Verification of Tail Distribution Bounds in a Theorem Prover

Osman Hasan and Sofiène Tahar

Department of Electrical and Computer Engineering,
Concordia University,
Montreal, Quebec, H3G 1M8, Canada

Abstract.
In the field of probabilistic analysis, bounding the tail distribution is a major tool for estimating the failure probability of systems. In this paper, we present the verification of Markov's and Chebyshev's inequalities for discrete random variables using the HOL theorem prover. The formally verified Markov and Chebyshev's inequalities allow us to precisely reason about tail distribution bounds for probabilistic systems within the core of a higher-order-logic theorem prover and thus prove to be quite useful for the analysis of systems used in safety-critical domains, such as space, medicine and military. For illustration purposes, we show how we can obtain bounds on the tail distribution of the Coupon Collector's problem in HOL.

Keywords: Chebyshev's Inequality; Higher-Order-Logic; HOL Theorem Prover; Markov's Inequality; Probabilistic Analysis.
PACS: 03B15, 03B35, 60A05, 68T15.

INTRODUCTION

Probability theory has become a tool of fundamental importance in modern science and engineering. The random and unpredictable elements are mathematically modeled by appropriate random variables and the performance and reliability issues are judged based on the corresponding probabilistic properties. Numbers such as mean or variance, which provide valuable information about random variables, are then used for decision making. One of the major advantages of using mean and variance is that they also allow us to obtain bounds on the tail distribution, which is the probability that a random variable assumes values that are far from its expectation or mean value. These tail bounds are usually calculated using the Markov's or the Chebyshev's inequalities [1]. Because of widespread interest in failure probabilities, these inequalities have now become one of the core techniques in modern probabilistic analysis.

The state-of-the-art in conducting probabilistic analysis is computer simulation [2], where the main idea is to approximately answer a query on a probability distribution by analyzing a large number of samples. The simulation approach is easy to use as most of the analysis can be automated and really shines in handling problems that cannot be solved analytically. On the other hand, the results are usually inaccurate and large problems cannot be handled because of enormous CPU time requirements. McCullough [3, 4] proposed a collection of intermediate-level tests for assessing the numerical reliability of simulation based probabilistic analysis tools and uncovered flaws in some of the mainframe statistical packages. An alternative is to use probabilistic model checking [5, 6], which is a formal state-based approach. Due to the formal nature of the models and analysis techniques, the results are always accurate but, like traditional model-checking, this approach is limited to systems that can be expressed as a probabilistic finite state machine and also suffers from the state-space explosion problem [7].

The inaccuracy of probabilistic analysis results and the inability to handle some specific cases poses a serious problem when a safety-critical section of the system is being analyzed. To overcome these limitations, we propose to use higher-order-logic theorem proving for the probabilistic analysis of safety-critical sections of the system. Higher-order logic is a system of deduction with a precise semantics and can be used for the development of almost all classical mathematical theories. Interactive theorem proving is the field of computer science and mathematical logic concerned with computer based formal proof tools that require some sort of human assistance. Due to the high expressive nature of the higher-order-logic and the inherent soundness of interactive theorem proving, this approach can be used to conduct error free probabilistic analysis at the cost of significant user interaction.

In order to build upon an existing higher-order-logic formalization of some probability theory [8], we have selected the HOL theorem prover [9] as our platform. HOL has already been successfully used to formalize discrete [8] and continuous [10] random variables and verify their probability distribution [11] and expectation properties [12]. In this paper, we extend the HOL libraries for probabilistic analysis with the proofs of the Markov's and the Chebyshev's inequalities for discrete random variables, which allows us to reason about tail distribution bounds within the HOL theorem prover and thus enhance the capabilities of HOL as a successful probabilistic analysis framework. We first present the HOL definitions for expectation, variance and standard deviation, which are in turn used to express Markov's and Chebyshev's inequalities in HOL. A summary of formal proofs for these properties is given in this paper and more details can be found in [13]. In order to illustrate the practical effectiveness of our approach, we then utilize the above results in HOL to obtain bounds on the tail distribution for the Coupon Collector's problem [14], which is a well known commercially used algorithm in computer science.

FORMALIZATION OF EXPECTATION, VARIANCE AND STANDARD DEVIATION

In [12], we tackled the verification of expectation properties for discrete random variables that attain values in positive integers only. In the current paper, rather than restricting ourselves to the expected value of a random variable, we consider the formalization of the expected value of a function of a discrete random variable, whereas the function accepts a positive integer and returns a real value. The main advantage of this new definition is that it allows us to formally define variance and standard deviation and thus in turn verify the Chebyshev's inequality for discrete random variables in HOL; a novelty that has not been available so far.

We first present the general idea of verifying random variables in HOL before going into the details of the formal definition of expectation. Random variables can be formalized in higher-order logic by thinking of them as deterministic functions with access to an infinite Boolean sequence $\mathbb{B}^{\infty}$; a source of infinite random bits [8]. These deterministic functions make random choices based on the result of popping the top most bit in the infinite Boolean sequence and may pop as many random bits as they need for their computation. When the sampling algorithms for random variables terminate, they return the result along with the remaining portion of the infinite Boolean sequence to be used by other programs. For example, a $Bernoulli(\frac{1}{2})$ random variable that returns 1 or 0 with equal probability $\frac{1}{2}$ can be modeled as a lambda abstraction function as follows

```
⊢ bit = λs. (if shd s then 1 else 0, stl s)
```

where *s* is the infinite Boolean sequences and `shd` and `stl` are the sequence equivalents of the list operation '*head*' and '*tail*'. Using the formalization of the mathematical measure theory in HOL, a probability function $\mathbb{P}$ is defined that accepts a set of infinite Boolean sequences and returns a *real* number between 0 and 1. The domain of $\mathbb{P}$ is the set $\mathscr{E}$ of events of the probability. Both $\mathbb{P}$ and $\mathscr{E}$ are defined using the Carathéodory's Extension theorem, which ensures that $\mathscr{E}$ is a σ-algebra: closed under complements and countable unions. The formalized $\mathbb{P}$ and $\mathscr{E}$ can be used to prove probabilistic properties for random variables such as

```
⊢ ℙ{s| fst (bit s) = 1} = 1/2
```

where `fst` selects the first component of a pair and $\{x|C(x)\}$ represents a set of all *x* that satisfy condition *C* in HOL.

The above infrastructure can now be used to formally define a HOL function that returns the expected value of a function of a discrete random variable.

Definition 1 $\vdash \forall\ \texttt{f R. expec\_fn f R} = \sum_{n=0}^{\infty}\ \texttt{(f n)}\ \mathbb{P}\texttt{\{s| fst (R s) = n\}}$

The infinite summation of a real sequence used in the above definition is formally defined in the HOL *real* number theory [15]. Definition 1 includes as a special case the identity function, which covers the formalization of the expected value of a random variable that attains values in the positive integers only.

Definition 2 $\vdash \forall\ \texttt{R. expec R = expec\_fn \& R}$

where the operator & is used in HOL to convert a variable of data type *positive integer* to *real*. The above definitions of expectation can be used to define functions for variance and standard deviation in HOL.

Definition 3 $\vdash \forall\ \texttt{R. variance R = expec\_fn (}\lambda\texttt{n. (\&n - expec R)}^2\texttt{) R}$
Definition 4 $\vdash \forall\ \texttt{R. std\_dev R} = \sqrt{\texttt{variance R}}$

where the functions for square and square root operations have been used from the HOL *real* number theory [15].

MARKOV'S INEQUALITY

Markov's inequality utilizes the definition of expectation to obtain a weak tail bound and can be expressed in HOL for a measurable discrete random variable with a well-defined expectation as follows.

Theorem 1 $\vdash \forall\ \texttt{R a.(0<a)} \Rightarrow \mathbb{P}\{\texttt{s|}\ \&\texttt{(fst (R s))} \geq \texttt{a}\} \leq \texttt{(expec R)/a}$

Proof: Rewriting with Definition 2 and simplifying using the *real* arithmetic properties in HOL, we obtain

$$\lceil a\rceil \mathbb{P}\{s|fst(R\ s) \geq \lceil a\rceil\} \leq \sum_{n=0}^{\infty}(n\mathbb{P}\{s|fst(R\ s)=n\} \tag{1}$$

where, $\lceil a\rceil$ denotes the ceiling of a real number a. Using the *set* theory, the additive law of probability and the properties of infinite summation of a *real* sequence in HOL, Equation 1 can be rewritten as follows.

$$\sum_{n=\lceil a\rceil}^{\infty}\lceil a\rceil \mathbb{P}\{s|fst(R\ s)=n\} \leq (\sum_{n=0}^{\lceil a\rceil}(n\mathbb{P}\{s|fst(R\ s)=n\} + \sum_{n=\lceil a\rceil}^{\infty}(n\mathbb{P}\{s|fst(R\ s)=n\}) \tag{2}$$

Equation 2 can now be verified since, the expression $\sum_{n=0}^{\lceil a\rceil}(n\mathbb{P}\{s|fst(R\ s)=n\}$ is greater than or equal to 0 and the expression $\sum_{n=\lceil a\rceil}^{\infty}(n\mathbb{P}\{s|fst(R\ s)=n\}$ is greater than or equal to $\sum_{n=\lceil a\rceil}^{\infty}\lceil a\rceil \mathbb{P}\{s|fst(R\ s)=n\}$.

CHEBYSHEV'S INEQUALITY

The expectation and variance can be used to derive a significantly stronger tail bound known as the Chebyshev's inequality, which can be expressed in HOL for a measurable discrete random variable with well-defined first and second moments and greater than 0 standard deviation as follows.

Theorem 2 $\vdash \forall\ \texttt{R a.(0<a)} \Rightarrow \mathbb{P}\{\texttt{s|}\ |\&\texttt{(fst (R s)) - expec R}| \geq \texttt{a(std\_dev R)}\} \leq \texttt{1/a}^2$

Proof: Using the properties of *absolute* function and *real* numbers and the additive law of probability, we get

$$a^2\sigma^2(\mathbb{P}\{s|X\ s \leq (\mu - a\sigma)\} + \mathbb{P}\{s|X\ s \geq (\mu + a\sigma)\}) \leq \sigma^2 \tag{3}$$

where $X = (\lambda s.fst(R\ s))$ and μ and σ denote the expectation and standard deviation of the random variable R, respectively. Using the *real number*, *set* and *probability* theories in HOL, the LHS of Equation 3 can be simplified as

$$a^2\sigma^2\sum_{n=0}^{\lceil \mu-a\sigma\rceil}\mathbb{P}\{s|X\ s=n\} + a^2\sigma^2\mathbb{P}\{s|X\ s=(\mu-a\sigma)\} + a^2\sigma^2\sum_{n=\lceil \mu+a\sigma\rceil}^{\infty}\mathbb{P}\{s|X\ s=n\} \tag{4}$$

whereas the RHS can be simplified using Definition 3 and properties of summation of a *real* sequence as follows

$$\sum_{n=0}^{\lceil \mu-a\sigma\rceil}(n-\mu)^2\mathbb{P}\{s|Xs=n\} + \sum_{n=\lceil \mu-a\sigma\rceil}^{\lceil \mu+a\sigma\rceil-\lceil \mu-a\sigma\rceil}(n-\mu)^2\mathbb{P}\{s|X\ s=n\} + \sum_{n=\lceil \mu+a\sigma\rceil}^{\infty}(n-\mu)^2\mathbb{P}\{s|X\ s=n\} \tag{5}$$

The three terms in Equation 4 can now be proved to be less than or equal to the respective three terms in Equation 5, which concludes the proof for Theorem 2.

APPLICATION: COUPON COLLECTOR'S PROBLEM

The Coupon Collector's problem [14] is motivated by "*collect all n coupons and win*" contests. Assuming that a coupon is drawn independently and uniformly at random from n possibilities, how many times do we need to draw new coupons until we find them all? In order to formalize the Coupon Collector's problem in HOL, let X be the number of trials until at least one of every type of coupon is obtained. Now, if X_i is the number of trials required to obtain the i^{th} coupon, while we had already acquired $i-1$ distinct coupons, then clearly $X = \sum_{i=1}^{n} X_i$. The advantage of breaking the random variable X into the sum of n random variables is that each X_i can be modeled as a Geometric random variable [16], which is a commonly used discrete random variable.

We presented a formalization of the Coupon Collector's problem in HOL as a function `cc` in [12]. The function `cc` accepts the number of distinct coupons, say k, and returns the summation of k Geometric random variables. Now, using the definitions given in the current paper and the linearity of expectation and variance properties in HOL, we are able to verify the expected value and a variance bound for the Coupon Collector's problem.

Theorem 3. $\vdash \forall$ `n. expec (cc (n+1)) = &(n+1)` $\sum_{i=0}^{n+1}$ `1/&(i+1)`

Theorem 4. $\vdash \forall$ `n. variance (cc (n+1))` $\leq$ `&(n+1)`^{2} $\sum_{i=0}^{n+1}$ `1/&(i+1)`^{2}

Next, using the formalized Markov's and Chebyshev's inequalities, presented in the current paper, along with some *real* arithmetic properties in HOL, we can also verify the following bounds for the Coupon Collector's problem.

Theorem 5. $\vdash \forall$ `n a.(0<a)` $\wedge$ `(0 < variance (cc (n+1)))` $\Rightarrow$
$\mathbb{P}$`{s| &(fst (cc (n+1) s))` $\geq$ `a}` $\leq$ `&(n+1)` $\sum_{i=0}^{n+1}$ `1/&(i+1)/a` $\wedge$
$\mathbb{P}$`{s| |&(fst (cc (n+1) s)) - expec (cc (n+1))|` $\geq$ `a}` $\leq$ `&(n+1)`^{2} $\sum_{i=0}^{n+1}$ `1/&(i+1)`^{2}`/a`^{2}

Thus, we have been able to conduct precise probabilistic analysis of the Coupon Collector's problem in HOL, which is something that cannot be achieved by any existing computer based probabilistic analysis tool.

CONCLUSIONS

In this paper, we presented the formal definitions of expectation, variance and standard deviation for discrete random variables in higher-order-logic. Building on these definitions, we verified the Markov's and Chebyshev's inequalities in HOL, which can be used to obtain bounds on the tail distribution of a random variable. To the best of our knowledge, this is the first time that these inequalities have been verified using a mechanical theorem prover and the results are found to be in good agreement with existing theoretical paper-and-pencil counterparts.

Based on our verification experience, we can say that formalizing mathematics in a mechanical system is a tedious and time consuming task as one often has to deal with proof steps that are usually ignored by many authors of mathematical texts. In return, we get the reliability and precision that is very useful for analyzing safety-critical systems. Besides this, theorem proving may be gainful in classical mathematical research too as it can help in coping with the explosion in mathematical knowledge and preserving mathematics from corruption.

REFERENCES

1. P. Billingsley, *Probability and Measure*, John Wiley, 1995.
2. P. Bratley, B. Fox, and L. Schrage, *A Guide to Simulation*, Springer-Verlag, 1987.
3. B. McCullough, *The American Statistician* **52**, 358–366 (1998).
4. B. McCullough, *The American Statistician* **53**, 149–159 (1999).
5. C. Baier, B. Haverkort, H. Hermanns, and J. Katoen, *IEEE Trans. on Software Engineering* **29**, 524–541 (2003).
6. J. Rutten, M. Kwaiatkowska, G. Normal, and D. Parker, *Mathematical Techniques for Analyzing Concurrent and Probabilisitc Systems*, vol. 23 of *CRM Monograph Series*, American Mathematical Society, 2004.
7. E. Clarke, O. Grumberg, and D. Peled, *Model Checking*, The MIT Press, 2000.
8. J. Hurd, *Formal Verification of Probabilistic Algorithms*, PhD Thesis, University of Cambridge, Cambridge, UK (2002).
9. J. Harrison, K. Slind, and R. Arthan, "HOL.," in *The Seventeen Provers of the World*, Springer, 2006, vol. 3600 of *LNCS*, pp. 11–19.
10. O. Hasan, and S. Tahar, "Formalization of the Continuous Probability Distributions," in *Conference on Automated Deduction*, Springer, 2007, vol. 4603 of *LNAI*, pp. 3–18.
11. O. Hasan, and S. Tahar, "Verification of Probabilistic Properties in HOL using the Cumulative Distribution Function," in *Integrated Formal Methods*, Springer, 2007, vol. 4591 of *LNCS*, pp. 333–352.
12. O. Hasan, and S. Tahar, "Verification of Expectation Properties for Discrete Random Variables in HOL," in *Theorem Proving in Higher-Order Logics*, Springer, 2007, vol. 4732 of *LNCS*, pp. 119–134.
13. O. Hasan, and S. Tahar, Formal Verification of Expectation and Variance for Discrete Random Variables, Technical Report, Concordia University, Montreal, Canada (June 2007; http://hvg.ece.concordia.ca/Publications/TECH_REP/FVEVDR_TR07).
14. M. Mitzenmacher, and E. Upfal, *Probability and Computing*, Cambridge University Press, 2005.
15. J. Harrison, *Theorem Proving with the Real Numbers*, Springer-Verlag, 1998.
16. R. Khazanie, *Basic Probability Theory and Applications*, Goodyear, 1976.