Verification of Probabilistic Properties in HOL Using the Cumulative Distribution Function

Osman Hasan and Sofiène Tahar

Dept. of Electrical & Computer Engineering, Concordat University
1455 de Maisonneuve W., Montreal, Quebec, H3G 1M8, Canada
{o_hasan,tahar}@ece.concordia.ca

Abstract. Traditionally, computer simulation techniques are used to perform probabilistic analysis. However, they provide inaccurate results and cannot handle large-scale problems due to their enormous CPU time requirements. To overcome these limitations, we propose to complement simulation based tools with higher-order-logic theorem proving so that an integrated approach can provide exact results for the critical sections of the analysis in the most efficient manner. In this paper, we illustrate the practical effectiveness of our idea by verifying numerous probabilistic properties associated with random variables in the HOL theorem prover. Our verification approach revolves around the fact that any probabilistic property associated with a random variable can be verified using the classical Cumulative Distribution Function (CDF) properties, if the CDF relation of that random variable is known. For illustration purposes, we also present the verification of a couple of probabilistic properties, which cannot be evaluated precisely by the existing simulation techniques, associated with the Continuous Uniform random variable in HOL.

Keywords: Interactive Theorem Proving, Higher-Order-Logic, Probabilistic Systems, Cumulative Distribution Function, HOL.

1 Introduction

Probabilistic analysis has become a tool of fundamental importance to virtually all engineers and scientists as they often have to deal with systems that exhibit significant random or unpredictable elements. The main idea behind probabilistic analysis is to model these uncertainties by random variables and then judge the performance and reliability issues based on the corresponding probabilistic properties.

Random variables are basically functions that map random events to numbers. Every random variable gives rise to a probability distribution, which contains most of the important information about this random variable. The probability distribution of a random variable can be uniquely described by its Cumulative Distribution Function (CDF), which is sometimes also referred to as the probability distribution function. The CDF of a random variable $R$, $F_R(x)$, represents

© Springer-Verlag Berlin Heidelberg 2007
the probability that the random variable $R$ takes on a value that is less than or equal to a real number $x$

$$F_R(x) = Pr(R \leq x)$$

(1)

where $Pr$ denotes the probability. The CDF of a random variable contains complete information about the probability model of the random variable and one of its major significance is that it can be used to characterize both discrete and continuous random variables. A distribution is called discrete if its CDF consists of a sequence of finite or countably infinite jumps, which means that it belongs to a random variable that can only attain values from a certain finite or countably infinite set. Discrete random variables can also be characterized by their *probability mass function* (PMF), which represents the probability that the given random variable $R$ is exactly equal to some value $x$, i.e., $Pr(R = x)$. A distribution is called continuous if its CDF is continuous, which means that it belongs to a random variable that ranges over a continuous set of numbers that contains all real numbers between two limits. A Continuous random variable can also be characterized by its *probability density function* (PDF), which represents the slope of its CDF, i.e., $\frac{dF_R(x)}{dx}$. Besides characterizing both discrete and continuous random variables, the CDF also allows us to determine the probability that a random variable falls in any arbitrary interval of the real line. Because of these reasons, the CDF is regarded as one of the most useful characteristic of random variables in the field of probabilistic analysis where the main goal is to determine the probabilities for various events.

Today, simulation is the most commonly used computer based probabilistic analysis technique. Most simulation softwares provide a programming environment for defining functions that approximate random variables for probability distributions. The random elements in a given system are modeled by these functions and the system is analyzed using computer simulation techniques, such as the Monte Carlo Method \cite{17}, where the main idea is to approximately answer a query on a probability distribution by analyzing a large number of samples. The inaccuracy of the probabilistic analysis results offered by simulation based techniques poses a serious problem in highly sensitive and safety critical applications, such as space travel, medicine and military, where a mismatch between the predicted and the actual system performance may result in either inefficient usage of the available resources or paying higher costs to meet some performance or reliability criteria unnecessarily. Besides the inaccuracy of the results, another major limitation of simulation based probabilistic analysis is the enormous amount of CPU time requirement for attaining meaningful estimates. This approach generally requires hundreds of thousands of simulations to calculate the probabilistic quantities and becomes impractical when each simulation step involves extensive computations.

In order to overcome the limitations of the simulation based approaches, we propose to use higher-order logic interactive theorem proving \cite{9} for probabilistic analysis. Higher-order logic is a system of deduction with a precise semantics and can be used for the development of almost all classical mathematics theories. Interactive theorem proving is the field of computer science and mathematical