due to computing confidence intervals via approaches that rely on the central limit theorem, while the inefficiency is notably due to the widespread use of the Okamoto bound [70] for estimating probabilities.

Our contribution is a comprehensive treatment of the problem of efficiently obtaining sound SMC results when estimating probabilities as well as expected rewards. We review the statistical methods available for probabilities in Sec. 3, which forms the basis for Table 1. For expected rewards, in Sec. 4, we provide a novel fully sound approach for the instantaneous and cumulative cases, prove that sound SMC is possible for reachability rewards in theory, and give a practically useful method. We implemented our methods and recommendations in the MODES SMC tool (Sec. 5) to experimentally confirm our findings in Sec. 6.

SMC for reachability **probabilities** comes down to estimating binomial proportions, a well-studied problem in statistics. Sound methods for **expected rewards**, on the other hand, have been an open problem. Here, we need to estimate the mean x of the path reward distribution μ, whose shape is unknown and which can have unbounded support. For this problem, no PAC statistical methods exist. Thus, to obtain a sound SMC approach for expected rewards, we must (1) use structural information to soundly reduce to case of bounded support $[a, b]$ to then (2) employ an appropriate statistical method for this case.

We review the methods available for Step 2 in Sec. 4.1, recommending the use of the Dvoretzky-Kiefer-Wolfowitz inequality (DKW) [39]. This inequality provides a very strong and versatile result that allows the derivation of useful confidence intervals for the mean even for conservative values for a and b, yet has been curiously ignored in the SMC community so far.

For Step 1, we distinguish two cases in Sec. 4.2: First, for (step- or time-bounded) **cumulative** and **instantaneous rewards**, we can derive safe and practical values for a and b given an upper bound on r_{max}, the highest reward assigned to any state, which can typically be obtained from the model's syntax. For (unbounded) **reachability rewards**, we introduce *bounding sets* that provide

a means to ignore very large path rewards while introducing an error of at most $\varepsilon' < \varepsilon$. We prove that a bounding set can be obtained for every finite discrete-time Markov chain (DTMC), given only bounds on certain parameters of the DTMC which can typically be derived syntactically, too. Yet, the resulting bounding set is not practical, thus serving only as a proof of the possibility of sound SMC for reachability rewards. In practice, we propose to use the DKW to obtain guaranteed *lower bounds* that provably converge to x as $k \to \infty$.

Our focus is on *estimating* probabilities and *undiscounted* expected rewards given either k or an *absolute* error ε. We briefly comment on *hypothesis testing* where appropriate, referring the reader to dedicated works on hypothesis testing in SMC like Reijsbergen et al.'s [76] for more details, cautioning that they may not emphasise soundness. Undiscounted rewards are standard in verification while discounting is ubiquitous in machine learning. It is easy to obtain good bounds $[a, b]$ on discounted rewards and thus apply the methods we review in Sec. 4.1 efficiently. For rare events [78] or very large expected rewards, one may want to specify a *relative error* $\varepsilon \cdot x$; we mention some methods specific for this case.

We consider SMC as in the original papers by Younes and Simmons [83] and Hérault et al. [49], motivated by the state space explosion problem which PMC faces for finite-but-large models of realistic applications, and the lack of scalable PMC approaches for non-Markovian models like stochastic automata (IOSA) [36] or HPnGs [68]. Thus, we sample runs from a (mostly) black-box model using $\mathcal{O}(1)$ memory to estimate global quantities of interest. This is in contrast to "model-based SMC" [1,8,65], which aims to apply PMC-like methods to black-box systems with simulation access by *learning* a model, in particular its transition probabilities, requiring memory quadratic in the number of states.

Related soundness. The formal methods community values trustworthy results with clear guarantees on possible analysis errors. For example, after finding that the common stopping criterion of the value iteration algorithm can lead to arbitrarily wrong results [20,44], the problem was addressed in many settings [6,9, 14,40,47,54,75]. Yet, in SMC, the issue of soundness has received little attention. Only recently, a survey of sound and unsound methods for estimating probabilities appeared [65], which our recommendations in Sec. 3 are based on.

2 Background

We write $\mathbb{1}_{pred}$ for the indicator function of *pred*: $\mathbb{1}_{pred}(x) = 1$ if $pred(x)$ else 0. A *probability distribution* over a countable set S is a function $\mu\colon S \to [0,1]$ such that $\sum_{s \in S} \mu(s) = 1$. Its *support* is $spt(\mu) \stackrel{\text{def}}{=} \{\, s \in S \mid \mu(s) > 0 \,\}$.

Models. The assumption of SMC is that models are given in a higher-level formalism—like HPnGs [68], IOSA [36], Jani [25], Modest [19,45], or the Prism language [55]—that allows behaviours to be randomly sampled without having to create an in-memory state space. Their semantics are some form of Markov process; we focus on the special case of DTMCs to simplify the presentation. All our methods immediately apply in the general setting or can be extended.

Definition 1. *A* discrete-time Markov chain *(DTMC) is a tuple* $\langle S, R, T, s_I \rangle$ *of a finite set of* states S*, a* reward *function* $R\colon S \to \mathbb{R}_{\geqslant 0}$*, an* initial state $s_I \in S$*, and a* transition *function* $T\colon S \to \mathit{Dist}(S)$ *mapping each state to a probability distribution over successor states. A (finite)* path π *(π_{fin}) is (a prefix of) an infinite sequence* $\pi = s_0\, s_1 \ldots \in S^\omega$ *such that* $s_0 = s_I$ *and* $\forall i\colon T(s_i)(s_{i+1}) > 0$.

A DTMC induces a probability measure $\mathbb{P}$ over sets of paths that, intuitively, corresponds to multiplying the probabilities along the path (see e.g. [13, Chp. 10]). Abusing notation, we also use π or π_{fin} to refer to the set of a path's states. We write $\pi[i]$ for s_i, the path's $(i+1)$-th state, and $\mathit{idx}(\pi, S') = \min\{i \in \mathbb{N} \mid s_i \in S'\}$ for the index of the first state in $S' \subseteq S$ on π, with $\mathit{idx}(\pi, S') = \infty$ if $\pi \cap S' = \varnothing$. We write $r_{max} \stackrel{\text{def}}{=} \max\{\, R(s) \mid s \in S \,\}$ for a DTMC's maximum reward and $p_{min} \stackrel{\text{def}}{=} \min(\{\, T(s)(s') \mid s, s' \in S \,\} \setminus \{\, 0 \,\})$ for its minimum probability. We assume that, from only the higher-level formalism's syntax, we can efficiently obtain bounds $\overline{|S|} \geq |S|$, $\overline{r}_{max} \geq r_{max}$, and $0 < \underline{p}_{min} \leq p_{min}$.

Properties. Every property to be model-checked can be cast as the expected value $\mathbb{E}(X)$ w.r.t. $\mathbb{P}$ of a random variable X that maps paths to values in $\mathbb{R}_{\geqslant 0}$. We consider the following kinds of properties, some of which take a step bound $c \in \mathbb{N}$ or a set of *goal states* $G \subseteq S$ specified as part of the model:

$$P_{\diamond G} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ \mathbb{1}_{G \cap \pi \neq \varnothing}) \qquad \textit{(reachability probability)}$$

$$P^{\leq c}_{\diamond G} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ \mathbb{1}_{\mathit{idx}(\pi, G) \leq c}) \qquad \textit{(bounded reach. probability)}_{\perp}$$

$$E_{\leq c} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ \textstyle\sum_{i=0}^{c} R(\pi[i])) \qquad \textit{(cumulative reward)}_{\perp}$$

$$E_{\diamond G} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ \textstyle\sum_{i=0}^{\mathit{idx}(\pi, G)} R(\pi[i])) \qquad \textit{(reachability reward)}$$

$$E^{\leq c}_{\diamond G} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ \textstyle\sum_{i=0}^{\min\{\mathit{idx}(\pi, G),\, c\}} R(\pi[i])) \qquad \textit{(bounded and reach. reward)}_{\perp}$$

$$E_{=c} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ R(\pi[c])) \qquad \textit{(instantaneous reward)}_{\perp}$$

$$E_{=G} \stackrel{\text{def}}{=} \mathbb{E}(\lambda\, \pi.\ R(\pi[\mathit{idx}(\pi, G)])) \qquad \textit{(reach-instant reward)}$$

Rewards are obtained upon entering states. $E_{\diamond G}$ and $E_{=G}$ are defined to be ∞ if $\mathbb{P}(\{\, \pi \mid \mathit{idx}(\pi, G) = \infty \,\}) > 0$ [41]. The properties marked $_{\perp}$ are *bounded*; all others are *unbounded*. The former are typical for SMC, required by e.g. Plasma Lab [60, Table 1] and Smc Storm [56, Sect. 2.1]. Unbounded reachability probabilities and rewards, on the other hand, are standard in PMC and dominate model collections like the Quantitative Verification Benchmark Set (QVBS) [48].

Statistical model checking. At its core, SMC is Monte Carlo simulation [2,52,59]: randomly generate a predetermined number k of paths, or *simulation runs*, that give rise to samples $X_1, \ldots, X_k$ of the random variable X; then compute the empirical mean $\hat{X} \stackrel{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{k} X_i$, and perform a *statistical evaluation* to obtain a confidence interval $I = [l, u] \ni \hat{X}$ at predetermined *confidence level* γ.

Simulation. How to obtain simulation runs is specific to each higher-level formalism. We only assume that a method $\mathit{sample}(M, \mathit{prop})$ exists that, given a

model M, implements the random variable X of property *prop*, i.e. that (pseudo-)randomly generates a path π_{fin} through M's semantics according to $\mathbb{P}$ that is long enough to evaluate X and returns $X(\pi_{fin})$. For bounded properties, the "long enough" criterion is straightforward: just generate paths of length c.

To end a simulation run for $P_{\Diamond G}$, we must determine whether it entered a bottom strongly connected component (BSCC) without goal states. For $E_{\Diamond G}$ and $E_{=G}$ to be finite, we must determine whether a non-goal BSCC *exists*. BSCCs can be detected statistically by sampling given some structural information (such as $\underline{p}_{min}$) [7]. Yet this requires storing a set of visited states that can be as large as S, breaking the $\mathcal{O}(1)$ memory property of SMC. Additionally, some fraction of $1-\gamma$ must be devoted to all these tests (see e.g. [35]). However, most verification models—such as those in the QVBS—are structured so that (i) for reachability probabilities, all BSCCs contain only one state, and (ii) the goal state sets in reachability rewards are reached with probability 1, which allows for an efficient but limited stopping criterion. We follow this assumption in this paper.

Statistical evaluation. If we repeat the SMC procedure m times to obtain confidence intervals $I_1, \ldots, I_m$, we find some of them might be incorrect, i.e. $\mathbb{E}(X) \notin I_i$ for some i; occasionally obtaining an "incorrect" result is the nature of a statistical approach based on sampling. The (a priori) probability for a correct result is the *coverage probability* $p_{cov}(k) = \lim_{m\to\infty} \frac{cov_m}{m}$, where cov_m denotes the amount of correct confidence intervals. We call an SMC procedure **sound** if it is guaranteed to provide *probably approximately correct* (PAC) results: Given k and confidence level γ, it has $p_{cov}(k) \geq \gamma$ while producing intervals of width $|I| \leq 2\varepsilon$. Then, the midpoint of this interval is ε-close to the true mean of X with high probability.

Sound SMC results are obtained by employing an appropriate *statistical (evaluation) method* (SM) that relates k (or the concrete X_i), γ, and ε to ensure the PAC requirement, with two values given and the third under control of the SM. We consider two settings: The **fixed-k** setting, where γ and k are given while ε is determined by the SM, and the **sequential** setting, where γ and ε are given so that the SM must determine k. In the latter, k can precomputed from γ and ε, or it can be determined by a *truly sequential* SM that continuously checks whether enough samples have been gathered to be γ-confident of an interval I with $|I| \leq 2\varepsilon$. We always assume γ to be given, a typical value being $\gamma = 0.95$.

3 Sound SMC for Probabilities

For probabilities, i.e. $P_{\Diamond G}$ and $P^{\leq c}_{\Diamond G}$, each simulation run is a Bernoulli trial with outcome $X_i \in \{0, 1\}$. Thus, the SM samples from a binomial distribution with success probability p. After k samples, it observed $k_s = \sum_{i=1}^{k} X_i$ successes and has empirical mean $\hat{p} = \frac{k_s}{k} = \hat{X}$. Constructing a γ-confidence interval around $\hat{p}$ is a well-studied problem in statistics, resulting in many SMs for this task. We often abbreviate $\delta = 1 - \gamma$ for readability.

Meggendorfer et al. [65, Sec. 3] survey SMs in the context of "model-based SMC" for Markov decision processes, where individual transition probabilities are

estimated to "learn" the model. Methods for this specific case also apply to SMC for reachability (and other "qualitative" 0/1 properties) in DTMCs. Hence, we recap their survey of SMs, extending it with examples and plots. Moreover, [65] only considers the fixed-k setting, whereas we also discuss the sequential setting and hypothesis testing. Our survey is the basis for the tool comparison in Sec. 5, where we show that existing tools use unsound and/or inefficient methods.

3.1 Unsound Methods

Denote by $p_{cov}(k, p)$ the coverage probability that the SM at hand attains given success probability p. Many of the commonly used SMs for binomial proportions only guarantee an *average* coverage probability of γ, i.e. $\int_0^1 p_{cov}(k, p)\,\mathrm{d}p \geq \gamma$. This is not in line with the frequentist definition of a confidence interval and **not** sufficient for sound SMC, producing too many incorrect results for certain values of p. We instead require that $\inf_{p=0}^{1} p_{cov}(k, p) \geq \gamma$.

As per [65], unsound methods include those based on the central limit theorem (CLT), in particular the textbook Wald interval, the Wilson score interval, the Agresti-Coull interval [3], the Arcsine interval, and the Logit interval.

The Wilson score interval with continuity correction (Wilson/CC) [67] complements the CLT by adding adjustment terms to improve coverage. However, Newcombe already observed slightly below-nominal coverage [67, Table II], and [65] confirms the insufficient coverage for high confidence levels and p close to 0 or 1.

Sequential setting. Given ε instead of k, Chow and Robbins [33] show that the coverage of constructing a Wald interval after every sample and terminating once this interval has half-width $\leq \varepsilon$ goes to γ as $\varepsilon \to 0$. For any concrete $\varepsilon > 0$, however, coverage $\geq \gamma$ may not be achieved and thus this procedure is not sound for SMC. Reijsbergen et al. [76] adapt it to perform hypothesis testing with some extra parameters that reduce, but do not eliminate, the chance for incorrect results. The sequential methods proposed by Chen [30] are *empirically* sound, i.e. they appear to produce sound results in practice, but soundness is not proven for p close to $\frac{1}{2}$, and even for p away from $\frac{1}{2}$ only as a one-sided version.

3.2 Sound Methods

Okamoto bound. In 1959, Okamoto [70] proved that, for binomial proportions, $\mathbb{P}(\hat{p} - p \geq \varepsilon) \leq \mathrm{e}^{-2k\varepsilon^2}$. We want $\delta \leq \mathbb{P}(|\hat{p} - p| \geq \varepsilon)$, giving

$$\frac{\delta}{2} \leq \mathrm{e}^{-2k\varepsilon^2} \quad \Leftrightarrow \quad \varepsilon \geq \sqrt{\frac{\ln \frac{2}{\delta}}{2k}} \quad \Leftrightarrow \quad k \geq \frac{\ln \frac{2}{\delta}}{2\varepsilon^2}$$

by distributing δ symmetrically. Thus the interval $I_{Oka} = [\hat{p} - \varepsilon, \hat{p} + \varepsilon]$ always has coverage $\geq \gamma$ when ε, k, and δ satisfy the above inequalities. This bound is also referred to as Hoeffding bound [50] after his more general inequality, see Sec. 4.1.

Clopper-Pearson interval. The "exact" binomial interval by Clopper and Pearson [34] guarantees coverage $\geq \gamma$ for all p. One of several ways to compute it is

$$I_{CP} = [B(\delta/2, k_s, k - k_s + 1), B(1 - \delta/2, k_s + 1, k - k_s)]$$

where $B(p, \alpha, \beta)$ is the p-quantile of the Beta(α, β) distribution.

Blyth-Still-Casella and Wang. The approaches of Blyth-Still-Casella [29] and Wang [81] are also sound and produce shortest intervals in a specific sense, but are intricate to implement and computationally very expensive.

Sequential setting. The Okamoto bound provides ε, k, or δ given the other two; thus it applies to the sequential setting, too. For the Clopper-Pearson interval, we use the recent result that its number of required samples is maximal when $\hat{p} = \frac{1}{2}$ [65]. Based on this worst case, we can precompute the smallest k where the interval width is $\leq 2\varepsilon$ and perform a fixed-k evaluation.

3.3 Discussion

Recipes for sequential SMs. The minimum k may depend on p; e.g. for Clopper-Pearson, lower k suffice for p close to 0 or 1. A truly sequential method could exploit this. Any fixed SM can be converted to truly sequential in the Chow-Robbins style by checking after every sample if half-width $\leq \varepsilon$ is met, resulting in methods like "sequential Clopper-Pearson". They however are **not sound** as in general sample mean $\hat{p}$ and precision $|I|$ are correlated [51].

We may first spend a fraction of the "error budget" δ to get a rough interval estimate of p, and then calculate the number of samples required (given the remaining part of δ) based on the worst case in this interval, e.g. the value closest to $\frac{1}{2}$ for Clopper-Pearson as in [21]. Jégourel et al.'s two-step approach [51] similarly uses the Massart bound which improves on Okamoto's if p is known to be away from $\frac{1}{2}$. While **sound** and better than precomputation, these are two-step (generalisable to n-step), not truly sequential, approaches.

Frey [42] calculates a δ^* a priori so that, if a confidence level of $\gamma^* = 1 - \delta^*$ is used in each iteration of a Chow-Robbins-style sequentialisation of a sound fixed SM, the overall coverage probability **sound**ly comes out to γ. However, computing such a δ^* becomes very hard already for small k (around 100-1000). The AdaSelect [38] and EBStop [66] algorithms are sequential methods for the relative error setting, recently generalised by Parmentier and Legay [71].

Example 1 (Soundness). To evaluate SMs for probabilities, we can directly compute their coverage probabilities for binomial distributions (see [27, App. C]). To give a visual comparison of coverage probabilities highlighting the concern for soundness, we fix confidence level $\gamma = 0.9$ and calculate the coverage probabilities for various methods in Fig. 1. The top row shows $p_{cov}(50, p)$ as achieved by the unsound Wald, the sound-but-inefficient Okamoto, and the sound-and-recommended Clopper-Pearson interval. Indeed, the Wald interval does not attain coverage ≥ 0.9 for many values of p, while the others do. Similarly, the bottom row concerns the truly sequential setting with $\varepsilon = 0.05$ and shows that the coverage probabilities for the unsound Chow-Robbins and sequential Clopper-Pearson methods are sometimes below γ. For Chen's method, we confirm its empirical but unproven soundness.

Example 2 (Sample Efficiency). We also observe that Okamoto significantly overshoots the desired confidence, which increases the number of samples it

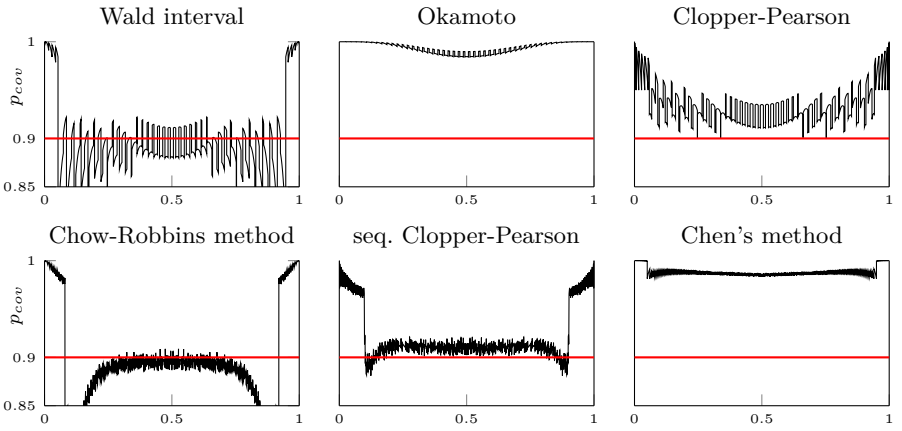

Fig. 1. Coverage for fixed (top, $k = 50$) and sequential methods (bottom, $\varepsilon = 0.05$).

requires. Indeed, I_{CP} always needs fewer samples than I_{Oka} experimentally [65, Sec. 3.3]. For example, with $\gamma = 0.95$ and $\varepsilon = 0.01$, we get a minimum k of 18 445 for Okamoto, independent of p or $\hat{p}$. For Clopper-Pearson, we get a worst-case k of 9 701; for k given and p closer to 0 or 1, we would in turn get much smaller ε.

Hypothesis testing. All methods that produce sound confidence intervals $I = [l, u]$ can be turned into sound hypothesis tests for deciding whether $p \sim p_t$ for a threshold p_t and $\sim \in \{\leq, \geq\}$: assuming $\sim$ is $\leq$, answer *yes* if $u \leq p_t$, *no* if $l \geq p_t$, and *unknown* otherwise. A dedicated and efficient method for hypothesis testing is the sequential probability ratio test (SPRT) [80]. It is **sound** if we consider its *indifference region* (an interval $[p - \varepsilon_i, p + \varepsilon_i]$ where the SPRT is allowed to give wrong answers) to fulfil the role of the ε error in our PAC requirement.

Our recommendation is to implement the Clopper-Pearson interval for $P_{\diamond G}$ and $P^{\leq c}_{\diamond G}$ in the fixed and sequential settings as it is proven sound *and* sample-efficient. In the sequential setting, a two-step approach can be considered. The Okamoto bound, employed by *most* tools using a sound method (Sec. 5), needs too many samples and produces overly conservative intervals: it should not be used for estimating probabilities. We highlight that our recommendations are independent of the underlying system dynamics and thus apply to SMC in general.

4 Sound SMC for Expected Rewards

For unbounded expected rewards, each simulation run is a sample from an unknown *path reward distribution* μ with outcomes in $[0, \infty)$. Given k samples, we want a PAC guarantee for the expected value $E_{\diamond G}$. In general, *no* SM can guarantee coverage for unknown distributions with unbounded support. Intuitively, k gives the SM an indication of how likely it is to have missed some paths; with

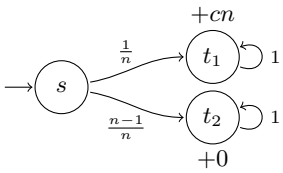

Fig. 2. High reward with low probability

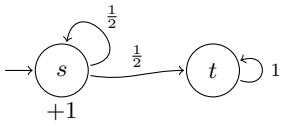

Fig. 3. Unbounded path rewards

bounded support (e.g. for probabilities), this allows to quantify the uncertainty and thus ε. With unbounded support, outcomes with extremely low probability can dominate the expectation if they are even more extremely large. We discuss the general case in [27, App. A] and here provide an illustrative example.

Example 3. The DTMC in Fig. 2 has $E_{=\{t_1,t_2\}} = c$. If k is significantly lower than n, however, the SM likely only sees paths to t_2 and—if it is not sound—returns a confidence interval with an upper bound far below c.

When $spt(\mu) \subseteq [a, b]$, a number of sound SMs exists, which we survey below. Then, in Sec. 4.2, we avoid the general impossibility of the unbounded case in two ways. First, we exploit structural information about the DTMC to reduce to the bounded case. Second, we introduce a novel perspective by considering a new notion of statistically converging lower bounds.

4.1 Statistical Methods for Bounded Distributions

The textbook confidence interval for the mean of an unknown distribution is the *normal interval*: $I_{Norm} = \hat{X} \pm z_\delta \hat{\sigma}/\sqrt{k}$, where $\hat{\sigma}$ is the empirical standard deviation and z_δ the $(1 - \frac{\delta}{2})$-quantile of the standard normal distribution. (Obtaining z_δ via the Student's-t distribution with $k - 1$ degrees of freedom instead may work better for smaller k.) While asymptotic statement of Chow and Robbins about the normal interval also holds in the general, non-binomial setting, these methods are **not sound**. For example, on the DTMC of Fig. 2 with $n = 1000$, $c = 1$ and $k = 500$, we experimentally found I_{Norm} for $\gamma = 0.95$ to have a coverage probability of only $\approx 0.39 \ll 0.95$. Knowing bounds $[a, b]$ on the distribution's support, however, the **sound** methods we list below in this section are available.

Hoeffding's inequality. The Okamoto bound of Sec. 3.2 is a special case of Hoeffding's inequality, which actually bounds the sum of independent (not necessarily i.d.) bounded random variables [50]. It states that

$$\mathbb{P}(\hat{X} - \mathbb{E}(X) \geq \varepsilon) \leq \mathrm{e}^{-\frac{2k\varepsilon^2}{(b-a)^2}}$$

and accordingly $\varepsilon \geq (b-a)\sqrt{(\ln 2/\delta)/2k}$ for the two-sided case by distributing δ equally. Note that Chernoff bounds [32] can be used to derive this inequality.

Bennett's and Bernstein's inequalities. Bennett's inequality can provide tighter bounds on a sum of random variables than Hoeffding's by taking the variance σ^2 into account [16]. However, not knowing the distribution, we do not know σ^2 either. We could insert bounds, a simple one being $\sigma^2 \leq \frac{1}{4}(b-a)^2$. Then, however,

Bennet's inequality is strictly worse than Hoeffding's [65, App. B]. Bernstein's inequality [17,18] is a relaxation of Bennet's that is easier to compute, but yields even wider intervals. Thus, in our setting, Hoeffding's inequality is preferable.

Dvoretzky-Kiefer-Wolfowitz(-Massart) inequality (DKW). The DKW [39, 62] relates the cumulative distribution function (cdf) $F(x) = \mathbb{P}(X \leq x)$ of the unknown distribution μ to the empirical cdf $\hat{F}(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{X_i \leq x}$ as follows:

$$\mathbb{P}\Big(\sup_{x\in\mathbb{R}} |\hat{F}(x) - F(x)| > \varepsilon\Big) \leq 2e^{-2k\varepsilon^2}.$$

DKW is about thresholds, i.e. in our setting the probability of exceeding a certain reward. It characterizes a confidence band in which the real cdf lies with high probability. This can be used to derive bounds on the expected value by computing the expected values of the best- and worst-case cdfs within the confidence band. Formally, let C be a confidence band containing an uncountable set of cdfs; then with probability at least $1 - 2e^{-2k\varepsilon^2}$ we have

$$\min_{\underline{F}\in C} \mathbb{E}(Y \mid Y \sim \underline{F}) \leq \mathbb{E}(X) \leq \max_{\overline{F}\in C} \mathbb{E}(Y \mid Y \sim \overline{F}).$$

The cdfs minimising or maximising the expectation can be easily computed, as they are the upper and lower bound of the confidence band, respectively:

$$\underline{F}(x) = \min\left\{1, \hat{F}(x) + \sqrt{\tfrac{1}{2k}\ln\tfrac{2}{\delta}}\right\} \qquad \overline{F}(x) = \max\left\{0, \hat{F}(x) - \sqrt{\tfrac{1}{2k}\ln\tfrac{2}{\delta}}\right\}$$

Fig. 4 illustrates the DKW[8] for $[a, b] = [-3, 3]$, with F the smooth orange line, $\hat{F}$ the light blue step function in the center, and the outer purple step functions being $\underline{F}$ (to the left, with a higher probability for smaller outcomes) and $\overline{F}$ (to the right). All steps of $\underline{F}$ are the same as those of $\hat{F}$ except that we "map" the largest $\sqrt{(\ln 2/\delta)/2k}$ fraction of steps to the lower bound (i.e. at $x = a$). Similarly, $\overline{F}$ shifts probability mass into the upper bound b. In the worst case, the expectations of $\underline{F}$ or $\overline{F}$ coincide with Hoeffding, but provide a tighter confidence interval when few samples have an extremal value of a or b. Applying DKW is therefore especially advantageous when one of the a priori bounds a and b is very loose and samples are far above/below it. The best case is obtained when all samples coincide, where the width of confidence interval halves when using DKW as opposed to Hoeffding. In any case, the DKW interval is always contained in the Hoeffding interval.

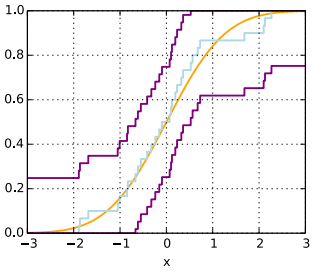

Fig. 4. The DKW cdfs

Proposition 1. *For given confidence level γ and set of samples from a distribution with bounds $[a, b]$, let $[l_d, u_d]$ and $[l_h, u_h]$ be confidence intervals given by DKW and Hoeffing's inequality, respectively. Then $l_h \leq l_d < u_d \leq u_h$ and $\frac{u_h - l_h}{u_d - l_d} \leq 2$.*

DKW is also used by Phan et al. [58,72] with a view towards machine learning applications, who attribute its application to expected rewards to Anderson [4].

[8] Fig. 4 is based on file commons.wikimedia.org/wiki/File:DKW_bounds.svg (CC0).

Sequential setting and hypothesis testing. Hoeffding's inequality applies in the sequential setting in the same way as the Okamoto bound. For DKW, we could precompute k based on the worst case, but this coincides with Hoeffding. As mentioned, the Chow-Robbins scheme remains applicable and unsound. The AdaSelect algorithm we mentioned in Sec. 3.3 also works soundly for bounded distributions when a relative error is desired. For hypothesis testing, the SPRT in principle also applies to bounded distributions, and will in general perform better than the DKW for testing a single threshold. The latter's advantage is that it provides an entire confidence *band* around the cdf and thereby allows deriving the expected reward as well as probability bounds on *all* reward thresholds *at once*. In this way, the DKW can also be used to tackle quantile problems.

Our recommendation for estimating a bounded distribution is to use DKW in the fixed-k setting and resort to Hoeffding's inequality when given ε.

4.2 Bounding Expected Rewards

For a full sound SMC procedure for expected rewards, it remains to find the bounds $[a, b]$ on the path rewards. As rewards are non-negative, $a = 0$ is a safe lower bound (though larger a may give lower ε or k), leaving b to be determined.

Instantaneous and cumulative rewards. We know $\overline{r}_{max}$, an upper bound on the maximum state reward (Sec. 2). For instantaneous reward properties $E_{=c}$ and $E_{=G}$, a path's reward is at most $\max_{s \in S} R(s)$, making $b = \overline{r}_{max}$ the tightest safe upper bound we can give. For step-bounded reward properties $E_{\leq c}$ and $E^{\leq c}_{\Diamond G}$, we can upper-bound the reward of a path $\pi = s_0 \ldots$ by $b = (c{+}1) \cdot \overline{r}_{max} \geq \sum_{i=0}^{c} R(s_i)$. Exploiting specific structures in higher-level languages may yield tighter bounds.

Reachability rewards. For unbounded reachability rewards $E_{\Diamond G}$, we need to bound the accumulated path reward until visiting a state in G, i.e. $PR(\pi) = \sum_{i=0}^{idx(\pi,G)} R(\pi[i])$. We assume $\mathbb{E}(PR)$ to be finite, but PR can still be unbounded:

Example 4. Consider the very simple DTMC in Fig. 3. We have $E_{\Diamond G} = \sum_{i=1}^{\infty} i \cdot (1/2)^i = 2$, but the reward of a single path is unbounded, since every reward $v \in \mathbb{N}$ is obtained with positive probability $(1/2)^v$.

This is not a degenerate case, but occurs whenever there exists a cycle with non-zero rewards. Consequently, we cannot directly apply the SMs from Sec. 4.1. Nevertheless, we can give meaningful estimations, by requiring additional knowledge or by relaxing the constraints on the result.

Bounding large values. As we cannot bound PR, we aim to bound the effect that large values have on $\mathbb{E}(PR)$. Let us work in a general setting, as follows:

Definition 2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X \colon \Omega \to \mathbb{R}_{\geqslant 0}$ a random variable with finite expectation, and $B \in \mathcal{F}$. We call B an* ε-bounding set *if (i) $\int_{\overline{B}} X \, \mathrm{d}\mathbb{P} < \varepsilon$ (with $\overline{B} = \Omega \setminus B$), and (ii) $\forall \omega \in B \colon X(\omega) \in [a, b]$.*

(Concretely, X could be PR.) If B is a bounding set, then we can rewrite

$$\mathbb{E}(X) = \int_{\Omega} X \,\mathrm{d}\mathbb{P} = \int_{B} X \,\mathrm{d}\mathbb{P} + \int_{\overline{B}} X \,\mathrm{d}\mathbb{P} < \int_{B} X \,\mathrm{d}\mathbb{P} + \varepsilon.$$

Observe that $\int_B X \,\mathrm{d}\mathbb{P} = \mathbb{E}(X^B)$, where $X^B(\omega) = X(\omega)$ if $X \in B$ else 0. Since we chose B such that $X(\omega) \in [a, b]$ for all $\omega \in B$, X^B clearly is bounded and we can apply our previous methods to obtain a statistical estimate of X^B. Inserting in the above equation then yields a bound on the overall expectation of X.

One possible choice for B would be $\{\, \omega \mid |X(\omega)| < t \,\}$ for a sufficiently large t. Such a t exists for any random variable with finite expectation by positivity and additivity of $\mathbb{P}$ (when $\mathbb{E}(X) = \int_\Omega X \,\mathrm{d}\mathbb{P} < \infty$ we necessarily have $\lim_{t\to\infty} \int_{\{|X|>t\}} X \,\mathrm{d}\mathbb{P} = 0$). However, without further assumptions we cannot derive such a t or any other kind of bounding set just by sampling. Thus, in the following we exploit that DTMCs give some structure to the random variable.

Geometric path lengths. While the value of PR in Ex. 4 can be arbitrarily large, as long as the expectation $\mathbb{E}(PR)$ is finite, this only happens with vanishingly low probabilities. This is the case for DTMCs (and many other Markov systems, see Remark 1) in general: Intuitively, the number of steps until G is reached is (roughly) geometrically distributed.

Lemma 1. *Let $\varepsilon > 0$. Choose q such that*

$$\overline{|S|} \cdot \overline{r}_{max} \cdot (1 - (\underline{p}_{min})^{\overline{|S|}})^q \cdot (q - q(\underline{p}_{min})^{\overline{|S|}} + 1) \cdot (\underline{p}_{min})^{-\overline{|S|}} < \varepsilon.$$

Then $B = \Diamond^{\leq q \cdot \overline{|S|}} G = \{\, \pi \mid idx(\pi, G) \leq q \cdot \overline{|S|} \,\}$ is a bounding set.

Proof (Sketch). Every state s has a path of length at most $|S|$ to the goal G by assumption. Such a path has probability at least $(p_{\min})^{|S|}$ and reward at most $|S| \cdot r_{\max}$. Considering $|S|$ steps as an "episode", we can geometrically lower bound the probability to reach G after q episodes, and use this to upper bound the reward. See [27, App. B.1] for the full proof.

Corollary 1. *Given bounds $\overline{|S|} \geq |S|$, $0 < \underline{p}_{min} \leq p_{\min}$, and $\overline{r}_{max} \geq r_{\max}$, we can give PAC guarantees on $E_{\Diamond G}$.*

Remark 1. Due to the worst-case over-approximation involved, q is extremely large even for very small DTMCs and is thus not a practical solution. For example, a DTMC with 5 states, $p_{min} = 0.05$, and $r_{max} = 1$ with a desired bound $\varepsilon = 1$ requires $q > 10^8$. The value $(p_{\min})^{-|S|}$ is closely related to the *mixing time* of a DTMC, see e.g. [61], which, in our setting, intuitively upper bounds the time until a goal state is reached with high probability. While often a coarse bound, there exists DTMCs for which it is tight [44, Fig. 3]. Determining better bounds on the mixing time (and thus q) requires knowledge of the DTMC's state space and transitions, which SMC explicitly does not have access to. For Markovian systems other than DTMCs, the geometric path lengths construction can also be used, provided we can obtain similar bounds; we conjecture that a sufficient condition is that the system is finite and goal states are reached almost surely.

Lower bounds. The inability to practically bound b means that we cannot gain confidence in an *upper* bound on $E_{\diamond G}$. However, since rewards are non-negative, we cannot miss any extreme "negative" events. Thus we at least want to derive meaningful *lower bounds*. Since 0 trivially is a correct lower bound, we need a definition of bounds being "close" to the true value. We propose the novel definition of *limit-PAC* lower bounds: they are not only (i) sound, but additionally require that (ii) given enough samples, they (unknowingly) become ε-close.

Definition 3. *Let X be a random variable. A procedure $\mathcal{A}$ yields* limit-PAC lower bounds *on $\mathbb{E}(X)$ if, for any confidence γ, the following two conditions hold: (i) For a collection of independent samples Ξ drawn from X, we have $\mathbb{P}(\mathcal{A}(\Xi,\gamma) \leq \mathbb{E}(X)) \geq \gamma$. (ii) For any precision $\varepsilon > 0$, there exists a threshold k_0 such that for a collection of independent samples Ξ drawn from X with $|\Xi| \geq k_0$, we have $\mathbb{P}(\mathbb{E}(X) - \varepsilon \leq \mathcal{A}(\Xi,\gamma) \leq \mathbb{E}(X)) \geq \gamma$.*

Remark 2. Classical procedures such as normal intervals do not provide limit-PAC bounds. While they may satisfy condition (ii) and provide enough coverage to satisfy condition (i) in the limit, they can be unsound for many sample sets Ξ that are not "sufficiently close to the limit."

We describe a procedure "DKW-$\mathbb{E}$-Lower" which provides limit-PAC lower bounds: For a given set of samples Ξ with $k = |\Xi|$, set $\chi_k = \sqrt{(\ln 2/1 - \gamma)/2k}$ (using χ instead of ε to avoid a clash of notation) and compute the empirical average over Ξ, however setting the largest χ_k fraction of samples to 0. This is equivalent to computing the expectation of the minimising cdf $\underline{F}(x)$ provided by the DKW (with width χ_k), as explained in Sec. 4.1.

Theorem 1. *For any non-negative, finite-expectation random variable X, DKW-$\mathbb{E}$-Lower gives limit-PAC lower bounds on $\mathbb{E}(X)$.*

Proof (sketch, full proof in [27, App. B.2]). Condition (i) holds by the DKW with coverage $\geq \gamma$ due to our choice of χ_k. For condition (ii), note that for every $\varepsilon/2$, we can find some bounding set $\{ X > t \}$. Then for large enough Ξ, the difference between the actual expected value and the output of DKW-$\mathbb{E}$-Lower on $[0,t]$ can be bounded by $\varepsilon/2$ as in the bounded case (see Sec. 4.1), and on $[t,\infty)$ the difference is also bounded by $\varepsilon/2$ by definition of bounding sets.

Corollary 2. *DKW-$\mathbb{E}$-Lower gives limit-PAC lower bounds on $E_{\diamond G}$.*

Note that Theorem 1 directly extends to any random variable with a known lower bound, i.e. $X \in [a,\infty)$, by considering $X' = X - a$. Then $X' \geq 0$ and any limit-PAC estimation of X' also yields one for X, as $\mathbb{E}(X) = \mathbb{E}(X') + a$. Similarly, for $X \in (-\infty, a]$ we can give limit-PAC *upper* bounds.

5 Implementation

State of the art. In Table 1, we collect the results of an extensive survey of the SMs used by default in all current SMC tools we are aware of. It is based on

Table 1. Default statistical methods used in state-of-the-art SMC tools

Tool		For probabilities $p \in [0,1]$		For rewards $r \in [a,b]$	
Name	Ref.	fixed k	seq. ε	fixed k	seq. ε
C-SMC	[31]	**Okamoto**	—	—	—
Cosmos	[15]	**Clopper-Pearson**	Chow-Robbins	**Hoeffding**	Chow-Robbins
Fig	[22]*	Wilson w/o CC	seq. Student's-t	—	—
Hypeg	[74]	—	seq. Student's-t	—	—
modes (prev.)	[24]*	**Okamoto**	Chen	Normal	Chow-Robbins
MultiVeSta	[43]	—	Chow-Robbins	—	Chow-Robbins
Plasma Lab	[60]*	**Okamoto**	**Okamoto**	**Hoeffding**	**Hoeffding**
Prism	[55]*	Student's-t	seq. Student's-t	Student's-t	seq. Student's-t
Sbip	[69]	Normal	**Okamoto**	—	—
Smc Storm	[56]	—	Chen	—	Chow-Robbins
Uppaal Smc	[37]*	**Clopper-Pearson**	seq. Clopper-P.	Student's-t	—
modes v3.1.281		**Clopper-Pearson**	**Clopper-Pearson**	**DKW**	**Hoeffding**

the information available in their tool papers (column *Ref.*); for those marked "*", we also tested a current version or consulted its documentation[9] for more accurate information. The "seq." prefix for a method indicates a Chow-Robbins-like procedure using an interval different from Wald's/the normal approximation-based one. We highlight the provably sound methods in boldface. Entries "—" indicate that the tool does not appear to support that setting.

We see that, in the fixed setting for probabilities, 5 of 8 tools choose a sound method, although three of those use the inefficient Okamoto bound; in the sequential setting, only 2 of 10 tools use a sound (but inefficient) method. For expected rewards, Cosmos and Plasma Lab apply Hoeffding's inequality when there is an obvious upper bound b, with Cosmos using information from its higher-level formalism (e.g. a system's finite capacity bound when estimating the average number of clients) for this purpose. In the general setting, in particular for their respective variants of reachability rewards, Cosmos will build a normal interval instead while Plasma Lab will return the estimate $\hat{X}$ only, without error bounds. Overall, *no tool* implements a sound *and* efficient method for probabilities in the sequential setting, nor for rewards in the fixed setting; those tools that use Hoeffding for rewards only do so for very specific cases.

Sound SMC in modes. We have implemented the recommendations we make w.r.t. SMs for probabilities and the new methods we propose for soundly handling expected rewards in the newest version of the modes statistical model checker as shown in the last row of Table 1. In particular, modes uses the k-precomputation based on the Clopper-Pearson interval in the sequential setting for probabilities, and the DKW in the fixed setting for expected rewards, improving upon the state of the art in soundness and sample efficiency. modes supports $P^{\leq c}_{\diamond G}$, $P_{\diamond G}$,

[9] We used Fig 1.3, the previous version of modes from the Modest Toolset v3.1.265, Prism 4.8.1, Uppaal Smc 5.0.0 with its online documentation as of 2024-10-09, and the Plasma Lab 1.4.4 documentation from the Web Archive as of 2019-11-01.

$E^{\leq c}_{\diamond G}$, and $E_{\diamond G}$ properties. For $E^{\leq c}_{\diamond G}$ properties, it computes the upper bound as $b = (c+1) \cdot \overline{r}_{max}$. For $E_{\diamond G}$, it uses our new DKW-$\mathbb{E}$-Lower method by default. MODES also implements the Wilson/CC, Wald/normal, and Student's-t intervals, the Okamoto bound, Chen's methods, the Chow-Robbins approach, and the SPRT. Via a command-line parameter, the user can provide a preference list of these methods; for each property being analysed, MODES chooses the first in the list that can be applied to it. By default, it prefers sound over unsound and then efficient over less efficient methods, resulting in the first choices as in Table 1.

6 Experimental Evaluation

To evaluate SMs for probabilities, we can directly work with the binomial distribution as in Ex. 1. With expected rewards, however, the shape of the (unknown) reward distribution matters. We thus use our implementation in MODES on models from the QVBS [48] to evaluate the coverage probability, performance, and effectiveness of the methods we propose in Sec. 4 in a realistic setting. The code and scripts for reproduction are available online [28].

Experimental setup. We used MODES version 3.1.273[10]. We chose all DTMC and Markov decision process (MDP) models from the QVBS that contain an expected-reward property (except those that just ask for an expected number of transitions), excluding only the artificial *haddad-monmege* model plus *bluetooth* and *oscillators*, which MODES cannot handle for technical reasons (the former having multiple initial states, which MODES does not support[11], and the latter's syntax being too large to parse and compile[12]). We turn the MDP into DTMC by applying the PRISM language's DTMC semantics, which resolves all nondeterministic choices uniformly at random. The models are parametrised; we use up to four parameter valuations each, including the smallest and largest ones included in QVBS. A triple $\langle \textit{model}, \textit{parameter values}, \textit{property} \rangle$ is a *benchmark instance*.

We consider all $E^{\leq c}_{\diamond G}$ and $E_{\diamond G}$ properties included with the models, the only $E^{\leq c}_{\diamond G}$ property being in the *resource-gathering* model. To be able to study the DKW and Hoeffding methods, we manually turn all $E_{\diamond G}$ into $E^{\leq c}_{\diamond G}$ by experimentally determining a small c that does not change the value of the property up to the third significant digit.[13] This in essence constitutes a manually-derived bounding

[10] MODES 3.1.273 implements all methods as described in Sec. 5, but uses Wilson/CC for probabilities by default. Version 3.1.281 defaults to Clopper-Pearson as in Table 1.

[11] Supporting multiple initial states, while natural for a PCTL model checker like PRISM, would require an SMC tool to perform a separate analysis starting from each initial state and thus defeat the scalability of SMC.

[12] The *oscillators* model explicitly encodes a large flat state space in the higher-level PRISM language's syntax, overwhelming MODES' parser that assumes its input to *compactly* encode a potentially large state space for SMC to sample runs from.

[13] We obtain reference results from the QVBS, if available, or via SMC with $k = 5 \cdot 10^6$. We determine a c as follows: Use SMC with $k = 10^6$ and some step bound c to obtain a value. If the three most significant digits of this value are equal to the reference result, stop and report c. Otherwise, increase c and repeat. All choices of c greater

set except we truncate rewards instead of setting them to 0. [27, App. D.1] lists the resulting 44 instances (including the values of c).

Coverage probabilities. While the normal interval and Chow-Robbins are unsound, it was not clear if this manifests on real models under typical k and ε. To investigate this, we implemented an empirical coverage test inside MODES: Given a benchmark instance, a confidence γ, step bound k, and a number m, it executes SMC with fixed k for m times, each time computing a γ-confidence interval. It counts w, the number of times the reference result was $w$rong, i.e. not in the computed confidence interval. Thus, we obtain the empirical coverage probability as $p_{cov} = \frac{w}{m}$, and compute a "meta" confidence interval $[l_{cov}, u_{cov}]$ around it using Clopper-Pearson.

In Table 2, we report the result of using this empirical coverage test choosing $\gamma = 0.95$, $k = 1000$, and $m = 5000$. We report the number of benchmark instances where the respective SM attained insufficient coverage ($p_{cov} < \gamma$) in a statistically significant way ($u_{cov} < \gamma$), as well as the minimum $\min p_{cov}$ and average $\varnothing\, p_{cov}$ of the 44 coverage probabilities. Detailed results are in [27, App. D.2]. Hoeffding's inequality and the DKW produced only sound results as expected, although Hoeffding timed out (>10 minutes) 39 times in the sequential setting. The unsound methods produce incorrect results much more often than they claim, with the insufficiency even being statistically significant in almost a quarter of the benchmark instances.

Table 2. Coverage over 44 instances ($\gamma = 0.95$)

SM	$u_{cov}<\gamma$	$p_{cov}<\gamma$	$\min p_{cov}$	$\varnothing\, p_{cov}$
Normal	10	31	0.908	0.946
Student's-t	9	32	0.902	0.947
Hoeffding (k)	0	0	1	1
DKW	0	0	0.999	1.000
Chow-Robbins	16	24	0.723	0.937
Hoeffding (ε)	0	0	1	1

Performance. We next evaluate the performance of the two sound SMs available when $[a, b]$ is known. The runtime spent on the calculations involved with the SMs that we consider is negligible compared to that for generating sample paths. We thus compare the performance of Hoeffding's inequality and the DKW via the half-width ε of the interval returned given fixed $k = 500\,000$. The results are shown as a scatter plot in Fig. 5, where every point $\langle x, y \rangle$ (blue for DTMCs, orange for MDPs) is the result of one benchmark instance, stating that using Hoeffding's inequality resulted in $\varepsilon = x$ while the DKW gave $\varepsilon = y$. Note the logarithmic scale; points on the dotted diagonals mark 2× differences. We see that, as expected, the DKW consistently produces smaller intervals; the geometric mean of the ratios $\frac{\varepsilon \text{ for Hoeffding}}{\varepsilon \text{ for DKW}}$ over all 44 instances is 1.72, close to the

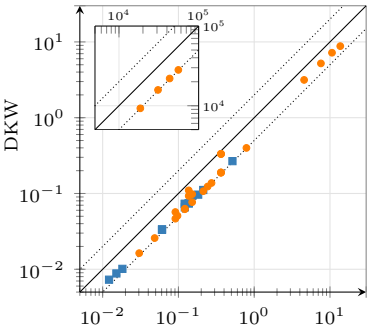

Fig. 5. Sound ε given k

than the output of this procedure allow to approximate the result with precision 10^{-3}, in particular since in all QVBS models the target state is reached with probability 1.

theoretical maximum of 2 (see Proposition 1). Our upper bounds b computed as per Sec. 4.2 are rather loose, benefiting DKW and resulting in very asymmetric DKW intervals with a lower bound close to the true value and an upper bound similar to Hoeffding's.

Effectiveness. For unknown b as in $E_{\diamond G}$ properties, we test how quickly our novel DKW-$\mathbb{E}$-Lower method converges to the (usually unknown) true value by applying it to our 44 benchmark instances for $k = 10^i$ with $i \in \{2, 3, 4, 5, 6\}$. All results are in [27, App. D.3] and show behaviour similar to the three benchmark instances of Fig. 6 (from top to bottom: $\langle$*coupon*, $\langle 15, 4, 5\rangle$, *exp_draws*$\rangle$, $\langle$*resource-gathering*, $\langle 1300, 100, 100\rangle$, *expgold*$\rangle$, and $\langle$*egl*, $\langle 5, 8\rangle$, *messagesB*$\rangle$). The distance between the lower bound and the true value in most cases decreases by a factor between 2 and 3 on each step. Over all instances and steps, on average (geometric mean) the distance decreases by a factor of 2.6, which gives an indication of the practical convergence rate of the DKW-$\mathbb{E}$-Lower method.

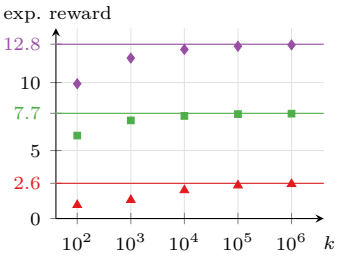

Fig. 6. DKW-$\mathbb{E}$-Lower

7 Conclusion

We raise attention to the issue of soundness in SMC given a state of the art where many tools use unsound statistical methods. For estimating probabilities, several sound methods exist, which have recently been compared in [65]. We summarised them as a reference for the SMC practitioner, and expanded upon [65] by looking into the sequential setting as well as adding coverage probability plots that highlight the level of (un)soundness at a glance and providing an overview of the methods employed by tools. For expected-reward properties, only two tools had (ad-hoc and inefficient) sound methods so far; we contribute a recommendation for the—apparently little-known—DKW and a thorough treatment of the problem of bounding the path reward distribution. While our proof that sound SMC is possible for reachability rewards is currently of theoretical use only, we expect our notion of bounding sets to be crucial for future practical solutions based on the identification of specific structural features of a model's state space or higher-level description. On the practical side, we formalised the notion of *limit-PAC* procedures, which we instantiate by the DKW-$\mathbb{E}$-Lower method that we show to give close bounds in practice. As immediate future work, our results can be extended to estimating rare event probabilities, where samples are in $[0, 1]$ or potentially unbounded depending on the rare event simulation method used. Our contributions should transfer to continuous-time Markov chains straightforwardly.

Data availability statement. The models, tools, and scripts to reproduce our experimental evaluation are archived and available at DOI 10.5281/zenodo.14743520 [28].

References

1. Agarwal, C., Guha, S., Kretínský, J., Muruganandham, P.: PAC statistical model checking of mean payoff in discrete- and continuous-time MDP. In: CAV (2). Lecture Notes in Computer Science, vol. 13372, pp. 3–25. Springer (2022). https://doi.org/10.1007/978-3-031-13188-2_1
2. Agha, G., Palmskog, K.: A survey of statistical model checking. ACM Trans. Model. Comput. Simul. **28**(1), 6:1–6:39 (2018). https://doi.org/10.1145/3158668
3. Agresti, A., Coull, B.A.: Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician **52**(2), 119–126 (1998). https://doi.org/10.2307/2685469
4. Anderson, T.W.: Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. Bulletin of The International and Statistical Institute **43**, 249–251 (1969)
5. Ashok, P., Butkova, Y., Hermanns, H., Kretínský, J.: Continuous-time Markov decisions based on partial exploration. In: Lahiri, S.K., Wang, C. (eds.) 16th International Symposium on Automated Technology for Verification and Analysis (ATVA). Lecture Notes in Computer Science, vol. 11138, pp. 317–334. Springer (2018). https://doi.org/10.1007/978-3-030-01090-4_19
6. Ashok, P., Chatterjee, K., Daca, P., Kretínský, J., Meggendorfer, T.: Value iteration for long-run average reward in Markov decision processes. In: CAV (1). Lecture Notes in Computer Science, vol. 10426, pp. 201–221. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9_10
7. Ashok, P., Daca, P., Kretínský, J., Weininger, M.: Statistical model checking: Black or white? In: Margaria, T., Steffen, B. (eds.) 9th International Symposium on Leveraging Applications of Formal Methods (ISoLA). Lecture Notes in Computer Science, vol. 12476, pp. 331–349. Springer (2020). https://doi.org/10.1007/978-3-030-61362-4_19
8. Ashok, P., Kretínský, J., Weininger, M.: PAC statistical model checking for markov decision processes and stochastic games. In: CAV (1). Lecture Notes in Computer Science, vol. 11561, pp. 497–519. Springer (2019). https://doi.org/10.1007/978-3-030-25540-4_29
9. Azeem, M., Evangelidis, A., Kretínský, J., Slivinskiy, A., Weininger, M.: Optimistic and topological value iteration for simple stochastic games. In: ATVA. Lecture Notes in Computer Science, vol. 13505, pp. 285–302. Springer (2022). https://doi.org/10.1007/978-3-031-19992-9_18
10. Baier, C.: Probabilistic model checking. In: Esparza, J., Grumberg, O., Sickert, S. (eds.) Dependable Software Systems Engineering, NATO Science for Peace and Security Series – D: Information and Communication Security, vol. 45, pp. 1–23. IOS Press (2016). https://doi.org/10.3233/978-1-61499-627-9-1
11. Baier, C., de Alfaro, L., Forejt, V., Kwiatkowska, M.: Model checking probabilistic systems. In: Clarke, E.M., Henzinger, T.A., Veith, H., Bloem, R. (eds.) Handbook of Model Checking, pp. 963–999. Springer (2018). https://doi.org/10.1007/978-3-319-10575-8_28
12. Baier, C., Haverkort, B.R., Hermanns, H., Katoen, J.P.: Performance evaluation and model checking join forces. Commun. ACM **53**(9), 76–85 (2010). https://doi.org/10.1145/1810891.1810912
13. Baier, C., Katoen, J.P.: Principles of model checking. MIT Press (2008)

14. Baier, C., Klein, J., Leuschner, L., Parker, D., Wunderlich, S.: Ensuring the reliability of your model checker: Interval iteration for Markov decision processes. In: CAV (1). Lecture Notes in Computer Science, vol. 10426, pp. 160–180. Springer (2017). https://doi.org/10.1007/978-3-319-63387-9_8
15. Ballarini, P., Barbot, B., Duflot, M., Haddad, S., Pekergin, N.: HASL: A new approach for performance evaluation and model checking from concepts to experimentation. Performance Evaluation **90**, 53–77 (2015). https://doi.org/10.1016/j.peva.2015.04.003
16. Bennett, G.: Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association **57**(297), 33–45 (1962). https://doi.org/10.1080/01621459.1962.10482149
17. Bernstein, S.: On a modification of Chebyshev's inequality and of the error formula of Laplace. Ann. Sci. Inst. Sav. Ukraine, Sect. Math **1**(4), 38–49 (1924)
18. Bernstein, S.: Theory of Probability. 2 edn. (1934)
19. Bohnenkamp, H.C., D'Argenio, P.R., Hermanns, H., Katoen, J.P.: MoDeST: A compositional modeling formalism for hard and softly timed systems. IEEE Trans. Software Eng. **32**(10), 812–830 (2006). https://doi.org/10.1109/TSE.2006.104
20. Brázdil, T., Chatterjee, K., Chmelik, M., Forejt, V., Kretínský, J., Kwiatkowska, M.Z., Parker, D., Ujma, M.: Verification of Markov decision processes using learning algorithms. In: Cassez, F., Raskin, J.F. (eds.) 12th International Symposium on Automated Technology for Verification and Analysis (ATVA). Lecture Notes in Computer Science, vol. 8837, pp. 98–114. Springer (2014). https://doi.org/10.1007/978-3-319-11936-6_8
21. Bu, H., Sun, M.: Clopper-pearson algorithms for efficient statistical model checking estimation. IEEE Transactions on Software Engineering (01), 1–20 (2024). https://doi.org/10.1109/TSE.2024.3392720
22. Budde, C.E.: FIG: the Finite Improbability Generator v1.3. SIGMETRICS Perform. Evaluation Rev. **49**(4), 59–64 (2022). https://doi.org/10.1145/3543146.3543160
23. Budde, C.E.: Using statistical model checking for cybersecurity analysis. In: Skarmeta, A.F., Canavese, D., Lioy, A., Matheu, S.N. (eds.) First International Workshop on Digital Sovereignty in Cyber Security: New Challenges in Future Vision (CyberSec4Europe). Communications in Computer and Information Science, vol. 1807, pp. 16–32. Springer (2022). https://doi.org/10.1007/978-3-031-36096-1_2
24. Budde, C.E., D'Argenio, P.R., Hartmanns, A., Sedwards, S.: An efficient statistical model checker for nondeterminism and rare events. Int. J. Softw. Tools Technol. Transf. **22**(6), 759–780 (2020). https://doi.org/10.1007/S10009-020-00563-2
25. Budde, C.E., Dehnert, C., Hahn, E.M., Hartmanns, A., Junges, S., Turrini, A.: JANI: Quantitative model and tool interaction. In: Legay, A., Margaria, T. (eds.) 23rd International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS). Lecture Notes in Computer Science, vol. 10206, pp. 151–168 (2017). https://doi.org/10.1007/978-3-662-54580-5_9
26. Budde, C.E., Hartmanns, A., Klauck, M., Kretínský, J., Parker, D., Quatmann, T., Turrini, A., Zhang, Z.: On correctness, precision, and performance in quantitative verification – qcomp 2020 competition report. In: Margaria, T., Steffen, B. (eds.) 9th International Symposium on Leveraging Applications of Formal Methods (ISoLA). Lecture Notes in Computer Science, vol. 12479, pp. 216–241. Springer (2020). https://doi.org/10.1007/978-3-030-83723-5_15
27. Budde, C.E., Hartmanns, A., Meggendorfer, T., Weininger, M., Wienhöft, P.: Sound statistical model checking for probabilities and expected rewards. CoRR **abs/2411.00559** (2024). https://doi.org/10.48550/arXiv.2411.00559

28. Budde, C.E., Hartmanns, A., Meggendorfer, T., Weininger, M., Wienhöft, P.: Sound statistical model checking for probabilities and expected rewards (experimental reproduction package) (2025). https://doi.org/10.5281/zenodo.14743520
29. Casella, G.: Refining binomial confidence intervals. Canadian Journal of Statistics **14**(2), 113–129 (1986). https://doi.org/https://doi.org/10.2307/3314658
30. Chen, J.: Properties of a new adaptive sampling method with applications to scalable learning. Web Intell. **13**(4), 215–227 (2015). https://doi.org/10.3233/WEB-150322
31. Chenoy, A., Duchene, F., Given-Wilson, T., Legay, A.: C-SMC: A hybrid statistical model checking and concrete runtime engine for analyzing C programs. In: Laarman, A., Sokolova, A. (eds.) 27th International Symposium on Model Checking Software (SPIN). Lecture Notes in Computer Science, vol. 12864, pp. 101–119. Springer (2021). https://doi.org/10.1007/978-3-030-84629-9_6
32. Chernoff, H.: A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. The Annals of Mathematical Statistics **23**(4), 493–507 (1952). https://doi.org/10.1214/aoms/1177729330
33. Chow, Y.S., Robbins, H.: On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. The Annals of Mathematical Statistics **36**(2), 457–462 (1965). https://doi.org/10.1214/aoms/1177700156
34. Clopper, C., Pearson, E.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**(4), 404–413 (1934). https://doi.org/10.1093/biomet/26.4.404
35. Daca, P., Henzinger, T.A., Kretínský, J., Petrov, T.: Faster statistical model checking for unbounded temporal properties. ACM Trans. Comput. Log. **18**(2), 12:1–12:25 (2017). https://doi.org/10.1145/3060139
36. D'Argenio, P.R., Monti, R.E.: Input/output stochastic automata with urgency: Confluence and weak determinism. In: Fischer, B., Uustalu, T. (eds.) 15th International Colloquium on Theoretical Aspects of Computing (ICTAC). Lecture Notes in Computer Science, vol. 11187, pp. 132–152. Springer (2018). https://doi.org/10.1007/978-3-030-02508-3_8
37. David, A., Larsen, K.G., Legay, A., Mikučionis, M., Poulsen, D.B.: Uppaal SMC tutorial. Int. J. Softw. Tools Technol. Transf. **17**(4), 397–415 (2015). https://doi.org/10.1007/s10009-014-0361-y
38. Domingo, C., Gavaldà, R., Watanabe, O.: Adaptive sampling methods for scaling up knowledge discovery algorithms. In: Discovery Science. Lecture Notes in Computer Science, vol. 1721, pp. 172–183. Springer (1999). https://doi.org/10.1007/3-540-46846-3_16
39. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. The Annals of Mathematical Statistics **27**(3), 642–669 (1956). https://doi.org/10.1214/aoms/1177728174
40. Eisentraut, J., Kelmendi, E., Kretínský, J., Weininger, M.: Value iteration for simple stochastic games: Stopping criterion and learning algorithm. Inf. Comput. **285**(Part), 104886 (2022). https://doi.org/10.1016/j.ic.2022.104886
41. Forejt, V., Kwiatkowska, M.Z., Norman, G., Parker, D.: Automated verification techniques for probabilistic systems. In: Bernardo, M., Issarny, V. (eds.) 11th International School on Formal Methods for the Design of Computer, Communication and Software Systems (SFM). Lecture Notes in Computer Science, vol. 6659, pp. 53–113. Springer (2011). https://doi.org/10.1007/978-3-642-21455-4_3
42. Frey, J.: Fixed-width sequential confidence intervals for a proportion. The American Statistician **64**(3), 242–249 (2010), https://www.jstor.org/stable/20799919

43. Gilmore, S., Reijsbergen, D., Vandin, A.: Transient and steady-state statistical analysis for discrete event simulators. In: IFM. pp. 145–160. Springer (2017). https://doi.org/10.1007/978-3-319-66845-1_10
44. Haddad, S., Monmege, B.: Interval iteration algorithm for mdps and imdps. Theor. Comput. Sci. **735**, 111–131 (2018). https://doi.org/10.1016/J.TCS.2016.12.003
45. Hahn, E.M., Hartmanns, A., Hermanns, H., Katoen, J.P.: A compositional modelling and analysis framework for stochastic hybrid systems. Formal Methods Syst. Des. **43**(2), 191–232 (2013). https://doi.org/10.1007/S10703-012-0167-Z
46. Hartmanns, A., Hermanns, H.: The Modest Toolset: An integrated environment for quantitative modelling and verification. In: Ábrahám, E., Havelund, K. (eds.) 20th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS). Lecture Notes in Computer Science, vol. 8413, pp. 593–598. Springer (2014). https://doi.org/10.1007/978-3-642-54862-8_51
47. Hartmanns, A., Kaminski, B.L.: Optimistic value iteration. In: Lahiri, S.K., Wang, C. (eds.) 32nd International Conference on Computer Aided Verification (CAV). Lecture Notes in Computer Science, vol. 12225, pp. 488–511. Springer (2020). https://doi.org/10.1007/978-3-030-53291-8_26
48. Hartmanns, A., Klauck, M., Parker, D., Quatmann, T., Ruijters, E.: The quantitative verification benchmark set. In: Vojnar, T., Zhang, L. (eds.) 25th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS). Lecture Notes in Computer Science, vol. 11427, pp. 344–350. Springer (2019). https://doi.org/10.1007/978-3-030-17462-0_20
49. Hérault, T., Lassaigne, R., Magniette, F., Peyronnet, S.: Approximate probabilistic model checking. In: Steffen, B., Levi, G. (eds.) 5th International Conference on Verification, Model Checking, and Abstract Interpretation (VMCAI). Lecture Notes in Computer Science, vol. 2937, pp. 73–84. Springer (2004). https://doi.org/10.1007/978-3-540-24622-0_8
50. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58**(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830
51. Jégourel, C., Sun, J., Dong, J.S.: Sequential schemes for frequentist estimation of properties in statistical model checking. ACM Trans. Model. Comput. Simul. **29**(4), 25:1–25:22 (2019). https://doi.org/10.1145/3310226
52. Kretínský, J.: Survey of statistical verification of linear unbounded properties: Model checking and distances. In: ISoLA (1). Lecture Notes in Computer Science, vol. 9952, pp. 27–45 (2016). https://doi.org/10.1007/978-3-319-47166-2_3
53. Kretínský, J., Meggendorfer, T.: Of cores: A partial-exploration framework for Markov decision processes. Log. Methods Comput. Sci. **16**(4) (2020), https://lmcs.episciences.org/6833
54. Kretínský, J., Meggendorfer, T., Weininger, M.: Stopping criteria for value iteration on stochastic games with quantitative objectives. In: 38th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2023, Boston, MA, USA, June 26-29, 2023. pp. 1–14. IEEE (2023). https://doi.org/10.1109/LICS56636.2023.10175771

55. Kwiatkowska, M.Z., Norman, G., Parker, D.: PRISM 4.0: Verification of probabilistic real-time systems. In: Gopalakrishnan, G., Qadeer, S. (eds.) 23rd International Conference on Computer Aided Verification (CAV). Lecture Notes in Computer Science, vol. 6806, pp. 585–591. Springer (2011). https://doi.org/10.1007/978-3-642-22110-1_47
56. Lampacrescia, M., Klauck, M., Palmas, M.: Towards verifying robotic systems using statistical model checking in STORM. In: Steffen, B. (ed.) Second International Conference on Bridging the Gap Between AI and Reality (AISoLA). Lecture Notes in Computer Science, vol. 15217, pp. 446–467. Springer (2024). https://doi.org/10.1007/978-3-031-75434-0_28
57. Lanotte, R., Merro, M., Zannone, N.: Impact analysis of coordinated cyber-physical attacks via statistical model checking: A case study. In: Huisman, M., Ravara, A. (eds.) 43rd IFIP WG 6.1 International Conference on Formal Techniques for Distributed Objects, Components, and Systems (FORTE). Lecture Notes in Computer Science, vol. 13910, pp. 75–94. Springer (2023). https://doi.org/10.1007/978-3-031-35355-0_6
58. Learned-Miller, E.G., Thomas, P.S.: A new confidence interval for the mean of a bounded random variable. CoRR **abs/1905.06208** (2019), https://arxiv.org/abs/1905.06208
59. Legay, A., Lukina, A., Traonouez, L.M., Yang, J., Smolka, S.A., Grosu, R.: Statistical model checking. In: Steffen, B., Woeginger, G.J. (eds.) Computing and Software Science – State of the Art and Perspectives, Lecture Notes in Computer Science, vol. 10000, pp. 478–504. Springer (2019). https://doi.org/10.1007/978-3-319-91908-9_23
60. Legay, A., Sedwards, S., Traonouez, L.M.: Plasma Lab: A modular statistical model checking platform. In: Margaria, T., Steffen, B. (eds.) 7th International Symposium on Leveraging Applications of Formal Methods (ISoLA). Lecture Notes in Computer Science, vol. 9952, pp. 77–93 (2016). https://doi.org/10.1007/978-3-319-47166-2_6
61. Levin, D.A., Peres, Y.: Markov chains and mixing times, vol. 107. American Mathematical Soc. (2017)
62. Massart, P.: The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. The Annals of Probability **18**(3), 1269–1283 (1990). https://doi.org/10.1214/aop/1176990746
63. Mazurek, F., Tschand, A., Wang, Y., Pajic, M., Sorin, D.J.: Rigorous evaluation of computer processors with statistical model checking. In: 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). pp. 1242–1254. ACM (2023). https://doi.org/10.1145/3613424.3623785
64. Meggendorfer, T., Weininger, M.: Playing games with your PET: Extending the partial exploration tool to stochastic games. In: Gurfinkel, A., Ganesh, V. (eds.) Computer Aided Verification - 36th International Conference, CAV 2024, Montreal, QC, Canada, July 24-27, 2024, Proceedings, Part III. Lecture Notes in Computer Science, vol. 14683, pp. 359–372. Springer (2024). https://doi.org/10.1007/978-3-031-65633-0_16
65. Meggendorfer, T., Weininger, M., Wienhöft, P.: What are the odds? Improving the foundations of statistical model checking. CoRR **abs/2404.05424** (2024). https://doi.org/10.48550/arXiv.2404.05424
66. Mnih, V., Szepesvári, C., Audibert, J.Y.: Empirical Bernstein stopping. In: Cohen, W.W., McCallum, A., Roweis, S.T. (eds.) 25th International Conference on Machine Learning (ICML). ACM International Conference Proceeding Series, vol. 307, pp. 672–679. ACM (2008). https://doi.org/10.1145/1390156.1390241

67. Newcombe, R.G.: Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in medicine **17**(8), 857–872 (1998)
68. Niehage, M., Pilch, C., Remke, A.: Simulating hybrid Petri nets with general transitions and non-linear differential equations. In: 13th EAI International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS). pp. 88–95. ACM (2020). https://doi.org/10.1145/3388831.3388842
69. Nouri, A., Mediouni, B.L., Bozga, M., Combaz, J., Bensalem, S., Legay, A.: Performance evaluation of stochastic real-time systems with the $\mathcal{S}$BIP framework. International Journal of Critical Computer-Based Systems **8**(3-4), 340–370 (2018). https://doi.org/10.1504/IJCCBS.2018.096439
70. Okamoto, M.: Some inequalities relating to the partial sum of binomial probabilities. Annals of the Institute of Statistical Mathematics **10**(1), 29–35 (1959)
71. Parmentier, M., Legay, A.: Adaptive stopping algorithms based on concentration inequalities. In: Steffen, B. (ed.) Second International Conference on Bridging the Gap Between AI and Reality (AISoLA). Lecture Notes in Computer Science, vol. 15217, pp. 336–353. Springer (2024). https://doi.org/10.1007/978-3-031-75434-0_23
72. Phan, M., Thomas, P.S., Learned-Miller, E.G.: Towards practical mean bounds for small samples. In: Meila, M., Zhang, T. (eds.) 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8567–8576. PMLR (2021), https://proceedings.mlr.press/v139/phan21a.html
73. Pilch, C., Edenfeld, F., Remke, A.: HYPEG: Statistical model checking for hybrid Petri nets: Tool paper. In: Marin, A., Houdt, B.V., Casale, G., Petriu, D.C., Rossi, S. (eds.) 11th EAI International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS). pp. 186–191. ACM (2017). https://doi.org/10.1145/3150928.3150956
74. Pilch, C., Remke, A.: Statistical model checking for hybrid petri nets with multiple general transitions. In: DSN. pp. 475–486. IEEE Computer Society (2017). https://doi.org/10.1109/DSN.2017.41
75. Quatmann, T., Katoen, J.P.: Sound value iteration. In: Chockler, H., Weissenbacher, G. (eds.) 30th International Conference on Computer Aided Verification (CAV). Lecture Notes in Computer Science, vol. 10981, pp. 643–661. Springer (2018). https://doi.org/10.1007/978-3-319-96145-3_37
76. Reijsbergen, D., de Boer, P., Scheinhardt, W.R.W., Haverkort, B.R.: On hypothesis testing for statistical model checking. Int. J. Softw. Tools Technol. Transf. **17**(4), 377–395 (2015). https://doi.org/10.1007/S10009-014-0350-1
77. Roberts, R., Lewis, B., Hartmanns, A., Basu, P., Roy, S., Chakraborty, K., Zhang, Z.: Probabilistic verification for reliability of a two-by-two network-on-chip system. In: Lluch-Lafuente, A., Mavridou, A. (eds.) 26th International Conference on Formal Methods for Industrial Critical Systems (FMICS). Lecture Notes in Computer Science, vol. 12863, pp. 232–248. Springer (2021). https://doi.org/10.1007/978-3-030-85248-1_16
78. Rubino, G., Tuffin, B. (eds.): Rare Event Simulation using Monte Carlo Methods. Wiley (2009). https://doi.org/10.1002/9780470745403
79. Sebastio, S., Vandin, A.: MultiVeStA: statistical model checking for discrete event simulators. In: Horváth, A., Buchholz, P., Cortellessa, V., Muscariello, L., Squillante, M.S. (eds.) 7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS). pp. 310–315. ICST/ACM (2013). https://doi.org/10.4108/ICST.VALUETOOLS.2013.254377
80. Wald, A.: Sequential Tests of Statistical Hypotheses. The Annals of Mathematical Statistics **16**(2), 117–186 (1945). https://doi.org/10.1214/aoms/1177731118

81. Wang, W.: An iterative construction of confidence intervals for a proportion. Statistica Sinica **24**(3), 1389–1410 (2014), https://www.jstor.org/stable/24310993
82. Younes, H.L.S., Kwiatkowska, M.Z., Norman, G., Parker, D.: Numerical vs. statistical probabilistic model checking. Int. J. Softw. Tools Technol. Transf. **8**(3), 216–228 (2006). https://doi.org/10.1007/S10009-005-0187-8
83. Younes, H.L.S., Simmons, R.G.: Probabilistic verification of discrete event systems using acceptance sampling. In: Brinksma, E., Larsen, K.G. (eds.) 14th International Conference on Computer Aided Verification (CAV). Lecture Notes in Computer Science, vol. 2404, pp. 223–235. Springer (2002). https://doi.org/10.1007/3-540-45657-0_17
84. Zuliani, P.: Statistical model checking for biological applications. Int. J. Softw. Tools Technol. Transf. **17**(4), 527–536 (2015). https://doi.org/10.1007/S10009-014-0343-0