Error Control for Probabilistic Model Checking*

Håkan L. S. Younes

Computer Science Department, Carnegie Mellon University,
Pittsburgh, PA 15213, USA

Abstract. We introduce a framework for expressing correctness guarantees of model-checking algorithms. The framework allows us to qualitatively compare different solution techniques for probabilistic model checking, both techniques based on statistical sampling and numerical computation of probability estimates. We provide several new insights into the relative merits of the different approaches. In addition, we present a new statistical solution method that can bound the probability of error under any circumstances by sometimes reporting undecided results. Previous statistical solution methods could only bound the probability of error outside of an "indifference region".

1 Introduction

Probabilistic model checking, based on the model-checking paradigm pioneered by Clarke and Emerson [4], is a technique for automated verification of stochastic processes. Given a model of a stochastic process, for example a Markov chain, the model-checking task is to determine whether the model satisfies some property Φ. For instance, consider a queuing system with random (according to some distribution) arrivals and departures. We may ask whether the probability is at most 0.5 that the queue will become full in the next hour of operation. This is an example of a probabilistic time-bounded property. Techniques for verifying such properties are the focus of this paper.

Algorithms for probabilistic model checking of time-bounded properties come in two flavors: *numerical* [3, 12] and *statistical* [18, 8, 14, 16]. The former rely on numerical algorithms for probability computations, while the latter use statistical sampling and discrete-event simulation to assess the validity of probabilistic properties. Some insights into the relative merits of the two approaches are given by Younes et al. [17]. Yet, a direct comparison is difficult because numerical and statistical techniques provided quite different correctness guarantees. Furthermore, conflicting claims have been made about the benefits of competing statistical solution methods. Hérault et al. [8] state that their solution method, based on statistical *estimation*, is better than the method of Younes and Simmons [18], based on *hypothesis testing*, because the sample size of the former method is known exactly. Sen et al. [14] provide empirical data that seem to

* Supported in part by the US Army Research Office (ARO), under contract no. DAAD190110485.

suggest that hypothesis testing with *fixed-size samples* consistently outperforms *sequential* hypothesis testing (the latter being advocated by Younes et al. [17]).

This paper is an attempt to set the record straight regarding the relative merits of different solution methods for probabilistic model checking. We establish a framework for expressing the correctness guarantees of model-checking algorithms (Sect. 3). Section 4 shows how to connect the truncation error, ϵ, of numerical methods with the parameter δ (the half-width of the "indifference region") of statistical methods. We conclude that numerical and statistical solution methods can, indeed, be interpreted as solving the same problem. Statistical solution methods are simply randomized algorithms for the same problems that numerical methods solve. We are also able to prove that statistical estimation, when used for probabilistic model checking, reduces to hypothesis testing with fixed-size samples. It follows that Younes and Simmons' solution method *never* needs to use a larger sample size than Hérault et al.'s estimation-based method, and it will often use a much smaller sample size to achieve the same correctness guarantees. Our framework for error control also helps us understand the results of Sen et al., which seem to contradict results presented by Younes [16].

The second contribution of this paper is a new statistical method for probabilistic model checking. Current statistical solution methods provide bounds for the probability of error only when a formula holds (or does not hold) *with some margin*. Our new method bounds the probability of error *under all circumstances*. This is accomplished by permitting an *undecided* result. Sen et al. [14] have previously toyed with the idea of undecided results for statistical solution methods, but only for nested probabilistic operators and without providing any mechanisms for bounding the probability of producing an undecided result (or even an incorrect result, for that matter). Section 5 shows, for the first time, how to bound the probability of undecided and incorrect results for *any formula*, including conjunctions of probabilistic statements and nested probabilistic statements. Section 6 discusses the computational complexity of statistical solution methods in general. A brief empirical evaluation of the new statistical solution method is provided in Sect. 7.

2 Probabilistic Model Checking

This section describes stochastic discrete-event systems, which is the class of models that we consider for probabilistic model checking. A logic, UTSL, for expressing properties of such models is introduced. We describe the semantics of UTSL and of UTSL$_\delta$, the latter being a relaxation of the former logic that permits practical model-checking algorithms.

2.1 Stochastic Discrete-Event Systems

A *stochastic discrete-event system* is any stochastic process that can be thought of as occupying a single state for a duration of time before an *event* causes an instantaneous state transition to occur. The canonical example is a queuing

system, with the state being the number of items currently in the queue. The state changes at the occurrence of an arrival or departure event.

The evolution of a stochastic discrete-event system over time is captured by a *trajectory*. The trajectory of a stochastic discrete-event system is piecewise constant and can be represented as a sequence $\sigma = \{\langle s_0, t_0\rangle, \langle s_1, t_1\rangle, \ldots\}$, with $s_i \in S$ and $t_i > 0$. Let

$$T_i = \begin{cases} 0 & \text{if } i = 0 \\ \sum_{j=0}^{i-1} t_i & \text{if } i > 0 \end{cases} , \tag{1}$$

so that T_i is the time at which state s_i is entered and t_i is the duration of time for which the process remains in s_i before an event triggers a transition to state s_{i+1}. It is assumed that $\lim_{i\to\infty} T_i < \infty$. This implies that only a finite number of events can trigger in a finite interval of time, which is a reasonable assumption for any physical process (cf. [1]).

A *measurable* stochastic discrete-event system is a triple $\mathcal{M} = \langle S, T, \mu\rangle$, where S is the state space, T is the time domain ($\mathbb{Z}^*$ for discrete-time models and $[0, \infty)$ for continuous-time models), and μ is a probability measure over sets of trajectories with *common prefix*. A prefix of $\sigma = \{\langle s_0, t_0\rangle, \langle s_1, t_1\rangle, \ldots\}$ is a sequence $\sigma_{\leq\tau} = \{\langle s'_0, t'_0\rangle, \ldots, \langle s'_k, t'_k\rangle\}$, with $s'_i = s_i$ for all $i \leq k$, $\sum_{i=0}^{k} t'_i = \tau$, $t'_i = t_i$ for all $i < k$, and $t'_k < t_k$. Let $Path(\sigma_{\leq\tau})$ denote the set of trajectories with common prefix $\sigma_{\leq\tau}$. This set must be measurable for probabilistic model checking to have meaning and its measure is determined by μ. The exact definition of μ depends on the structure of the process. Baier et al. [3] provide a definition for continuous-time Markov chains and Younes [16] discusses the construction of a probability space for trajectories of stochastic discrete-event systems in general.

2.2 UTSL: The Unified Temporal Stochastic Logic

A stochastic discrete-event system is a triple $\langle S, T, \mu\rangle$. We assume a factored representation of S, with a set of state variables SV and a value assignment function $V(s, x)$ providing the value of $x \in SV$ in state s. The domain of x is the set $D_x = \bigcup_{s\in S} V(s, x)$ of possible values that x can take on. We define the syntax of UTSL for a factored stochastic discrete-event system $\mathcal{M} = \langle S, T, \mu, SV, V\rangle$ as

$$\Phi ::= x \sim v \mid \neg\Phi \mid \Phi \wedge \Phi \mid \mathcal{P}_{\bowtie\theta}[\Phi\ \mathcal{U}^I\ \Phi] \ ,$$

where $x \in SV$, $v \in D_x$, $\sim \in \{\leq, =, \geq\}$, $\theta \in [0, 1]$, $\bowtie \in \{\leq, \geq\}$, and $I \subset T$. Additional UTSL formulae can be derived in the usual way. For example, $\bot \equiv (x = v) \wedge \neg(x = v)$ for some $x \in SV$ and $v \in D_x$, $\top \equiv \neg\bot$, $\Phi \vee \Psi \equiv \neg(\neg\Phi \wedge \neg\Psi)$, $\Phi \rightarrow \Psi \equiv \neg\Phi \vee \Psi$, and $\mathcal{P}_{<\theta}[\varphi] \equiv \neg\mathcal{P}_{\geq\theta}[\varphi]$.

The standard logic operators have their usual meaning. $\mathcal{P}_{\bowtie\theta}[\varphi]$ asserts that the probability measure over the set of trajectories satisfying the path formula φ is related to θ according to $\bowtie$. Path formulae are constructed using the temporal path operator $\mathcal{U}^I$ ("until"). The path formula $\Phi\ \mathcal{U}^I\ \Psi$ asserts that Ψ becomes true $t \in I$ time units into the future while Φ holds continuously prior to t. The

validity of a UTSL formula is inductively defined as follows:

$$\begin{aligned}
\mathcal{M}, \{\langle s_0, t_0\rangle, \ldots, \langle s_k, t_k\rangle\} &\models x \sim v && \text{if } V(s_k, x) \sim v \\
\mathcal{M}, \sigma_{\leq\tau} &\models \neg\Phi && \text{if } \mathcal{M}, \sigma_{\leq\tau} \not\models \Phi \\
\mathcal{M}, \sigma_{\leq\tau} &\models \Phi \wedge \Psi && \text{if } (\mathcal{M}, \sigma_{\leq\tau} \models \Phi) \wedge (\mathcal{M}, \sigma_{\leq\tau} \models \Psi) \\
\mathcal{M}, \sigma_{\leq\tau} &\models \mathcal{P}_{\bowtie\theta}[\varphi] && \text{if } \mu(\{\sigma \in \mathit{Path}(\sigma_{\leq\tau}) \mid \mathcal{M}, \sigma, \tau \models \varphi\}) \bowtie \theta
\end{aligned}$$

$$\begin{aligned}
\mathcal{M}, \sigma, \tau \models \Phi\ \mathcal{U}^I\ \Psi \quad \text{if } \exists t \in I.\big((\mathcal{M}, \sigma_{\leq\tau+t} \models \Psi) \\
\wedge \forall t' \in T.\big((t' < t) \rightarrow (\mathcal{M}, \sigma_{\leq\tau+t'} \models \Phi)\big)\big)
\end{aligned}$$

The semantics of $\Phi\ \mathcal{U}^I\ \Psi$ requires that Φ holds continuously, i.e. at every point in time, along a trajectory until Ψ is satisfied. For Markov chains, it is sufficient to consider time points at which state transitions occur. The semantics of UTSL therefore coincides with the semantics for Hansson and Jonsson's [7] PCTL interpreted over discrete-time Markov chains and Baier et al.'s [3] CSL interpreted over continuous-time Markov chains. For non-Markovian models, however, the validity of Φ or Ψ may vary over time in the same state if these formulae contain probabilistic operators. Because of this, the statistical solution method for probabilistic model checking presented in this paper is restricted to Markov chains for properties with nested probabilistic operators. Without nesting, the method does not rely on this restriction.

We typically want to know whether a property Φ holds for a model $\mathcal{M}$ if execution starts in a specific state s. A *model-checking problem* $\langle \mathcal{M}, s, \Phi\rangle$ has an affirmative answer if and only if $\mathcal{M}, \{s, 0\} \models \Phi$.

2.3 UTSL$_\delta$: UTSL with Indifference Regions

Consider the model-checking problem $\langle \mathcal{M}, s, \mathcal{P}_{\bowtie\theta}[\varphi]\rangle$ and let p be the probability measure for the set of trajectories that start in s and satisfy φ. If p is "sufficiently close" to θ, then it is likely to make little difference to a user whether or not $\mathcal{P}_{\bowtie\theta}[\varphi]$ is reported to hold by a model-checking algorithm.

To formalize this idea, we introduce UTSL$_\delta$ as a relaxation of UTSL. With each formula of the form $\mathcal{P}_{\bowtie\theta}[\varphi]$, we associate an indifference region centered around θ with half-width δ. If $|p - \theta| < \delta$, then the truth value of $\mathcal{P}_{\bowtie\theta}[\varphi]$ is undetermined for UTSL$_\delta$; otherwise, it is the same as for UTSL.

The formal semantics of UTSL$_\delta$ is given by a satisfaction relation $\models^\delta_\top$ and an unsatisfaction relation $\models^\delta_\bot$. For standard logic formulae, $\models^\delta_\top$ replaces $\models$ and $\models^\delta_\bot$ replaces $\not\models$. For probabilistic formulae we have the following rules:

$$\begin{aligned}
\mathcal{M}, \sigma_{\leq\tau} &\models^\delta_\top \mathcal{P}_{\geq\theta}[\varphi] && \text{if } \mu(\{\sigma \in \mathit{Path}(\sigma_{\leq\tau}) \mid \mathcal{M}, \sigma, \tau \models^\delta_\top \varphi\}) \geq \theta + \delta \\
\mathcal{M}, \sigma_{\leq\tau} &\models^\delta_\bot \mathcal{P}_{\geq\theta}[\varphi] && \text{if } \mu(\{\sigma \in \mathit{Path}(\sigma_{\leq\tau}) \mid \mathcal{M}, \sigma, \tau \models^\delta_\bot \varphi\}) \geq 1 - (\theta - \delta) \\
\mathcal{M}, \sigma_{\leq\tau} &\models^\delta_\top \mathcal{P}_{\leq\theta}[\varphi] && \text{if } \mu(\{\sigma \in \mathit{Path}(\sigma_{\leq\tau}) \mid \mathcal{M}, \sigma, \tau \models^\delta_\top \varphi\}) \leq \theta - \delta \\
\mathcal{M}, \sigma_{\leq\tau} &\models^\delta_\bot \mathcal{P}_{\leq\theta}[\varphi] && \text{if } \mu(\{\sigma \in \mathit{Path}(\sigma_{\leq\tau}) \mid \mathcal{M}, \sigma, \tau \models^\delta_\bot \varphi\}) \leq 1 - (\theta + \delta)
\end{aligned}$$

A model-checking problem $\langle \mathcal{M}, s, \Phi\rangle$ may very well belong to neither of the two relations $\models^\delta_\top$ and $\models^\delta_\bot$, in which case the problem is considered "too close to call".

3 Error Control

This section discusses error control for model-checking algorithms in general terms. The discussion establishes ideal conditions for the correctness guarantees of a model-checking algorithm. These conditions are used as a point of reference in later sections when we discuss error control in practical algorithms for probabilistic model checking.

Given a model-checking problem $\langle \mathcal{M}, s, \Phi \rangle$ and a model-checking algorithm $\mathcal{A}$, let $\mathcal{M}, s \vdash_\top \Phi$ represent the fact that Φ is accepted as true by $\mathcal{A}$ and $\mathcal{M}, s \vdash_\bot \Phi$ that Φ is rejected as false by $\mathcal{A}$ (for the remainder of the paper we will leave out $\mathcal{M}$ from relations for the sake of brevity). Ideally, we would like the probability to be low that $\mathcal{A}$ produces an incorrect answer. More precisely, the probability of a false negative should be at most α and the probability of a false positive at most β, as expressed by the following conditions:

$$\Pr[s \vdash_\bot \Phi \mid s \models \Phi] \leq \alpha \tag{2}$$

$$\Pr[s \vdash_\top \Phi \mid s \not\models \Phi] \leq \beta \tag{3}$$

In addition, the probability should be low that $\mathcal{A}$ does not produce a definite answer. Let $s \vdash_\perp \Phi$ denote that $\mathcal{A}$ is *undecided*. We add

$$\Pr[s \vdash_\perp \Phi] \leq \gamma \tag{4}$$

to represent this requirement. Finally, $\mathcal{A}$ should always terminate with one of the three possible answers (accept, reject, or undecided):

$$\Pr[(s \vdash_\top \Phi) \vee (s \vdash_\bot \Phi) \vee (s \vdash_\perp \Phi)] = 1 \tag{5}$$

A model-checking algorithm that satisfies (2) through (5) is guaranteed to produce a correct answer with probability at least $1 - \alpha - \gamma$ when Φ holds and $1-\beta-\gamma$ when Φ does not hold. To make these probabilities high, α, β, and γ need to be low. If all three parameters are zero, then $\mathcal{A}$ is a deterministic algorithm for probabilistic model checking. If both $\alpha + \gamma$ and $\beta + \gamma$ are less than 0.5, but non-zero, then $\mathcal{A}$ is a randomized algorithm for probabilistic model checking.

Unfortunately, it is generally not possible, in practice, to satisfy all four conditions with low values for all three parameters. Next, we will discuss how these conditions are relaxed by current solution methods, and then we will present a new statistical solution method based on an alternative relaxation.

4 Current Solution Methods

Current solution methods, both numerical and statistical, can be seen as relying on a relaxation of (2) and (3) to become tractable. The reference point for error is changed from UTSL to UTSL$_\delta$ semantics, replacing (2) and (3) with:

$$\Pr[s \vdash_\bot \Phi \mid s \mathrel{|\!\approx^\delta_\top} \Phi] \leq \alpha \tag{6}$$

$$\Pr[s \vdash_\top \Phi \mid s \mathrel{|\!\approx^\delta_\bot} \Phi] \leq \beta \tag{7}$$

4.1 Statistical Hypothesis Testing

The predominant statistical solution method for verifying $\mathcal{P}_{\bowtie\theta}[\varphi]$ in a single state s is based on statistical *hypothesis testing*. This method was first proposed by Younes and Simmons [18] and further refined by Younes [16]. The approach always produces a definite result ($\gamma = 0$). This ensures a high probability of a correct answer when $s \models^{\delta}_{\top} \Phi$ or $s \models^{\delta}_{\bot} \Phi$ holds.

Let Φ be $\mathcal{P}_{\geq\theta}[\varphi]$, let p be the probability measure of the set of trajectories that start in s and satisfy φ, and let X_i be Bernoulli variates with $\Pr[X_i = 1] = p$. To verify Φ we test the hypothesis $H_0 : p \geq \theta + \delta$ against the alternative hypothesis $H_1 : p \leq \theta - \delta$ based on observations of X_i (the result of verifying φ over a sample trajectory starting in s). Note that H_0 corresponds to $s \models^{\delta}_{\top} \Phi$ and H_1 corresponds to $s \models^{\delta}_{\bot} \Phi$. If we take acceptance of H_0 to mean acceptance of Φ as true and acceptance of H_1 to mean rejection of Φ as false, then we can use *acceptance sampling* to verify Φ. Acceptance sampling is a well-established technique for statistical hypothesis testing. An acceptance sampling test with *strength* $\langle\alpha, \beta\rangle$ guarantees that H_1 is accepted with probability at most α when H_0 holds and H_0 is accepted with probability at most β when H_1 holds. Hence, we can use such a test to satisfy (6) and (7) for the verification of Φ.

Any acceptance sampling test with the prescribed strength can be used. A straightforward approach is to use a fixed number of observations $x_1, \ldots, x_n$ of the Bernoulli variates $X_1, \ldots, X_n$ and pick a constant c. If $\sum_{i=1}^{n} x_i$ is greater than c, then H_0 is accepted, otherwise H_1 is accepted. The pair $\langle n, c\rangle$ is called a *single sampling plan* [5]. The sum of n Bernoulli variates with parameter p has a binomial distribution with cumulative distribution function

$$F(c; n, p) = \sum_{i=0}^{c} \binom{n}{i} p^i (1-p)^{n-i} \quad . \tag{8}$$

Using a single sampling plan $\langle n, c\rangle$ we accept hypothesis H_1 with probability $F(c; n, p)$ and hypothesis H_0 with probability $1 - F(c; n, p)$. To achieve strength $\langle\alpha, \beta\rangle$ we need to choose n and c so that $F(c; n, \theta+\delta) \leq \alpha$ and $1-F(c; n, \theta-\delta) \leq \beta$. For optimal performance we choose n and c so that n is minimized. There is no closed-form solution for n, in general. Younes [16] describes an algorithm based on binary search that finds an optimal single sampling plan.

The sample size for a single sampling plan is fixed and therefore independent of the actual observations made. It is often possible to reduce the *expected* sample size required to achieve a desired test strength by taking the observations into account as they are made. This is called *sequential* acceptance sampling. Wald's [15] *sequential probability ratio test* (SPRT) is a particularly efficient sequential test. The reduction in expected sample size, compared to a single sampling plan, is often substantial, although there is no fixed upper bound on the sample size. The SPRT is carried out as follows. At the mth stage, i.e. after making m observations $x_1, \ldots, x_m$ we calculate the quantity

$$f_m = \prod_{i=1}^{m} \frac{\Pr[X_i = x_i \mid p = p_1]}{\Pr[X_i = x_i \mid p = p_0]} = \frac{p_1^{d_m}(1-p_1)^{m-d_m}}{p_0^{d_m}(1-p_0)^{m-d_m}} \quad , \tag{9}$$

where $d_m = \sum_{i=1}^{m} x_i$. Hypothesis H_0 is accepted if $f_m \leq \beta/(1-\alpha)$, and hypothesis H_1 is accepted if $f_m \geq (1-\beta)/\alpha$. Otherwise, additional observations are made until either termination condition is satisfied.

4.2 Statistical Estimation

An alternative statistical solution method, based on *estimation* instead of hypothesis testing, has been developed by Lassaigne and Peyronnet [13]. Hérault et al. [8] provide more details of this approach.

As before, let Φ be $\mathcal{P}_{\geq\theta}[\varphi]$ and p the probability measure of the set of trajectories that start in s and satisfy φ. This approach uses n observations $x_1, \ldots, x_n$ to compute an estimate of p: $\tilde{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The estimate is such that

$$\Pr\big[|\tilde{p} - p| < \delta\big] \geq 1 - \alpha \ . \tag{10}$$

Using a result derived by Hoeffding [10, Theorem 1], it can be shown that

$$n = \left\lceil \frac{1}{2\delta^2} \log \frac{2}{\alpha} \right\rceil \tag{11}$$

is sufficient to satisfy (10). If we accept Φ as true when $\tilde{p} \geq \theta$ and reject Φ as false otherwise, then it follows from (10) that the answer is correct with probability at least $1-\alpha$ if either $s \models^{\delta}_{\top} \Phi$ or $s \models^{\delta}_{\bot} \Phi$ holds. Consequently, the verification procedure satisfies (6) and (7) with $\beta = \alpha$. As with the solution method based on hypothesis testing, a definite answer is always generated ($\gamma = 0$).

To compare the estimation-based approach with the approach based on hypothesis testing, let $c = \lfloor n\theta + 1 \rfloor$ and $d = n\tilde{p} = \sum_{i=1}^{n} x_i$. It should be clear that $\tilde{p} \geq \theta \iff d > c$. This means that the estimation-based approach can be interpreted as a single sampling plan $\langle n, c \rangle$. It follows that the approach proposed by Younes and Simmons [18], when using a single sampling plan, *will always be at least as efficient as the estimation-based approach*. Typically, it will be more efficient because: (i) the sample size is derived using the true underlying distribution, (ii) c is not restricted to be $\lfloor n\theta + 1 \rfloor$, and (iii) $\beta \neq \alpha$ can be accommodated. The last property, in particular, is important when dealing with conjunctive and nested probabilistic statements. The advantage of hypothesis testing is demonstrated in Table 1. Note, also, that the SPRT often can be used to improve efficiency even further for the approach based on hypothesis testing.

4.3 Numerical Transient Analysis

To verify the formula $\mathcal{P}_{\bowtie\theta}[\varphi]$ in some state s we can compute p—the probability measure of the set of trajectories that start in s and satisfy φ—numerically and test if $p \bowtie \theta$ holds.

For time-bounded properties ($\varphi = \Phi\ \mathcal{U}^{[0,\tau]}\ \Psi$), which are the focus of this paper, such numerical computation is primarily feasible for Markov chains. Let $\mathcal{M}$ be a continuous-time Markov chain. First, as initially proposed by Baier

Table 1. Sample sizes for estimation and optimal single sampling plan ($\delta = 10^{-2}$)

θ	α	β	n_{est}	n_{opt}	$n_{\text{est}}/n_{\text{opt}}$
0.5	10^{-2}	10^{-2}	26,492	13,527	1.96
0.5	10^{-8}	10^{-2}	95,570	39,379	2.43
0.5	10^{-8}	10^{-8}	95,570	78,725	1.21
0.9	10^{-2}	10^{-2}	26,492	4,861	5.45
0.9	10^{-8}	10^{-2}	95,570	13,982	6.84
0.9	10^{-8}	10^{-8}	95,570	28,280	3.38

et al. [2], the problem is reduced to *transient analysis* of a modified Markov chain $\mathcal{M}'$, where all states in $\mathcal{M}$ satisfying $\neg\Phi \vee \Psi$ have been made absorbing. Now, p is equal to the probability of occupying a state satisfying Ψ at time τ in model $\mathcal{M}'$. This probability can be computed using a technique called *uniformization*, originally proposed by Jensen [11]. Let $\mathbf{Q}$ be the generator matrix of $\mathcal{M}'$, $q = \max_i -q_{ii}$, and $\mathbf{P} = \mathbf{I} + \mathbf{Q}/q$. Then p can be expressed as follows:

$$p = \boldsymbol{\mu}_0 \cdot \sum_{k=0}^{\infty} e^{-q\cdot\tau} \frac{(q\cdot\tau)^k}{k!} \mathbf{P}^k \cdot \boldsymbol{\chi}_\Psi \tag{12}$$

Here, $\boldsymbol{\mu}_0$ is a 0-1 row vector with a 1 in the column for the initial state s and $\boldsymbol{\chi}_\Psi$ is a 0-1 column vector with a 1 in each row corresponding to a state that satisfies Ψ.

In practice, the infinite summation in (12) is truncated by using the techniques of Fox and Glynn [6], so that the truncation error is bounded by ϵ. If $\tilde{p}$ is the computed probability, then $\tilde{p} \leq p \leq \tilde{p}+\epsilon$. It follows that by accepting $\mathcal{P}_{\bowtie\theta}[\varphi]$ as true if $\tilde{p} + \epsilon/2 \bowtie \theta$ and rejecting the formula as false otherwise, the numerical solution method satisfies (6) and (7) with $\delta = \epsilon/2$ and $\alpha = \beta = 0$. As with the statistical solution methods, a definite answer is always given ($\gamma = 0$). This shows that numerical and statistical solution methods for probabilistic model checking can, indeed, be viewed as solving the same problem, i.e. UTSL$_\delta$ model checking rather than UTSL model checking. Statistical solution methods are truly randomized algorithms for UTSL$_\delta$ model checking.

When using uniformization to verify $\mathcal{P}_{\geq\theta}[\Phi\ \mathcal{U}^{[0,\tau]}\ \Psi]$, it is actually possible to know when we cannot make an informed decision. If we accept the formula as true when $\tilde{p} \geq \theta$, reject it as false when $\tilde{p} + \epsilon < \theta$, and report "undecided" otherwise, then (2) and (3) can be satisfied with $\alpha = \beta = 0$. This alternative implementation of the numerical solution method no longer satisfies (4). That condition is replaced by $\Pr[s \vdash_\perp \Phi \mid (s \models \Phi) \vee (s \mathrel{\not\approx}^\delta_\perp \Phi)] = 0$, with $\delta = \epsilon$, for $\mathcal{P}_{\geq\theta}[\varphi]$ without nested probabilistic operators, and

$$\Pr[s \vdash_\perp \Phi \mid (s \mathrel{\not\approx}^\delta_\top \Phi) \vee (s \mathrel{\not\approx}^\delta_\perp \Phi)] = 0 \tag{13}$$

for an arbitrary formula Φ. The use of undecided results with numerical methods for probabilistic model checking has been suggested by Hermanns et al. [9], although it is not clear if any tool implements this approach. The leading tool for probabilistic model checking, PRISM [12], does not produce undecided results.

5 Statistical Solution Method with Undecided Results

Existing statistical solution methods provide no meaningful error bounds if neither $s \not\models^{\delta}_{\top} \Phi$ nor $s \not\models^{\delta}_{\bot} \Phi$ holds. This section presents a new statistical solution method that satisfies (2) and (3), so whenever a definite result is given the probability of error is bounded. We accomplish this by allowing an undecided result with some probability. The goal is to replace (4) with

$$\Pr[s \vdash_{\mathtt{I}} \Phi \mid (s \not\models^{\delta}_{\top} \Phi) \vee (s \not\models^{\delta}_{\bot} \Phi)] \leq \gamma \ . \tag{14}$$

5.1 Probabilistic Operator without Nesting

Let Φ be $\mathcal{P}_{\geq\theta}[\varphi]$ without nested probabilistic operators ($\mathcal{P}_{\leq\theta}[\varphi]$ is analogous). To satisfy (2), (3), and (14) simultaneously using a sample of size n we pick two constants c_0 and c_1 such that $0 \leq c_1 < c_0 < n$ and the following conditions hold:

$$F(c_1; n, \theta) \leq \alpha \tag{15}$$

$$1 - F(c_1; n, \theta - \delta) \leq \gamma \tag{16}$$

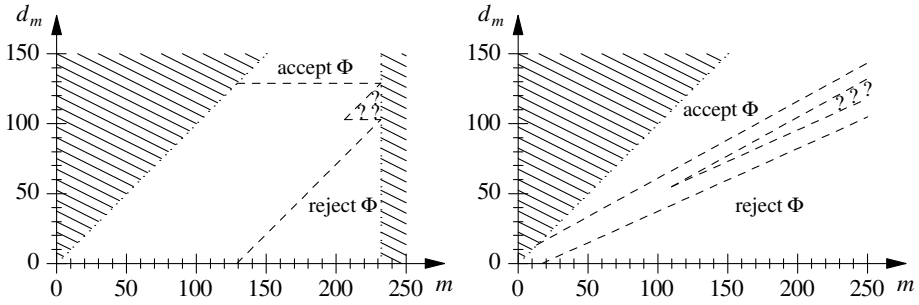
$$1 - F(c_0; n, \theta) \leq \beta \tag{17}$$

$$F(c_0; n, \theta + \delta) \leq \gamma \tag{18}$$

Let $d = \sum_{i=1}^{n} x_i$. We accept Φ as true if $d > c_0$, we reject Φ as false if $d \leq c_1$, otherwise ($c_1 < d \leq c_0$) the result is undecided.

The procedure just given can be interpreted as using *two simultaneous* acceptance sampling tests. The first is used to tests $H_0^{\bot} : p \geq \theta$ against $H_1^{\bot} : p \leq \theta - \delta$ with strength $\langle \alpha, \gamma \rangle$. The second is used to tests $H_0^{\top} : p \geq \theta + \delta$ against $H_1^{\top} : p \leq \theta$ with strength $\langle \gamma, \beta \rangle$. $H_0^{\top}$ represents acceptance of Φ as true, $H_1^{\bot}$ represents rejection of Φ as false, and the remaining two hypotheses represent an undecided result. Combining the results from both tests, Φ is accepted as true if both $H_0^{\top}$ and $H_0^{\bot}$ are accepted, Φ is rejected as false if both $H_1^{\top}$ and $H_1^{\bot}$ are accepted, otherwise the result is undecided. Of course, this means that we do not need to use hypothesis testing with fixed-size samples. We could use any acceptance sampling plans with the prescribed strengths and combine their results as specified. In particular, we could use the SPRT to reduce the expected sample size.

Graphical representations of two acceptance sampling tests with undecided results are shown in Fig. 1 for $\theta = 0.5$, $\delta = 0.1$, $\alpha = 0.04$, $\beta = 0.08$, and $\gamma = 0.1$. The horizontal axis represents the number of observations and the vertical axis represents the number of positive observations. Figure 1(a) represents a sequential version of a single sampling plan with $n = 232$, $c_0 = 128$, $c_1 = 102$. The line $d_m = 129$ is the boundary for acceptance of Φ. There is a line for rejection of Φ and two lines defining the boundary of the region that represents an undecided result. Figure 1(b) shows the corresponding decision boundaries for the SPRT.



(a) Sequential single sampling plan (b) SPRT

Fig. 1. Graphical representation of acceptance sampling tests

5.2 Composite Formulae

For a negation $\neg\Phi$ we have $s \vdash_{\mathtt{T}} \neg\Phi \iff s \vdash_{\mathtt{T}} \Phi$. Hence, if we can satisfy (14) for Φ, then we have the same bound, γ, on the probability of an undecided result for the negation of Φ. The roles of α and β are reversed for negation (cf. Younes and Simmons [18] and Younes [16]).

For a conjunction $\Phi \wedge \Psi$ we get the following general bound on the probability of an undecided result (see Appendix A for proof):

$$\begin{aligned}&\Pr[s \vdash_{\mathtt{T}} \Phi \wedge \Psi \mid (s \not\approx^{\delta}_{\top} \Phi \wedge \Psi) \vee (s \not\approx^{\delta}_{\bot} \Phi \wedge \Psi)] \\ &\quad \leq \max(\gamma_\Phi + \gamma_\Psi, \gamma_\Phi + \beta_\Phi, 2\gamma_\Psi + \beta_\Psi)\end{aligned} \tag{19}$$

In practice, the dependence on β_Φ and β_Ψ can be disregarded. We have β_Φ in (19) because $\Pr[s \vdash_{\top} \Phi \mid s \not\approx^{\delta}_{\bot} \Phi] \leq \Pr[s \vdash_{\top} \Phi \mid s \not\models \Phi] \leq \beta_\Phi$ (similarly for β_Ψ), but $\Pr[s \vdash_{\top} \Phi \mid s \not\approx^{\delta}_{\bot} \Phi]$ is typically negligible compared to $\Pr[s \vdash_{\top} \Phi \mid s \not\models \Phi]$. Let $\gamma' = \gamma_\Phi = \gamma_\Psi$. Then (19) can, *for all practical purposes*, be replaced by

$$\Pr[s \vdash_{\mathtt{T}} \Phi \wedge \Psi \mid (s \not\approx^{\delta}_{\top} \Phi \wedge \Psi) \vee (s \not\approx^{\delta}_{\bot} \Phi \wedge \Psi)] \leq 2\gamma' \ . \tag{20}$$

Consequently, if we want to ensure at most a γ probability of an undecided result for $\Phi \wedge \Psi$, and we use the same bound for both conjuncts, then we can use $\gamma/2$ when verifying Φ and Ψ. For a conjunction of size n, the symmetric bound for each conjunct could be set to γ/n.

To satisfy (2) we should choose α_Φ and α_Ψ such that $\alpha_\Phi + \alpha_\Psi \leq \alpha$ (cf. Younes and Simmons [18][1]):

$$\begin{aligned}&\Pr[(s \vdash_{\bot} \Phi) \vee (s \vdash_{\bot} \Psi) \mid (s \models \Phi) \wedge (s \models \Psi)] \\ &\quad \leq \Pr[s \vdash_{\bot} \Phi \mid s \models \Phi] + \Pr[s \vdash_{\bot} \Psi \mid s \models \Psi] \leq \alpha_\Phi + \alpha_\Psi\end{aligned} \tag{21}$$

[1] Younes [16] gives the bound $\min(\alpha_\Phi, \alpha_\Psi)$, but this is a bound only *for each individual way* of rejecting a conjunction as false. The result due to Younes and Simmons [18] and reproduced here bounds the probability of rejecting a conjunction *in any way*.

Similar to γ, we can use α/n when verifying the parts of a conjunction of size n. Unlike γ, however, this does not involve any approximation. To satisfy (3), it suffices to use the same error bound, β, for the individual conjuncts:

$$\begin{aligned}&\Pr[(s \vdash_\tau \Phi) \wedge (s \vdash_\tau \Psi) \mid (s \not\models \Phi) \vee (s \not\models \Psi)] \\ &\quad \leq \max(\Pr[s \vdash_\tau \Phi \mid s \not\models \Phi], \Pr[s \vdash_\tau \Psi \mid s \not\models \Psi]) \leq \max(\beta_\Phi, \beta_\Psi) \end{aligned} \tag{22}$$

5.3 Nested Probabilistic Statements

We use acceptance sampling to verify probabilistic statements. The observations that are used by the acceptance sampling test correspond to the verification of a path formula, φ, over sample trajectories. If φ contains probabilistic statements, then the observations may be incorrect or undecided. We assume that φ can be verified with parameters α_φ, β_φ, and γ_φ. This can be accomplished by treating the path formula as a large disjunction of conjunctions, as described by Younes and Simmons [18, p. 231] and Younes [16, p. 78].

It remains to show how to use the verification results for φ to verify a probabilistic statement, $\Phi = \mathcal{P}_{\geq\theta}[\varphi]$, so that (2), (3), and (14) are satisfied. This can be accomplished by a single sampling plan with n, c_0, and c_1 chosen to satisfy the following conditions (see Appendix B for proof):

$$F(c_1; n, \theta(1-\alpha_\varphi)) \leq \alpha \tag{23}$$

$$1 - F(c_1; n, (\theta-\delta) + (1-(\theta-\delta))(1-\gamma_\varphi-\beta_\varphi)) \leq \gamma \tag{24}$$

$$1 - F(c_0; n, \theta + (1-\theta)\beta_\varphi) \leq \beta \tag{25}$$

$$F(c_0; n, (\theta+\delta)(1-\gamma_\varphi-\alpha_\varphi)) \leq \gamma \tag{26}$$

This assumes that Φ is accepted as true when more than c_0 positive observations are made, Φ is rejected as false when at most c_1 observations are non-positive (i.e., negative *or* undecided), and the result is undecided otherwise.

Compared to (15) through (18) for acceptance sampling without nested probabilistic operators, the only difference is that the probability thresholds have been modified. The indifference regions of the two acceptance sampling tests have been made narrower to account for the possibility of erroneous or undecided observations. We can use the same modification with the SPRT.

It should be noted that α_φ, β_φ and γ_φ can be chosen independently of α, β, and γ. The choice of parameters for the verification of φ is restricted only by the following conditions:

$$(\theta-\delta) + (1-(\theta-\delta))(1-\gamma_\varphi-\beta_\varphi) < \theta(1-\alpha_\varphi) \tag{27}$$

$$\theta + (1-\theta)\beta_\varphi < (\theta+\delta)(1-\gamma_\varphi-\alpha_\varphi) \tag{28}$$

The choice of α_φ, β_φ, and γ_φ can have a significant impact on performance (cf. the discussion by Younes [16] regarding the impact of observation error on performance for the standard statistical solution method).

6 Complexity of Statistical Solution Methods

The time complexity of any statistical solution method for probabilistic model checking can be understood in terms of two main factors: the sample size and the length of sample trajectories. The sample size depends on the method used for verifying probabilistic statements and the desired strength. The length of trajectories depends on model characteristics and the property that is being verified. An additional factor is simulation effort, which can be both model and implementation dependent.

Consider the formula $\mathcal{P}_{\bowtie\theta}[\Phi\ \mathcal{U}^{[0,\tau]}\ \Psi]$ without nested probabilistic operators. Let q be the expected number of state transitions per time unit, let m be the simulation effort per state transition, and let N be the sample size. The time complexity of statistical probabilistic model checking for the given formula is $O(q \cdot \tau \cdot m \cdot N)$. The sample size, N, is the only factor that varies between different statistical solution methods, regardless of implementation details.

If we use a single sampling plan with strength $\langle\alpha, \beta\rangle$ and indifference region of half-width δ, then N is roughly proportional to $\log \alpha$ and $\log \beta$ and inversely proportional to δ^2 [16, p. 23]. We have shown in this paper that the approach based on statistical estimation described by Hérault et al. [8] never uses a smaller sample size than a single sampling plan, given the same parameters, and often uses a much larger sample size. Using the SPRT instead of a single sampling plan can reduce the expected sample size by orders of magnitude in most cases, although the SPRT is not guaranteed always to be more efficient (this is well known in the statistics literature; Younes [16] provides examples of this in the context of model checking). The new statistical approach presented in this paper, which can produce undecided results, has the same time complexity as the old statistical solution method. Given the same α, β, and δ, the new method will require a larger sample size because it is based on acceptance sampling with indifference regions of half-width $\delta/2$, instead of δ for the old method.

Results presented by Sen et al. [14] make it seem as if single sampling plans consistently outperform the SPRT, but this is due to poorly designed experiments. Sen et al. manually selected the sample sizes for their single sampling plans, guided by a desire to achieve a low p-value (K. Sen, personal communication, May 20, 2004). The selected sample sizes are not sufficient, however, to achieve the same strength as used to produce the results for the SPRT reported by Younes et al. [17], on which they base their comparison. All their empirical evaluation really proves is that a smaller sample size results in shorter verification time—which should surprise no one—but the casual reader may be misled into believing that Sen et al. have devised a novel statistical solution method.

7 Empirical Evaluation

The performance of our new statistical solution method is similar to that of the previous statistical solution method, which has been studied by Younes et al. [17]. We limit the empirical evaluation in this paper to a brief study of the effect that the parameter γ has on performance.

Figure 2 plots the *expected* sample size, as a function of the (unknown) probability p that a path formula holds, for the SPRT and a *sequential* single sampling plan (SSSP) with different parameter choices ($\theta = 0.5$, $\delta = 0.1$, $\alpha_\Delta = 0.004$, $\alpha_\nabla = 0.04$, $\beta_\Delta = 0.008$, $\beta_\nabla = 0.08$, $\gamma_\Delta = 0.01$, and $\gamma_\nabla = 0.1$). The expected sample size is low outside of the indifference region (gray area), especially for the SPRT, and peaks in the indifference region. Note the drop in expected sample size at the threshold θ where an undecided result is given with high probability. The expected sample size, as a function of p, will be similar for other parameter values, with the SPRT almost always outperforming a (sequential) single sampling plan by a wide margin.

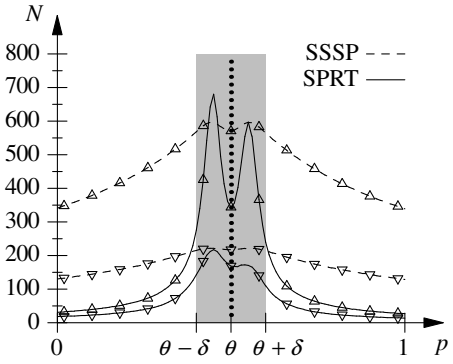


Fig. 2. Expected sample size

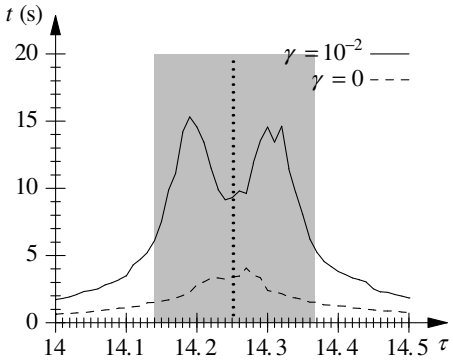


Fig. 3. Verification time

Now, consider the model-checking problem for an n-station symmetric polling system used by Younes et al. [17]. Each station has a single-message buffer and the stations are attended by a single server in cyclic order. The server begins by polling station 1. If there is a message in the buffer of station 1, the server starts serving that station. Once station i has been served, or if there is no message at station i when it is polled, the server starts polling station $i+1$ (or 1 if $i = n$). We verify the property $m_1{=}1 \rightarrow \mathcal{P}_{\geq 0.5}[\top\ \mathcal{U}^{[0,\tau]}\ \mathit{poll}_1]$, which states that if station 1 is full, then it is polled within τ time units with probability at least 0.5. We do so in the state where station 1 has just been polled and all buffers are full.

Figure 3 plots the verification time for the symmetric polling system problem ($n = 10$), as a function of the formula time bound τ, averaged over 100 runs. The plot shows the verification time for the new solution method with $\gamma = 10^{-2}$ (solid curve) and the old solution method without undecided results (dashed curve); $2\delta = 10^{-2}$ and $\alpha = \beta = 10^{-2}$ in both cases. The verification time is lower for the standard statistical solution method, but it produces more erroneous results. Table 2 shows the number of times a certain result is produced for seven different values of τ. The new statistical solution method does not produce an erroneous result in any of the experiments, while the error probability is high for the standard statistical solution method for values of τ close to 14.251 (where

the value of the verified property goes from false to true). Higher reliability in the results are obtained at the cost of efficiency.

Table 2. Result distribution *with* (bottom) and *without* (top) undecided results

result	**14.10**	**14.15**	**14.20**	**14.25**	**14.30**	**14.35**	**14.40**
accept	0	3	9	50	88	97	100
reject	100	97	91	50	12	3	0
accept	0	0	0	0	32	99	100
reject	100	99	42	1	0	0	0
undecided	0	1	58	99	68	1	0

8 Discussion

We have presented a framework for expressing correctness guarantees of model-checking algorithms. Using this framework, we have shown how current solution methods for probabilistic model checking are related. In particular, we have shown that Younes and Simmons' [18] statistical solution method based on hypothesis testing has clear benefits over Hérault et al.'s [8] estimation-based approach, and that numerical and statistical solution methods can be interpreted as solving the same *relaxed* model-checking problems. In addition, we have presented a new statistical solution method that bounds the probability of error under all circumstances. This is accomplished by permitting undecided results, and we have shown how to guarantee bounds for the probability of getting an undecided result for any formula.

References

1. Alur, R. and Dill, D. L. A theory of timed automata. *Theoretical Computer Science*, 126(2):183–235, 1994.
2. Baier, C., Haverkort, B. R., Hermanns, H., and Katoen, J.-P. Model checking continuous-time Markov chains by transient analysis. In *Proc. 12th International Conference on Computer Aided Verification*, volume 1855 of *LNCS*, pages 358–372. Springer, 2000.
3. Baier, C., Haverkort, B. R., Hermanns, H., and Katoen, J.-P. Model-checking algorithms for continuous-time Markov chains. *IEEE Transactions on Software Engineering*, 29(6):524–541, 2003.
4. Clarke, E. M. and Emerson, E. A. Design and synthesis of synchronization skeletons using branching time temporal logic. In *Proc. 1981 Workshop on Logics of Programs*, volume 131 of *LNCS*, pages 52–71. Springer, 1982.
5. Duncan, A. J. *Quality Control and Industrial Statistics*. Richard D. Irwin, fourth edition, 1974.
6. Fox, B. L. and Glynn, P. W. Computing Poisson probabilities. *Communications of the ACM*, 31(4):440–445, 1988.

7. Hansson, H. and Jonsson, B. A logic for reasoning about time and reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994.
8. Hérault, T., Lassaigne, R., Magniette, F., and Peyronnet, S. Approximate probabilistic model checking. In *Proc. 5th International Conference on Verification, Model Checking, and Abstract Interpretation*, volume 2937 of *LNCS*, pages 73–84. Springer, 2004.
9. Hermanns, H., Katoen, J.-P., Meyer-Kayser, J., and Siegle, M. A tool for model-checking Markov chains. *International Journal on Software Tools for Technology Transfer*, 4(2):153–172, 2003.
10. Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
11. Jensen, A. Markoff chains as an aid in the study of Markoff processes. *Skandinavisk Aktuarietidskrift*, 36:87–91, 1953.
12. Kwiatkowska, M., Norman, G., and Parker, D. Probabilistic symbolic model checking with PRISM: A hybrid approach. *International Journal on Software Tools for Technology Transfer*, 6(2):128–142, 2004.
13. Lassaigne, R. and Peyronnet, S. Approximate verification of probabilistic systems. In *Proc. 2nd Joint International PAPM-PROBMIV Workshop*, volume 2399 of *LNCS*, pages 213–214. Springer, 2002.
14. Sen, K., Viswanathan, M., and Agha, G. Statistical model checking of black-box probabilistic systems. In *Proc. 16th International Conference on Computer Aided Verification*, volume 3114 of *LNCS*, pages 202–215. Springer, 2004.
15. Wald, A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.
16. Younes, H. L. S. *Verification and Planning for Stochastic Processes with Asynchronous Events*. PhD thesis, Computer Science Department, Carnegie Mellon University, 2005. CMU-CS-05-105.
17. Younes, H. L. S., Kwiatkowska, M., Norman, G., and Parker, D. Numerical vs. statistical probabilistic model checking: An empirical study. In *Proc. 10th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, volume 2988 of *LNCS*, pages 46–60. Springer, 2004.
18. Younes, H. L. S. and Simmons, R. G. Probabilistic verification of discrete event systems using acceptance sampling. In *Proc. 14th International Conference on Computer Aided Verification*, volume 2404 of *LNCS*, pages 223–235. Springer, 2002.

A Proof: Conjunction

For a conjunction $\Phi \wedge \Psi$, there are two ways to get an undecided result: (i) $s \vdash_{\perp} \Phi$ and $s \nvdash_{\perp} \Psi$; (ii) $s \vdash_{\top} \Phi$ and $s \vdash_{\perp} \Psi$. We can bound the probability of each case occurring, given that $s \not\approx^{\delta}_{\top} \Phi \wedge \Psi$, $s \not\approx^{\delta}_{\perp} \Phi$, or $s \not\approx^{\delta}_{\perp} \Psi$ holds:

$$\begin{aligned}
&\Pr[(s \vdash_{\perp} \Phi) \wedge (s \nvdash_{\perp} \Psi) \mid (s \not\approx^{\delta}_{\top} \Phi) \wedge (s \not\approx^{\delta}_{\top} \Psi)] \\
&\quad \leq \min(\Pr[s \vdash_{\perp} \Phi \mid s \not\approx^{\delta}_{\top} \Phi], \Pr[s \nvdash_{\perp} \Psi \mid s \not\approx^{\delta}_{\top} \Psi]) \leq \min(\gamma_{\Phi}, 1) = \gamma_{\Phi} \quad (29) \\
&\Pr[(s \vdash_{\top} \Phi) \wedge (s \vdash_{\perp} \Psi) \mid (s \not\approx^{\delta}_{\top} \Phi) \wedge (s \not\approx^{\delta}_{\top} \Psi)] \\
&\quad \leq \min(\Pr[s \vdash_{\top} \Phi \mid s \not\approx^{\delta}_{\top} \Phi], \Pr[s \vdash_{\perp} \Psi \mid s \not\approx^{\delta}_{\top} \Psi]) \leq \min(1, \gamma_{\Psi}) = \gamma_{\Psi} \quad (30) \\
&\Pr[(s \vdash_{\perp} \Phi) \wedge (s \nvdash_{\perp} \Psi) \mid s \not\approx^{\delta}_{\perp} \Phi] \\
&\quad \leq \min(\Pr[s \vdash_{\perp} \Phi \mid s \not\approx^{\delta}_{\perp} \Phi], \Pr[s \nvdash_{\perp} \Psi]) \leq \min(\gamma_{\Phi}, 1) = \gamma_{\Phi} \quad (31) \\
&\Pr[(s \vdash_{\top} \Phi) \wedge (s \vdash_{\perp} \Psi) \mid s \not\approx^{\delta}_{\perp} \Phi] \\
&\quad \leq \min(\Pr[s \vdash_{\top} \Phi \mid s \not\approx^{\delta}_{\perp} \Phi], \Pr[s \vdash_{\perp} \Psi]) \leq \min(\beta_{\Phi}, 1) = \beta_{\Phi} \quad (32) \\
&\Pr[(s \vdash_{\perp} \Phi) \wedge (s \nvdash_{\perp} \Psi) \mid s \not\approx^{\delta}_{\perp} \Psi] \\
&\quad \leq \min(\Pr[s \vdash_{\perp} \Phi], \Pr[s \nvdash_{\perp} \Psi \mid s \not\approx^{\delta}_{\perp} \Psi]) \leq \gamma_{\Psi} + \beta_{\Psi} \quad (33) \\
&\Pr[(s \vdash_{\top} \Phi) \wedge (s \vdash_{\perp} \Psi) \mid s \not\approx^{\delta}_{\perp} \Psi] \\
&\quad \leq \min(\Pr[s \vdash_{\top} \Phi], \Pr[s \vdash_{\perp} \Psi \mid s \not\approx^{\delta}_{\perp} \Psi]) \leq \gamma_{\Psi} \quad (34)
\end{aligned}$$

Combining (29) and (30), (31) and (32), and (33) and (34) gives us:

$$\Pr[s \vdash_{\perp} \Phi \wedge \Psi \mid (s \not\approx^{\delta}_{\top} \Phi) \wedge (s \not\approx^{\delta}_{\top} \Psi)] \leq \gamma_{\Phi} + \gamma_{\Psi} \quad (35)$$

$$\Pr[s \vdash_{\perp} \Phi \wedge \Psi \mid s \not\approx^{\delta}_{\perp} \Phi] \leq \gamma_{\Phi} + \beta_{\Phi} \quad (36)$$

$$\Pr[s \vdash_{\perp} \Phi \wedge \Psi \mid s \not\approx^{\delta}_{\perp} \Psi] \leq 2\gamma_{\Psi} + \beta_{\Psi} \quad (37)$$

By taking the maximum of (35) through (37) we get (19). □

B Proof: Nested Probabilistic Statements

Let X_i, Y_i, $\hat{Z}_i$, and $\check{Z}_i$ be random variables such that, for any sample trajectory σ, we have the following:

$$\begin{aligned}
X_i = 1 &\iff \sigma, \tau \models \varphi & \qquad Y_i = 1 &\iff \sigma, \tau \not\approx^{\delta}_{\top} \varphi \\
X_i = 0 &\iff \sigma, \tau \not\models \varphi & \qquad Y_i = 0 &\iff \sigma, \tau \not\approx^{\delta}_{\perp} \varphi \\
\hat{Z}_i = 1 &\iff \sigma, \tau \nvdash_{\perp} \varphi & \qquad \check{Z}_i = 1 &\iff \sigma, \tau \vdash_{\top} \varphi \\
\hat{Z}_i = 0 &\iff \sigma, \tau \vdash_{\perp} \varphi & \qquad \check{Z}_i = 0 &\iff \sigma, \tau \nvdash_{\top} \varphi
\end{aligned}$$

Note that all but Y_i are Bernoulli variates. If we verify φ over σ with parameters α_φ, β_φ, and γ_φ, then the following conditions are guaranteed to hold:

$$\Pr[\hat{Z}_i = 0 \mid X_i = 1] \leq \alpha_\varphi \tag{38}$$

$$\Pr[\check{Z}_i = 1 \mid X_i = 0] \leq \beta_\varphi \tag{39}$$

$$\Pr[\hat{Z}_i = 1 \mid Y_i = 0] \leq \gamma_\varphi + \beta_\varphi \tag{40}$$

$$\Pr[\check{Z}_i = 0 \mid Y_i = 1] \leq \gamma_\varphi + \alpha_\varphi \tag{41}$$

Let $\Pr[X_i = 1] = p$, $\Pr[Y_i = 1] = p_\top$, and $\Pr[Y_i = 0] = p_\bot$. From (38) through (41) and the formula of total probability we can derive the following bounds:

$$p(1 - \alpha_\varphi) \leq \Pr[\hat{Z}_i] \leq 1 - p_\bot(1 - \gamma_\varphi - \beta_\varphi) \tag{42}$$

$$p_\top(1 - \gamma_\varphi - \alpha_\varphi) \leq \Pr[\check{Z}_i] \leq p + (1 - p)\beta_\varphi \tag{43}$$

Let $\check{z}_i$ denote an observation of $\check{Z}_i$ and $\hat{z}_i$ an observation of $\hat{Z}_i$. Given the sample $\check{z}_1, \ldots, \check{z}_n, \hat{z}_1, \ldots, \hat{z}_n$, compute $\check{d} = \sum_{i=1}^n \check{z}_i$ and $\hat{d} = \sum_{i=1}^n \hat{z}_i$. The probabilistic formula $\Phi = \mathcal{P}_{\geq\,\theta}[\varphi]$ is accepted as true in state s if $\check{d} > c_0$, Φ is rejected as false in s if $\hat{d} \leq c_1$, and the result is undecided otherwise.

Let $\Pr[\hat{Z}_i = 1] = \hat{p}$ and $\Pr[\check{Z}_i = 1] = \check{p}$. The probability of rejection is $F(c_1; n, \hat{p})$. Since $F(c; n, p)$ is a non-increasing function of p in the interval $[0, 1]$, we have $F(c_1; n, \hat{p}) \leq F(c_1; n, p(1 - \alpha_\varphi))$. The probability of rejection should be at most α when $p \geq \theta$, which gives us (23).

The probability of acceptance is $1 - F(c_0; n, \check{p})$. This is at most $1 - F(c_0; n, p + (1 - p)\beta_\varphi)$. The probability of acceptance should be at most β when $p < \theta$, which gives us (25).

Next, we have $1 - F(c_1; n, \hat{p}) \leq 1 - F(c_1; n, 1 - p_\bot(1 - \gamma_\varphi - \beta_\varphi))$. This should be at most γ when $p_\bot \geq 1 - (\theta - \delta)$, which gives us (24).

Finally, we have $F(c_0; n, \check{p}) \leq F(c_0; n, p_\top(1 - \gamma_\varphi - \alpha_\varphi))$. This should be at most γ when $p_\top \geq \theta + \delta$, which gives us (26). □