Formalization of Normal Random Variables in HOL

Muhammad Qasim[1(✉)], Osman Hasan[1], Maissa Elleuch[2,3], and Sofiène Tahar[1]

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada
{m_qasi,o_hasan,tahar}@ece.concordia.ca
[2] CES Laboratory, Sfax University, Sfax, Tunisia
maissa.elleuch@ceslab.org
[3] Digital Research Center of Sfax, Sfax, Tunisia

Abstract. Many components of engineering systems exhibit random and uncertain behaviors that are normally distributed. In order to conduct the analysis of such systems within the trusted kernel of a higher-order-logic theorem prover, in this paper, we provide a higher-order-logic formalization of Lebesgue measure and Normal random variables along with the proof of their classical properties. To illustrate the usefulness of our formalization, we present a formal analysis of the probabilistic clock synchronization in wireless sensor networks.

1 Introduction

Many engineering systems exhibit *normally distributed* elements of randomness. Some notable examples include noise in communication channels, lengths and weights of manufactured goods, message arrival times in communication networks, blood pressure readings of a general population, lifetimes of an electric bulb and maximum speed of a car. The importance of normal distribution is also evident from its relationship with the central limit theorem [2], which states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterations of independent random variables, each with a well-defined expected value and variance, is approximately normally distributed, regardless of the underlying distribution [20]. Therefore, if the sample size is large enough, the sample mean of other distributions may also be treated as normal.

Traditionally, paper-and-pencil based approaches are used for carrying out probabilistic analysis. This method, however, is prone to human error and is not scalable to deal with large systems. Similarly, simulation cannot provide accurate results due to approximations in numerical computations and its incompleteness, which is an outcome of enormous processing time requirements.

Given the safety-critical nature of present age engineering systems, these inaccuracies cannot be tolerated. Higher-order-logic theorem proving, which provides computerized mathematical proofs, can overcome the above-mentioned limitations and has been used to formalize probability theory [16], Markov

M. Kohlhase et al. (Eds.): CICM 2016, LNAI 9791, pp. 44–59, 2016.
DOI: 10.1007/978-3-319-42547-4_4

Chains [10,12] and discrete [8] and continuous [7] random variables. These foundations have been used to formally analyze many aspects of engineering applications, including the Stop-and-Wait protocol [9], wireless sensor networks [3], anonymity and confidentiality protocols [17], oil and gas pipelines [1], multiprocessor systems [13] and reconfigurable memory arrays [6]. However, to the best of our knowledge, no system, exhibiting the Normal random variables, has been reported in the literature. In Isabelle/HOL, there is a formalization of exponential, uniform and normal distributions [19], however, they lack the notion of probability density function and random variables, which play a vital role in analyzing real-world systems. To overcome this limitation, we ported Lebesgue-Borel measure from Isabelle/HOL [11] to HOL4 theorem prover and built upon Mhamdi's formalization of measure, Lebesgue and probability theories [16], available in the HOL4 theorem prover, to formalize probability density function and Normal random variables. We formally verify the correctness of our formalization of Normal random variables by verifying their various properties. These formalizations allow us to formally reason about the correctness of many engineering systems that involve Normal random variables. For illustration purposes, we present a formal analysis of the probabilistic clock synchronization in wireless sensor networks.

2 Preliminaries

2.1 Measure Theory

A measure assigns a number to a set corresponding to its size. Formally, a function defined on a set is a measure if it is positive and countably additive [16].

Definition 1 *(Measure Space).*
A triplet $(X, \mathcal{A}, \mu)$ *is a measure space* iff $(X, \mathcal{A})$ *is a* σ*-field and* $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ *(i.e.,* $\mathbb{R} \cup \{-\infty, +\infty\}$*) is a non-negative and countably additive measure function.*

```
⊢ measure_space (X,A,μ) =
  sigma_algebra (X,A) ∧ positive (X,A,μ) ∧ countably_additive (X,A,μ)
```

The pair $(X, \mathcal{A})$ is called a σ-field or a measurable space and $\mathcal{A}$ is called a sigma algebra over X or a set of measurable sets.

Definition 2 *(Sigma Algebra).*
Let $\mathcal{A}$ *be a collection of subsets (or subset class) of a space* X*.* $\mathcal{A}$ *defines a sigma algebra on* X iff $\mathcal{A}$ *contains the empty set* $\{\}$*, and is closed under countable unions and complementation within the space* X*.*

```
⊢ sigma_algebra (X,A) = subset_class X A ∧ (∀s. s ∈A ⇒ X DIFF s ∈A) ∧
  {} ∈ A ∧ (∀c. countable c ∧ c ⊆ A ⇒ BIGUNION c ∈A)
```

where `subset_class` and `countable` are defined as:

```
⊢ subset_class X A = ∀s. s ∈ A ⇒ s ⊆ X
⊢ countable s = ∃f. ∀x. x ∈s⇒ ∃(n:num). f n = x
```

For any collection G of subsets of X, there is at least one sigma algebra on X containing G, namely the powerset of X. The smallest sigma algebra on X containing G is an intersection of all those sigma algebras, and is called the sigma algebra on X generated by G. This notion is defined in HOL as:

```
⊢ sigma X G = (X, BIGINTER {s | G ⊆ s ∧ sigma_algebra (X,s)})
```

Some helper functions [16] for a σ-field or a measure space are

```
⊢ space (X,A) = X ∧subsets (X,A) = A
⊢ m_space (X,A,μ) = X ∧ measurable_sets (X,A,μ) = A ∧ measure (X,A,μ) = μ
```

For measurable functions, the inverse image of each measurable set is measurable.

Definition 3 *(Measurable Functions).*
Let $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be two measurable spaces. A function $f : X_1 \to X_2$ is called measurable with respect to $(\mathcal{A}_1, \mathcal{A}_2)$ (or $(\mathcal{A}_1, \mathcal{A}_2)$ measurable) iff $f^{-1}(A) \in \mathcal{A}_1$ for all $A \in \mathcal{A}_2$.

```
⊢ f ∈measurable a b =
 sigma_algebra a ∧ sigma_algebra b ∧ f ∈(space a → space b) ∧
  ∀s. s ∈ subsets b ⇒ PREIMAGE f s ∩ space a ∈subsets a
```

2.2 Lebesgue Integration Theory

Similar to the way in which step functions are used in the development of the Riemann integral, the Lebesgue integral makes use of a special class of functions called positive simple functions. In HOL [15] a positive simple function g is represented by the triplet (s, a, α) as a finite linear combination of indicator functions of measurable sets (a_i) that form a partition of the space X.

$$\forall x \in X,\ g(x) = \sum_{i \in s} \alpha_i I_{a_i}(x), \quad \alpha_i \geq 0 \tag{1}$$

where s is a set of partition tags, a_i is a sequence of measurable sets, α_i is a sequence of real numbers and I_{a_i} is an indicator function on a_i:

```
⊢ indicator_fn A = (λx. if x ∈ A then 1 else 0)
```

The Lebesgue integral is first defined for positive simple functions and then extended to non-negative functions.

Definition 4 *(Lebesgue Integral of Positive Simple Functions).*
Let $(X, \mathcal{A}, \mu)$ be a measure space. The integral of the positive simple function g with respect to the measure μ is defined as $\int_X g\, d\mu = \sum_{i \in s} \alpha_i \mu(a_i)$.

```
⊢ pos_simple_fn_integral m s a α = SIGMA (λi. αi * measure m (a i)) s
```

Definition 5 *(Lebesgue Integral of Non-Negative Measurable Functions).*
Let $(X, \mathcal{A}, \mu)$ be a measure space. The integral of a non-negative measurable function f is defined as $\int_X f\, d\mu = \sup \{\int_X g\, d\mu \mid g \leq f$ and g positive simple function$\}$.

```
⊢ pos_fn_integral m f = sup {r | ∃g. r ∈ psfis m g ∧ ∀x. g x ≤ f x}
```

where `r ∈ psfis m g` is equivalent to `r = pos_simple_fn_integral m s a` α and `g` is a positive simple function represented by (s, a, α).

2.3 Probability Theory

The probability space is defined in HOL [16] as a measure space, i.e., (Ω, F, p), where Ω is the sample space, F is a set of events and p is the probability measure such that $p(\Omega) = 1$. A random variable is defined as a measurable function.

Definition 6 *(Random Variable).*

```
⊢ random_variable X p s ⇔
   prob_space p ∧ X ∈ measurable (p_space p,events p) s
```

where `p_space` is a renaming of `m_space` and `events` is a renaming of `measurable_sets`. The probability distribution of a random variable X is defined as the function assigning to A the probability of the event $\{X \in A\}$.

$$\forall A \in \mathcal{B}(\overline{\mathbb{R}}),\ p(\{X \in A\}) = p(X^{-1}(A))$$

Definition 7 *(Probability Distribution).*

```
⊢ distribution p X = (λA. prob p (PREIMAGE X A ∩ p_space p))
```

3 Formalization of Lebesgue-Borel Measure

For evaluating an integral using the Lebesgue integral [16], a suitable Lebesgue measure is required. For this purpose, we have defined a Lebesgue measure based on the Gauge integral. Our formalization is greatly inspired from the formalizations of Lebesgue measure in Isabelle/HOL [11].

3.1 Gauge Integral

Definition 8 *(Gauge Integral).*
Let f:[a,b]→ ℝ be some function, and let y be some number. We say that y is the Gauge integral of f over i written $y = \int_i f(x)\,\mathrm{d}x$*, if for each number e > 0 there exists a Gauge d such that* $|\sum_p f - y| < e$*, where, p is a tagged division of i and p is δ-fine with respect to p.*

```
⊢ (f has_integral_compact_interval y) i = ∀e. 0 < e ⇒ ∃d. gauge d ∧
   ∀p. p tagged_division_of i ∧ d fine p ⇒
      abs (sum p (λ(x,k). content (k) * f(x)) - y) < e
```

An alternate definition of the Gauge integral that simplifies the proof steps for integration over intervals is given as:

```
⊢ (f has_integral y) i =
    if ∃a b. i = interval [a,b] then (f has_integral_compact_interval y) i
    else ∀e. 0 < e ⇒ ∃B. 0 < B ∧ ∀a b. ball (0,B) SUBSET interval [a,b] ⇒
      ∃z. ((λx. if x ∈ i then f x else 0) has_integral_compact_interval z)
           (interval [a,b]) ∧ abs (z - y) < e
```

The functional form of the above definition, using the Hilbert choice operator (@), is as follows,

```
⊢ integral i f = @y. (f has_integral y) i
```

3.2 Borel Measurable Sets

A collection of all borel measurable sets on the real line forms a sigma algebra, called the Borel sigma algebra. It allows us to prove various properties of measurable functions. The Borel sigma algebra is defined as the smallest sigma algebra generated by the open sets of the real line. Mhamdi [16] formalized Borel sigma algebra in the Measure theory as a sigma algebra generated by open intervals of extended real numbers $\overline{\mathbb{R}}$. Because the Gauge integral is formalized for real numbers $\mathbb{R}$ and we are working with Borel measurable functions, we had to formalize real valued Borel sigma algebra in addition to extended real valued Borel sigma algebra. We formalize the real valued Borel sigma algebra in HOL with the help of the `sigma` function, defined in Sect. 2.1.

```
⊢ borel = sigma UNIV {s | open s}
```

where `UNIV` is the universal set of real numbers $\mathbb{R}$ and `open` is defined as:

Definition 9 *(Open Set).*
A set s is called open if, given any point $x \in s$, there exists a real number $\epsilon > 0$ such that, given any point $y \in \mathbb{R}$ whose distance from x is smaller than ϵ, $y \in s$.
`⊢ open s = ∀x. x ∈ s ⇒ ∃ ϵ. ϵ> 0 ∧ ∀y. dist (y,x) < ϵ ⇒ y ∈s`

Using the above definition of `borel`, we proved that all open and closed sets are in Borel sigma algebra.

Theorem 1. *All open and closed sets of $\mathbb{R}$ are in $\mathcal{B}(\mathbb{R})$.*
`⊢ ∀s. {s | open s} ∈ subsets borel ∧{s | closed s} ∈subsets borel`

In order to reuse the proof steps of Mhamdi for proving various properties of our Borel sigma algebra, generated by open sets of real numbers $\mathbb{R}$, we proved that our Borel sigma algebra can also be generated by open intervals of real numbers $\mathbb{R}$.

Theorem 2. *$\mathcal{B}(\mathbb{R})$ is also generated by open intervals of real numbers.*
`⊢ borel = sigma UNIV (IMAGE (λ(a,b). interval (a,b)) UNIV)`

Real-Valued Borel Measurable Functions: For a function to be integrable over a Borel measurable set, it has to be Borel measurable, i.e., the inverse image of the function should belongs to the Borel sigma algebra.

Theorem 3. *If f and g are $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$ measurable and $c \in \overline{\mathbb{R}}$ then $c*f$, $|f|$, f^n, $f+g$, $f*g$ and $max(f,g)$ are $(\mathcal{A}, \mathcal{B}(\mathbb{R})$ measurable.*

```
⊢ ∀a f g h c. sigma_algebra a ∧
   f ∈measurable a Borel ∧g ∈measurable a Borel ⇒
   ((λx. c * f x) ∈measurable a Borel) ∧
   ((λx. abs(f x)) ∈measurable a Borel) ∧
   ((λx. f x pow n) ∈measurable a Borel) ∧
   ((λx. f x + g x) ∈measurable a Borel) ∧
   ((λx. f x * g x) ∈measurable a Borel) ∧
   ((λx. max (f x) (g x)) ∈measurable a borel)
```

Theorem 4. *Every continuous functions is $(\mathcal{B}(\mathbb{R}), \mathcal{B}(\overline{\mathbb{R}}))$ measurable.*

```
⊢ ∀g. g continuous UNIV(:real) ⇒g ∈ measurable borel Borel
```

Notice that `borel` is our Borel sigma algebra generated by open sets of real numbers $\mathbb{R}$ and `Borel` is the Borel sigma algebra of Mhamdi [16] generated by open intervals of extended real numbers $\overline{\mathbb{R}}$.

3.3 Lebesgue Measure

The Lebesgue measure is defined as the supremum of Gauge integrals of X_a for all intervals [-n,n] (or `line n`), where X_a is the indicator function of a set A. We define it as a triplet by pairing it with the Lebesgue space and Lebesgue measurable sets, i.e., all sets for which their indicator function is integrable with respect to the inertval [-n,n].

Definition 10 *(Lebesgue Measure).*

```
⊢ lebesgue = (univ(:real), {A | ∀n. indicator A integrable_on line n},
 (λA. sup {Normal (integral (line n) (indicator A)) | n IN univ(:real)}))
```

where the function `Normal` is used to map real numbers to their corresponding extended real numbers. We prove that Borel measurable sets are also Lebesgue measurable.

Theorem 5. *borel ⊂ lebesgue*

```
⊢ ∀s. s ∈subsets borel ⇒s ∈ measurable_sets lebesgue
```

3.4 Lebesgue-Borel Measure

A Lebesgue measure assigned to Borel measurable sets is called a Lebesgue-Borel measure. We work with the Lebesgue-Borel measure to leverage upon the available formally verified properties of Borel sigma algebra and Borel measurable functions. Thus, we define the triplet of Lebesgue-Borel measure by pairing Lebesgue measure with Borel space and Borel sigma algebra. Also, we prove that Lebesgue-Borel is a sigma finite measure.

Definition 11 *(Lebesgue-Borel Measure).*

```
⊢ lborel = (space borel, subsets borel, measure lebesgue)
```

Theorem 6. *Lebesgue-Borel measure is σ-finite.*
⊢ `sigma_finite_measure lborel`

where `sigma_finite_measure` is defined in HOL as:

```
⊢ sigma_finite_measure (X,A,u) =
   ∃s. countable s ∧ s SUBSET A ∧ (BIGUNION A = X) ∧
       (∀a. a ∈ A ⇒ (u a ≠ PosInf))
```

4 Formalization of Normal Random Variables

Like any other continuous distribution, normal distribution is generally defined by its probability distribution function (PDF) [20]:

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{(-\frac{(x-\mu)^2}{2\sigma^2})} \tag{2}$$

where μ represents its mean and σ is the standard deviation.

4.1 Radon Nikodym Theorem

The Radon-Nikodym derivative of a measure ν with respect to the measure μ is defined as a non-negative measurable function f, satisfying the following formula [5], for any measurable set A:

$$\int_A f \, d\mu = \nu(A) \tag{3}$$

```
⊢ RN_deriv m v = @f. f IN measurable (X,S) Borel ∧ ∀x ∈ X, 0 ≤ f x ∧
                     ∀a ∈ S, integral m (λx. f x × I_a x) = v a
```

The existence of the Radon-Nikodym derivative is guaranteed for absolutely continuous measures by the Radon-Nikodym theorem stating that if ν is absolutely continuous with respect to μ, then there exists a non-negative measurable function f satisfying Eq. (3) for any measurable set A. Mhamdi [16] proved the Radon Nikodym theorem for finite measures. Our main objective is to define the probability density function as a Radon Nikodym derivative of probability measure with respect to the Lebesgue-Borel measure. However, since the Lebesgue-Borel measure is not finite so we have to first generalize the Radon-Nikodym theorem for sigma finite measures.

Theorem 7. *Given a measurable space (X,S), if a measure ν on (X,S) is absolutely continuous with respect to a sigma-finite measure μ on (X,S), then there is a measurable function f, such that for any measurable subset $A \subset X$, $\int_A f \, d\mu = \nu(A)$.*

```
⊢ ∀u v X S. sigma_finite_measure (X,S,u) ∧
    measure_space (X,S,u) ∧ measure_space (X,S,v) ∧
    measure_absolutely_continuous (X,S,u) (X,S,v) ⇒
     ∃f. f ∈ measurable (X,S) Borel ∧ ∀x ∈ X, 0 ≤ f x ∧
        ∀a ∈ S, pos_fn_integral u (λx. f x × I_a x) = v a
```

where `measure_absolutely_continuous` is defined in HOL as:

Definition 12 *(Absolutely Continuous Measures).*
If u and v are two measures on a measure space (X,S), then v is absolutely continuous with respect to u if v(A) = 0 for any A ∈ S such that u(A) = 0.

```
⊢ ∀u v. measure_absolutely_continuous u v =
  ∀A. A ∈ measurable_sets u ∧ (measure v A = 0) ⇒ (measure u A = 0)
```

4.2 Probability Density Function

The distribution of a continuous random variable is usually defined by its PDF:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} p(x)\,dx$$

where $p(x)$ represents the PDF of the random variable x. Formally, the PDF can be defined as a Radon-Nikodym derivative. The distribution of random variables paired with Borel space and Borel sigma algebra gives the probability measure. The PDF of a random variable X is the derivative of the probability measure with respect to the Lebesgue-Borel measure.

Definition 13 *(Probability Density Function).*

```
⊢ PDF X p = RN_deriv lborel
  (space borel, subsets borel, measurable_distr p X)
```

where `measurable_distr` is the same as the distribution in the Probability theory but limited to sets measurable with respect to the Lebesgue-Borel measure. We introduced `measurable_distr` because it is not possible to find the distribution of non-measurable sets.

Definition 14 *(Measurable Distribution).*

```
⊢ measurable_distr p X =
  (λA. if A ∈ measurable_sets lborel then distribution p X A else 0)
```

With the help of the Radon-Nikodym Theorem, discussed in Sect. 4.1, the following properties of PDF were proved in HOL.

Theorem 8. *PDF of a random variable is always positive.*

```
⊢ ∀p X v. (v = (space borel, subsets borel, measurable_distr p X)) ∧
  measure_space v ∧ measure_absolutely_continuous v lborel ⇒
  ∀x. 0 ≤ PDF p X x
```

Theorem 9. *Integral of PDF over the whole space is equal to 1.*

```
⊢ ∀p X v. (v = (space borel, subsets borel, measurable_distr p X)) ∧
  prob_space v ∧ measure_absolutely_continuous v lborel ⇒
  (integral m (PDF p X) = 1)
```

4.3 Normal Random Variables

From Eq. (2), it is clear that the probability density of a Normal random variable, called normal density, is defined by its mean μ and variance σ^2.

Definition 15 *(Normal Density).*

```
⊢ normal_density μ σ x =
  1 / sqrt (2 * π * σ pow 2) * exp (- (x - μ) pow 2 / 2 * σ pow 2)
```

We verified the following useful properties of the normal density.

Theorem 10. *Normal density is always positive.*

```
⊢ ∀ μ σ x. 0 ≤ normal_density μ σ x
```

Theorem 11. *If $0 < \sigma$, then normal density is also greater than 0.*

```
⊢ ∀ μ σx. 0 < σ ⇒0 < normal_density μ σ x
```

Theorem 12. *Normal density is a Borel measurable function.*

```
⊢ ∀ μ σ. (λx. Normal (normal_density μ σ x)) ∈
  measurable (m_space lborel, measurable_sets lborel) Borel
```

where the function `Normal` is used to map real numbers to their corresponding extended real numbers. To prove various properties of Normal random variables, it is required to perform Lebesgue integration on normal density and since the Lebesgue Integral is defined for extended real valued functions, we have to use the function `Normal` in our formalization of normal density.

Now we formalize the probability that an event A (i.e., $P(X \in A)$) will occur for a Normal random variable X.

Definition 16 *(Normal Probability Measure).*

```
⊢ normal_pmeasure μ σA =
  if A ∈ measurable_sets lborel
   then pos_fn_integral lborel
    (λx. Normal (normal_density μ σ x) * indicator_fn A x) else 0
```

Our definition is limited to measurable functions since it is not possible to evaluate the integral of a function over non-measurable sets.

Definition 17 *(Normal Random Variable).*

```
⊢ normal_rv X p μ σ=
  random_variable X p borel ∧ (measurable_distr p X = normal_pmeasure μ σ)
```

The first conjunct indicates that X is a real random variable, i.e., it is measurable from the probability space to Borel space and the second conjunct ensures that it is a Normal random variable.

4.4 Properties of Normal Random Variables

In this section, we prove some interesting properties of Normal random variables. These properties are going to be very useful in minimizing the formal reasoning effort while conducting the formal analysis of real-world applications involving Normal random variables.

Theorem 13. *PDF of a Normal random variable is non-negative.*

```
⊢ ∀X p μ σ. normal_rv X p μ σ ⇒ ∀x. 0 ≤ PDF p X x
```

Theorem 14. *PDF interval over the whole space is equal to 1*

```
⊢ ∀X p μ σ. normal_rv X p μ σ ⇒ (integral lborel (PDF p X) = 1)
```

Theorem 15. *For a Normal random variable X,*

$$\int_{\mu-a}^{\mu} PDF\ p\ X\ dx = \int_{\mu}^{\mu+a} PDF\ p\ X\ dx$$

```
⊢ ∀X p μ σ a. normal_rv X p μ σ ⇒
    pos_fn_integral lborel
      (λx. PDF p X x * indicator_fn {x | μ-a ≤ x ∧ x ≤ μ} x) =
    pos_fn_integral lborel
      (λx. PDF p X x * indicator_fn {x | μ ≤ x ∧ x ≤ μ+a } x)
```

Theorem 16. *For a normal random variable X with p(x) = N(μ, σ),*

$$\int_{-\infty}^{\infty} p(x)\,dx = \int_{-\infty}^{\mu} p(x)\,dx + \int_{\mu}^{\infty} p(x)\,dx$$

```
⊢ ∀X p μ σ. normal_rv X p μ σ ∧
   (A = {x | x ≤ μ}) ∧ (B = {x | μ ≤ x}) ⇒
    pos_fn_integral lborel (λx. PDF p X x) =
    pos_fn_integral lborel (λx. PDF p X x * indicator_fn A x) +
    pos_fn_integral lborel (λx. PDF p X x * indicator_fn B x)
```

Theorem 17. *For a normal random variable X with p(x) = N(μ, σ),*

$$\int_{-\infty}^{\mu} p(x)\,dx = \int_{\mu}^{\infty} p(x)\,dx = \frac{1}{2}$$

```
⊢ ∀X p μ σ. normal_rv X p μ σ ∧A = {x | x ≤ μ} ∧B = {x | μ ≤x} ⇒
  (pos_fn_integral lborel (λx. PDF p X x * indicator_fn A x) = 1 / 2) ∧
  (pos_fn_integral lborel (λx. PDF p X x * indicator_fn B x) = 1 / 2)
```

Theorem 18. *If X is a Normal random variable with mean μ and standard deviation σ, then Y = b + a ∗ X is also a Normal random variable with mean b + a * μ and standard deviation | a | * σ.*

```
⊢ ∀X p μ Y a b. normal_rv X p μ σ ∧ (∀x. Y x = b + a * X x) ∧
      a ≠ 0 ∧ 0 < σ ∧ ⇒ normal_rv Y p (b + a * μ) (abs a * σ)
```

Theorem 19. *Convolution of Normal density with mean $\mu = 0$.*

```
⊢ ∀ σ1 σ2 p X Y x. 0 < σ1 ∧0 < σ2 ∧ normal_rv X p 0 σ1 ⇒
  pos_fn_integral lborel
   (λy. Normal (normal_density 0 σ1 (x - y) *
        Normal (normal_density 0 σ2 y))) =
  Normal (normal_density 0 (sqrt (σ1 pow 2 + σ2 pow 2)) x)
```

Theorem 20. *If $X \sim N(\mu 1, \sigma 1^2)$ and $Y \sim N(\mu 2, \sigma 2^2)$ are two independent Normal random variables, then $Z = X + Y$ is also normal with mean $(\mu 1 + \mu 2)$ and variance $(\sigma 1^2 + \sigma 2^2)$.*

```
⊢ ∀p X Y μ1 μ2 σ1 σ2. prob_space p ∧0 < σ1 ∧0 < σ2 ∧
  indep_var p borel_triplet X borel_triplet Y ∧
  normal_rv X p μ1 σ1 ∧ normal_rv Y p μ2 σ2 ⇒
  normal_rv (λx. X x + Y x) p (μ1 + μ2) (sqrt (σ1 pow 2 + σ2 pow 2))
```

where `borel_triplet` represents `(borel space, subsets borel, (λx. 0))`.

Theorem 21. *If $X_i \sim N(\mu_i, \sigma_i^2)$ is a finite set of independent Normal random variables, and $Z = \Sigma\ X_i$ then, $Z \sim N(\Sigma\ \mu_i, \Sigma\ \sigma_i^2)$.*

```
⊢ ∀p X μ σ I. prob_space p ∧ FINITE I ∧ ∧ I ≠ {} ∧
   indep_vars p (λi. borel_triplet) X I ∧ (∀i, i ∈ I ⇒ 0 < σ i) ∧
   (∀i, i ∈ I ⇒ normal_rv (X i) p (μ i) (σ i)) ⇒
   normal_rv (λx. sum I (λx. X i x)) p (sum I μ)
             (sqrt (sum I (λi. (σ i) pow 2)))
```

where `indep_vars` and `indep_sets` are defined as:

```
⊢ indep_vars p M X I =
   (∀i. i ∈ I ⇒
     random_variable (X i) p (m_space (M i), measurable_sets (M i))) ∧
   indep_sets p
     (λi. PREIMAGE X A INTER p_space p | A ∈ measurable_sets (M i)) I

⊢ indep_sets p F I = prob_space p ∧
   (∀i. i ∈ I ⇒ F i SUBSET events p) ∧
   (∀J. J SUBSET I ∧ J ≠ {} ∧ FINITE J ⇒)
      ∀A. A ∈ (Pi J F) ⇒ (prob p (BIGINTER A j| j ∈ J) =
         Normal (product J (λj. real (prob p (A j)))))
```

where `Pi J F` represents $\{f \mid \forall x.\ x \in J \Rightarrow f(x) \in F(x)\}$. Using above definition of `indep_vars`, two independent random variables are defined as:

```
⊢ indep_var p M_a A M_b B =
   indep_vars p (λi. if i = 0 then M_a else M_b)
                (λi. if i = 0 then A else B) UNIV
```

In the proof of above properties, the theories of Extended Real, Measure, Lebesgue Integral and Probability from HOL4 along with the theory of Lebesgue measure ported from Isabelle/HOL were used. Also, the tactics `SET_TAC` and

`Induct` (on Borel measurable functions) proved to be very useful and were ported from HOL Light and Isabelle/HOL theorem provers. The proof script of the formalization and verification of the notions presented in this paper required around 17500 lines of HOL4 code.

5 Application: Probabilistic Clock Synchronization in Wireless Sensor Networks

Wireless sensor networks involve highly accurate clock synchronization protocols, which require more processing and hence more energy consumption. Due to these unique characteristics, it is difficult to apply traditional approaches for clock synchronization. Elson *et al.* [4] presented an analytical way to convert service specifications to protocol parameters, called Reference Broadcast Sychronization (RBS). PalChaudhuri *et al.* [18] extended this work and provided probabilistic bounds on clock synchronization error for single and multi-hop networks. We conduct the formal analysis for both of these cases as an illustrative example.

The main cause of error in clock synchronization is the non-determinism in message delivery latency. The RBS protocol entails synchronizing a set of receivers with each other, in contrast to synchronizing with the sender. For this reason, the time required to build the message at the sender node and the waiting time required to get access to the transmission channel are identical for all receivers. While the time required for the message to reach the receiver and the processing time required at the receiver may vary.

5.1 Single-Hop Network

Elson *et al.* [4] discovered the distribution of the synchronization error among receivers. Multiple pulses are sent from the sender to the set of receivers. The difference in actual reception time at the receivers is plotted. As each of these pulses are independently distributed, the difference in reception times gives a normal distribution with zero mean. PalChaudhuri *et al.* [18] extended this work and provided probabilistic bounds on clock synchronization error. If the maximum error that is allowed between two sensors is ϵ_{max}, then the probability of synchronization with an error $\epsilon \leq \epsilon_{max}$ is given as

$$P(|\epsilon| \leq \epsilon_{max}) = \frac{\int_{-\epsilon_{max}}^{\epsilon_{max}} \exp^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \tag{4}$$

For n reference packets from the sender, the receivers exchange their observations. The slope of the skew between the receivers is found by a least square linear estimation using the n data points. The calculated slope of the skew has an associated error in it. This error is the difference in phase between the calculated slope and the actual slope. As the points have a normal distribution, this error can be calculated as

$$P(|\epsilon| \leq \epsilon_{max}) = 2\ erf\left(\frac{\sqrt{n}\epsilon_{max}}{\sigma}\right) \tag{5}$$

where ϵ is the synchronization error, i.e., difference in packet reception time between two sensors, ϵ_{max} is the maximum allowable error, n is the minimum number of synchronization messages to guarantee the specified error, σ^2 is the variation of the distribution and *erf* is the error function given as

$$erf\,(z) = \frac{\int_0^z \exp^{-\frac{x^2}{2}}\,\mathrm{d}x}{\sqrt{2\pi}} \tag{6}$$

Definition 18 *(Error Function).*

```
⊢ err_func z = pos_fn_integral lborel
   (λx. Normal (1 / sqrt (2 * π) * exp (-(x pow 2) / 2)) *
   indicator_fn {x | 0 ≤ x ∧ x ≤ z} x)
```

Now we formally verify the result of Eq. (5).

Theorem 22. *Probability of synchronization error for single hop network*

```
⊢ ∀p X μ σ n Emax. prob_space p ∧ (I = (1 .. n)) ∧
   (0 < σ) ∧ (0 < n) ∧ (∀i. i ∈ I ⇒ sync_error (X i) p μ σ) ∧
   (Z = (λx. sum I (λi. X i x) / n)) ∧ (μ = 0) ∧ 0 ≤ Emax ⇒
   (prob_sync_error p Z {x | abs (x) ≤ Emax} =
    2 * err_func (Emax * sqrt n / σ))
```

where `sync_error` is a Normal random variable, Z is the average error for n reference packets, `prob_sync_error p Z` represents the distribution of random variable `Z`, i.e., `measurable_distr p Z` and `Emax` is the maximum allowable synchronization error.

5.2 Multi-hop Network

For this protocol, the senders are considered at various levels. A sender which does not need any synchronization is called a sender at level 0. A sensor node which is within the broadcast region of a sender at level 0 can behave as a sender in order to synchronize sensor nodes, which are two hops away from the sender at level 0. Such a sender is called a sender at level 1. Receivers within the broadcast region of the sender at level 0 are synchronized using the same method discussed in the previous section. Once these receivers get synchronized, each receiver starts behaving as a sender at level 1. In the same manner, suitable time transformations can be performed all along the routing path of the message. We define the transformation for multi-hops in HOL as the sum of synchronization errors and find the maximum synchronization possible along with the probability that the error will stay within bounds for k hops.

Definition 19 *(Transformation).*

```
⊢ transformation X k = (λx. sum (1 .. k) (λi. X i x))
```

Theorem 23. *If Emax is the max allowable error for a single hop, then the maximum error between two sensor nodes, k hops apart, is k * Emax.*

```
⊢ ∀X Emax k. 0 ≤ Emax ⇒
  (∀x. (∀i. (X i) x ∈ {x:real | abs (x) ≤ Emax}) ⇒
  transformation X k x ∈ {x:real | abs (x) ≤ Emax * &k})
```

Theorem 24. *If we consider the error over a single hop to Emax then the error over k hops will be sqrt (k) * Emax.*

```
⊢ ∀p X μ σ k Emax.
   prob_space p ∧ (I = (1 .. n)) ∧ (0 < σ) ∧
   indep_vars p (λi. borel_triplet) X I ∧
   (0 < k) ∧ (∀i. i ∈ I ⇒ sync_error (X i) p μ σ) ∧
   (Z = (λx. sum I (λi. X i x))) ∧ (μ = 0) ∧ (0 ≤ Emax) ⇒
   (prob_sync_error p Z {x | abs (x) ≤ Emax * sqrt(k)} =
    prob_sync_error p (X k) {x | abs (x) ≤ Emax})
```

5.3 Discussion

In this case study, we were able to formally reason about the probabilities of clock synchronization error in single-hop and multi-hop wireless sensor networks with universally quantified variables for various design. This is a novelty which is not available in the simulation based approaches. This added benefit comes at the cost of a significant amount of time and effort spent, while formalizing the systems behavior, by the user. However, the formalization of Normal random variables, presented in Sect. 4 of this paper, greatly facilitated the reasoning process and the proof script corresponding to the application, which only consists of 500 lines of HOL4 code. Besides simulation and testing, the analysis of clock synchronization algorithms for WSN has been sometimes performed using timed automata model checking (e.g. [14,21,22]). However, both probability modeling and scalability in these works were very limited. For example, only a 7 node network was analysed in [14], which is very restricting for wireless sensor networks.

6 Conclusion

The analysis of engineering systems used in safety critical domains, such as transportation and medicine, is usually done using informal techniques. The unreliable results produced using such techniques may lead to heavy financial loss, or even the loss of human lives. Therefore, in this paper we propose to conduct the probabilistic analysis of engineering systems exhibiting normally distributed randomness using higher-order-logic theorem proving. To do so, we have provided a formalization of Normal random variables along with the mathematical notions required to formalize them. Compared to the standard techniques of computer simulation and paper-and-pencil analysis, our approach provides more accurate and trusted results by exploiting the soundness of theorem proving. It also allows to provide generic results instead of proving the properties for specific instances of the system. To prove the usefulness of our formalization, we conducted the formal analysis of the probabilistic clock synchronization in wireless sensor networks. This application highlight the feasibility and benefits of conducting a formal probabilistic analysis using a higher-order-logic theorem prover. Our HOL4 proof script is available for download at http://hvg.ece.concordia.ca/projects/prob-it/pr7.html, and thus can be used for further developments and analysis of different engineering systems.

Acknowledgement. This publication was made possible by NPRP grant # [5 - 813 - 1 134] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author[s].

References

1. Ahmed, W., Hasan, O., Tahar, S., Hamdi, M.S.: Towards the formal reliability analysis of oil and gas pipelines. In: Watt, S.M., Davenport, J.H., Sexton, A.P., Sojka, P., Urban, J. (eds.) CICM 2014. LNCS, vol. 8543, pp. 30–44. Springer, Heidelberg (2014)
2. Billingsley, P.: Probability and Measure. Wiley, New York (2012)
3. Elleuch, M., Hasan, O., Tahar, S., Abid, M.: Formal probabilistic analysis of detection properties in wireless sensor networks. Formal Aspects Comput. **27**(1), 79–102 (2015)
4. Elson, J., Girod, L., Estrin, D.: Fine-grained network time synchronization using reference broadcasts. ACM SIGOPS Oper. Syst. Rev. **36**(SI), 147–163 (2002)
5. Goldberg, R.R.: Methods of Real Analysis. Wiley, New York (1976)
6. 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)
7. Hasan, O., Tahar, S.: Formalization of continuous probability distributions. In: Pfenning, F. (ed.) CADE 2007. LNCS (LNAI), vol. 4603, pp. 3–18. Springer, Heidelberg (2007)
8. Hasan, O., Tahar, S.: Using theorem proving to verify expectation and variance for discrete random variables. Autom. Reasoning **41**(3–4), 295–323 (2008)
9. Hasan, O., Tahar, S.: Performance analysis and functional verification of the stop-and-wait protocol in HOL. Autom. Reasoning **42**(1), 1–33 (2009)
10. Hölzl, J.: Analyzing discrete-time Markov chains with countable state space in Isabelle/HOL (2013). http://home.in.tum.de/hoelzl/classifying/
11. Hölzl, J., Heller, A.: Three chapters of measure theory in Isabelle/HOL. In: van Eekelen, M., Geuvers, H., Schmaltz, J., Wiedijk, F. (eds.) ITP 2011. LNCS, vol. 6898, pp. 135–151. Springer, Heidelberg (2011)
12. Liu, L., Hasan, O., Tahar, S.: Formalization of finite-state discrete-time Markov chains in HOL. In: Bultan, T., Hsiung, P.-A. (eds.) ATVA 2011. LNCS, vol. 6996, pp. 90–104. Springer, Heidelberg (2011)
13. Liu, L., Hasan, O., Tahar, S.: Formal analysis of memory contention in a multiprocessor system. In: Iyoda, J., de Moura, L. (eds.) SBMF 2013. LNCS, vol. 8195, pp. 195–210. Springer, Heidelberg (2013)
14. McInnes, A.I.: Model-checking the flooding time synchronization protocol. In: International Conference on Control and Automation, pp. 422–429. IEEE (2009)
15. Mhamdi, T., Hasan, O., Tahar, S.: On the formalization of the Lebesgue integration theory in HOL. In: Kaufmann, M., Paulson, L.C. (eds.) ITP 2010. LNCS, vol. 6172, pp. 387–402. Springer, Heidelberg (2010)
16. Mhamdi, T., Hasan, O., Tahar, S.: Formalization of entropy measures in HOL. In: van Eekelen, M., Geuvers, H., Schmaltz, J., Wiedijk, F. (eds.) ITP 2011. LNCS, vol. 6898, pp. 233–248. Springer, Heidelberg (2011)
17. Mhamdi, T., Hasan, O., Tahar, S.: Evaluation of anonymity and confidentiality protocols using theorem proving. Formal Methods Syst. Des. **47**(3), 265–286 (2015)

18. PalChaudhuri, S., Saha, A.K., Johnson, D.B.: Adaptive clock synchronization in sensor networks. In: Information Processing in Sensor Networks, pp. 340–348. ACM (2004)
19. Isabelle/HOL Probability Distribution Repository (2016). https://isabelle.in.tum.de/dist/library/HOL/HOL-Probability/Distributions.html
20. Rice, J.A.: Mathematical Statistics and Data Analysis. Duxbury Press, Pacific Grove (1995)
21. Schuts, M., Zhu, F., Heidarian, F., Vaandrager, F.: Modelling clock synchronization in the Chess gMAC WSN protocol. In: Quantitative Formal Methods: Theory and Applications. EPTCS, vol. 13, pp. 41–54 (2009)
22. Zhang, F., Bu, L., Wang, L., Zhao, J., Chen, X., Zhang, T., Li, X.: Modeling and evaluation of wireless sensor network protocols by stochastic timed automata. Electron. Notes Theoret. Comput. Sci. **296**, 261–277 (2013)