Remaining Useful Life Estimation in Prognosis: An Uncertainty Propagation Problem

Shankar Sankararaman *
SGT Inc., NASA Ames Research Center, Moffett Field, CA 94035, USA

Kai Goebel†
NASA Ames Research Center, Moffett Field, CA 94035, USA

The estimation of remaining useful life is significant in the context of prognostics and health monitoring, and the prediction of remaining useful life is essential for online operations and decision-making. However, it is challenging to accurately predict the remaining useful life in practical aerospace applications due to the presence of various uncertainties that affect prognostic calculations, and in turn, render the remaining useful life prediction uncertain. It is challenging to identify and characterize the various sources of uncertainty in prognosis, understand how each of these sources of uncertainty affect the uncertainty in the remaining useful life prediction, and thereby compute the overall uncertainty in the remaining useful life prediction. In order to achieve these goals, this paper proposes that the task of estimating the remaining useful life must be approached as an uncertainty propagation problem. In this context, uncertainty propagation methods which are available in the literature are reviewed, and their applicability to prognostics and health monitoring are discussed.

I. Introduction

Prognostics involves the prediction of future performance of engineering systems, and in turn, predicting their remaining useful life. Prognostics is an important component of system health management and condition-based monitoring; it is important to continuously monitor the performance of the system, perform diagnosis (fault detection, isolation, and estimation), and quantify the remaining useful life in order to aid online decision-making. Sometimes, it may be challenging to perform health monitoring on the whole system due to its sheer complexity, and therefore, diagnosis and prognosis need to be performed on individual components which constitute the overall system. In this approach, mathematical models are developed for individual components, and then the component models are integrated to form the overall system. These models can be used in the health monitoring to guide in model-based diagnostics and prognostics. Alternatively, data-driven approaches are also available for health monitoring; here, several experiments are performed to collect data and this data, in turn, is used to learn about system performance.

This paper focuses on the topic of estimation of remaining useful life (RUL) in prognosis, and is applicable to both physics-based and data-driven approaches. In order to predict the RUL at any given time instant, there are three essential steps that need to be performed. The first step is to estimate the present health of the system at the given time instant. The second step is to predict the health of the system in future (continuously as a function of time) using a degradation model which may either be physics-based or data-driven. The third step is to define a threshold function which defines the end of life; this threshold function is a binary function and can be used to calculate the remaining useful life. The remaining useful life prediction is affected by several sources of uncertainty such as modeling errors, measurement errors, future loading uncertainty, etc., and it is important to accurately account for these sources of uncertainty while estimating the RUL. It is important to understand that the uncertainty in RUL is simply a dependent quantity, dependent on the aforementioned sources of uncertainty. The major goal of this paper is to establish the mathematical

*Post-Doctoral Research Scholar, NASA Ames Research Center, Moffett Field, CA 94035, USA.
†Intelligent Systems Division, NASA Ames Research Center, Moffett Field, CA 94035, USA.
relationship between the various sources of uncertainty and the uncertainty in the RUL prediction. In the process, this paper explains that the estimation of RUL needs to be approached as an uncertainty propagation problem in order to accurately quantify the uncertainty in RUL prediction, and understand how each of the different sources of uncertainty affects the uncertainty in RUL prediction.

The rest of the paper is organized as follows. Section II discusses the various sources of uncertainty in the context of prognostics, and explains how this uncertainty must be interpreted. Section III mathematically formulates the RUL prediction problem, establishes the relationship between RUL and the different sources of uncertainty, and illustrates how the RUL prediction problem becomes an uncertainty propagation problem. Section IV discusses the different types of methods for uncertainty propagation, and investigates their relevance in the context of prognostics and system health management.

II. Uncertainty in Prognostics

As mentioned earlier in the introduction, the presence of uncertainty has a significant impact on prognostics and the remaining useful life prediction. When the state estimates, future loading conditions, operating conditions, etc. are uncertain, the future states and the remaining useful life also become uncertain. While non-probabilistic methods\(^4\) such as Fuzzy logic, possibility theory, Dempster-Shafer theory, Evidence theory, etc. may have been used for the treatment of uncertainty, probabilistic methods have predominantly used for uncertainty representation in prognostics.\(^5,6,7\) Further, probabilistic approaches are contextually meaningful for uncertainty representation and quantification since they are consistent with decision-theory analysis. Therefore, the rest of this paper focuses only on probabilistic approaches for uncertainty quantification and propagation.

While the mathematical axioms and theorems of probability have been well-established the literature, there is considerable disagreement among researchers on the interpretation of probability. There are two major interpretations based on physical and subjective probabilities, respectively. Physical probabilities,\(^8\) also referred to objective or frequentist probabilities, are related to random physical systems such as rolling dice, tossing coins, roulette wheels, etc. Each trial of the experiment leads to an event (which is a subset of the sample space), and in the long run of repeated trials, each event tends to occur at a persistent rate, and this rate is referred to as the relative frequency. These relative frequencies are expressed and explained in terms of physical probabilities. Thus, physical probabilities are defined only in the context of random experiments. On the other hand, subjective probabilities\(^9\) can be assigned to any “statement”. It is not necessary that the concerned statement is in regard to an event which is a possible outcome of a random experiment. In fact, subjective probabilities can be assigned even in the absence of random experiments. The Bayesian methodology is based on subjective probabilities, which are simply considered to be degrees of belief and quantify the extent to which the statement is supported by existing knowledge and available evidence. Calvetti and Somersalo\(^10\) explain that “randomness” in the context of physical probabilities is equivalent to “lack of information” in the context of subjective probabilities. In this approach, even deterministic quantities can be represented using probability distributions which reflect the subjective degree of the analyst’s belief regarding such quantities.

This leads to the obvious question - is one particular interpretation more suitable to prognostics? In general, both interpretations may be suitable. However, in the particular context of condition-based monitoring or online health monitoring, there is only one system which is being monitored, and hence, at any time instant, there is no “physical randomness” associated with the system (from a frequentist point of view). Therefore, any quantity associated with a system, even though it may be uncertain, cannot be represented using a probability distribution, following the frequentist interpretation of probability. Nevertheless, system state estimation during health monitoring is commonly performed using particle filters and Kalman filters, and these approaches compute probability distributions for the state variables; therefore, the only possible explanation for such calculation is that the subjective (Bayesian) approach is being inherently used for uncertainty quantification. Such filtering approaches are known as “Bayesian tracking” methods not only because they make use of Bayes theorem, but also fall within the realm of subjective probability. This implies that the uncertainty estimated through the aforementioned filtering algorithms are simply reflective of the analyst’s degree of belief, and not related to actual physical probabilities. Having reviewed the physical meaning of uncertainty in the context of health monitoring, the commonly encountered sources of uncertainty in prognostics are listed below:

1. **Measurement Errors:** In the context of health monitoring or condition-based monitoring, measure-
ments from the system are continuously available. These measurements may be uncertain due to the presence of sensor bias and sensor noise.

2. **Model uncertainty errors**: The models used for representing the system behavior have different types of uncertainty associated with them. First, the model parameters may be uncertain, and second, the model form may not capture the true underlying behavior. This is sometimes expressed through the use of process noise, but this is not an accurate representation of the true model form error.

3. **Present State Estimate**: As stated earlier, state estimation is commonly performed through Bayesian tracking methods such as Kalman filter, particle filter, etc. These approaches account for the uncertainty in the model and measurements, and estimate the state of the system as a random variable, and therefore impart additional uncertainty to prognostic calculations.

4. **Future loading and operating conditions**: One important challenge in prognostics is to anticipate the future loading on the system, and predict the future operating conditions. It is practically impossible to be certain about such predictions, and there is always an element of subjectivity while assessing the uncertainty regarding such variables.

At this juncture, it must be acknowledged that accurate quantification of the various sources of uncertainty is very challenging in practical applications. Sometimes, the resultant uncertainty may be high and it may be desirable to reduce some of these uncertainties. While some of these uncertainties can be reduced by improving measurement techniques or modeling techniques, it is practically impossible to eliminate them altogether. However, representing them and accounting for them in prognostic calculations is extremely important, because it directly affects decision-making. In fact, there are several PHM approaches quantify risk based on uncertainty quantification in an algorithm's output and incorporate it into a corresponding cost-benefit equation through monetary concepts. Therefore, the next section focuses on how these sources of uncertainty affect prognostic calculations and RUL estimation.

### III. Remaining Useful Life Estimation in Prognosis

This section discusses uncertainty quantification in prognostics, with a focus on computing the RUL prediction. In prognostics, the remaining useful life at a generic time-instant \( t_P \) is a condition-based estimation of the usage time left until failure, using measurements of key variables and past usage information up to time \( t_P \). This process typically consists of forecasting the future state of health beyond \( t_P \) and identifying when the state of health will cross a failure threshold representative of a functional failure. In addition, RUL in prognostics considers future usage (loading and operating) conditions. As a result, the probability distribution of the states and the RUL prediction will continuously vary as a function of time (the time at which prediction is performed).

Typically, the procedure for RUL computation (at any generic time \( t_P \)) consists of three steps:

1. **Present State Estimation**
2. **Future State Prediction**
3. **RUL computation**

The first step involves estimating the state at the given time-instant \( t_P \). Conventionally this is performed using a filtering algorithm; while the Kalman filter can be used for linear models where the state variables are assumed to follow Gaussian (normal) distributions, particle filter is used when the state variables are assumed to follow non-normal distributions. For details of filtering approaches, refer to Kalman and Doucet et al. At the end of the first step, the state estimate is available in form of a probability distribution whose density function (PDF) is given by \( f_X(x) \). If there is more than one state, then this density function is the joint density function (denoted by \( f_X(x) \)) of all the states at the time instant \( t_P \). If Kalman filter is used, then this PDF is Gaussian, and if a particle filter is used, the distribution is typically non-parametric and expressed in terms of samples drawn from the probability distribution.

The second step is to forecast the future states using the health degradation model; this model is represented in terms of the state space equations, as follows:

\[
\dot{x}(t) = h(x(t), u(t), w(t))
\]
In the above equation, $u(t)$ and $w(t)$ denote the loading and model error (process noise) at any generic time instant $t$, and $h$ is the functional model that represents the state evolution. This model may be developed using first-principle physics or using data-driven approaches. Note that scalar state variables have been shown, only for the purpose of illustration; the prediction methodology can be extended to consider a vector of state variables. Using Eq. 1, it is possible to forecast the state value from the given time instant $t_P$ until any generic later time instant $t_l$ ($t_l \geq t_P$). The forecast value obviously depends on the loading values and model errors from time $t_P$ until time $t_l$. As stated earlier in Section II, these loading values and model errors are also uncertain quantities, and are assumed to have been characterized before RUL computation.

The third and final step of RUL computation is complicated from an analytical point of view. It is first necessary to define the end of life using a threshold function. This threshold function is evaluated at any generic time $t_l$; its output at any time instant is binary, indicating whether failure has occurred or not. For the sake of illustration, assume that there are no uncertainties, and all quantities (initial state $x_p$, and loading and model errors at all times) are deterministically known. Starting with the state value of $x_p$, the state value can be continuously forecasted until the first time instant when the end of life (as determined using the threshold function) is attained; this time instant corresponds to the end of life, and is denoted as $t_{EOL}$. It can be easily seen that $t_{EOL}$, when calculated at time $t_P$, is a function of:

1. State value at the time at which prediction needs to be made, i.e., at $t_P$. These state values are denoted by $x(t_P)$.
2. Loading values continuously from time $t_P$ until $t_{EOL}$; the operating conditions, if known, may also be included. Let $u$ denote this vector.
3. Model error values from time $t_P$ until $t_{EOL}$; let $w$ denote this vector.

The evaluation of end-of-life can be graphically represented as shown in Fig. 1, and expressed mathematically as:

$$EOL(t_P) = \Psi(x(t_P), u, w)$$

Then, the RUL at time $t_P$ is calculated as:

$$RUL(t_P) = EOL(t_P) - t_P$$

Note that $t_P$ is deterministic, since the time at which prediction needs to be performed is known. In fact, the above two equations can be combined, and the RUL can be directly expressed mathematically as:

$$RUL(t_P) = G(x(t_P), u, w)$$

Figure 1: Definition of $\Psi$

The evaluation of end-of-life can be graphically represented as shown in Fig. 1, and expressed mathematically as:

$$EOL(t_P) = \Psi(x(t_P), u, w)$$

Then, the RUL at time $t_P$ is calculated as:

$$RUL(t_P) = EOL(t_P) - t_P$$

Note that $t_P$ is deterministic, since the time at which prediction needs to be performed is known. In fact, the above two equations can be combined, and the RUL can be directly expressed mathematically as:

$$RUL(t_P) = G(x(t_P), u, w)$$
Thus, using Eq. 4, for every value of $x(t_P)$, $u$, and $w$, the value of RUL can be computed. However, the values of these variables are uncertain, and only probability distributions may be available for them. Since these variables are uncertain, RUL is also uncertain, and hence, the goal would be to compute the probability distribution of RUL. This probability distribution of RUL can be computed by “propagating” the uncertainty in $x(t_P)$, $u$, and $w$ through $G$ in Eq. 4. Hence, the estimation of RUL in prognosis is simply an uncertainty propagation problem, and well-established statistical tools for uncertainty propagation may be investigated for this purpose. The following section is devoted to this topic, and discusses the relevance of well-known uncertainty propagation methods in prognostics and health monitoring.

IV. Uncertainty Propagation Methods for RUL Estimation

Researchers in the areas of non-deterministic methods and uncertainty quantification techniques have developed different types of statistical methods for uncertainty propagation during the past 30 years. The most general case of uncertainty propagation considers the mathematical function given by:

$$Y = G(X_1, X_2...X_n)$$

(5)

It is clear that the above equation is very similar to Eq. 4 in Section III. Here, there are $n$ inputs given by $X_i$ ($i = 1$ to $n$), and the uncertainty in each input is given by the probability density function (PDF) $f_{X_i}(x_i)$ or the cumulative distribution function (CDF) $F_{X_i}(x_i)$. The joint PDF of all inputs is denoted as $f_X(x)$.

The goal in uncertainty propagation is to compute the uncertainty in $Y$, either in terms of the PDF $f_Y(y)$ or CDF $F_Y(y)$. The entire CDF $F_Y(y)$ can be calculated as:

$$F_Y(y) = \int_{y(X)<y} f_X(x)dx$$

(6)

It is harder to write a similar expression for PDF calculation, although the following equation attempts to.

$$f_Y(y) = \int f_Y(y|x) f_X(x)dx$$

(7)

In Eq. 7, the domain of integration is such that $f_X(x) \neq 0$. Note that Eq. 7 is not very meaningful because $y$ is single-valued given $x$, and hence $f_Y(y|x)$ is nothing but a Dirac delta function. Alternatively, the PDF can be calculated by differentiating the CDF, as:

$$f_Y(y) = \frac{dF_Y(y)}{dy}$$

(8)

The different methods which have been used by researchers for uncertainty quantification aim at solving the above equations in mathematically intelligent ways. These methods can be classified into two types – sampling-based and analytical methods; while some may calculate the PDF of $Y$, other methods calculate the CDF.

A. Sampling-based Methods

The most intuitive method for uncertainty propagation is to make use of Monte Carlo simulation (MCS). The basic underlying concept of Monte Carlo simulation is to generate a pseudo-random number which is uniformly distributed on the interval [0, 1]; then the CDF of $X$ is inverted to generate the corresponding realization of $X$. Following this procedure, several random realizations of $X$ are generated, and the corresponding random realizations of $Y$ are computed. Then the CDF $F_Y(y)$ is calculated as the proportion of the number of realizations where the output realization is less than a particular $y_c$. The generation of each realization requires one evaluation/simulation of $G$. Several thousands of realizations may often be needed to calculate the entire CDF, especially for very high/low values of $y$. Error estimates for the CDF, in terms of the number of simulations, are available in the literature. Alternatively, the entire PDF $f_Y(y)$ can be computed by constructing a histogram based on the available samples of $Y$, using kernel density estimation.

There are several variations of the basic Monte Carlo algorithm which are used by several researchers. Some of these approaches are listed below:
1. **Importance Sampling**: This algorithm does not generate random realizations of \( X \) from the original distribution. Instead, random realizations are generated from a proposal density function, statistics of \( Y \) are estimated and then corrected based on the original density values and proposal density values.

2. **Stratified Sampling**: In this sampling approach, the overall domain of \( X \) is divided into multiple sub-domains and samples are drawn from each sub-domain independently. The process of dividing the overall domain into multiple sub-domains is referred to as stratification. This method is applicable when subpopulations within the overall population are significantly different.

3. **Latin Hypercube Sampling**: This is a sampling method commonly used in design of computer experiments. When sampling a function of \( N \) variables, the range of each variable is divided into \( M \) equally probable intervals, thereby forming a rectangular grid. Then, sample positions are chosen such that there is exactly one sample in each row and exactly one sample in each column of this grid. Each resultant sample is then used to compute a corresponding realization of \( Y \), and thereby the PDF \( f_Y(y) \) can be calculated.

4. **Unscented Transform Sampling**: Unscented transform sampling\(^{19} \) is a sampling approach which focuses on estimating the mean and variance of \( Y \) accurately, instead of the entire probability distribution of \( Y \). Certain pre-determined sigma points are selected in the \( X \) — space and these sigma points are used to generate corresponding realizations of \( Y \). Using weighted averaging principles, the mean and variance of \( Y \) are calculated.

### B. Analytical Methods

A new class of methods was developed by reliability engineers in order to facilitate efficient, quick but approximate calculation of the CDF \( F_Y(y) \); the focus is not on the calculation of the entire CDF function but only to evaluate the CDF at a particular value \( (y_c) \) of the output, i.e. \( F_Y(Y = y_c) \).

The basic concept is to “linearize” the model \( G \) so that the the output \( Y \) can be expressed as a linear combination of the random variables. Further, the random variables are transformed into uncorrelated standard normal space and hence, the output \( Y \) is also a normal variable (since the linear combination of normal variables is normal). Therefore, the CDF value \( F_Y(Y = y_c) \) can be computed using the standard normal distribution function. The transformation of random variables \( X \) into uncorrelated standard normal space \( (U) \) is denoted by \( U = T(X) \), and the details of the transformation can be found in Haldar and Mahadevan.\(^{17} \)

Since the model \( G \) is non-linear, the calculated CDF value depends on the location of “linearization”. This linearization is done at the so-called most probable point (MPP) which is the shortest distance from origin to the limit state, calculated in the \( U \) — space. Then, the CDF is calculated as \( F_Y(y_c) = \Phi(-\beta) \), where \( \Phi \) denotes the standard normal CDF function, and \( \beta \) denotes the aforementioned shortest distance. The MPP and the shortest distance are estimated through a gradient-based optimization procedure. This optimization is solved using the well-known Rackwitz-Fiessler algorithm,\(^{20} \) which is in turn based on repeated linear approximation of the non-linear constraint \( G(x) - y_c = 0 \). This method is popularly known as the first-order reliability method (FORM). There are also several second order reliability methods (SORM) based on the quadratic approximation of the limit state.\(^{17,21,22,23} \)

The entire CDF can be calculated using repeated FORM analyses by considering different values of \( y_c \); for example, if FORM is performed at 10 different values of \( y_c \), the corresponding CDF values are calculated, and an interpolation scheme can be used to calculate the entire CDF, which can be differentiated to obtain the PDF. This approach is difficult because it is almost impossible to choose such multiple values of \( y_c \), because the range (i.e. extent of uncertainty) of \( Y \) is unknown. This difficulty is overcome by the use of an inverse FORM method\(^{24,25} \) where multiple CDF values are chosen and the corresponding values of \( y_c \) are calculated. This approach is simpler because it is easier to choose multiple CDF values since the range of CDF is known to be \([0, 1]\).  

### C. Discussion

Since sampling-based methods may require several thousands of “samples” or “particles” in order to accurately calculate the PDF or CDF, they are time consuming and hence, may not be suitable in the context of online prognostics and decision-making. Further, in general, sampling-based methods (other than the
unscented transform sampling approach) are not “deterministic methods”; in other words, every time a sampling-based algorithm is executed, it may result in a slight different PDF or CDF. The ability to produce a deterministic solution is sometimes an important criterion for existing verification, validation, and certification protocols in the aerospace domain.

On the other hand, analytical methods are not only computationally cheaper but also usually deterministic; in other words, they produce the same PDF or CDF every time the algorithm is executed. However, these analytical methods are still based on approximations, and not readily suitable to account for all types of uncertainty in prognosis. For example, consider the FORM method, which is solved using gradient-based optimization equations. If $t_{EOL} >> t_p$, then the number of elements in \( u \) and \( w \) may be of the order of a few hundreds or thousands, and hence, it is necessary to compute hundreds or thousands of derivatives of Eq. 5. In that case, the computational efficiency of the analytical approach is as good (or as bad) as sampling-based approaches. It is clear from the above discussion that, though uncertainty propagation methods may be available in the literature, it is challenging to make direct use of them for prognostics.

In addition to the above described methods, researchers have also advocated the use of surrogate models for uncertainty propagation. These surrogate models approximate the function \( G(x) \) using different types of basis functions such as radial basis, Gaussian basis, Hermite polynomials, etc. These surrogate models are inexpensive to evaluate and therefore, facilitate efficient uncertainty propagation. Future research will investigate the use of such surrogate models for uncertainty quantification in prognostics.

V. Conclusion

It is important to accurately estimate the remaining useful life (RUL) prediction in the context of prognostics and condition-based monitoring. This is a challenging problem since prognostics deals with future prediction which is affected by several sources of uncertainty. Therefore, a meaningful prognostics algorithm must be able to account for these sources of uncertainty and predict the uncertainty in future behavior as well as the remaining useful life prediction. Estimating the uncertainty in prognostics is important for decision-making as it can guide several activities such as fault mitigation, fault recovery, mission replanning, etc.

This paper discussed the various issues related to uncertainty quantification in prognostics and explained that it is useful to view the problem of estimating the uncertainty in prognostics as an uncertainty propagation problem. The uncertainty in present conditions and estimates of future conditions can be propagated through prediction model to quantify the uncertainty in the remaining useful life prediction. Certain fundamental principles of uncertainty propagation were explained and mathematical techniques were discussed. Different types of sampling methods and analytical approaches were outlined. Future research needs to delve deeper into these approaches and investigate their applicability to prognostics and system health management.
Acknowledgment

The work reported herein was in part funded by the NASA System-wide Safety Assurance Technologies (SSAT) project under the Aviation Safety (AvSafe) Program of the Aeronautics Research Mission Directorate (ARMD), and by the NASA Automated Cryogenic Loading Operations (ACLO) project under the Office of the Chief Technologist (OCT) of Advanced Exploration Systems (AES).

References


