Formal Reasoning about Expectation Properties for Continuous Random Variables

Osman Hasan[1], Naeem Abbasi[1], Behzad Akbarpour[1],
Sofiène Tahar[1], and Reza Akbarpour[2]

[1] ECE Department, Concordia University, Montreal, QC, Canada
{o_hasan,n_ab,behzad,tahar}@ece.concordia.ca
[2] Imaging Research Laboratories, Robarts Research Institute, London, ON, Canada
rakbarpour@robarts.ca

Abstract. Expectation (average) properties of continuous random variables are widely used to judge performance characteristics in engineering and physical sciences. This paper presents an infrastructure that can be used to formally reason about expectation properties of most of the continuous random variables in a theorem prover. Starting from the relatively complex higher-order-logic definition of expectation, based on Lebesgue integration, we formally verify key expectation properties that allow us to reason about expectation of a continuous random variable in terms of simple arithmetic operations. In order to illustrate the practical effectiveness and utilization of our approach, we also present the formal verification of expectation properties of the commonly used continuous random variables: Uniform, Triangular and Exponential.

1 Introduction

Probabilistic analysis is a tool of fundamental importance to virtually all scientists and engineers as they often have to deal with systems that exhibit random or unpredictable elements. Traditionally, computer simulation techniques [6] are used to perform probabilistic analysis. However, they provide less accurate results and cannot handle large-scale problems due to their enormous processing time requirements. Due to the recent increase in the usage of hardware and software systems in safety-critical applications, such as medicine and transportation, the precision and accuracy of their analysis has become imperative. Therefore, simulation should not be relied upon for the analysis of such systems.

To overcome the above mentioned limitations, it has been recently proposed to conduct probabilistic analysis of systems in a higher-order-logic theorem prover [11]. The main idea behind this approach is to formally specify the behavior of systems, with random or unpredictable components, in higher-order logic, while representing the random components as formalized random variables. The probabilistic and statistical properties of random variables are then used to formally reason about systems characteristics, such as downtime, availability, number of failures, capacity, and cost, in a theorem prover. The analysis carried out in this

A. Cavalcanti and D. Dams (Eds.): FM 2009, LNCS 5850, pp. 435–450, 2009.

way is free from any approximation issues or flaws due to the mathematical nature of the models and the inherent soundness of the theorem proving approach. The milestones achieved so far, in this endeavor of developing a complete theorem proving based probabilistic analysis framework that is capable of analyzing any hardware or software system, include the formalization of probability theory [15], the ability to formalize discrete and continuous random variables and verify their probabilistic properties [15,11] and the ability to verify statistical properties of discrete random variables [11]. Whereas, to the best of our knowledge, the formal reasoning about statistical properties regarding continuous random variables has not been tackled in the open literature so far.

In this paper, as a first step towards filling the above mentioned gap, we present an infrastructure that allows us to formally reason about the expectation properties of most of the commonly used continuous random variables in a higher-order-logic theorem prover. Expectation plays a major role in decision making as it tends to summarize the probability distribution characteristics of a random variable in a single number. Thus, the contribution of this paper paves the way to formally analyze many engineering and physical science systems with continuous random components in a theorem prover. Some of the interesting examples include the performance analysis of *computer arithmetic systems* like floating-point arithmetic [19], where the Uniform random variable can be used to model the roundoff error, algorithms that utilize continuous random variables, such as the *Balls and Bins with feedback* [16] and network protocols by modeling the request arrival rates by the exponential random variables.

The most commonly used definition of expectation, for a continuous random variable X, is the probability density-weighted integral over the real line [16].

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx \tag{1}$$

The function f in the above equation represents the probability density function (PDF) of X and the integral is the well-known Reimann integral. The above definition is only limited to continuous random variables that have a well-defined PDF. A more general, but not so commonly used, definition of expectation for a random variable X, defined on a probability space (Ω, Σ, P) [7], is as follows:

$$E[X] = \int_{\Omega} X dP \tag{2}$$

This definition utilizes the Lebesgue integral and is general enough to cater for both discrete and continuous random variables. The reason behind its limited usage in the probabilistic analysis domain is the complexity of solving the Lebesgue integral, which takes its foundations from the measure theory that most engineers and computer scientists are not familiar with.

The obvious advantage of using Equation (1) is the user familiarity with Reimann integral that facilitates the reasoning process regarding the expectation properties in the theorem proving based probabilistic analysis approach. On the other hand, it requires extended real numbers, $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, whereas

all the foundational work regarding theorem proving based probabilistic analysis has been built upon the standard real numbers $\mathbb{R}$, formalized by Harrison [10]. Thus, the formalization of the expectation definition, given in Equation (1), and making it compatible with the available formal probabilistic analysis infrastructure would require creating a new data type $\overline{\mathbb{R}}$, and re-verifying the already proven results in a theorem prover for this new data-type, which is a considerable amount of work. Now, the expectation definition, given in Equation (2), does not involve extended real numbers, as it accommodates infinite limits without any ad-hoc devices due to the inherent nature of the Lebesgue integral. It also offers a more general solution. The limitation, however, is the compromise on the interactive reasoning effort, as it is not a straightforward task for a user to build on this definition to formally verify the expectation of a random variable.

In this paper, we address the above mentioned limitation of using Lebesgue integration for defining expectation. Starting from Equation (2), we mainly utilize the properties of the Lebesgue integral to formally verify two simplified expressions for the expectation. The first one is for the case when the random variable X is bounded in the positive interval $[a, b]$

$$E[X] = \lim_{n\to\infty}\left[\sum_{i=0}^{2^n-1} a + \frac{i}{2^n}(b-a)P\left\{a + \frac{i}{2^n}(b-a) \le X < a + \frac{i+1}{2^n}(b-a)\right\}\right] \tag{3}$$

and the second one is for an unbounded positive random variable [7].

$$E[X] = \lim_{n\to\infty}\left[\sum_{i=0}^{n2^n-1} \frac{i}{2^n}P\left\{\frac{i}{2^n} \le X < \frac{i+1}{2^n}\right\} + nP(X \ge n)\right] \tag{4}$$

Both of the above expressions do not involve any concepts from Lebesgue integration theory and are based on the well-known arithmetic operations like summation, limit of a real sequence, etc. Thus, users can simply utilize them, instead of Equation (2), to reason about the expectation properties of their random variables and gain the benefits of the original Lebesgue based definition. It is also important to note that we have a different expression for the bounded case in order to facilitate the formal reasoning about the probability term, which becomes very challenging to reason about if the unbounded expectation equation is used for a bounded random variable.

To demonstrate the effectiveness of the above expressions, we utilize them for the formal verification of the expected values for the commonly used continuous random variables Uniform, Triangular and Exponential. Besides being illustrative examples, these results can be essentially utilized in conducting the formal performance analysis of many system that utilize these random variables.

The work described in this paper is done using the HOL theorem prover [8], which is based on higher-order logic. The main motivation behind this choice is the fact that most of the work that we build upon is developed in HOL, such as the formalization of the real number theory [10], probability theory [15], continuous random variables [11] and Lebesgue integration [4]. Though, it is

important to note here that the ideas presented in this paper are not specific to the HOL theorem prover and can be adapted to any other higher-order-logic theorem prover as well, such as Isabelle, Coq or PVS.

The rest of the paper is organized as follows: Section 2 provides a review of related work. Then, in Section 3, we present some foundations regarding higher-order-logic based probabilistic analysis approach, such as the formalization of probability theory, random variables and Lebesgue integration. Next, Section 4 outlines the formal proof details regarding Equations (3) and (4). We utilize these theorems to illustrate the formal reasoning process regarding the expectation properties of the above mentioned three continuous random variables in Section 5. In Section 6, we present the formal probabilistic analysis of rounding error in floating-point numbers, in order to demonstrate the usefulness of our results in the domain of probabilistic analysis. Finally, Section 7 concludes the paper.

2 Related Work

Early foundations of probabilistic analysis in a higher-order-logic theorem prover were laid down by Nędzusiak [17] and Bialas [3] when they proposed a formalization of measure and probability theories in higher-order logic. Hurd [15] implemented their work and developed a framework for the verification of probabilistic algorithms in the HOL theorem prover. Random variables are basically probabilistic algorithms and thus can be formalized and verified, based on their probability distribution properties, using the methodology proposed in [15]. In fact, building upon Hurd's formalization, most of the commonly used discrete [15] and continuous [11] random variables have been formalized. The above mentioned formalization of probability theory has also been used to formally reason about statistical properties, such as expectation and variance, of discrete random variables [11]. Due to the fact that the discrete random variables can only attain a countable number of values, the expectation in this case has been formally defined using a summation rather than integration. Obviously such a definition cannot be used with continuous random variables, which have an uncountable range. The probabilistic analysis foundations, mentioned above, have been successfully used to conduct precise probabilistic analysis of many systems, such as computation algorithms [15,11], real-time systems [11], communication protocols [13], wireless systems [14], and hardware components [12].

As mentioned in the last section, Lebesgue integration is the core concept in the definition of expectation. Richter [18] formalized a significant portion of the Lebesgue integration theory in higher-order logic using Isabelle/HOL. But, this formalization can only handle functions that map subsets of real numbers to real numbers. This limitation somewhat restricts the usage of this formalization to define the expectation, where the function that needs to be integrated is the random variable that in its most general form maps the subsets of an arbitrary sample space to real numbers. More recently, Coble [4] formalized the Lebesgue integration theory in HOL. This formalization overcomes the limitation of Richter's work as it allows integration over functions that are measurable

from a space of any arbitrary data-type to any subset of the real numbers. Coble's formalization of the Lebesgue integral has been used to formally define expectation of a random variable [4]. But in this formalization, some theorems have been verified under the assumption that measurable sets have to be equal to the power set of the sample space. This fact restricts Coble's formalization for sample spaces that do not contain any non-measurable subsets. Whereas, this condition is not satisfied for sample spaces for continuous random variables. Daumas *et. al.* [5] have also formalized some Lebesgue integration theory in the PVS theorem prover. The authors claim to have formally defined expectation based on this formalization, but no details were given in [5]. Moreover, to the best of our knowledge, no information regarding the utilization of this definition to formally reason about the expectation of continuous random variables has been provided in this work, which is the main contribution of our paper.

In this paper, we extend the measure theoretic formalization infrastructure, based on the works, presented in [15,11], available in the HOL theorem prover, with the ability to formally reason about expectation properties of *continuous random variables*. This would be a novelty that to the best of our knowledge has not been presented in the open literature so far. The main motivation behind using the measure theoretic approach instead of the one proposed by Audebaud [2] is to be able to utilize the Lebesgue integral, which has a foundational relationship with the measure theory. We utilize the Lebesgue integral formalization, presented in [4], for our work because it is available in the HOL theorem prover and is thus compatible with the other theories [15,11] that we build upon. Though, we make it general enough to tackle sample spaces for continuous random variables as well.

3 Preliminaries

In this section, we provide an overview of the higher-order-logic formalizations of probability theory, continuous random variables and Lebesgue integration theory. The intent is to introduce the main ideas along with some notation that is going to be used later in this paper.

3.1 Probability Theory and Random Variables in HOL

A *measure space* is defined as a triple (Ω, Σ, μ), where Ω is a set, called the *sample space*, Σ represents a σ-algebra of subsets of Ω and the subsets are usually referred to as *measurable sets*, and μ is a *measure* with domain Σ [7]. A *probability space* is a measure space (Ω, Σ, P) such that the measure, referred to as the probability and denoted by P, of the sample space is 1.

Hurd [15] formalized some measure theory to define a measure space as a pair (Σ, μ). Whereas the sample space, on which this pair is defined, is implicitly implied from the higher-order-logic definitions to be equal to the universal set of the appropriate data-type. Building upon this formalization, the probability space was also defined in HOL as a pair $(\mathcal{E}, \mathbb{P})$, where the domain of $\mathbb{P}$ is the set

$\mathcal{E}$, which is a set of subsets of infinite Boolean sequences $\mathbb{B}^\infty$. Both $\mathbb{P}$ and $\mathcal{E}$ are defined using the Carathéodory's Extension theorem, which ensures that $\mathcal{E}$ is a σ-algebra: closed under complements and countable unions.

Now, a random variable, which is one of the core concepts in probabilistic analysis, is fundamentally a probabilistic function and thus can be modeled in higher-order logic as a deterministic function, which accepts the infinite Boolean sequence as an argument. 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 functions terminate, they return the result along with the remaining portion of the infinite Boolean sequence to be used by other programs. Thus, a random variable which takes a parameter of type α and ranges over values of type β can be represented in HOL by the function $\mathcal{F}$.

$$\mathcal{F} : \alpha \to B^\infty \to \beta \times B^\infty$$

As an example, consider the Bernoulli($\frac{1}{2}$) random variable that returns 1 or 0 with equal probability $\frac{1}{2}$. It can be formalized in HOL as follows

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

It accepts an infinite Boolean sequence, where `shd` and `stl` are the sequence equivalents of the list operation *'head'* and *'tail'*. The formalized $\mathbb{P}$ and $\mathcal{E}$ can be used to verify the basic laws of probability as well as probabilistic properties regarding random variables in the HOL theorem prover. For example:

$\vdash \mathbb{P}$ {`s | fst (bit s) = 1`} = $\frac{1}{2}$

where the HOL function `fst` selects the first component of a pair and $\{x|C(x)\}$ represents a set of all x that satisfy the condition C. It is important to note here that, since the probability measure $\mathbb{P}$ is only defined on sets in $\mathcal{E}$, it is absolutely necessary to verify that the set that appears in a probabilistic property is in $\mathcal{E}$ before we can formally verify that property in HOL. For the above example, this condition translates to the verification of {`s | fst (bit s) = 1`} $\in$ $\mathcal{E}$.

The above approach has been successfully used to formalize and verify most of the commonly used discrete random variables [15]. The sampling algorithms for discrete random variables are either guaranteed to terminate or satisfy probabilistic termination, meaning that the probability that the algorithm terminates is 1. On the other hand, the formalization of continuous random variables involves non-terminating algorithms and hence require a different approach than discrete random variables.

Building upon the above mentioned probability theory framework, an approach for the formalization of continuous random variables has been presented in [11]. The main idea is based on the concept of the Inverse Transform Method (ITM) [6], according to which, the random variable X, for any continuous cumulative distribution function (CDF) F, can be defined as $X = F^{-1}(U)$, where F^{-1} is the inverse function of F, and U represents the Standard Uniform random

Table 1. Continuous Random Variables in HOL

Distribution	CDF	Formalized Random Variable
Uniform(a, b)	0 if $x \leq a$; $\frac{x-a}{b-a}$ if $a < x \leq b$; 1 if $b < x$.	$\vdash \forall s\ l.\ uniform\_rv\ a\ b\ s\ =$ $(b-a)(std\_unif\_rv\ s) + a$
Triangular$(0, a)$	0 if $x \leq 0$; $(\frac{2}{a}(x - \frac{x^2}{2a}))$ if $x < a$; 1 if $a \leq x$.	$\vdash \forall s\ a\ .\ triangular\_rv\ l\ s\ =$ $a(1 - \sqrt{1 - std\_unif\_rv\ s})$
Exponential(l)	0 if $x \leq 0$; $1 - e^{-lx}$ if $0 < x$.	$\vdash \forall s\ l.\ exp\_rv\ l\ s\ =$ $-\frac{1}{l} ln(1 - std\_unif\_rv\ s)$

variable. The formal proof of this proposition is based on the CDF characteristic of the Standard Uniform random variable and some of the CDF properties [11]. ITM allows us to formalize any continuous random variable, which has a well-defined CDF, in terms of a formalized Standard Uniform random variable (`std_unif_rv`). Based on this approach, the CDFs and higher-order-logic definitions of three continuous random variables are given in Table 1 [11]. In this paper, we will utilize formally verified expressions, corresponding to Equations (3) and (4), to verify the expectation relations for these random variables in Section 5.

3.2 Lebesgue Integration in HOL

Lebesgue integration is based on the concept of measure and is defined for a class of functions called *measurable functions*, which are well-behaved functions between measurable spaces. Coble [4] formalized the Lebesgue integration theory in HOL based on a generalized measure space $(S, \mathbb{S}, \lambda)$. It is important to note here that, unlike Hurd's formalization of the measure space, we do have the flexibility to choose any sample space S in this case. The higher-order-logic definition of the Lebesgue integral utilizes the concepts of *indicator function* and *positive simple-function* [7]. The indicator function is defined as follows for a set A

$$\mathbb{I}_A(a) = \begin{cases} 1 & \text{if } a \in A; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Whereas, a function g is said to be a positive simple-function for the measure space $(S, \mathbb{S}, \lambda)$ iff it can be expressed as follows

$$g = \sum_{i=0}^{n} c_i \mathbb{I}_{a_i} \tag{6}$$

where c_i is a sequence of positive real values and a_i is a sequence of disjoint measurable sets such that $\bigcup_{i=0}^{n} a_i$ forms a partition of S. Now the integral for such a positive-simple function g can be defined as follows.

$$\int_S g \; d\lambda = \sum_{i=0}^{n} c_i(\lambda a_i) \tag{7}$$

The next step towards the formal definition of the Lebesgue integral is to define the integral for a positive function f that is measurable from $(S, \mathbb{S})$ to $(S', \mathbb{S}')$

$$\int_S f \; d\lambda = sup\left\{ \int_S g d\lambda \;\middle|\; (\forall x.g(x) \leq f(x)) \right\} \tag{8}$$

where g is a positive-simple function w.r.t the measure space $(S, \mathbb{S}, \lambda)$.

The Lebesgue integral of a real-valued measurable function from $(S, \mathbb{S})$ to $(S', \mathbb{S}')$ can now be formalized in terms of Equation (8) as follows

$$\int_S f \; d\lambda = \int_S f^+ \; d\lambda - \int_S f^- \; d\lambda \tag{9}$$

where $f(x) = f^+(x) - f^-(x)$ and f^+ and f^- are the positive and negative portions of f, respectively, and are both positive functions. It is also important to note that the integral of f is well-defined iff both f^+ and f^- are measurable from $(S, \mathbb{S})$ to $(S', \mathbb{S}')$ and their integrals do not both diverge to infinity.

Besides the formalization of the above definitions, many useful properties regarding the Lebesgue integral have also been verified in [4] as higher-order-logic theorems. For example, we utilize the following convergence of a positive measurable function to the Lebesgue integral property.

$$\begin{aligned} \big(\forall x \in S.(\forall n \; x.f_n(x) \leq f(x)) \wedge \; \big(\lim_{n\to\infty} f_n(x) = f(x)\big) \; \wedge \; \big(\lim_{n\to\infty} \int_S f_n \; d\lambda = r\big) \\ \Rightarrow \int_S f \; d\lambda = r \end{aligned} \tag{10}$$

The function f, in the above equation, is a positive real-valued function that is measurable from $(S, \mathcal{P}(S))$ to $(S', \mathcal{P}(S'))$, where $\mathcal{P}(A)$ denotes the power set of the set A. Whereas, the sequence f_n is a monotonically increasing sequence of positive simple-functions. It is important to note here that this theorem and many others in Coble's work [4] have been verified for the case when the measurable sets $\mathbb{S}$ is equal to the power set of the sample space S. This restricts the usage of these theorems to sample spaces for which all possible subsets are measurable. This condition is not satisfied for sample spaces that are used to model continuous random variables.

4 Verification of Expectation Relations

In this section, we utilize the probability and Lebesgue integration theories, described in the previous section, to formally verify the expectation relations for

the bounded and unbounded random variables, given in Equations (3) and (4), respectively.

The first step in this regard is to formally define the expectation in terms of the Lebesgue integral. For this purpose, we utilize the definition of Lebesgue integral, given in Equation (9), as follows:

Definition 1. *Expectation of a Random Variable*

$$\vdash \forall\ \texttt{f.}\ \texttt{expec}\ (\mathcal{U}, \mathcal{E}, \mathbb{P})\ \texttt{f} = \int_{\mathcal{U}} \texttt{f}\ d\mathbb{P}$$

The function `expec` accepts a probability space, $(\mathcal{U}, \mathcal{E}, \mathbb{P})$, and a random variable f that maps infinite Boolean sequences to real numbers. It is important to note that by using Hurd's formalization of the probability space $(\mathcal{U}, \mathcal{E}, \mathbb{P})$, where $\mathcal{U}$ represents the universal set of all Boolean sequences, as outlined in Section 3, we can utilize the above definition to reason about expectation of random variables formalized in [15,11]. Though, we had to generalize the Lebesgue integration theorems, proposed in Coble's work [4]. Since, the existing theorems are based on the assumption $\mathbb{S} = \mathcal{P}(S)$, which is not true for our probability space $(\mathcal{U}, \mathcal{E}, \mathbb{P})$, where the power set of the set of all Boolean sequences do contain non-measurable sets as has been formally verified in [15]. Our more generalized version of these theorems are based on the assumption that $\mathbb{S} = \{x | (x \in \mathcal{P}(S)) \wedge (x\ is\ measurable)\}$, which is obviously true for our probability space $(\mathcal{U}, \mathcal{E}, \mathbb{P})$.

4.1 Bounded Random Variables

The expectation property, given in Equation (3), can be expressed as a higher-order-logic theorem using Definition 1 as follows:

Theorem 1. *Expectation of Bounded Random Variables*

$$\vdash \forall\ \texttt{a b f.}\ (\texttt{0} \leq \texttt{a}) \wedge (\texttt{a} < \texttt{b}) \wedge (\forall\ \texttt{s.}\ \texttt{a} \leq \texttt{f s} \leq \texttt{b}) \wedge (\forall\ \texttt{x y.}\ \texttt{x} < \texttt{y} \Rightarrow \{\texttt{s} \mid \texttt{x} \leq \texttt{f s} < \texttt{y}\} \in \mathcal{E}) \Rightarrow$$

$$\left(\texttt{expec}\ (\mathcal{U}, \mathcal{E}, \mathbb{P})\ \texttt{f} = \lim_{\texttt{n} \to \infty} \left[\sum_{\texttt{i}=0}^{2^{\texttt{n}}-1} (\texttt{a} + \tfrac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b} - \texttt{a}))\mathbb{P}\left\{\texttt{s} \;\middle|\; \texttt{a} + \tfrac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b} - \texttt{a}) \leq \texttt{f s} < \texttt{a} + \tfrac{\texttt{i+1}}{2^{\texttt{n}}}(\texttt{b} - \texttt{a})\right\}\right]\right)$$

The first three assumptions ensure that the random variable f is bounded in the positive interval $[a, b]$. Whereas, the fourth assumption ensures that the set involved in this verification is measurable.

In order to utilize any definition or property of Lebesgue integration theory with the above theorem, we first need to show that the triple $(\mathcal{U}, \mathcal{E}, \mathbb{P})$ is a measure space with a positive measure. We verified these conditions based on the corresponding theorems available in Hurd's formalization of the probability space $(\mathcal{E}, \mathbb{P})$ along with the definition of measure in [4] under the given assumptions.

Since our random variable f is a positive-valued real number, we do not have the term involving the f^- term in the Lebesgue integral definition and thus, for this specific case, Equations (8) and (9) become equivalent. This allows us to

use the convergence of a positive measurable function to the Lebesgue integral property, given in Equation (10), to reason about Theorem 1. Using Modus Ponens (MP) rule, we can split the proof goal of Theorem 1 to the following five subgoals, corresponding to the monotonicity and positive simple-function requirement on f_n and the three assumptions of Equation (10):

$$\texttt{mono\_increasing}\left[\sum_{\texttt{i}=0}^{2^{\texttt{n}}-1}(\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a}))\mathbb{I}_{\left\{\texttt{s}\;\middle|\;\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\leq \texttt{f}\ \texttt{s}<\texttt{a}+\frac{\texttt{i}+1}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\right\}}(\texttt{x})\right] \tag{11}$$

$$\left(\forall \texttt{i}.(\texttt{i}<2^{\texttt{n}})\Rightarrow\left\{\texttt{s}\;\middle|\;\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\leq \texttt{f}\ \texttt{s}<\texttt{a}+\frac{\texttt{i}+1}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\right\}\in\mathcal{E}\right)\wedge \left(\forall \texttt{i}.0\leq \texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\right)\wedge\left(\texttt{FINITE}\{\texttt{i}|\texttt{i}<2^{\texttt{n}}\}\right) \tag{12}$$

$$\left[\sum_{\texttt{i}=0}^{2^{\texttt{n}}-1}(\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a}))\mathbb{I}_{\left\{\texttt{s}\;\middle|\;\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\leq \texttt{f}\ \texttt{s}<\texttt{a}+\frac{\texttt{i}+1}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\right\}}(\texttt{x})\right]\leq \texttt{f}(\texttt{x}) \tag{13}$$

$$\lim_{\texttt{n}\to\infty}\left[\sum_{\texttt{i}=0}^{2^{\texttt{n}}-1}(\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a}))\mathbb{I}_{\left\{\texttt{s}\;\middle|\;\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\leq \texttt{f}\ \texttt{s}<\texttt{a}+\frac{\texttt{i}+1}{2^{\texttt{n}}}(\texttt{b}-\texttt{a})\right\}}(\texttt{x})\right]=\texttt{f}(\texttt{x}) \tag{14}$$

$$\exists \texttt{y}.\lim_{\texttt{n}\to\infty}\left[\sum_{\texttt{i}=0}^{2^{\texttt{n}}-1}(\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}\texttt{b}-\texttt{a})\mathbb{P}\left\{\texttt{s}\;\middle|\;\texttt{a}+\frac{\texttt{i}}{2^{\texttt{n}}}\texttt{b}-\texttt{a}\leq \texttt{fs}<\texttt{a}+\frac{\texttt{i}+1}{2^{\texttt{n}}}\texttt{b}-\texttt{a}\right\}\right]=\texttt{y} \tag{15}$$

The monotonically increasing property in the first subgoal is verified based on the facts that (1) the indicator function is 1 in only one interval or for one particular value of i and (2) as the argument of the sequence increases, i.e., n, the intervals become finer and thus the resulting value of the sequence increasingly gets closer to the value of $f\ x$. The second subgoal corresponds to the pre-conditions for the positive simple-function function f_n and consists of three subgoals. These three subgoals are discharged based on the fourth assumption of Theorem 1, arithmetic reasoning and set theory principles, respectively. The third subgoal is true as there is only one i, say i', for which the real value of $f\ x$ falls in the interval $[a+\frac{i}{2^n}(b-a), a+\frac{i+1}{2^n}(b-a))$ out of the 2^n possible values for i. Thus the indicator function is 1 for this particular i only and 0 otherwise, meaning that the summation is equal to $(a+\frac{i'}{2^n}(b-a))$. Now, substituting this value for the summation in the third subgoal along with the fact that $f\ x$ lies in the interval $[a+\frac{i'}{2^n}(b-a), a+\frac{i'+1}{2^n}(b-a))$ leads to its verification. The fourth subgoal is discharged based on reasoning similar to the previous subgoal, the monotonicity of the given sequence and the definition of the limit of a real sequence. Finally,

the real sequence in the fifth subgoal is verified to be convergent by verifying that it is monotonic and that the probability term in the sequence is non-zero for only one particular value of i. The sequence thus has an upper bound b since the value of i is always less than 2^n and the maximum value for the probability term is 1. The verification of these five subgoals also concludes the verification of Theorem 1.

4.2 Unbounded Random Variables

The expectation property, given in Equation (4), can be expressed as a higher-order-logic theorem using Definition 1 as follows:

Theorem 2. *Expectation of Unbounded Random Variables*

$$\begin{array}{l} \vdash \forall\, \mathtt{f}.\ (\forall\, \mathtt{s}.\ 0 \leq \mathtt{f\ s}) \wedge (\forall\, \mathtt{x}.\ \{\mathtt{s} \mid \mathtt{f\ s} \geq \mathtt{x}\} \in \mathcal{E}) \\ \quad (\forall\, \mathtt{x\ y}.\ \mathtt{x} < \mathtt{y} \Rightarrow \{\mathtt{s} \mid \mathtt{x} \leq \mathtt{f\ s} < \mathtt{y}\} \in \mathcal{E}) \Rightarrow \\ \left(\mathtt{expec}\ (\mathcal{U}, \mathcal{E}, \mathbb{P})\ \mathtt{f} = \right. \\ \left. \lim\limits_{\mathtt{n} \to \infty} \left[\sum_{\mathtt{i}=0}^{\mathtt{n}2^{\mathtt{n}}-1} \left(\frac{\mathtt{i}}{2^{\mathtt{n}}}\right) \mathbb{P}\left\{\mathtt{s} \;\middle|\; \frac{\mathtt{i}}{2^{\mathtt{n}}} \leq \mathtt{f\ s} < \frac{\mathtt{i}+1}{2^{\mathtt{n}}}\right\} + \mathtt{n}\mathbb{P}\left\{\mathtt{s} \;\middle|\; \mathtt{f\ s} \geq \mathtt{n}\right\}\right]\right) \end{array}$$

The first assumption ensures that the random variable f is positive. The second and third guarantee that the sets that arise in this verification are measurable events. The summation range has been extended to $[0, n2^n - 1]$ so that the first probability term in the above theorem covers the interval $[0, n)$. While, the second probability term covers the rest of the positive unbounded interval.

The verification steps for Theorem 2 are very similar to the ones for Theorem 1. The major step is to split this goal into subgoals using Equation (10). These subgoals are then verified using arithmetic reasoning, set theory principles and the fact that the events in the two probability terms of the proof goal are disjoint, which means that one of the probability term is always equal to 0.

Our verification results matched the paper-and-pencil analysis counterpart for Theorem 2, which is available in [7], and confirmed the correctness of Theorem 1, which we had worked out ourselves and were not able to find it in any published texts. Besides checking for correctness for these mathematical relationships, the major motivation behind their verification is to utilize them to reason about the expected values of continuous random variables and thus in turn use these results for conducting formal probabilistic analysis of systems.

5 Expectation of Continuous Random Variables

To illustrate the effectiveness of the expectation relations, proved in the previous section, we now utilize them to verify the expectation of three continuous random variables, i.e., Uniform, Triangular and Exponential.

5.1 Uniform Random Variable

The expectation relation for the continuous Uniform random variable bounded in the interval $[a, b]$ can be formalized as follows:

Theorem 3. *Expectation of the Uniform(a,b) Random Variable*

$$\vdash \forall\ \mathtt{a\ b.\ (0 \leq a) \wedge (a < b) \Rightarrow} \left(\mathtt{expec}\ (\mathcal{U}, \mathcal{E}, \mathbb{P})\ \mathtt{(uniform\_rv\ a\ b) = } \frac{a+b}{2}\right)$$

In order to utilize Theorem 1 to reason about the correctness of the above theorem, we first verify that the Uniform random variable satisfies all pre-conditions, given in Theorem 1, based on the theorems given in [11]. Next, we rewrite the probability term in Theorem 1, using the CDF for the Uniform random variable, given in Table 1, to simplify our proof goal as follows:

$$\lim_{\mathtt{n}\to\infty}\left[\sum_{\mathtt{i}=0}^{2^{\mathtt{n}}-1}(\mathtt{a}+\frac{\mathtt{i}}{2^{\mathtt{n}}}(\mathtt{b}-\mathtt{a}))\left(\frac{\mathtt{a}+\frac{\mathtt{i}+1}{2^{\mathtt{n}}}(\mathtt{b}-\mathtt{a})-\mathtt{a}}{\mathtt{b}-\mathtt{a}}-\frac{\mathtt{a}+\frac{\mathtt{i}}{2^{\mathtt{n}}}(\mathtt{b}-\mathtt{a})-\mathtt{a}}{\mathtt{b}-\mathtt{a}}\right)\right]=\frac{\mathtt{a}+\mathtt{b}}{2} \tag{16}$$

The above subgoal can now be discharged using arithmetic reasoning, along with the properties of summation of a real sequence and the limit of a real sequence. This also concludes the verification of Theorem 3.

5.2 Triangular Random Variable

The expectation relation for the continuous Triangular random variable bounded in the interval $[0, b]$ can be formalized as follows:

Theorem 4. *Expectation of the Triangular(b) Random Variable*

$$\vdash \forall\ \mathtt{b.\ (0 < b) \Rightarrow} \left(\mathtt{expec}\ (\mathcal{U}, \mathcal{E}, \mathbb{P})\ \mathtt{(triangular\_rv\ b) = } \frac{b}{3}\right)$$

The verification steps are similar to the ones for Theorem 3 and are primarily based on Theorem 1 and the CDF of the Triangular random variable.

5.3 Exponential Random Variable

The expectation for the continuous Exponential random variable, which is unbounded at the upper end, i.e., defined in $[0, \infty)$, can be formalized as follows:

Theorem 5. *Expectation of the Exponential(l) Random Variable*

$$\vdash \forall\ \mathtt{a.\ (0 < a) \Rightarrow} \left(\mathtt{expec}\ (\mathcal{U}, \mathcal{E}, \mathbb{P})\ \mathtt{(exp\_rv\ a) = } \frac{1}{\mathtt{a}}\right)$$

Due to its unbounded nature, we use Theorem 2 to reason about the expectation of Exponential random variable. Now, after rewriting the probability term and some arithmetic simplification, we get the following subgoal:

$$\lim_{\mathtt{n}\to\infty}\left[\left(1-\mathtt{e}^{-\frac{\mathtt{a}}{2^{\mathtt{n}}}}\right)\left(\sum_{\mathtt{i}=0}^{\mathtt{n}2^{\mathtt{n}}-1}\frac{\mathtt{i}}{2^{\mathtt{n}}}\mathtt{e}^{-\mathtt{a}\frac{\mathtt{i}}{2^{\mathtt{n}}}}\right)+\mathtt{n}\mathtt{e}^{-\mathtt{an}}\right]=\frac{1}{\mathtt{a}} \tag{17}$$

which can be broken into the following two subgoals.

$$\lim_{n\to\infty}\left(ne^{-an}\right) = 0 \tag{18}$$

$$\lim_{n\to\infty}\left[\left(\frac{1-e^{-\frac{a}{2^n}}}{2^n}\right)\left(\sum_{i=0}^{n2^n-1} i(e^{-\frac{a}{2^n}})^i\right)\right] = \frac{1}{a} \tag{19}$$

We proceed with the verification of the first subgoal by rewriting the exponential term e^{-an} as $(1+x)^{-n}$, where $x > 0$. Next, we verify that the term $(1+x)^n$ is greater than $1+nx+\frac{1}{2}n(n-1)x^2$, for all values of n, as the latter represents a truncated form of its Binomial expansion. This fact leads us to verify that the value of the real sequence $(\lambda n.n(1+x)^{-n})$ will be less than the real sequence $(\lambda n.n(\frac{1}{2}n(n-1)x^2)^{-1})$ for all values of n. This reasoning allows us to discharge the first subgoal, given in Equation (18), as the limit value of the real sequence $(\lambda n.n(\frac{1}{2}n(n-1)x^2)^{-1}) = (\lambda n.\frac{2}{x^2(n-1)})$ is 0.

In order to simplify the verification of the second subgoal, given in Equation (19), we first evaluate the summation term by verifying the summation of a finite arithmetic-geometric series in HOL.

$$\sum_{k=0}^{n} kq^k = \frac{q}{(1-q)^2}(1-q^n) - \frac{nq^{n+1}}{1-q} \tag{20}$$

The above relationship allows us to rewrite the second subgoal as follows:

$$\lim_{n\to\infty}\left(\frac{e^{-\frac{a}{2^n}}(1-e^{-an})}{2^n(1-e^{-\frac{a}{2^n}})} - ne^{-an}\right) = \frac{1}{a} \tag{21}$$

Now, Equation (18) and the already proved fact that the limit value of the real sequence $(\lambda n.e^{-ln})$ is 0 allows us to simplify the above subgoal as follows.

$$\lim_{n\to\infty}\left(\frac{e^{-\frac{a}{2^n}}}{2^n(1-e^{-\frac{a}{2^n}})}\right) = \frac{1}{a} \tag{22}$$

We reason about the correctness of the above limit by first evaluating the following limit relationship.

$$\lim_{x\to 0}\left(\frac{xe^{-ax}}{(1-e^{-ax})}\right) = \frac{1}{a} \tag{23}$$

The proof of the above equation is primarily based on the L'Hopital's Rule, which we also verified in HOL as part of this project. Now, the variable x in Equation (23) can be specialized to $\frac{1}{2^n}$. This expression along with the definitions of limit of a real sequence and the limit of a function when its arguments approaches a real value leads to the verification of the remaining subgoal, given in Equation (22). This also concludes the proof of Theorem 5.

The verification of the above three expectation properties does not involve any reasoning based on the Lebesgue integral. As a consequence, the verification process, which just took around 80 man hours with approximately 3500 lines of

HOL code, was very straightforward and quick in comparison to the verification of Theorems 1 and 2, which took around 350 man-hours and approximately 5000 lines. This clearly demonstrates the strength of our work, which is to provide the ability to build upon Theorems 1 and 2 and reduce the interactive reasoning efforts regarding the expectation properties of continuous random variables. Also, our theorems are quite general and can be built upon to reason about expected values of many other random variables as well, such as the Rayleigh and Pareto.

6 Round-Off Error in Floating-Point Representation

Algorithms involving floating-point numbers are extensively used these days in almost all digital equipment ranging from computer and digital processing to telecommunication systems. Due to their complexity and wide spread usage in safety critical domains, formal methods are generally preferred over traditional testing to ensure correctness of floating-point algorithms. A classical work in this regard is Harrison's error analysis of floating-point arithmetic in higher-order logic [9]. Harrison presents a formalization of floating point numbers, verification of upper bounds on the error in representing a real number with floating-point system and the error in floating-point arithmetic operations. Even though this analysis is very useful in identifying the worst case conditions, it doest not reflect upon the typical or average errors. In fact, the assumed worst case conditions rarely occur in practice. So the error analysis, based under these worst-case conditions can improperly suggest that the performance of the algorithm is poor.

In paper-and-pencil analyses, probabilistic techniques are thus utilized in the error analysis of floating-point algorithms [19]. The main idea behind this probabilistic approach is to model the error in a single floating-point number by an appropriate random variable and utilize this information to judge the expected value of error while representing a real number in floating-point system. This expected value of error can then be used to find the expected value of error in different floating-point arithmetic operations.

The above mentioned probabilistic analysis involves reasoning about the expectation value of a continuous random variable, since the error between a real number and its corresponding floating-point representation is continuous in nature. Thus, our proposed infrastructure can be directly utilized to conduct such analysis, something that to the best of our knowledge was not possible before.

We built upon Harrison's error bounds for floating-point representations of *big* ($|x| \in [2^k, 2^{k+1})$, *small* ($|x| \in [\frac{1}{2^{k+1}}, \frac{1}{2^k}] : k < 126$), and *tiny* ($|x| \in [0, \frac{1}{2^{126}}]$) real numbers [9]. The error is defined as the difference between the real value of the floating-point representation and the actual value of the corresponding real number ($\texttt{error(x)} = \texttt{float(x)} - \texttt{x}$), with round-to-nearest rounding mode. Based on this definition, upper bounds on the absolute value of error are verified to be equal to $\frac{2^k}{2^{24}}$, $\frac{1}{2^{k+1}2^{24}}$ and $\frac{1}{2^{150}}$, for the three cases above, respectively.

Assuming any value of error to be equally likely [19], we constructed formal probabilistic models for representing the above mentioned rounding errors using Uniform random variables defined in the intervals $[0, \frac{2^k}{2^{24}}]$, $[0, \frac{1}{2^{k+1}2^{24}}]$ and

$[0, \frac{1}{2^{150}}]$, respectively. Theorem 3 was then used to verify the expectation values of these floating-point errors using the HOL theorem prover.

Theorem 6. *Expectation of Floating-Point Errors*

$$\vdash \forall\ \mathtt{k\ x.}\ \left(\mathtt{expec(uniform\_rv\ 0}\ \tfrac{2^k}{2^{24}}\mathtt{)} = \tfrac{2^{k-1}}{2^{24}}\right) \wedge \left(\mathtt{expec(uniform\_rv\ 0}\ \tfrac{1}{2^{k+1}2^{24}}\mathtt{)} = \tfrac{1}{2^{k+1}2^{25}}\right) \wedge \left(\mathtt{expec(uniform\_rv\ 0}\ \tfrac{1}{2^{150}}\mathtt{)} = \tfrac{1}{2^{151}}\right)$$

The above theorem plays a pivotal role in the statistical error analysis of floating-point arithmetic. Based on these average values of error in a single floating-point number, the average errors in floating point operations, like addition, subtraction and multiplication, that involve multiple floating-point numbers, can be evaluated. Similarly, this information can be further utilized in conducting the statistical error analysis of basic digital signal processing (DSP) systems by building on top of the DSP verification framework in HOL [1], which as of now does not include any probabilistic and statistical considerations.

7 Conclusions

In this paper, we have presented an infrastructure to reason about expectation properties of continuous random variables using a higher-order-logic theorem prover. This capability allows us to conduct formal statistical analysis of systems with continuous random components, a novelty, which is not supported by most of the existing probabilistic analysis tools.

We built upon a formalized Lebesgue integration theory to define expectation and based on this definition we verified two alternate expectation relations. These relations do not involve any concepts from the mathematically complex Lebesgue integration theory and thus facilitate reasoning about expected values of continuous random variables significantly. We utilized these relations to verify the expected values of the extensively used continuous random variables Uniform, Triangular and Exponential. To the best of our knowledge, this is the first time that the formal reasoning about the expectation of these continuous random variables has been presented in a higher-order-logic theorem prover.

Our formally verified expectation relations are valid for discrete random variables as well, due to the generic nature of the Lebesgue integral. In fact, we plan to link these relations to the summation based definition of expectation [11] in order to come up with a unified reasoning framework for both discrete and continuous random variables. Also, the presented results can be extended to be used for random variables that are not positive functions, since the Lebesgue integral allows integration over negative functions, as can be observed from Equation (9). Other interesting future research directions, that benefit from this work, include the formal reasoning frameworks for variance properties and tail distribution bounds and the ability to reason about statistical properties of systems that involve multiple continuous random variables.

References

1. Akbarpour, B., Tahar, S.: An Approach for the Formal Verification of DSP Designs using Theorem Proving. IEEE Transactions on CAD of Integrated Circuits and Systems 25(8), 1141–1457 (2006)
2. Audebaud, P., Paulin-Mohring, C.: Proofs of Randomized Algorithms in Coq. In: Uustalu, T. (ed.) MPC 2006. LNCS, vol. 4014, pp. 49–68. Springer, Heidelberg (2006)
3. Bialas, J.: The σ-Additive Measure Theory. J. of Formalized Mathematics 2 (1990)
4. Coble, A.: On Probability, Measure, and Integration in HOL4. Technical Report, Computing Laboratory, University of Cambridge, UK (2009), `http://www.srcf.ucam.org/~arc54/techreport.pdf`
5. Daumas, M., Martin-Dorel, E., Lester, D., Truffert, A.: Stochastic Formal Correctness of Numerical Allgorithms. In: First NASA Formal Methods Symposium, pp. 136–145 (2009)
6. Devroye, L.: Non-Uniform Random Variate Generation. Springer, Heidelberg (1986)
7. Galambos, J.: Advanced Probability Theory. Marcel Dekker Inc., New York (1995)
8. Gordon, M.J.C., Melham, T.F.: Introduction to HOL: A Theorem Proving Environment for Higher-Order Logic. Cambridge University Press, Cambridge (1993)
9. Harrison, J.: Floating Point Verification in HOL Light: The Exponential Function. Technical Report 428, Computing Laboratory, University of Cambridge, UK (1997)
10. Harrison, J.: Theorem Proving with the Real Numbers. Springer, Heidelberg (1998)
11. Hasan, O.: Formal Probabilistic Analysis using Theorem Proving. PhD Thesis, Concordia University, Montreal, QC, Canada (2008)
12. Hasan, O., Abbasi, N., Tahar, S.: Formal Probabilistic Analysis of Stuck-at Faults in Reconfigurable Memory Arrays. In: Leuschel, M., Wehrheim, H. (eds.) IFM 2009. LNCS, vol. 5423, pp. 277–291. Springer, Heidelberg (2009)
13. Hasan, O., Tahar, S.: Performance Analysis of ARQ Protocols using a Theorem Prover. In: Proc. International Symposium on Performance Analysis of Systems and Software, pp. 85–94. IEEE Computer Society, Los Alamitos (2008)
14. Hasan, O., Tahar, S.: Performance Analysis of Wireless Systems using Theorem Proving. In: Proc. First International Workshop on Formal Methods for Wireless Systems, Toronto, ON, Canada, pp. 3–18 (2008)
15. Hurd, J.: Formal Verification of Probabilistic Algorithms. PhD Thesis, University of Cambridge, Cambridge, UK (2002)
16. Mitzenmacher, M., Upfal, E.: Probability and Computing. Cambridge University Press, Cambridge (2005)
17. Nedzusiak, A.: σ-fields and Probability. J. of Formalized Mathematics 1 (1989)
18. Richter, S.: Formalizing Integration Theory, with an Application to Probabilistic Algorithms. Diploma Thesis, Technische Universität München, Department of Informatics, Germany (2003)
19. Widrow, B.: Statistical Analysis of Amplitude-quantized Sampled Data Systems. AIEE Trans. (Applications and Industry) 81, 555–568 (1961)