A Particle Filtering-based Framework for On-line Fault Diagnosis and Failure Prognosis

A Dissertation Proposal
Presented to
The Academic Faculty
By

Marcos E. Orchard

In Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy in Electrical Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
August 2006
# Table of Contents

**Summary** .............................................................................................................................. VI

**Section 1.1 Introduction** ........................................................................................................ 1
  1.1.1 Motivation and Objectives ......................................................................................... 1
  1.1.2 Method Overview .................................................................................................... 4
  1.1.3 Contributions ........................................................................................................ 7

**Section 1.2 Origin and History of the Problem** ................................................................. 9
  1.2.1 Scope and Aim of the section .................................................................................. 9
  1.2.2 Nonlinear Bayesian Filtering ................................................................................... 11
  1.2.3 Sequential Monte Carlo Methods: Particle Filtering ........................................... 15
    1.2.3.1 Importance Sampling Scheme ........................................................................ 16
    1.2.3.2 Resampling step: the SIR Particle Filter ....................................................... 20
  1.2.4 Improved Sequential Monte Carlo Methods ........................................................... 24
    1.2.4.1 Optimal Importance Density Function ......................................................... 24
    1.2.4.2 Regularized Particle Filter ............................................................................ 27
    1.2.4.3 Model Parameter Estimation using Particle Filters ...................................... 30
  1.2.5 Particle Filtering in Real-Time Diagnosis Applications .......................................... 33
    1.2.5.1 Variable Resolution Particle Filters ............................................................ 36
    1.2.5.2 Risk Sensitive Particle Filters ..................................................................... 37
  1.2.6 Particle Filtering in Real-Time Prognosis Applications ........................................... 38

**Section 1.3 Proposed Research** ........................................................................................... 42
  1.3.1 Particle Filtering for Diagnosis in Stochastic Nonlinear Systems .......................... 42
  1.3.2 Particle Filtering for Prognosis: Generation of Long Term Prediction in Stochastic Nonlinear Systems ................................................................. 43
  1.3.3 The Inner and Outer Loops for Diagnosis and Prognosis ......................................... 44

**Section 2.1 Preliminary Research** ....................................................................................... 47
  2.1.1 Particle Filtering for Diagnosis in Stochastic Nonlinear Systems .......................... 47
    2.1.1.1 Detection of Cracks in Blades of a Turbine Engine ...................................... 48
    2.1.1.2 Detection of Degradation Level for a Battery .............................................. 52
    2.1.1.3 Detection of Cracks in a UH-60 Planetary Gear Plate .................................. 54
  2.1.2 Particle Filtering for Prognosis in Stochastic Nonlinear Systems ............................ 57
    2.1.2.1 First Prognosis Level: Generation of Long-Term Predictions .................... 60
      2.1.2.1.1 P-step ahead Long-Term predictions: First Approach ...................... 60
      2.1.2.1.2 P-step ahead Long-Term predictions: Second Approach ............... 62
      2.1.2.1.3 P-step ahead Long-Term predictions: Third Approach ............... 65
2.1.2.2 Second Prognosis Level: Estimation and Statistical Characterization of the Remaining Useful Life (RUL) of Equipment ........................................ 66
2.1.3 Illustrative Example: RUL Statistical Characterization............................ 69
2.1.4 Case Study: Analysis of Crack Growth in a Turbine Engine Blade ........ 74
2.1.5 Case Study: UH-60 Planetary Gear Plate. Analysis of Axial Crack Growth ........................................................................................................ 78
2.1.6 Discussion & Conclusion.......................................................................... 91

SECTION 2.2 REMAINING RESEARCH ................................................................... 94
2.2.1 The Inner and Outer Loops for Diagnosis and Prognosis......................... 94

SECTION 2.3 FACILITIES NEEDED ..................................................................... 97
2.3.1 Computational Requirements................................................................. 97
2.3.2 Failure Ground Truth Data................................................................. 98

REFERENCES ............................................................................................................. 100
DISCLAIMER STATEMENT

Section 1 (background/literature survey) of this proposal was prepared without input from my research advisor or any other person. While technical writing guides may have been referred to, I did not solicit or receive assistance from any other person while preparing this portion of this document.

Signed _________________________ (student)  ________________________ (print name)

Approved: ______________________ (research advisor)  ______________________ (print name)
LIST OF FIGURES

Figure 2.1.1-1. Particle filter-based FDI module. Crack in a turbine engine blade........... 49
Figure 2.1.1-2. Detail of detection results for a crack in a turbine engine blade............. 51
Figure 2.1.1-3. Particle filter-based FDI module. Battery degradation problem........... 53
Figure 2.1.1-4. Particle filter-based FDI module. Results for battery degradation problem
........................................................................................................................................ 53
Figure 2.1.1-5. Particle filter-based FDI module. Cracked plate problem................... 56
Figure 2.1.2-1. Correction algorithm for RUL expectation........................................... 68
Figure 2.1.3-1. Result comparison for RUL statistical characterization....................... 71
Figure 2.1.4-1. Picture of turbine engine under study .................................................. 74
Figure 2.1.4-2. Prognosis results for crack growth in a turbine engine blade.............. 77
Figure 2.1.5-1. ANSYS model of the planetary gear plate, showing crack location..... 79
Figure 2.1.5-2. Loading profile diagram versus GAG cycles................................. 80
Figure 2.1.5-3. Deterministic bounds for crack length evolution vs. GAG cycles....... 82
Figure 2.1.5-4. Particle-filtering-based approach for prognosis for the study of crack
  growth in a planetary gear plate...................................................................................... 86
Figure 2.1.5-5. Prediction results for a single hazard threshold .................................. 88
Figure 2.1.5-6. Time-varying model parameter vs. GAG cycles.............................. 89
Figure 2.1.5-7. Prediction results for a unique hazard zone at 6.2”............................. 90
Figure 2.2.1-1. Application example for inner and outer loops in failure prognosis..... 95
Figure 3.2.1-1. Combined approach for block sampling strategies Error! Bookmark not defined.

LIST OF TABLES

Table 2.1.5-1. Prediction results for particle-filtering-based approach for prognosis .... 85
Summary

The object of the proposed research is to implement an on-line particle-filtering-based framework for fault diagnosis and failure prognosis in nonlinear, non-Gaussian systems. The methodology assumes the definition of a set of fault indicators, which are appropriate for monitoring purposes, the availability of real-time process measurements, and the existence of empirical knowledge (or historical data) to characterize both nominal and abnormal operational conditions.

As it will be shown, the incorporation of particle filtering (PF) techniques in the proposed scheme not only allows for the implementation of real-time algorithms, but also provides a solid theoretical framework to handle the problem of fault detection, identification, and failure prognosis. Founded on the concept of sequential importance sampling (SIS) and Bayesian theory, these methods seek to approximate the conditional state probability distribution $p(x_t | y_{[t]})$ by a swarm of points called “particles” and a set of weights representing discrete probability masses. Particles can be easily generated and recursively updated given a nonlinear process model, a measurement model and a set of available measurements $Y = \{Y_t , t \in \mathbb{N}\}$.

Two autonomous modules have been considered in the proposed approach. On one hand, a fault detection and identification (FDI) module uses a hybrid state-space model of the plant and a particle filtering algorithm to calculate the probability of any given fault condition in real time, simultaneously providing information about type II
detection errors and other critical confidence statistics. Once a fault mode is statistically confirmed, current estimates for the state probability density function (pdf) are used as initial conditions in the prognostic module, hence allowing swift transitions between FDI and prognostic routines.

The failure prognosis module, on the other hand, computes the remaining useful life (RUL) pdf of the faulty subsystem in real-time, using a particle-filtering-based algorithm that consecutively updates the current state estimate for a nonlinear state-space model (with unknown time-varying parameters) and predicts the evolution in time of the fault indicator pdf. The outcome of the prognosis module provides information about precision and accuracy of long-term predictions, RUL expectations, 95% confidence intervals, and other hypothesis tests for the failure condition under study. Finally, inner and outer correction loops (learning schemes) are used to periodically improve the parameters that characterize the performance of FDI and/or prognosis algorithms. Data from a seeded fault test for a UH-60 planetary gear plate is used to validate all proposed approaches.

Expected contributions of this research include the establishment of a general methodology for reliable real-time FDI and prognosis in nonlinear processes with unknown model parameters, the definition of appropriate procedures to generate dependable statistics about fault conditions, and the characterization of specific ways to utilize information from real-time measurements to improve the precision and accuracy of the predictions for the state probability density function (pdf).
Section 1.1 Introduction

1.1.1 Motivation and Objectives

Fault diagnosis and failure prognosis for complex systems have become key issues in a world where the economic impact of reliability related issues and the cost-effective operation of critical assets is steadily increasing. On one hand, failure diagnosis involves the detection of a fault in the system and the identification of its cause and location, which are indeed critical tasks in any health management system (HMS). Prognosis – as a natural extension to the fault detection and identification (FDI) problem – intends to characterize the evolution in time of the detected insipient failure condition, thus allowing the estimation of the remaining useful life (RUL) for affected subsystems or components.

Several examples can be cited here to illustrate the range of applications for these types of algorithms: electro-mechanical systems, continuous-time manufacturing processes, structural damage analysis, and even fault tolerant software architectures. Most of them have in common the fact that they are highly complex, nonlinear, and affected by large-grain uncertainty.

The task can be particularly difficult when the system under study is operating in real-time, especially for prognosis. Most of the approaches currently available in the reliability arena involve intensive computations and the processing of large amounts of historical data. More importantly, the obtained results do not necessarily include
knowledge about the physics of the system and there is little room left for on-line updates in the predicted RUL when the system is behaving differently from what is expected. Learning paradigms — so useful in the control field — are scarcely applied for prognostic purposes, thus limiting the implementation of automatic contingency management (ACM) systems or other automated corrective schemes. In addition, even in the case when efficient FDI algorithms are implemented, there are no unified approaches that can perform the transition from FDI results to prognostic modules.

In that sense, the research work proposed here intends to establish a general framework to deal with the problems of real-time fault detection, identification and failure prognosis via the utilization of particle filtering (PF) techniques, an emerging methodology for sequential signal processing that is very suitable in the case when the system is nonlinear or in the presence of non-Gaussian process/observation noise. To accomplish this objective, three main research tasks have to be simultaneously achieved.

The first main objective is to implement an on-line particle-filtering-based framework for fault detection and identification (FDI) in nonlinear, non-Gaussian systems. This architecture should be able to pinpoint both the presence and nature of a fault condition in real time, given a set of measurements and a characterization of the plant behavior under nominal operational conditions (baseline data). Furthermore, it is expected that this scheme will provide the means for a swift transition between the FDI and prognostic applications.
The second main research objective focuses on the use of a particle-filtering-based framework for on-line failure prognosis in nonlinear, non-Gaussian systems. It is expected that this implementation will statistically characterize the remaining useful life (RUL) of a subsystem or piece of equipment affected by a fault condition, i.e., to estimate the probability density function of the subsystem RUL. A set of measurements will be used to improve current estimates, and nonlinear state-space models with unknown time varying parameters will define the evolution in time of the fault indicator. The outcome of the prognosis module, namely the RUL pdf, will be available and updated in real time, providing information about statistical confidence intervals, expectations, and other hypothesis tests for the failure condition under study.

The last objective for the proposed research work is to establish learning schemes where the information acquired on-line, from measurements, is transformed into a set of corrections for the parameters that characterize the performance of FDI and/or prognosis algorithms. According to the nature of the parameters that these schemes modify, these loops will be classified into two different categories: inner and outer correction loops.

Data from a seeded fault test for a UH-60 planetary gear plate will be used to validate all proposed approaches. Other academic and illustrative examples will also be implemented to illustrate the advantages and disadvantages of the proposed methodology with respect to the current state of the art in the field.
1.1.2 Method Overview

Particle filtering (PF) is an emerging and powerful methodology for sequential signal processing with a wide range of applications in science and engineering [1]. It has captured the attention of many researchers in various communities, including those in signal processing, statistics, and econometrics.

Founded on the concept of sequential importance sampling (SIS) and the use of Bayesian theory, particle filtering is very suitable in the case when the system is nonlinear or in the presence of non-Gaussian process/observation noise as in engines, gas turbines, gearboxes and the like, where the nonlinear nature and ambiguity of the rotating machinery world is significant when operating under fault conditions. Furthermore, particle filtering allows information from multiple measurement sources to be fused in a principled manner, which is an attribute of decisive significance for fault detection/diagnostic purposes.

The underlying principle of the methodology is the approximation of the conditional state probability distribution $p(x_{0:t}, y_{0:t})$ by a swarm of points (particles) containing samples from the space of the unknowns and a set of weights – associated with them – representing discrete probability masses. Particles can be easily generated and recursively updated [1], [2], [10] given a nonlinear process model (which describes the evolution in time of the system under analysis), a measurement model, a set of
available measurements $Y = \{Y_t, t \in \mathbb{N}\}$, and an initial estimation for the state probability density function (pdf), $p(x_0)$.

The capability of obtaining an estimate of the state pdf in real time is directly applied in the design of the proposed real-time FDI particle-filtering-based framework, where a hybrid-state dynamic model is used to represent the behavior of the system under analysis. Discrete-valued states are used to choose amongst several possible nonlinear models representing the expected behavior of the system under no-fault or faulty modes. Additionally, a set of continuous-valued states (common to every operational mode) relates the acquired measurements with the most relevant variables of the process. Particle weights associated with the discrete-valued states estimate the probability of occurrence for each faulty mode (given the current set of measurements), while the ones related to continuous-valued states simultaneously provide a pdf estimate suitable for the use of FDI hypothesis testing routines. In addition, obtained estimates may be used as initial conditions for prognostic modules, which are launched after a fault condition is statistically confirmed.

Prognosis, on the other hand, is essentially understood as the generation of long-term (multi-step) predictions describing the evolution in time of a fault indicator, with the purpose of estimating the RUL of a failing component/subsystem. In this sense, prognosis entails large-grain uncertainty, since it cannot use information from future measurements to correct for model inaccuracies. Hence, in this case, research effort has focused on the development and testing of a novel real-time particle filter-based framework for
prognosis, capable of dealing with these issues. This approach intends to reduce the uncertainty associated with long-term predictions using a two-level procedure that has been developed and subsequently tested for particular applications.

In a first prognosis level, $p$-step ahead predictions are generated on the basis of an a priori state estimate, adjusting their associated probabilities according to the noise model structure. A second prognosis level uses these predictions and the definition of critical thresholds to estimate the RUL pdf, also referred to as the time-to-failure (TTF) pdf, and simultaneously implements a correction model (outer correction loop) to compensate for all main error sources in the absence of future measurements. Once the probability density function of the RUL is estimated, other important attributes – such as expectations and 95% confidence intervals – may be computed.

Both of the proposed frameworks, FDI and prognosis, allow the inclusion of customer specifications (required statistical confidence in fault detection, minimum prediction window for prognosis, etc.) in a simple and direct way. Moreover, all the outcomes are easily provided to plant operators through real-time updated graphs, thus ensuring timely corrective actions to avoid catastrophic failures.
1.1.3 Contributions

Several approaches have been proposed for the development and implementation of a real-time particle-filtering-based framework for FDI and failure prognosis, all of them based on the fact that the current state pdf estimate may be used to determine the operational condition of the system and/or to predict the progression of a fault indicator, given a dynamic state model and process measurements.

Concerning the general framework for FDI proposed in this section, preliminary results indicate that the particle-filtering-based methodology has been successful and very efficient in pinpointing abnormal conditions in real time for a variety of cases (change in growth dynamics, reaching a pre-determined threshold, etc.), given the definition of a hybrid nonlinear state-space model for the system, a characterization of the plant behavior under nominal operational conditions (baseline data), and the availability of real-time measurements. Furthermore, the proposed scheme provides the tools to simultaneously calculate the type II detection error, execute classical FD statistical hypothesis tests (e.g., Fisher’s Discriminant Ratio, Scheffé’s test, etc.), and perform swift transitions between FDI and prognostic-oriented applications.

Regarding prognosis, a novel real-time approach capable of estimating the probability of failure for future time instants (RUL pdf) has been developed. This methodology combines state pdf estimates, long-term predictions, and empirical knowledge about critical conditions for the system (also referred to as the hazard zones).
to provide information about time-to-failure (TTF) expectations, statistical confidence intervals, and other hypothesis test. Particularly, it has been shown that a combination of resampling schemes in long-term predictions and Epanechnikov kernels helps to reduce the impact of model errors, and simultaneously offers a balanced answer in terms of accuracy and precision in RUL estimates. In addition, it was shown that an approach based solely on the expectation of the long-term prediction also provides acceptable results, and moreover, it is suitable for on-line applications with limited computational resources. Two successful case studies are presented to illustrate the performance of this simple methodology (SIR particle filter and an expectation-based long-term prediction generation) with real failure data. Both studies provide excellent insight about how model inaccuracies and/or customer specifications (hazard zone definition or desired prediction window) may affect the algorithm performance.

Finally, two outer correction loops have been successfully implemented to update parameters of great significance in the overall performance of FDI and/or prognosis algorithms. The first one corresponds to an autoregressive correction algorithm utilized to improve accuracy in RUL expectations. The second, used in the analysis of the crack growth on planetary gear plates, updates the nonlinear mapping that relates the feature value with a preliminary noisy estimate of the crack length. Both of them illustrate how the algorithm accuracy may be significantly enhanced when several learning loops – combining model-based and data driven techniques – are working in parallel within the prognosis framework.
Section 1.2 Origin and History of the Problem

1.2.1 Scope and Aim of the section

The performance and efficiency of any model-based approach for fault detection and failure prognosis will rely, to a great extent, on the dynamic model ability of mimicking the behavior of the process under study. Linear and Gaussian dynamic models may accomplish this task satisfactorily when either the process complexity is reduced or when the time framework intended for long-term predictions is shortened.

Most of the time, though, real processes do require the inclusion of nonlinear dynamics or non-Gaussian stochastic components for an accurate description, especially when the time horizon required for the generation of dependable prognosis results is long enough to make evident any deficiencies/shortcomings introduced through linearization methodologies.

Nonlinear Bayesian estimation techniques and sequential Monte Carlo (SMC) methods provide a solid and consistent theoretical framework to handle most of the difficulties mentioned above and, as such, these topics have been the subject of a broad and intensive amount of research over the past years in many diverse disciplines, including economics, biostatistics, and even statistical signal processing problems in the engineering domain such as time series analysis, radar and sonar target tracking, and communications [9].
Hence, before highlighting the contributions that these techniques may offer in the area of FDI and prognostics, it is sensible and imperative to present a comprehensive theoretical review about the state of the art in SMC methods, also referred to as particle filters, with particular emphasis on state-space (SS) estimation and model parameter identification.

With the intention of achieving this objective, the present section has been structured as follows. Subsection 1.2.2 presents both the definition and the general formulation for the nonlinear Bayesian filtering problem, as well as an introductory notion about how SMC methods can help to actually solve it. Subsection 1.2.3 focuses on theoretical aspects behind the implementation of SMC methods, and in particular of the sequential importance sampling resampling (SIR) particle filter, including the main limitations for these types of algorithms. In addition, subsection 1.2.3 presents a summary of the most recent improvements done in the field, with emphasis on those that can be useful to deal with applications related to FDI and/or model parameter estimation.

Finally, subsections 1.2.4 and 1.2.5 provide a general overview of the state of the art for the application of particle filtering algorithms in the field of real-time diagnosis and prognosis. These methodologies and published results – in areas such as robotics, automation, and artificial intelligence – will become the foundation for the novel approaches proposed.
1.2.2 Nonlinear Bayesian Filtering

Nonlinear filtering is defined as the process of estimating at least the first two moments of a state vector governed by a dynamic nonlinear, non-Gaussian state-space model from noisy observation data [15]. Although in principle the estimation procedure may be implemented on continuous-time systems, the present research will be solely focused on discrete-time systems since in most of the applications relevant to FDI and prognosis the streaming measurement data is sent and received through digital devices.

Within a Bayesian general formulation for the dynamic state estimation problem, the main goal of a nonlinear filtering procedure is to generate an estimate of the posterior probability density function (pdf) for the state, based on the set of received measurements [15]. Since such an estimate for the state vector is required almost every time measurement data are received, it makes sense to use a recursive strategy to update the estimation results. Such strategy will avoid the problem of massive data storage and/or recalculation of the whole state trajectory in time.

Mathematically, let an unobserved process \( X = \{X_t, t \in \mathbb{N}\} \) be an \( \mathbb{R}^n \)-valued Markov process characterized both by its initial distribution \( p(x_0) \) and the transition probability \( p(x_t | x_{t-1}) \). Moreover, let \( p(x_t | x_{t-1}) \) be defined by (1.2.01), the distribution of the random variable \( X_t | X_{t-1} \).

\[
X_t | X_{t-1} = x_{t-1} \sim f_t(\cdot | x_{t-1}) \tag{1.2.01}
\]
Noisy observations \( Y = \{Y_t, t \in \mathbb{N}\} \) are accessible and assumed to be conditionally independent given the process \( X = \{X_t, t \in \mathbb{N}\} \). Equation (1.2.02) defines the distribution of \( Y_t | X_t \) and, hence, of the marginal distribution \( p(y_t | x_t) \).

\[
Y_t | X_t = x_t \sim g_t(\cdot | x_t) \tag{1.2.02}
\]

Let \( x_{0,t} \triangleq \{x_0, \ldots, x_t\} \) and \( y_{1,t} \triangleq \{y_1, \ldots, y_t\} \) denote, respectively, as the signal and the observations up to time \( t \). It is of interest to estimate the posterior distribution \( p(x_{0,t} | y_{1,t}) \), the marginal distribution \( p(x_t | y_{1,t}) \) (also referred to as the filtering distribution) and the expectations [9]:

\[
I(f_i) = E_{p(x_{0,t}, y_{1,t})}[f_i(x_{0,t})] \triangleq \int f_i(x_{0,t}) p(x_{0,t} | y_{1,t}) dx_{0,t} \tag{1.2.03}
\]

For any function \( f_i : \mathbb{R}^n \to \mathbb{R}^n \) integrable with respect to \( p(x_{0,t} | y_{1,t}) \), such as the conditional mean of \( x_t \) (1st moment, where \( f_i(x_{0,t}) = x_{0,t} \)) or the conditional variance of \( x_t \) (2nd moment, where \( f_i(x_{0,t}) = x_t x_t^T - E_{p(x_t|y_{1,t})}[x_t]E_{p(x_t|y_{1,t})}[x_t]^T \)). This task can be basically achieved by performing two sequential steps, namely prediction and filtering [1]. On one hand, prediction uses both the knowledge of the previous state estimate and the process model to generate the a priori state pdf estimate for the next time instant, as shown in (1.2.04).
\[ p(x_{0t} \mid y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{0t-1} \mid y_{1:t-1}) \, dx_{0t-1} \quad (1.2.04) \]

On the other hand, the filtering step (1.2.05) considers the current observation \( y_t \), the a priori state pdf, the likelihood function \( p(y_t \mid x_t) \), and Bayes formula to generate the posterior state pdf (1.2.05).

\[
p(x_{0t} \mid y_{1:t}) = \frac{p(y_{1:t} \mid x_{0t}) p(x_{0t})}{p(y_{1:t})} = \frac{p(y_{1:t} \mid x_{0t}) p(x_{0t})}{p(y_{1:t} \mid y_{1:t-1})} \prod_{i=1}^{t-1} p(y_{i} \mid x_{i-1}) p(x_{0i-1}) \frac{p(y_t \mid x_{0t-1}) p(x_{0t-1})}{p(y_t \mid y_{1:t-1})} \quad (1.2.05)\
\]

Furthermore, as (1.2.06) shows, the filtering step may be implemented by using the recursion formula:

\[
p(x_{0t} \mid y_{1:t}) = \frac{p(y_{1:t} \mid x_t) p(x_t)}{p(y_{1:t} \mid y_{1:t-1})} = \frac{p(y_{1:t} \mid x_t) p(x_t \mid x_{0t-1}, y_{1:t-1})}{p(y_{1:t} \mid y_{1:t-1})} \prod_{i=1}^{t-1} p(y_{i} \mid x_{i-1}) p(x_{0i-1}) \frac{p(y_t \mid x_{0t-1}) p(x_{0t-1})}{p(y_t \mid y_{1:t-1})} \quad (1.2.06)\
\]

where \( p(y_t \mid y_{1:t-1}) = \int p(y_t \mid x_t) p(x_t \mid y_{1:t-1}) \, dx_t \).
The recursive computation of the *posterior* state pdf given by (1.2.04) and (1.2.06) is more conceptual than practical, however, since the integrals in (1.2.04) and (1.2.06) do not have an analytical solution in most cases [1], [10]. Similarly, the computation of any expectation, as in (1.2.03), involves the solution of an integration process, which usually will not have a closed-form.

This is because, since the mid-1960s, numerous investigators have attempted to arrive at approximations able to minimize the variance of these integral estimates. Some examples of these methodologies are [1], [9]: the Extended Kalman Filter (Anderson and Moore 1979, Jazwinski 1970), the Gaussian sum filter (Sorenson and Alspach 1971), and grid-based methods (Handschin and Mayne 1969), which in fact are able to provide the optimal recursion for the *filtering distribution* when the state-space is discrete and consists of a finite number of states [1].

Numerical methods became increasingly interesting for the scientific community in the late 1980s, when the impressive increase in computational capabilities made it possible to compute and obtain results from Monte Carlo-based algorithms in a reasonable amount of time, especially if efficient sampling methods were used, such as SMC (particle filters), the formulation of which is now explained in detail.
1.2.3 Sequential Monte Carlo Methods: Particle Filtering

Consider a sequence of probability distributions \( \{\pi_t(x_{0:t})\}_{t \geq 1} \), where it is assumed that we can evaluate \( \pi_t(x_{0:t}) \) pointwise up to a normalizing constant. SMC methods, also referred to as PF, are a class of algorithms designed to approximately obtain samples sequentially from \( \{\pi_t\} \), i.e., to generate a collection of \( N \gg 1 \) weighted random samples \( \{w_t^{(i)}, x_{0:t}^{(i)}\}_{i=1}^{N} \), \( w_t^{(i)} \geq 0, \forall t \geq 1 \), satisfying (1.2.07) [2].

\[
\sum_{i=1}^{N} w_t^{(i)} \varphi_t(x_{0:t}^{(i)}) \rightarrow \int \varphi_t(x_{0:t}) \pi_t(x_{0:t}) dx_{0:t} \quad (1.2.07)
\]

where \( \varphi_t \) is any \( \pi_t \)-integrable function.

In the particular case of the Bayesian filtering problem, the target distribution \( \pi_t(x_{0:t}) = p(x_{0:t} \mid y_{1:t}) \) is the posterior pdf of \( X_{0:t} \), given a realization of the noisy observations \( Y_{1:t} = y_{1:t} \). Furthermore, from (1.2.01) and (1.2.02), the posterior may be also written as [10]

\[
\pi_t(x_{0:t}) = p(x_0) \prod_{k=1}^{t} f_k(x_k \mid x_{k-1}) g_k(y_k \mid x_k). \quad (1.2.08)
\]

The most basic SMC algorithm, the SIR particle filter (a.k.a. bootstrap filter), solves the problem stated in (1.2.07) using a sequential importance sampling resampling (SISR) scheme, which is described next.
1.2.3.1 Importance Sampling Scheme

Assume that, at time \( t-1 \), a set of \( N \) paths \( \{x_{t-1,i}\}_{i=1}^{N} \) distributed according to \( \pi_{t-1}(x_{t-1}) \) is available, i.e., it is possible to approximate \( \pi_{t-1}(x_{t-1}) \) with the empirical distribution

\[
\pi_{t-1}^N(x_{t-1}) = \frac{1}{N} \sum_{i=1}^{N} \delta(x_{t-1} - x_{t-1,i}). \tag{1.2.09}
\]

The objective is to efficiently obtain a set of \( N \) new paths (particles) \( \{x_{t,i}\}_{i=1}^{N} \) that are approximately distributed according to \( \pi_t(\tilde{x}_{t,i}) \). To generate this new set of particles, let the current paths \( x_{t-1,i} \) be extended by using a kernel function \( q_t(\tilde{x}_{t,i} | x_{t-1,i}) \).

Although the new paths generated by this method will distribute according to (1.2.10), it is desirable for them to distribute as close as possible to \( \pi_t(\tilde{x}_{t,i}) \) [2].

\[
q_t(\tilde{x}_{t,i}) = \int_{\mathcal{X}} q_t(\tilde{x}_{t,i} | x_{t-1,i}) \pi_{t-1}(x_{t-1}) dx_{t-1} \tag{1.2.10}
\]

To generate consistent estimates for (1.2.03), it is necessary to correct for the differences that, in practice, will exist between the distributions \( q_t(\tilde{x}_{t,i}) \), also referred to as the importance density function, and \( \pi_t(\tilde{x}_{t,i}) \). Importance sampling deals with this problem by setting the values of the \( N \) weights \( \{w_{t,i}\}_{i=1}^{N} \) equal to the ratio (Radon-Nikodym derivative)
Following this reasoning, and considering the importance sampling identity (1.2.12), it is possible to prove that a Monte Carlo approximation for \( \pi_t(x_{0z}) \) is given by (1.2.13) [2].

\[
\pi_t(\tilde{x}_{0z}) = \frac{w(\tilde{x}_{0z}) q_t(\tilde{x}_{0z})}{\int_{\mathcal{X}^{t+1}} w(\tilde{x}_{0z}) q_t(\tilde{x}_{0z}) d\tilde{x}_{0z}}
\]  

(1.2.12)

\[
\hat{\pi}_t^N(x_{0z}) = \sum_{i=1}^{N} w_{0z}^{(i)} \delta(x_{0z} - \tilde{x}_{0z}^{(i)})
\]  

(1.2.13)

where \( w_{0z}^{(i)} \propto w_{0z}(\tilde{x}_{0z}^{(i)}) \) and \( \sum_{i=1}^{N} w_{0z}^{(i)} = 1 \).

The efficiency of the procedure described here improves as the variance of the importance weights is minimized, which basically means that the importance distribution is close to the actual target distribution. Although the overall procedure seems to be simple, it requires evaluating (1.2.10) in a closed form, or equivalently, evaluating \( q_t(\tilde{x}_{0z}) \) up to a normalizing constant, which usually is impossible. Nevertheless, there exists an important case where interesting conclusions can be attained.

Consider the case when the importance distribution function is of the form (1.2.14), which is equivalent to setting \( \tilde{x}_{0z} = (x_{0z-1}, \tilde{x}_t) \), i.e., the current path is not modified up to time \( t-1 \), when extending its dimension. Then, by substituting both
(1.2.10) and the recursion formula (1.2.06) in (1.2.11), and taking into account the fact that \( \pi_{t-1}(x_{0:t-1}) = p(x_{0:t-1} | y_{1:t-1}) \), it is possible to write (1.2.15).

\[
q_t(\tilde{x}_t \mid x_{0:t-1}) = \delta(\tilde{x}_{0:t-1} - x_{0:t-1}) \cdot q_t(\tilde{x}_t \mid x_{0:t-1}) \tag{1.2.14}
\]

\[
w(\tilde{x}_{0:t}) = \frac{\pi_t(\tilde{x}_{0:t})}{q_t(\tilde{x}_{0:t})} = \frac{\pi_t(\tilde{x}_{0:t})}{\pi_{t-1}(x_{0:t-1}) \cdot q_t(\tilde{x}_t \mid x_{0:t-1})} \\
\propto \frac{\pi_{t-1}(x_{0:t-1}) \cdot p(y_t \mid \tilde{x}_t) \cdot p(\tilde{x}_t \mid x_{0:t-1})}{\pi_{t-1}(x_{0:t-1}) \cdot q_t(\tilde{x}_t \mid x_{0:t-1})} \cdot q_t(\tilde{x}_t \mid x_{0:t-1}) \tag{1.2.15}
\]

Expression (1.2.15) is of great significance in the implementation of the sequential algorithm. It provides not only a theoretical framework to find an optimal importance distribution according to a predefined minimization criterion, but it also sets the foundation for the most basic SMC implementation, the sequential importance sampling (SIS) particle filter.

The SIS particle filter is implemented as follows. Within the nonlinear Bayesian filtering framework, set the importance distribution as the a priori pdf for the state, i.e.,

\[
q_t(\tilde{x}_{0:t} \mid x_{0:t-1}) = p(\tilde{x}_t \mid x_{t-1}) = f_t(\tilde{x}_t \mid x_{t-1}) .
\]

Thus, equation (1.2.15) is reduced to

\[
w(\tilde{x}_{0:t}) \propto p(y_t \mid \tilde{x}_t) = g_t(y_t \mid \tilde{x}_t). \tag{1.2.16}
\]

That is, the weights for the newly generated particles are directly evaluated from the likelihood of the new observation. Although this procedure is simple, it presents
severe degeneracy problems when implemented, especially if the dimensionality of the problem is large, since the variance of the weights can only increase over time [1], an issue that is solved via the implementation of an importance resampling scheme, which will be discussed in the sequel.

Although the derivation of (1.2.16) assumed that (1.2.09) holds, it is fairly easy to generalize these results when only \( q_{t-1}(x_{0_{t-1}}) \) is available. For this, let us assume that (1.2.14) holds and that the importance density function at time \( t \) admits \( q_{t-1}(x_{0_{t-1}}) \) as marginal distribution at time \( t-1 \):

\[
q_t(x_{0_t}) = q_{t-1}(x_{0_{t-1}}) \cdot q_t(\tilde{x}_t | x_{0_{t-1}}). \tag{1.2.17}
\]

Therefore, (1.2.15) can be written as

\[
w(\tilde{x}_{0_t}) = \frac{\pi_t(\tilde{x}_{0_t})}{q_t(\tilde{x}_{0_t})} = \frac{\pi_t(\tilde{x}_{0_t})}{q_{t-1}(x_{0_{t-1}}) \cdot q_t(\tilde{x}_t | x_{0_{t-1}})} \times \frac{\pi_{t-1}(x_{0_{t-1}}) \cdot p(y_t | \tilde{x}_t) \cdot p(\tilde{x}_t | x_{0_{t-1}})}{q_{t-1}(x_{0_{t-1}})} \times \frac{w(x_{0_{t-1}}) \cdot p(y_t | \tilde{x}_t) \cdot p(\tilde{x}_t | x_{0_{t-1}})}{q_t(\tilde{x}_t | x_{0_{t-1}})} \tag{1.2.18}
\]

Finally, in the particular case when \( q_t(x_{0_t} | x_{0_{t-1}}) = p(\tilde{x}_t | x_{0_{t-1}}) = f_t(\tilde{x}_t | y_{t-1}) \), (1.2.19) holds and the weights for the next iteration step can be computed from (1.2.20).
The choice of the importance density function is critical in the performance of the particle filter scheme and hence, it should be considered as a parameter in the filter design. It is important to note that the update equation (1.2.20) is not always the best option that can be implemented in a nonlinear filtering framework. In that sense, several approaches geared to improve the performance of the algorithm, which are mainly based on the minimization of the evolution of the weight variance over time, have been suggested by different authors. The ones that are more relevant to the objectives of the present research are discussed in section 1.2.4.

1.2.3.2 Resampling step: the SIR Particle Filter

One of the main difficulties that must be addressed in the implementation of SIS particle filters is the degeneracy problem in the particle population. The degeneracy phenomenon consists in the fact that, as the algorithm evolves in time, the weight variances increase [8] and the importance weight distribution becomes progressively more skewed, at the point where (after a few iterations) all but one particle will have a negligible weight [1], [2], [10].
As a result, the approximation of the target distribution becomes very poor and a number of computational resources would be spent trying to update particles with minimum relevance in FDI or prognosis routines. Since the degeneracy in the particle population is directly related to the variance of the importance weights, several authors have proposed measuring it by using an estimate $\hat{N}_{\text{eff}}$ of the effective sample size $N_{\text{eff}}$ [10], [19], [22].

$$N_{\text{eff}} = \frac{N}{1 + \text{var}_{\pi}(w_{0 \tau})}, \quad \hat{N}_{\text{eff}} = \frac{1}{\sum_{i=1}^{N} (w_{t}^{(i)})^2}$$ (1.2.21)

Whenever $\hat{N}_{\text{eff}} \leq N_{\text{thres}}$, a fixed threshold, a resampling algorithm [1], [10], [28] is performed to eliminate particles with small weights and concentrate the computational efforts on those having large ones. The procedure itself generates a new set of particles $\{\hat{x}_{0 \tau}^{(i)}\}_{t=1,\ldots,N}$ by resampling $N$ times from (1.2.13) in such a way that $P(\hat{x}_{0 \tau}^{(i)} = x_{0 \tau}^{(j)}) = w_{t}^{(j)}$.

Thus, for each particle, a number of offspring $N_{t}^{(i)} \in \mathbb{N}(i = 1, \ldots, N)$ is created such that $\sum_{i=1}^{N} N_{t}^{(i)} = N$. After the resampling procedure, the new particle population $\{\hat{x}_{0 \tau}^{(i)}\}_{t=1,\ldots,N}$ is an i.i.d. sample of the empirical distribution (1.2.22) and therefore the weights are reset to $w_{t}^{(j)} = N^{-1}$.

$$\hat{\kappa}_{t}^{N}(x_{0 \tau}) = \frac{1}{N} \sum_{i=1}^{N} N_{t}^{(i)} \delta(x_{0 \tau} - \hat{x}_{0 \tau}^{(i)}) = \frac{1}{N} \sum_{i=1}^{N} \delta(x_{0 \tau} - \hat{x}_{0 \tau}^{(i)})$$ (1.2.22)
The following algorithm summarizes the procedure [10]:

**Sequential Importance Sampling Resampling (SIR) Particle Filter**

1. **Importance Sampling**
   - For \( i = 1, \cdots, N \), sample \( \tilde{x}_t^{(i)} \sim \pi(x_t | \tilde{x}_{0:t-1}^{(i)}, y_{0:t}) \) and set \( \bar{x}_{0t}^{(i)} = (x_{0:t-1}^{(i)}, \tilde{x}_t^{(i)}) \).
   - Evaluate the importance weights
     \[
     w(x_{0t}^{(i)}) = w_{0t-1}^{(i)} \cdot \frac{p(y_t | \tilde{x}_t^{(i)}) p(x_t^{(i)} | x_{0t-1}^{(i)})}{q_t(x_t^{(i)} | x_{0t-1}^{(i)})}
     \]
     \[
     w_{0t}^{(i)} = \frac{w(x_{0t}^{(i)})}{\sum_{i=1}^{N} w(x_{0t}^{(i)})}
     \]

2. **Resampling Algorithm**
   - If \( \hat{N}_{\text{eff}} \geq N_{\text{thres}} \)
     - \( \bar{x}_{0t} = \tilde{x}_{0t}^{(i)} \) for \( i = 1, \cdots, N \)
   - otherwise
     - For \( i = 1, \cdots, N \), sample an index \( j(i) \) distributed according to a discrete distribution satisfying \( P(j(i) = l) = w_l^{(i)} \) for \( l = 1, \cdots, N \).
     - For \( i = 1, \cdots, N \), \( \bar{x}_{0t}^{(i)} = \tilde{x}_{0t}^{(j(i))} \) and \( \bar{w}_t^{(i)} = N^{-1} \)

The side effect of a resampling technique is that, theoretically, the simulated paths are no longer statistically independent. Although this condition may suggest that
convergence results for the SIS algorithm are no longer valid, [3] and [6] have already established a solid theoretical foundation that proves convergence even in the SIR case.

Besides the algorithm presented here, there are other versions intended to perform particle resampling, including multinomial sampling [13], residual resampling [23]-[24], and minimum variance resampling [18]. All of these algorithms ensure that $E[N_t^{(i)}] = N \cdot w_{0z}^{(i)}$, although they differ in the $\text{var}(N_t^{(i)})$. Residual sampling is of particular interest since it is computationally cheaper than the classical SIR technique and it offers a smaller variance for $N_t^{(i)}$, the resulting number of offspring from particle $(i)$.  

Residual resampling involves mainly three steps [23], [32], the first being to compute \( \tilde{N}_t^{(i)} = \lfloor N \cdot w_{0z}^{(i)} \rfloor \). Secondly, a SIR procedure is performed to select the remaining \( \tilde{N} = N - \sum_{i=1}^{N} \tilde{N}_t^{(i)} \) particles, assigning them the new weights \( w_{0z}^{(i)} = \tilde{N}^{-1} \left( w_{0z}^{(i)} N - \tilde{N}_t^{(i)} \right) \). The newly generated samples are then added to the current $\tilde{N}_t^{(i)}$, which completes the third and last step. For this procedure, \( \text{var}(N_t^{(i)}) = \tilde{N} w_{0z}^{(i)} \left( 1 - w_{0z}^{(i)} \right) \).

Other approaches introduce the use of Markov Chain Monte Carlo (MCMC) methods [11], with the most famous being the Metropolis-Hastings (MH) algorithm and the Gibbs sampler (which is a particular case of MH) as an additional step after the resampling procedure [22].
Although these methods perform very well in off-line applications, they usually are not suitable for on-line or recursive estimation because their considerable computational requirements [12], [15]. For that particular reason, these techniques will not be discussed here.

Regardless of the method used to solve the degeneracy problem explained above, SIR Particle Filtering is a very useful and easily implementable approach that may be applied whenever we face the problem of nonlinear filtering with a reduced dimensionality of the state-space vector. Nevertheless it is important to consider the contributions from many authors in order to improve the efficiency of the algorithm, especially by defining a proper importance density function. The most relevant results are described next.

1.2.4 Improved Sequential Monte Carlo Methods

1.2.4.1 Optimal Importance Density Function

As mentioned in subsections 1.2.3.1 and 1.2.3.2, the efficiency of the importance sampling procedure improves considerably when the variance of the importance weights – conditional upon the simulated trajectory \( x_{tr-1}^{(i)} \) and the observations \( y_{tr} \) – is minimized, hence reducing the effect of the degeneracy problem.

The conditional variance can be easily computed from (1.2.23), obtaining (1.2.25).
\[
\text{var}_{q(\tilde{x}_{t-1}, y_t)}[w(\tilde{x}_{t-1}^{(i)})] = \left(\frac{\gamma}{\gamma - 1}\right)^2 \left[ \int \frac{p(y_t | \tilde{x}_t) \cdot p(\tilde{x}_t | x_{t-1}^{(i)})}{q(\tilde{x}_t | x_{t-1}^{(i)}, y_t)} \, d\tilde{x}_t - p^2(y_t | x_{t-1}^{(i)}) \right]
\] (1.2.25)

It is straightforward to check that the importance weight variance (1.2.25) is zero when \( q(\tilde{x}_t | x_{t-1}^{(i)}, y_t) = p(\tilde{x}_t | x_{t-1}^{(i)}, y_t) \), the optimal importance density function. This function was first introduced by [37] and utilized more recently in [5] and [19]. If used, it leads to the importance weight update equation:

\[
w(\tilde{x}_t^{(i)}) = w_{t-1}^{(i)} \cdot p(y_t | x_{t-1}^{(i)}) \quad \text{and} \quad w_{t}^{(i)} = \frac{w(\tilde{x}_t^{(i)})}{\sum_{i=1}^{N} w(\tilde{x}_t^{(i)})}. 
\] (1.2.26)

There are two major implementation problems related to the use of the optimal importance density function. First, it must be possible to draw samples from the distribution \( p(\tilde{x}_t | x_{t-1}^{(i)}, y_t) \). But more important, it requires the computation of the probability integral \( p(y_t | x_{t-1}^{(i)}) = \int p(y_t | \tilde{x}_t) \cdot p(\tilde{x}_t | x_{t-1}^{(i)}) \, d\tilde{x}_t \), which does not have an analytic closed-form in most cases. It is important to note, however, that there is an important class of problems where these difficulties may be overcome: the Gaussian state-space model with nonlinear transition equation:

\[
x_t = f(x_{t-1}) + \omega_t, \quad \omega_t \sim \mathcal{N}(0, \Sigma_\omega), \\
y_t = C \cdot x_t + v_t, \quad v_t \sim \mathcal{N}(0, \Sigma_v) 
\] (1.2.27)
where \( f: \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_y} \) is a real valued function, \( C \in \mathbb{R}^{n_x \times n} \) is a constant matrix, and the additive noises \( \omega_t \) and \( v_t \) are mutually independent i.i.d. Gaussian sequences. In fact, for this particular case, the importance weight update equation is as follows:

\[
\begin{align*}
\hat{w}(x_t^{(i)}) &= w_{t-1}^{(i)} \cdot \exp\left\{ -\frac{1}{2} \left[ (y_t - C \cdot f(x_{t-1}^{(i)}))^T \cdot \left( \Sigma_\omega + C \Sigma_C C^T \right)^{-1} \cdot (y_t - C \cdot f(x_{t-1}^{(i)})) \right]\right\} \\
\text{and} \quad w_t^{(i)} &= \frac{\hat{w}(x_t^{(i)})}{\sum_{i=1}^N \hat{w}(x_t^{(i)})} \quad (1.2.28)
\end{align*}
\]

Suboptimal procedures based on linearization techniques (similar to the ones used in the Extended Kalman filter) have been proposed in [10] for cases when the observation equation is of the form \( y_t = g(x_t) + v_t, \quad v_t \sim \mathcal{N}(0, \Sigma_v) \). The linearization procedure then ensures asymptotic convergence in the SMC algorithm when approximating the observation equation by

\[
y_t = g(f(x_{t-1})) + \left. \frac{\partial g(x_t)}{\partial x_t} \right|_{x_t = f(x_{t-1})} \cdot (x_t - f(x_{t-1})) + v_t. \quad (1.2.29)
\]

Other authors have proposed techniques that do not necessarily rely on linearization procedures. Among the most significant ones, are the following:

- **Auxiliary particle filter (ASIR)** [28]: It modifies the particle population at time \( t-1 \) according to the current measurement \( y_t \), before applying the transition kernel. By doing so, it ensures that the newly created particles at
time $t-1$, and hence at time $t$ as well, are most likely to be close to the true state. It often provides better state estimates and is less sensitive to outliers if the process noise is small or whenever the likelihood is situated in one of the prior tails. On the other hand, the ASIR particle filter will degrade its performance if the process noise is large.

- **Rao-Blackwellised particle filter (RBPF)** [4], [7]: It reduces the size of the state-space by marginalizing out some variables analytically in such a way that the posterior pdf of the state can be written as

  $$p(x_{0:t}, z_{0:t} | y_{1:t}) = p(x_{0:t} | z_{0:t}, y_{1:t}) p(z_{0:t} | y_{1:t}).$$

  It is useful when trying to estimate the posterior distribution of a continuous-valued state $x_t$ as a stochastic finite mixture of Gaussians.

- **Unscented particle filter (UPF)** [32]: It considers the implementation of an unscented Kalman filter (UKF) to approximate the optimal importance density function as a Gaussian pdf, thus minimizing errors in the covariance matrix estimate and higher order moments of the posterior, e.g., kurtosis. In general, UPF offers better performance when the likelihood is peaked or when measurement data contains outliers.

**1.2.4.2 Regularized Particle Filter**

Although resampling schemes help to reduce the degeneracy problem, they cannot completely avoid the loss of diversity among particles since the samples are still
drawn from a discrete distribution instead of a continuous one. To solve this issue, a modified approach – the regularized particle filter (RPF) – was first proposed in [26] and then reviewed in [1].

In a general perspective, the RPF is almost identical to the SIR particle filter, except for the resampling step. Whereas in the latter the resampling step is based on the discrete approximation (1.2.13) for the posterior density \( p(\tilde{x}_{0:1} | y_{1:1}) \), the former considers a continuous approximation given by

\[
p(\tilde{x}_{0:1} | y_{1:1}) \approx p^N(\tilde{x}_{0:1} | y_{1:1}) = \sum_{i=1}^{N} w_{0:1}^{(i)} K_h(x_{0:1} - \tilde{x}_{0:1}^{(i)})
\]

where \( K_h(\cdot) \) is a rescaled kernel density \( K(\cdot) \), \( h > 0 \) is the kernel bandwidth (scalar), and \( n_s \) is the dimension of the state vector. The kernel function is a symmetric pdf such that

\[
\int xK(x) \, dx = 0 \quad \text{and} \quad \int \|x\|^2 K(x) \, dx < \infty.
\]

Both the kernel and the bandwidth are selected to minimize (1.2.31).

\[
MISE(p^N) = E \left[ \int \left( p^N(\tilde{x}_{0:1} | y_{1:1}) - p(\tilde{x}_{0:1} | y_{1:1}) \right)^2 \, d\tilde{x}_{0:1} \right]
\]

In the special case of classically equally weighted samples, \( w_{0:1}^{(i)} = N^{-1}, \ i = 1, \ldots, N \), the optimal choice of the kernel is the Epanechnikov kernel [26].
where \( c_n \) is the volume of the unit sphere in \( \mathbb{R}^n \). Furthermore, if the density is Gaussian with unit covariance matrix, the optimal bandwidth is given by

\[
h_{opt} = A \cdot N^{-\frac{1}{n+4}}
\]

\[
A = \left( 8 c_n^{-1} \cdot (n_x + 4) \cdot (2\sqrt{\pi})^n \right)^{-\frac{1}{n+4}}
\]  

(1.2.33)

In the case of an arbitrary underlying density, two approximations are made. The first assumes that the density is Gaussian, and the second considers its covariance matrix \( S_t \) equal to the empirical covariance matrix of \( \tilde{x}_{0:t} \). However, even in a general case, a suboptimal filter may be obtained by replacing the second step of the SIR particle filter implementation with the following algorithm.

**RPF Resampling Step**

If \( \hat{N}_{eff} \leq N_{thres} \)

- Calculate \( S_t \), the empirical covariance matrix of \( \{\tilde{x}_{0:t}^{(i)}, w_{0:t}^{(i)}\}_{i=1}^N \)
- Compute \( D_t \) such that \( DD_t^T = S_t \)
• For \( i = 1, \cdots, N \), sample an index \( j(i) \) distributed according to a discrete distribution satisfying \( P(j(i) = l) = w_l^{(i)} \) for \( l = 1, \cdots, N \).

• For \( i = 1, \cdots, N \), \( \tilde{x}_0^{(i)} = \tilde{x}_0^{(i)} \) and \( \tilde{w}_i = N^{-1} \).

• For \( i = 1, \cdots, N \), draw \( \varepsilon^i \sim K \), the Epanechnikov kernel and \( \tilde{x}_0^{(i)*} = \tilde{x}_0^{(i)} + h_{opt} D_i \varepsilon^i \).

Although the complexity RPF algorithm increases considerably when the dimensionality of the state vector is large, it is well suited when the process noise is assumed to be independent and uncorrelated. In practice, the performance of RPF is better than the SIR particle filter when the process noise is small, although the samples are no longer guaranteed to asymptotically approximate the posterior pdf \( p(\tilde{x}_{0|t} | y_{1|t}) \).

### 1.2.4.3 Model Parameter Estimation using Particle Filters

A framework for fixed model parameter estimation can be obtained as an extension of the classical particle filter approach. Consider, for this effect, a combined sample \( \{\tilde{x}_0^{(i)}, \theta_t^{(i)}\}_{i=1}^N \) and associated weights \( \{w_t^{(i)}\}_{i=1}^N \), representing the posterior \( p(x_{0|t}, \theta | y_{1|t}) \) for both the state and model parameter vector. It is important to note that the suffix \( t \) in \( \theta_t^{(i)} \) indicates that the particle corresponds to the posterior pdf at time \( t \) and does not imply that \( \theta \) is a time-varying parameter. From Bayes’ theorem follows the following proportionality relationship for the posterior at time \( t \):

\[
p(x_{0|t}, \theta | y_{1|t}) \propto p(y_{1|t} | x_{0|t}, \theta) p(x_{0|t}, \theta | y_{1|t-1}) \\
\propto p(y_{1|t} | x_{0|t}, \theta) p(x_{0|t} | \theta, y_{1|t-1}) p(\theta | y_{1|t-1}) \tag{1.2.34}
\]
Nevertheless, it is not clear in (1.2.34) how to define \( p(\theta \mid y_{t_{z-1}}) \). In that sense, two approaches are found in the literature proposing ways to deal with this issue. Both of them have to deal with the issue of particle degeneracy, which is especially acute when using particle filter-based techniques for the estimation of fixed model parameters as part of an extended state vector [25].

“Artificial evolution” in [13] incorporates small random perturbations to all the parameter particles before evolving to the next time instant as in (1.2.35), viewing the parameters as if they were, in fact, time-evolving. This approach, however, has some drawbacks, the major one being that the parameters are, by assumption, fixed and hence there is an inherent loss of information in both the time and measurement updates.

\[
\begin{align*}
\theta_t &= \theta_{t-1} + \zeta_{t-1} \\
\zeta_t &\sim \mathcal{N}(0, W_t)
\end{align*}
\]  

(1.2.35)

On the other hand, [35] and [36] present kernel smoothing methods that gave a theoretical foundation for effective importance sampling techniques for \( p(\theta \mid y_{t_{z-1}}) \). Assuming a set of samples \( \{\theta_{t-1}^{(i)}\}_{i=1}^N \) and weights \( \{w_{t_{z-1}}^{(i)}\}_{i=1}^N \) that approximate \( p(\theta \mid y_{t_{z-1}}) \), define \( \overline{\theta}_{t-1} \) and \( V_{t-1} \) as the Monte Carlo mean and variance for \( p(\theta \mid y_{t_{z-1}}) \), respectively. The smooth kernel density is then given by

\[
p(\theta \mid y_{t_{z-1}}) \approx \sum_{i=1}^N w_{t_{z-1}}^{(i)} \mathcal{N} \left( \theta \mid m_{t_{z-1}}^{(i)}, h^2 V_{t_{z-1}} \right)
\]  

(1.2.36)
where \( \mathcal{N}(\cdot | m, S) \) is a multivariable normal density with mean \( m \) and covariance matrix \( S \), and \( h > 0 \) is a smoothing parameter. On the other hand, \( m^{(i)}_{t-1} \) are specified by the following rule:

\[
m^{(i)}_{t-1} = a \theta^{(i)}_{t-1} + (1 - a) \bar{\theta}_{t-2}
\]

\[
a = \sqrt{1 - h^2}
\]  

(1.2.37)

By condensing the elements presented in [35]-[36], [25] offers a general algorithm suitable for fixed model parameter estimation with an APF implementation.

### Combined State and Fixed Model Parameter Estimation

- For \( i = 1, \cdots, N \), compute

\[
\mu^{(i)}_t = E[x_i | x^{(i)}_{t-1}, \theta^{(i)}_{t-1}]
\]

\[
m^{(i)}_t = a \theta^{(i)}_{t-1} + (1 - a) \bar{\theta}_{t-2}
\]

- Sample an auxiliary integer variable \( k \in \{1, \cdots, N\} \) such that:

\[
Pr(k = i) \propto w^{(i)}_{t-1} \cdot p(y_t | \mu^{(i)}_t, m^{(i)}_{t-1})
\]

- Sample a new parameter vector \( \theta^{(k)}_t \) from the \( k^{th} \) component of the kernel density.

\[
\theta^{(k)}_t \sim \mathcal{N}(\cdot | m^{(k)}_{t-1}, h^2 V_{t-1})
\]

- Sample \( x^{(k)}_t \sim p(\cdot | x^{(i)}_{t-1}, \theta^{(i)}_t) \)

- Evaluate the importance weight

\[
w^{(k)}_t \propto \frac{p(y_t | x^{(i)}_t, \theta^{(i)}_t)}{p(y_t | \mu^{(i)}_t, m^{(i)}_{t-1})}
\]
1.2.5 Particle Filtering in Real-Time Diagnosis Applications

Particle filtering, as any Bayesian technique for state estimation, has a direct application in the arena of fault detection and identification (FDI). Indeed, once the current state of the system is approximately well known, it is easy to compare the process behavior with patterns regarding normal or faulty conditions.

Several authors have made use of the capabilities of SIS to complement or improve the efficiency of classical FDI approaches. Consider, for instance, the combination of particle filters and a model-based residual analysis for fault detection (FD) introduced in [17]. In that case, the PF state estimates are used to compute a residual (also referred to as the one-step prediction error), which basically indicates how far the system is from its expected (or desired) behavior. The residual signal is then employed to compute the likelihood of a faulty condition, given the assumption of additive Gaussian measurement noise.

Although this approach may be followed to improve a number of statistical tests currently available for fault detection, most of these tests rely on Gaussian additive noise assumptions to establish a closed form for the evaluation of the likelihood or for purposes of residual statistical analysis. However, many complex processes may not hold to this assumption and therefore the application of similar FD schemes may lead to inaccurate conclusions.
To solve this problem, a slightly different FDI approach is followed in [21]. Here, $M+1$ models are provided to describe the system behavior under $M$ different fault conditions ($M = 0$ representing the normal mode) and a particle filtering routine is performed to obtain the state pdf estimate. The joint likelihood of the observations, conditional on each hypothesized model, is then computed as the sum of $M$ logarithms of the Likelihood Ratio (LLR) for $H_m$ ($m = 1, 2, \cdots, M$) versus $H_0$ (normal operation). This paper also includes the concept of testing several plant models in parallel, which will prove to be extremely useful for the FDI approach proposed here.

This methodology, although very useful in certain application domains, does not explicitly provide statistical confidence levels or an estimate for the type II detection error, two of the most important customer specifications desired in a FDI routine. In fact, the FDI decision basically depends here on a fixed threshold for the sum of LLR (which will definitely vary for different applications) and no suggestions are made about how to include the already mentioned customer specifications in the FDI module design.

A different approach for FDI – based on the capability of particle filters to discriminate between discrete states – is followed in [20] with excellent results in the domain of propulsion systems. Hybrid dynamic models are used here to represent the operation of the plant for a set of known fault modes. Discrete states in each particle represent the operational mode, while continuous-valued states describe the evolution in time of process variables. The mode of the system is computed at every time step, by considering the one found most likely within the particle population.
A similar approach is utilized in [7] (Rao-Blackwellised particle filter), where the continuous-state models associated with each operational mode are considered to be linear with Gaussian additive noise. Accordingly, particle filters are only used here to estimate the probabilities of each fault mode, not for continuous-valued state estimation purposes (where a simple implementation of a bank of Kalman filters suffices). The assumption of linear models, though, may be too restrictive for a general FDI framework and therefore it may not always be applicable.

Approaches introduced in [20] and [7] make use of particle filtering not only as a tool for state estimation, but also as a means of obtaining the probability of a determined fault mode in a system. This attribute is also found in other interesting results published in the literature and it is of paramount importance for the present research work, since it sets the foundations for including customer specifications in the design.

In that sense, two applications of particle filter algorithms for FDI purposes are of particular interest. These approaches, which are based on the concept of hybrid dynamic models and the inclusion of risk functions for the allocation of particles among discrete states, are now described in detail.
1.2.5.1 Variable Resolution Particle Filters

The variable resolution particle filter (VRPF [33]-[34]) incorporates the concept of “abstract particles” in Markov Chain processes, where each particle may represent a single state or a set of similar states. This algorithm has the advantage that only a limited number of particles is needed to represent large portions of the state-space when measurements indicate that the likelihood is low. Moreover, once the likelihood of an abstract particle increases, it is possible to specialize it to represent more specific individual states.

A bias-variance trade-off is performed to determine the appropriate resolution for the abstract states, since the loss $l$ from a particle-based approximation of the true distribution is directly related to these terms:

$$l = E\left[p(x_t | y_t) - \tilde{p}^N(x_t | y_t)\right]^2 = b\{\tilde{p}^N(x_t | y_t)\}^2 + \text{var}\{\tilde{p}^N(x_t | y_t)\} \quad (1.2.38)$$

where $b\{\tilde{p}^N(x_t | y_t)\}$ is the bias and $\text{var}\{\tilde{p}^N(x_t | y_t)\}$ is the variance of the pdf estimates.

Now, consider a set of “physical” states $\{X^d\}_{d=1}^{d_t}$ in the Markov Chain with an associated stationary distribution $P(X^d)$, where $\sum_{d=1}^{d_t} P(X^d) = 1$, and a set of abstract states $\{S^k\}_{k=1}^{M}$, which may include both physical and other abstract states. As the resolution in the Markov Chain model decreases (more physical states are included in the “abstract states”), the variance of the estimate also decreases, whereas its bias is enlarged. In that sense, the VRPF decides to abstract to a coarser resolution abstract state $S^j$ if the state-
space is currently at the resolution of the states $S^i$, and the loss associated with resolution $S^j$ is less than the loss of all its children [34].

\[
    b\{S^i\}^2 + \text{var}\{S^i\} < \sum_{S^j \in \text{children}(S^i)} [b\{S^j\}^2 + \text{var}\{S^j\}] \quad (1.2.39)
\]

Interesting applications of the VRPF have been found in the arena of fault diagnosis in rovers and robots. In particular, a state-space model that includes both discrete and continuous states has been used to identify faults in six-wheel robots [33]. This approach considered a VRPF to manage the posterior pdf for the discrete states, while a one-step look ahead UKF was implemented to approximate the optimal importance and deal with the estimates for the continuous states.

This methodology, also referred to as VUF, has given excellent results in terms of reducing the number of particles needed to represent a system with numerous fault modes, although it requires the use of additional memory space and computational resources to store and process all sigma points and covariance matrices associated with each of the particles.

### 1.2.5.2 Risk Sensitive Particle Filters

The risk-sensitive particle filter (RSPF) [31] incorporates a cost model in the importance distribution to generate more particles in high-risk regions of the state-space [33]. Mathematically, the importance distribution is set as
where $d_i$ is a set of discrete-valued states representing fault modes, $x_i$ is a set of continuous-valued states that describe the evolution of the system given those operational conditions, $r(d_i)$ is a positive risk function that is dependent on the fault mode, and $\gamma_t$ is a normalizing constant. This methodology has proven to be very helpful in improving the tracking of states that are critical to the performance of a six-wheel robot [33]. An important drawback of this approach, though, is that it needs the inclusion of exogenous models to evaluate and estimate the risk associated with every fault mode, a task that may prove to be difficult to implement.

1.2.6 Particle Filtering in Real-Time Prognosis Applications

Prognosis may be understood as the result of the procedure where long-term (multi-step) predictions – describing the evolution in time of a fault indicator – are generated with the purpose of estimating the remaining useful life (RUL) of a failing component/subsystem.

Failure prognosis plays an important role in achieving a reliable and cost effective operation. Certainly, this is of great interest in many industrial processes, including mechanical systems (e.g. automotive and aircrafts), power systems, continuous-time processes (e.g. pulp, paper and hot steel mills) and other discrete-time processes such as semiconductor manufacture and food products.
Several approaches related to prognosis may be found in literature. Few of them, however, offer appropriate tools for real-time estimation of the RUL as a continuous function of time.

Consider, for instance, the most popular prognosis approach currently used in reliability studies: parameter estimation in Weibull-based risk probability density distributions. Generally speaking, two major types of reliability degradation modeling can be distinguished, both of them particularly well suited for the task due to the ability of Weibull functions to adjust its shape and include time-dependant fault mechanisms.

On one hand, the graphical reliability degradation modeling approach (based on statistical models and broadly used in practice) selects risk function parameters so that degradation data is satisfactorily represented at each observation time. Parameter estimation is performed through a two step off-line statistical procedure that greatly depends on the amount of degradation data and is subject to accumulation of computational errors [16]. Also, if the faulty system is behaving differently from what it was observed in past experiences, the method is unable to correct and adequate prognosis results, leading to inexact conclusions.

On the other hand, the degradation path curve approach intends to approximate a degradation trajectory versus time, using known physics models or a statistically designed path curve. Although this methodology offers better results that the graphical approach, it lacks of learning or adaptation mechanisms and it must be performed off-line.
since it is extremely computationally intensive. Recent improvements focus on the use of maximum likelihood estimation (MLE) techniques and truncated Weibull pdf’s, obtaining good results in the case of fatigue crack growth data [16], although this only applies when degradation data follow a two-parameter Weibull distribution and it does not solve the problem of real-time applications.

In that sense, the most comprehensive effort in establishing an on-line prognosis framework can be found in applications associated with the use of filtering techniques for the study of fatigue crack dynamics [29]. The filtering concept enhances the deterministic crack growth modeling standpoint based on the application of Paris’ equation, and keeps a close relationship with the physics of the problem. Efforts have been conducted to employ Markov processes and Extended Kalman Filters (EKF) to estimate the first two moments of a Gaussian state pdf of the system, also assuming independence between measurement noise and uncertainties in material properties. The obtained Gaussian pdf is afterwards projected in time and used to test $M$ disjoint statistical hypothesis, which divide the feasible crack length domain, thus obtaining $M$ different probability distributions describing the time evolution of the likelihood for each hypothesis [29].

Although the previous approach sets the foundations for other filtering-based real-time prognosis methodologies, there are still some unsolved issues. Firstly, the assumption of a Gaussian (or log-normal) pdf is not always held in nonlinear processes and therefore the projection in time of the filtered state pdf may lead to problems both in terms of accuracy and precision. Secondly, the fragmentation of the crack length domain
in $M$ disjoint regions leading to a set of $M$ hypothesis is not always the best prognosis procedure, since different crack length may affect the system’s performance in dissimilar ways. For example, the probability of failure is not uniform for all crack lengths; there are in fact some regions where a failure condition is more frequently detected. Hence, it would be desirable to somehow collect all the information contained in the $M$ hypothesis into a single prognosis outcome for the operator, also considering the probability of failure for each partition of the crack length domain. Thirdly, the methodology presented in [29] has only been tested for a particular application and it does not offer an alternative way to be implemented in processes with unknown time-varying model parameters. Last but not least, no indication is made in [29] about how to propagate a non-Gaussian state pdf in time through nonlinear transition models.

Regarding particle filters, most authors have visualized this technique (and other nonlinear filtering approaches) as a tool for detection, but not for prognosis. The former, mainly because there are no clear indications about how to project the particle population in time, while keeping the assumptions about model nonlinearities and non-Gaussian noise structures. In specific applications, such as chaos prediction, it has been suggested to assume absence of both process and measurement noise for prediction purposes [14], thus obtaining a long-term prediction with minimum variance. The particle population is then used as initial conditions for deterministic models in order to be used for decision theory, risk calculations and other statistical approaches. The implications of these assumptions, though, could be significant in real processes, especially in presence of vibration signals and therefore they must be evaluated with care.
Section 1.3  Proposed Research

Research work proposed hereby intends to establish a general framework to deal with the problems of fault detection, identification and failure prognosis in real time applications via the utilization of particle filtering techniques. In order accomplish this objective, three main research tasks have to be simultaneously achieved. Each one of these research objectives is now described in detail.

1.3.1 Particle Filtering for Diagnosis in Stochastic Nonlinear Systems

The first main objective aims for the implementation of an on-line particle-filtering-based framework for fault detection and identification (FDI) in nonlinear, non-Gaussian systems. This architecture should be able to pinpoint both the presence and nature of a fault condition in real time, given a set of measurements and a characterization of the plant behavior under nominal operational conditions (baseline data). Furthermore, it is expected for this scheme to provide the means for a swift transition between the FDI and prognostic oriented applications.

For this purpose, a hybrid-state particle filter will be implemented [33]. Discrete-valued states will be used to choose amongst several possible nonlinear models representing the expected behavior of the system under no-fault or faulty modes. Additionally, a set of continuous-valued states common to every operational mode will relate the acquired measurements with the most relevant variables of the process.
Particle weights associated with the discrete-valued states will estimate the probability of occurrence for each faulty mode (given the current set of measurements), while the ones related to continuous-valued states will provide a pdf estimate suitable for the use of FDI hypothesis testing routines (e.g. Fisher’s Discriminant Ratio, Scheffé’s test and other threshold-based techniques) and/or to be used as initial conditions in prognosis modules after a fault condition is statistically confirmed.

1.3.2 Particle Filtering for Prognosis: Generation of Long Term Prediction in Stochastic Nonlinear Systems

The second main research objective refers to the use of a particle-filtering-based framework for on-line failure prognosis in nonlinear, non-Gaussian systems. It is expected for this implementation to statistically characterize the RUL of a subsystem or component affected by a fault condition (which has been already detected), i.e. to estimate the probability density function of the subsystem RUL. As in every filtering application, a set of measurements (vibration data or feature data) will provide real time information about the state of the system and will be used to improve current estimates.

The proposed implementation will consider nonlinear state-space models, with unknown time varying parameters to be also estimated with the help of the particle filtering module. Resulting estimates for both model parameters and state probability density functions will be used on-line to generate long term predictions for the evolution in time of the fault indicators.
Empirical knowledge about critical conditions for the system will be included in the formulation of the prognosis problem in the form of thresholds for the main fault dimensions, also referred to as the hazard zones. In the case of real applications, it is expected for the hazard zones to be statistically determined on the basis of historical failure data.

Information present in long term predictions for fault indicators, in conjunction with the definition of hazard zones, will be translated into a single prognosis output: the probability of failure for any future time instant, namely the RUL pdf. This outcome will be available and updated in real time, providing information about statistical confidence intervals, expectations and other hypothesis tests for the failure condition under study. The proposed approach will be tested and validated using a stream of real failure data from a cracked planetary gear plate.

1.3.3 The Inner and Outer Loops for Diagnosis and Prognosis

One last objective for the proposed research work is to establish learning schemes where the information acquired on-line, from measurements, is transformed into a set of corrections for the parameters that characterize the performance of FDI and/or prognosis algorithms. In that sense, and according to the nature of the parameters that these schemes modify, two different kinds of correction loops may be distinguished: inner and outer correction loops.
*Inner correction loops* intend to use recent information to alter the value of unknown variables or parameters directly related to the state estimation problem, although assuming that the influence from exogenous variables is well known and that the structure of the nonlinear process model and/or noise parameters are accurate. On-line *inner correction loops* are inherent to the proposed particle filtering approach for FDI and/or failure prognosis, since the use of Bayesian approaches directly involves updates in the estimates (both unknown time varying model parameters and state pdf estimates) which depend directly on process measurements.

*Outer correction loops*, on the other hand, are capable of compensating for changes in exogenous variables (such as the loading profile in the analysis of cracked planetary gear plates), model structure inaccuracies and even imprecision in the initial assumptions about noise parameters (for example, the assumption of zero mean measurement noise). Implementation of *outer correction loops* necessarily implies the use of new algorithms that should run in parallel with the particle filtering module, in an on-line fashion.

Therefore, given the significant improvement that these algorithms may offer to the overall methodology in terms of both accuracy and efficiency, part of the proposed research work will be devoted to the implementation and testing of novel on-line approaches for *outer correction loops* in both FDI and prognosis modules.
In the case of FDI modules, outer loops will be focused on managing the size of the particle population that is needed to track unusual failure modes. The concept is founded on the idea of “abstract states,” introduced in [34]. Once the proper particle population size and abstract resolution level have been determined, risk sensitive importance distributions will be used to allocate the particles.

In prognosis applications, on the other hand, outer correction loops will use autoregressive models and records of past estimates for the system RUL expectation (generated by the inner correction loop) to compensate for errors in the structure of process model. It is expected that the combination of model-based (PF) and data driven techniques (parameter estimation for corrective autoregressive models) will significantly improve the prognosis algorithm accuracy.

Advanced stages in the research work will also intend to rectify initial assumptions in process/measurement noise parameters, through similar approaches. This would enable, for example, to modify the mean or standard deviation of a Gaussian measurement noise model according to the information collected from real-time data.
Section 2.1 Preliminary Research

2.1.1 Particle Filtering for Diagnosis in Stochastic Nonlinear Systems

A fault detection and identification (FDI) procedure may be interpreted as the fusion and utilization of the information present in a feature vector (measurements), with the objective of determining the operational condition (state) of a system and the causes for deviations from particularly desired behavioral patterns. From a nonlinear Bayesian state estimation standpoint, this task may be accomplished by the use of a particle filter-based module built upon the nonlinear dynamic state model (2.1.01) [33].

\[
\begin{align*}
    x_d(t+1) &= f_b(x_d(t) + n(t)) \\
    x_c(t+1) &= f_i(x_d(t), x_c(t), \omega(t)) \\
    \text{Features}(t) &= h_i(x_d(t), x_c(t), v(t)) 
\end{align*}
\]

(2.1.01)

Where \( f_b, f_i \) and \( h_i \) are non-linear mappings, \( x_d(t) \) is a collection of Boolean states associated with the presence of a particular operational condition in the system (normal operation, fault type #1, #2, etc.), \( x_c(t) \) is a set of continuous-valued states that describe the evolution of the system given those operational conditions, \( n(t) \) is zero-mean i.i.d. uniform white noise and \( \omega(t), v(t) \) are non-Gaussian distributions that characterize the process and feature noise signals, respectively.

A particle filter approach based on model (2.1.01) allows to statistically characterizing the evolution in time of the state pdf, as new feature data are received, via
the modification of the probability masses associated with each particle. Specifically, the output of the FDI module – defined as the current expectation of each Boolean state – provides a recursively updated estimate of the probability for each fault condition considered in the analysis. Once these on-line estimates are generated, any external criteria may be used to define thresholds for purposes of activating alarm indicators or prognostic modules.

In fact, pdf estimates for the system continuous-valued states (computed at the moment of fault detection) may be also used as initial conditions in the failure prognostic routines, giving an excellent insight about the inherent uncertainty in the prediction problem. As a result, a swift transition between the two modules (FDI and prognostics) may be performed, and moreover, reliable prognosis can be achieved within a few cycles of operation after the fault is declared. This characteristic is, in fact, one of the main advantages that the proposed particle filter-based FDI framework may offer.

The proposed approach has been successfully implemented on three different applications, as part of a preliminary research. For purposes of a better understanding, each of these applications will be now explained in depth.

2.1.1.1 Detection of Cracks in Blades of a Turbine Engine

Consider the case where the proposed methodology is applied for the detection of cracks in the blades of a turbine engine. Light probes on both the leading and the trailing edge of the blades have been installed in order to provide with the time-of-arrival (TOA)
for each blade. Some pre-processing techniques were needed in order to generate a feature that can be used for detection purposes, the tangential blade position (TBP). Figure 2.1.1-1 shows a schematic that illustrates the pre-processing steps, which basically start with the least square estimation of two parameters, namely A and B, that relate the inter-blade spacing (IBS) with the square of the existing normalized rpm value.

![Figure 2.1.1-1. Particle filter-based FDI module. Crack in a turbine engine blade.](image)

In addition to the vibration-based feature mentioned above, and after an exhaustive FRANC-3D structural analysis about the stresses on turbine blades undergoing a crack, it has been determined that (2.1.02) represents a model suitable for describing the growth of a crack on any blade under nominal load conditions.
\[
\frac{dL}{dn} = \frac{1}{6\alpha \cdot L^3(n) + p(L(n))} \quad (2.1.02)
\]

Where \( L \) is the length of the crack (in inches), \( n \) is the number of stress cycles applied to the material, \( \alpha \) is a model parameter to be estimated and \( p(L(n)) \) is a known fourth order polynomial determined with the help of the FRANC-3D structural model.

Assuming the existence of a 1-to-1 nonlinear mapping \( h(\cdot) \) between the feature information and the actual size of the crack in the blade, it is possible to implement a nonlinear model suitable for a particle filter-based FDI framework. The model is shown in (2.1.03), being \( \beta \) a known model parameter and where \( \omega(t) \) and \( v(t) \) have been selected as zero mean Gaussian noises for simplicity.

\[
\begin{bmatrix}
  x_{d,1}(t+1) \\
  x_{d,2}(t+1)
\end{bmatrix} = f_b\left( \begin{bmatrix}
  x_{d,1}(t) \\
  x_{d,2}(t)
\end{bmatrix} + n(t) \right)
\]

\[
x_c(t+1) = \left[ (1 + \beta) x_c(t) \right] \cdot x_{d,2}(t) + \omega(t)
\]

Feature\((t) = x_c(t) + v(t)
\]

\[
f_b(x) = \begin{cases} 
  \begin{bmatrix} 1 & 0 \end{bmatrix}^T, & \text{if } \|x - [1 & 0]^T\| \leq \|x - [0 & 1]^T\| \\
  \begin{bmatrix} 0 & 1 \end{bmatrix}^T, & \text{else} 
\end{cases}
\]

\[
\begin{bmatrix}
  x_{d,1}(0) \\
  x_{d,2}(0)
\end{bmatrix} = \begin{bmatrix} 1 \\
  0
\end{bmatrix}
\]
The FDI module itself discriminate between two Boolean states: absence of crack in any blade (*normal condition*) or presence of crack (*fault condition*) in at least one of them. As one value of the tangential blade position feature is calculated per blade and per cycle, the module is also able to pinpoint the blade which has been affected by a crack condition (*fault identification*). Results of the detection module for the case of one particular blade are shown in Figure 2.1.1-2.

![FDI results: Blade #12](image)

**Figure 2.1.1-2.** Detail of detection results for a crack in a turbine engine blade.

Although it is possible to observe some changes in the probability of failure condition around the 230th cycle of operation, it is clear that only after the 280th cycle the
evidence in the feature is strong enough to ensure the existence of a crack. After that particular time instant, the pdf estimate for the state $x_c(t)$ – together with $h(\cdot)$ – may be used for prognosis purposes.

### 2.1.1.2 Detection of Degradation Level for a Battery

A similar performance is obtained when applying this methodology to isolate the time instant where the degradation level of the battery, which is directly related to the value of its internal resistance, is significant. A good feature to be used in this case may be the amount of energy delivered during a short amount of time, i.e. $\int_{t_0}^{t_0+\Delta} v(t) \cdot i(t) dt$, which can be mapped through a 1-to-1 nonlinear relationship with the value of the internal resistance for the battery. Also in this case, the Boolean state $x_d(t)$ discriminates between two possible conditions for the battery: healthy or faulty, where the latter has been arbitrarily defined as the moment when the estimated internal resistance is $5 \, \text{m\Omega}$. The nonlinear model used for FDI purposes in this type of application may be significantly simplified, since it considers the existence of pre-defined thresholds for a state variable whose expectation increases monotonically in time, see (2.1.04), where $\beta$ is a known model parameter and where $\alpha(t)$ and $\nu(t)$ have been selected as zero mean Gaussian noises for simplicity.

\[
R(t+1) = (1 + \beta) \cdot R(t) + \omega(t)
\]

\[
x_d(t+1) = \begin{cases} 
1, \text{ if } R(t+1) \geq 5 \\
0, \text{ else}
\end{cases}
\]

\[
y(t) = R(t) + \nu(t)
\]

(2.1.04)
Under this approach, the probability of failure is basically given by the sum of the weights of those particles which consider an internal resistance value greater than the given threshold of 5 (mΩ). Figure 2.1.1-3 and 2.1.1-4 show some results obtained when a battery with an initial internal resistance of 3 (mΩ) is considered. The detection routine shows that the probability of failure changes drastically after 72 weeks of operation.

**Figure 2.1.1-3.** Particle filter-based FDI module. Battery degradation problem

**Figure 2.1.1-4.** Particle filter-based FDI module. Results for battery degradation problem
2.1.1.3 Detection of Cracks in a UH-60 Planetary Gear Plate

The study of the growth of an axial crack in a UH-60 gear plate, and the structural model details that should be considered before implementing any on-line, model-based approach for fault analysis will be topic of an elaborate discussion in section (2.1.5). For the time being, consider the case of a seeded fault test on a planetary gear plate. During this test, a cyclic load profile is being applied to the plate to analyze how it affects the growth of an existent axial crack. It is assumed that it is possible to compute a noisy estimate of the crack length based on the processing of real-time vibration data. In particular, the main objective of this case study is to determine when this crack increases its axial growth rate. Customer specifications include early detection of changes in the growth rate and the desired statistical confidence level.

Thus, we will distinguish between two main operational conditions: the normal condition will reflect the fact that the crack is growing very slowly or not growing at all, meanwhile the faulty condition indicates an abrupt change in the growing rate. Consequently, nonlinear model (2.1.05) may be used to detect the faulty condition with the help of a Particle Filter-based FDI module, where $\beta$ is a time-varying model parameter dependant from the loading profile which is being applied to the gearbox and where $\alpha(t)$ and $\nu(t)$ have been selected as zero mean Gaussian noises for simplicity. In this case, besides the detection of the faulty condition, it is desired to obtain some measure of the statistical confidence of the alarm signal. For that purpose, let us assume that enough historical data has been collected to characterize the pdf of the noisy crack estimate as a Normal distribution $\mathcal{N}(\mu, \sigma^2)$. 
\[
\begin{align*}
\begin{bmatrix} x_{d,1}(t+1) \\ x_{d,2}(t+1) \end{bmatrix} &= f_b \left( \begin{bmatrix} x_{d,1}(t) \\ x_{d,2}(t) \end{bmatrix} + n(t) \right) \\
x_i(t+1) &= x_i(t) + \beta \cdot x_i(t) \cdot x_{d,2}(t) + \omega(t) \\
y(t) &= x_i(t) + v(t)
\end{align*}
\]

(2.1.05)

\[
\begin{bmatrix} x_{d,1}(0) \\ x_{d,2}(0) \\ x_i(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 3.4 \end{bmatrix}
\]

One way to generate an indicator of statistical confidence for the detection procedure is by considering the sum of the weights of all particles $i$ such that

\[x_c^{(i)}(T) \geq z_{1-\alpha, \mu, \sigma^2},\]

where $\alpha$ is the desired test confidence and $T$ is the detection time. If additional information is required, it is possible to compute the value of the Fisher’s Discriminant Ratio, given in this particular case by (2.1.06). Once computed, either of the above mentioned indicators may help to define an appropriate detection threshold for the problem under study.

\[
F_{index}(T) = \frac{\left| \mu - \sum_{i=1}^{N} w_r^{(i)} \cdot x_c^{(i)}(T) \right|^2}{\sigma^2 + \sum_{i=1}^{N} w_r^{(i)} \cdot \left( x_c^{(i)}(T) - \sum_{j=1}^{N} w_r^{(j)} \cdot x_c^{(j)}(T) \right)^2}
\]

(2.1.06)
Figure 2.1.1-5 shows the results obtained when applying the proposed FDI approach to the problem of crack growth detection in the planetary gear plate. The vertical line that discriminates between the two pdf's in Figure 2.1.1-5 is fixed by the desired type I detection error (probability of false positives), considering the data used as a baseline for detection purposes. By comparing the trend of the vibration-based crack estimate over time, it is clear that no significant increment in the crack length happened before the 100th GAG (ground-air-ground) cycle. The algorithm only needs 35 additional GAG cycles to detect a change in the growth rate of the crack, with a confidence level of approximately 70% (type II error $\approx 30\%$).

![PF Detection Routine: GAG =135](image1)

![Probability of Failure](image2)

![Fisher Discriminant Ratio](image3)

**Figure 2.1.1-5.** Particle filter-based FDI module. Cracked plate problem
Figure 2.1.1-5 shows, in fact, three indicators that are simultaneously computed. The first indicator, depicted as a function of time, shows the probability of a determined failure mode within the system, and it is based on the estimate of the discrete-valued states in model (2.1.05), $x_{d,1}$ and $x_{d,2}$ respectively. FDI results are obtained whenever this probability reaches the desired confidence level for a particular mode. If more information is needed, the value of the Fisher’s Discriminant Ratio or the type II detection error (2\textsuperscript{nd} and 3\textsuperscript{rd} indicators respectively) may be considered.

It must be noted that, in this approach, no particular specification about the detection threshold has to be made prior to the actual experiment. Customer specifications are translated into acceptable margins for the type I and II errors in the detection routine. The algorithm itself will indicate when the type II error (probability of false negatives) has decreased to the desired level. Additional research efforts will be spent in order to define general guidelines for the use of these statistical indicators in any application.

2.1.2 Particle Filtering for Prognosis in Stochastic Nonlinear Systems

Prognosis may be essentially understood as the generation of long-term (multi-step) predictions describing the evolution in time of a particular signal of interest or fault indicator, with the purpose of estimating the remaining useful life (RUL) of a failing component/subsystem. Since prognosis intends to project the current system condition in
time – using a state dynamic model in the absence of future measurements – it necessarily entails large-grain uncertainty. For this reason, any accurate and precise prognosis scheme should unavoidably consider fault indicators (and other critical state variables) as random processes in such a way that, once their probability distributions are estimated, other important attributes – such as confidence intervals – may be computed.

Real-time data (measurements or features) provided by sensors monitoring key fault parameters, suggest a possible solution to the prognosis problem based on recursive Bayesian estimation techniques. Under this approach, long-term predictions for the fault indicators are generated using dynamic growth models, while accurate real-time state estimates are used as initial conditions for those models. If the incipient failure is detected and isolated at early stages, it is reasonable to assume that sensor data will be available for a certain time window allowing for adjustments in the predictions, i.e. improvements in the model parameter estimates, so that prognosis will provide more accurate and precise results. At the end of the observation time window, however, the prediction outcome should be delivered to the user (operator, maintainer), who must decide which corrective actions to take in order to avoid a catastrophic event.

Particle filtering, as it has been previously mentioned, is particularly useful when dealing with difficult nonlinear and/or non-Gaussian processes and thus, it is very suitable as a Bayesian tool for prognosis purposes [27]. Sequential importance sampling helps to reduce the number of samples required to approximate distributions with appropriate precision, being faster and more computationally efficient than classical
Monte Carlo methods. Moreover, particle filtering allows information from multiple measurement sources to be fused in a principled manner.

Prognosis, though, is a problem that goes beyond the scope of filtering applications since it involves future time horizons. Hence, if PF-based algorithms are to be used, it is necessary to propose a procedure that has the capability to project the current particle population in time in absence of new observations, adjusting weights if necessary. Furthermore, since the main source for uncertainty is related to the fact that both the process and measurement models (including noise definition and statistics) are subject to errors, it is not reasonable to expect an improvement in prognosis accuracy only by perfecting the state and model parameter estimation technique. With the purpose of solving these issues, a two-level procedure has been developed and subsequently tested for particular applications. This procedure intends to reduce the uncertainty associated with long-term predictions by using the current state pdf estimation, the process noise model, and a record of corrections made to previous computed predictions.

In a first prognosis level, $p$-step ahead predictions are generated on the basis of an a priori estimate, adjusting their associated probabilities according to the noise model structure. A second prognosis level uses these predictions and the definition of critical thresholds to estimate the RUL pdf, also referred to as the time-to-failure (TTF) pdf, and simultaneously implements a correction model (outer correction loop) to compensate for all main error sources. A detailed description of each level is now presented.
2.1.2.1 First Prognosis Level: Generation of Long-Term Predictions

The first prognosis level is related to the generation of a $p$-step ahead long term prediction for the state pdf of a dynamic system, which can be obtained in a recursive manner using both the model update equation (1.2.01) and the current state estimate, as shown in (2.1.07).

$$
\tilde{p}(x_{t+p} \mid y_{1:t}) = \int \tilde{p}(x_t \mid y_{1:t}) \prod_{j=t+1}^{t+p} p(x_j \mid x_{j-1}) dx_{t+1:t+p-1}
$$

$\approx \sum_{i=1}^{N} w_i^{(i)} \int \cdots \int p(x_{t+1} \mid x_{t+1}^{(i)}) \prod_{j=t+2}^{t+p} p(x_j \mid x_{j-1}) dx_{t+1:t+p-1}
$ (2.1.07)

The evaluation of these integrals, though, may be difficult and/or may require significant computational effort, even in the case when a PF algorithm is used to approximate the state pdf for subsequent time instants. In order to simplify and solve this problem, three main approaches are now presented and explained in detail.

2.1.2.1.1 $p$-step ahead Long-Term predictions: First Approach.

This first approach predicts the evolution in time of each particle by successively taking the expectation of the model update equation (1.2.01) for every future time instant, considering the state value associated to that particle as initial condition, as shown in (2.1.08).

$$
\tilde{x}_{t+p}^{(i)} = E[f_{t+p}(\tilde{x}_{t+1:t+p-1}^{(i)}, \omega_{t+p})] \quad ; \quad \hat{x}_t^{(i)} = \tilde{x}_t^{(i)}
$$

(2.1.08)
The weight of every particle should be modified (at each prediction step) in order to take into account the fact that the noise and process nonlinearities could change the shape of the state pdf as time passes. However, as the weight update procedure is needed as part of a prediction problem, it can not depend on the acquisition of new measurements. In order to solve this difficulty, a procedure based on the use of the process noise model has been implemented.

The procedure is as follows: consider the predicted conditional pdf

\[ \hat{p}(x^{(i)}_{t+k} | \hat{z}^{(i)}_{t+k-1}) , \]

which describes the state distribution at the future time instant \( t+k \ (k=1,\cdots,p) \) when the particle \( \hat{z}^{(i)}_{t+k-1} \) is used as initial condition. Assuming that the initial weights \( \{w^{(i)}_t\}_{i=1}^N \) are a good representation for the current state pdf, then it is possible to approximate the predicted state pdf at time \( t+k \), by using the law of total probabilities and the particle weights at time \( t+k-1 \), as shown in (2.1.09).

\[
\hat{p}(x_{t+k} | \hat{x}_{t+k-1}) \approx \sum_{i=1}^N w^{(i)}_{t+k-1} \cdot \hat{p}(x^{(i)}_{t+k} | \hat{z}^{(i)}_{t+k-1}) ; \ k = 1,\cdots,p \quad (2.1.09)
\]

As it was mentioned before, (2.1.09) assumes that the updated weights \( w^{(i)}_{t+k-1} \) represent an accurate sampled version for \( \hat{p}(x_{t+k-1} | x_{0_2}, y_{1_2}) \), namely the predicted state pdf for the previous time instant. If in addition the domain of the particle population \( \{x^{(i)}_{t+k}\}_{i=1}^N \) is considered to be a representative subset for the domain for the state random
variable \( x_{t+k} \), then the following algorithm may be used to assign adequate values to each particle weight at the prediction time \( t+k \):

### Weight update for Long-Term Prediction

- Construct a partition of the random variable domain by defining:
  \[ d_{t+k}^{(1)} = -\infty; \quad d_{t+k}^{(N+1)} = \infty \]
  \[ d_{t+k}^{(j)} = \frac{1}{2} (\hat{x}_{t+k}^{(j)} + \hat{x}_{t+k}^{(j-1)}), \quad j = 2, \ldots, N \]

- Generate the updated particle weights by computing:
  \[ w_{t+k}^{(i)} = \int_{\hat{x}_{t+k}^{(i)}}^{\hat{x}_{t+k}^{(i+1)}} \hat{p}(x_{t+k} | \hat{x}_{0:t+k-1}, y_{t\tau}) dx_{t+k} \]

The proposed method is easy to implement as long as the process noise is uncorrelated (diagonal covariance matrix for \( \omega(t) \)). If that assumption does not hold, though, the integral associated with the weight update step may not be solvable.

### 2.1.2.1.2 P-step ahead Long-Term predictions: Second Approach.

The second approach for long-term prediction intends to avoid the computational effort implied in the update of the particle weights for future time instants, especially if the prediction time horizon is large. In that sense, instead of recalculating the particle weights, uncertainty for future transitions is incorporated by simply resampling the predicted state pdf (2.1.09).
Thus, the information about the distribution of the state for future time instants is now given by the position of the particles, not by the particle weight value. The implementation of this methodology, however, must ensure that the resampled population is representative of (2.1.09). A computationally affordable solution for this predicament is proposed, based on the assumption of uncorrelated process noise (diagonal covariance matrix for $\omega(t)$) and the use of kernel transitions to describe the state pdf before the resampling step, as it is also done in the case of the regularized particle filter (RPF).

Consider, in that sense, a discrete approximation (2.1.10) for the predicted state pdf (2.1.09), where $K(\cdot)$ is a kernel density function, which may correspond to the process noise pdf, a Gaussian kernel or a rescaled version of the Epanechnikov kernel (1.2.32) [26].

$$\hat{p}(x_{t+k} | \hat{x}_{t+k-1}) \approx \sum_{i=1}^{N} w_{t+k-1}^{(i)} K\left(x_{t+k} - E\left[x_{t+k} | \hat{x}_{t+k-1}^{(i)}\right]\right) \quad (2.1.10)$$

It is reasonable to try to represent the uncertainty present in (2.1.10), instead of just projecting the conditional expectations of the state variables. One way to achieve this task is to generate a new population of equally weighted particles for the time instant $t+k$, $1 \leq k \leq p$, performing an inverse transform resampling [30] procedure for the particle population. This method obtains samples distributing according to (2.1.10), selecting $N$ realizations of $u^{(i)} \sim U(0,1)$ and interpolating a value for $\hat{x}_{t+k}^{(i)}$ from the
cumulative state distribution \( F(X_{t+k} \leq x_{t+k}) = \int_{-\infty}^{x_{t+k}} \hat{p}(x_{t+k} | \hat{x}_{t+t+k-1}) dx_{t+k} \) in accordance with 
\[
\hat{x}^{(i)}_{t+k} = F^{-1}(u^{(i)}) .
\]

The inherent randomness present in the inverse transform resampling method, however, may lead to unrepresented areas in the domain of the cumulative state distribution function, situation which is difficult to correct in long term predictions since there are no measurements available that may be used for this purpose. Thus, and in order to overcome this difficulty, a two-step procedure is proposed.

The first step in the resampling strategy performs a simplified version of the inverse transform resampling procedure, which will focus in representing the growth of uncertainty present in (2.1.10). Samples distributing according to (2.1.10) are obtained by selecting \( u^{(i)} = \frac{i}{N+1} \) \((i:1, \cdots, N)\), and interpolating a value for \( \hat{x}^{(i)}_{t+k} \) from the cumulative state distribution \( F(X_{t+k} \leq x_{t+k}) = \int_{-\infty}^{x_{t+k}} \hat{p}(x_{t+k} | \hat{x}_{t+t+k-1}) dx_{t+k} \) in accordance with 
\[
\hat{x}^{(i)}_{t+k} = F^{-1}(u^{(i)}) .
\]

In order to avoid loss of diversity among particles, an additional step inspired in the RPF is performed. In that sense, it is assumed that the state covariance matrix \( \hat{S}_{t+k} \) equal to the empirical covariance matrix of \( \hat{x}_{t+k} \) and that a set of equally weighted samples for \( \hat{x}_{t+k} \) is available, in such a way that the efficiency in the use of Epanechnikov kernels for pdf approximation is maximized [26].
In consequence, considering all of the above, the regularization algorithm applied for long term predictions is as follows:

**Long Term Predictions: Second Approach**

- Apply modified inverse transform resampling procedure. For \( i = 1, \cdots, N \), \( w_{t+k}^{(i)} = N^{-1} \)
- Calculate \( \hat{S}_{t+k} \), the empirical covariance matrix of \( \{ E\left[ x_{t+k}^{(i)} \mid \hat{z}_{t+k-1}^{(i)} \right], w_{t+k}^{(i)} \}_{i=1}^{N} \)
- Compute \( \hat{D}_{t+k} \) such that \( \hat{D}_{t+k} \hat{D}_{t+k}^T = \hat{S}_{t+k} \)
- For \( i = 1, \cdots, N \), draw \( \varepsilon \sim K \), the Epanechnikov kernel and assign 
  \[ \tilde{x}_{t+k}^{(i)*} = \tilde{x}_{t+k}^{(i)} + h_{t+k}^{opt} \hat{D}_{t+k} \varepsilon \]
  where \( h_{t+k}^{opt} \) is computed as in (1.2.33)

It is important to notice that the assumption of uncorrelated process noise is only included for the sake of reducing the computational effort of the resampling procedure. In fact, there are no theoretical restrictions for the application of this methodology in the presence of correlated process noise.

### 2.1.2.1.3 P-step ahead Long-Term predictions: Third Approach.

The third and last approach for long-term prediction presented in this section is, in fact, simpler in terms of computational effort than the previous ones. Whereas the first approach defines an update equation for the particle weights and the second approach intends to apply resampling steps for future time instants, the third approach states that the error that can be generated by considering the particle weights invariant for future time instants is negligible with respect to other sources of error that may appear in
practical applications, such as model inaccuracies or even in the assumptions made for process and measurement noise parameters [10].

Therefore – from this standpoint – (2.1.08) is considered sufficient to extend the trajectories $\hat{x}_{a+z_k}^{(i)}$, while the current particle weights are propagated in time without changes. The computational burden of this method is significantly reduced and, as it will be shown in simulation results, the method still offers a satisfactory view about how the system behaves for most practical applications.

2.1.2.2 Second Prognosis Level: Estimation and Statistical Characterization of the Remaining Useful Life (RUL) of Equipment

The final outcome for any prognosis algorithm is an estimate for the system RUL pdf, which is intrinsically entangled with the probability of failure at future time instants. This probability can be obtained from long-term predictions, when the empirical knowledge about critical conditions for the system is included in the form of thresholds for main fault indicators, also referred to as the hazard zones.

In real applications, it is expected for the hazard zones to be statistically determined on the basis of historical failure data, defining a critical pdf with lower and upper bounds for the fault indicator ($H_{lb}$ and $H_{up}$, respectively).

Since the hazard zone specifies the probability of failure for a fixed value of the fault indicator, and the weights $\left\{ w_{i+z_k}^{(j)} \right\}_{i=1}^{N}$ represent the predicted probability for the set of
predicted paths, then it is possible to compute the probability of failure at any future time instant (namely the RUL pdf) by applying the law of total probabilities, as shown in (2.1.11). Once the RUL pdf is computed, combining the weights of predicted trajectories with hazard zone specifications, it is well known how to obtain prognosis confidence intervals, as well as the RUL expectation.

\[
\hat{p}_\text{TTF}(\text{ttf}) = \sum_{i=1}^{N} \Pr\left(\text{Failure} \mid X = \hat{x}_{\text{ttf}}^{(i)}, H_{lb}, H_{up}\right) \cdot w_{\text{ttf}}^{(i)} \quad (2.1.11)
\]

Expression (2.1.11) provides a solution for the RUL pdf estimation problem which is suitable for on-line applications. As it depends on the predicted trajectory weights, though, it is subject to uncertainty and it may be sensitive to modeling errors. Moreover, uncertainty inherent to RUL expectations will increase as the prediction horizon grows. This issue is of especial interest in prognosis, since the estimation of the RUL must be done immediately after the fault condition has been detected, and hence most of the prediction horizons involve considerable periods of time.

In particular, and in order to reduce the uncertainty inherent to a Particle Filter-based failure prognosis and improve the accuracy of the RUL expectation, an additional outer correction loop has been included as part of the proposed second prognosis level, see Figure 2.1.2-1.

This outer loop is basically a data-driven learning paradigm. It computes a correction term \( C_n \) consisting of the difference between the current RUL expectation and
the one that was computed in the previous iteration of the prognosis algorithm. Once $p$ correction terms are obtained, a linear autoregressive model is built to establish a relationship between all past correction terms.

The obtained linear autoregressive model is then used to generate an estimate for all future corrections that would be applied to the current RUL expectation, if both process and measurement noises were to be wide sense stationary (WSS). In simple words, the proposed outer correction loop intends to capture the pattern of past measurement-driven prediction updates inside a simple model, which can be used afterwards to estimate and correct for the accuracy of the current prediction.

$$T_k^{(c)} = T_k - \sum_{l=k+1}^{k+p} C_l$$

**Figure 2.1.2-1.** Correction algorithm for RUL expectation

The learning scheme proposed here is just an example about how the combination of model-based (Particle Filtering) and data driven techniques (linear autoregressive
correction models) in an outer correction loop can significantly improve the prognosis algorithm accuracy.

Other approaches may be also implemented as a manner of incorporating information from past instances: additional outer correction loops may also help to reduce prediction uncertainty by modifying both the structure and parameters of process noise in the dynamic model. These topics will be considered for future research efforts.

2.1.3 Illustrative Example: RUL Statistical Characterization

Subsection 2.1.2 introduces several approaches that can be considered to solve the problem of RUL statistical estimation by means of a particle-filtering-based methodology. Whereas some of them are intended to generate a more accurate representation of the RUL probability density function, others aim to simplify the computational burden so it can be easily implemented in on-line applications. In order to properly understand the main advantages and disadvantages that these methods may offer, it is recommended to compare their performances in an illustrative example. For this purpose, consider the problem of RUL estimation in a process for which the evolution in time of a known failure condition (for instance, a crack in a material) is described by the nonlinear model (2.1.12).
\[
\begin{align*}
    x_1(k+1) &= x_1(k) + 3 \cdot 10^{-4} \left( 0.05 + 0.1 \cdot x_2(k) \right)^3 + \omega(k) \\
    x_2(k+1) &= x_2(k) + 1 \\
    y(k) &= x_1(k) + v(k)
\end{align*}
\]
(2.1.12)

\[
\omega(k) \sim \text{Gamma}(0.15, 0.3)
\]

\[
v(k) \sim \frac{1}{4} N(-0.5, 0.25) + \frac{3}{4} N(0.5, 0.25)
\]

Furthermore, in order to analyze the effect that inaccuracies and model errors imply in RUL estimates, let us assume that noise is believed to be Gaussian. In that case, the first two moments of both process and observation noises may be then estimated using historical data, obtaining (2.1.13).

\[
\omega(k) \sim N(0.045, 0.1162), \quad v(k) \sim N(0.25, 0.5)
\]
(2.1.13)

The hazard zone, which in real applications must be defined on the base of customer specifications or ground truth failure data, is defined here as a normal probability density function with parameters \(\mu = 9.0\) and \(\sigma = 0.3\). The main objective is to generate a 95\% confidence interval for the RUL of the process, 40 cycles after the fault condition is detected.

In addition to the techniques described in subsection 2.1.2, an Extended Kalman Filter (EKF)–based prognosis procedure has also been considered as a means for both comparison and performance evaluation for the proposed particle-filtering-based techniques.
All results are summarized in Figure 2.1.3-1, where the green and magenta lines represent, respectively, the noisy measurements and the process output estimation obtained from an SIR particle filter, and where the blue line shows the actual evolution of the failure condition for future time instants (information that is unknown at the moment where the RUL estimation is performed).

![Particle Filters: Non-Linear System State Estimation](image)

**Figure 2.1.3-1.** Result comparison for RUL statistical characterization

Figure 2.1.3-1 provides valuable information that may be used to evaluate the capability of each algorithm to predict the evolution in time of the state probability
distribution, particularly when some performance metrics (such as precision and accuracy) are invoked to assess the algorithm performance.

For instance, consider the RUL (also referred to as TTF) pdf estimates obtained by using the PF–based first approach for long term predictions and the ones obtained with the EKF–based procedure. It is possible to notice an important difference in terms of the uncertainty of the prediction (precision), even though both techniques coincide on the expectation for the RUL of the process (accuracy).

Concerning this last issue, it must be noted that differences in precision are mainly caused by the fact that the EKF assumes a normal pdf for the state, regardless of the likelihood of the estimation, whereas the PF is able to discard possible outcomes for the state if the probability masses associated with them are negligible.

The major problem with both techniques, however, seems to be related to the accuracy of the algorithm, i.e. the capability of estimating the expectation of the RUL for the process. In the case of the EKF–based procedure, this issue is related to the fact that long term predictions greatly depend on the current expectation of the state, assumption that may not be representative of the non-linear process behavior in a long term horizon when model errors are included. Similar problems affect the PF–based first approach for long term predictions, since the algorithm depends importantly on those particles which have a higher likelihood, given the current observation data.
On the other hand, as it is also shown in Figure 2.1.3-1, the PF–based second approach for long term predictions is capable to overcome the bias introduced by model errors, due to its ability to represent the state probability space. The combination of resampling techniques and Epanechnikov kernels for pdf approximation in long term predictions is able to simultaneously reduce the impact of model inaccuracies and provide a balanced result in terms of accuracy and precision in the RUL estimate. Furthermore, the actual fault indicator (unknown when the long term predictions were performed) reaches the previously defined hazard zone inside the 95% confidence interval, confirming the validity of the RUL pdf estimate.

Results for the PF-based third approach for long term predictions are not shown in Figure 2.1.3-1 since they are quite similar to the ones obtained with the first approach. It is important to note, however, that the computational burden is considerably less in the case of the third approach and, thus, it seems to be more suitable than the first one for on-line applications.

Finally, it must be noted that when the outer loop correction scheme — introduced as part of the second prognosis level — is applied to the PF-based second approach for long term prediction generation, it allows improving the estimate of the RUL expectation to the extent that the corrected estimate time-to-failure (TTF) coincides with the time instant where the actual failure growth reaches the mean of the hazard zone (9"), see Figure 2.1.3-1.
2.1.4 Case Study: Analysis of Crack Growth in a Turbine Engine Blade

The implementation and testing of the proposed particle-filtering-based methodology for fault prognosis on real process data, and the subsequent assessment of the obtained results, has been amongst the most important contributions intended for this preliminary research work. With this purpose, two different sets of data have been used to define study cases related to the analysis of the growth of a crack in a rotatory piece of equipment.

Figure 2.1.4-1. Picture of turbine engine under study

For the first case study, consider the problem discussed in subsection 2.1.1.1 where the objective is to analyze the growth in a crack on a turbine engine blade, see Figure 2.1.4-1. As it was previously mentioned, light probes on both the leading and the trailing edge of the blades have been installed in order to provide with the time-of-arrival (TOA) for each blade, and this information has been processed to come up with a feature
directly related to the size of the crack in the blade, namely the tangential blade position (TBP).

Although subsection 2.1.1.1 was primarily focused in the use of this feature for fault detection purposes, the obtained results may be also used to set up a fault prognosis framework where not only the size of the crack will be estimated, but also a 95% confidence interval for the RUL of the piece of equipment may be computed.

With this purpose, as it was also mentioned in subsection 2.1.1.1, an exhaustive turbine structural analysis was conducted in a FRANC-3D environment to determine a model capable of describing the growth of a crack in one blade under nominal load conditions. The result is shown in model (2.1.14) where \( L \) is the length of the crack (in inches), \( n \) is the number of stress cycles applied to the material, \( \alpha \) is a random variable with known first two moments, \( p(L(n)) \) is a known fourth order polynomial determined with the help of FRANC-3D structural model, and \( \omega(n) \) and \( \nu(n) \) are i.i.d. white noises.

\[
\frac{dL}{dn} = \frac{1}{6\alpha \cdot L^2(n) + p(L(n))} + \omega(n) \quad (2.1.14)
\]

\[
y(n) = h^{-1}(L(n)) + \nu(n)
\]

The measurement equation is considered to be linear on \( h^{-1}(\cdot) \), where \( h(\cdot) \) is a known nonlinear mapping between the crack size and feature value. Besides the fact that
the state equation in non-linear, the noise signal $\omega(n)$ is non-negative, and thus non-Gaussian.

Model (2.1.14) is suitable for the generation of long term predictions and hence, for the implementation of a particle-filtering-based framework for prognosis. Moreover, a similar realization has been already used for fault detection purposes, giving evidence that the existence of a crack in a blade is ensured after the 280\textsuperscript{th} cycle of operation. Thus, FDI results can be included as initial conditions for prognosis routines by assigning the resulting pdf estimate for the state $x_c(t)$ at $t = 280$, from model (2.1.03), as the initial particle population for the state $x_{c,1}(t)$ of the model (2.1.15).

$$
\begin{align*}
    x_{c,1}(t+1) &= \left(1 + \frac{1}{6 \cdot 6 \cdot x_{c,2}(t) \cdot x_{c,1}^2(t) + p(x_c(t))}\right) x_{c,1}(t) + \omega_1(t) \\
    x_{c,2}(t+1) &= x_{c,2}(t) + \omega_2(t) \\
    [x_{c,1}(280)] &= \begin{bmatrix} E[x_c(280)] \\ E[\alpha] \end{bmatrix} \\
    y(t) &= h^{-1}(x_{c,1}(t)) + v(t)
\end{align*}
$$

(2.1.15)

Preliminary studies about the value of the parameter $\alpha$ may be conducted for nominal load conditions to determine a proper value for the initial condition of the second state $x_{c,2}(t)$. A hazard zone around 0.3 (inches) was defined according to customer specifications. Results obtained by the application of a SIR particle filter for state
estimation, and the implementation of the third approach for long term prediction generation (discussed in subsection 2.1.2.1.3) are shown in Figure 2.1.4-2.

Both the RUL pdf estimate – also referred to as the time-to-failure (TTF) pdf – depicted in Figure 2.1.4-2, and the long term prediction bounds have been computed considering only 40 cycles of data after the detection time instant, a population of 20 particles for the algorithm, and model (2.1.15). Obtained results are excellent in terms of accuracy for both the estimated expected failure time and its 95% confidence interval, offering also a prediction time window of approximately 300 cycles, which is reasonable for corrective actions before the crack turns into a catastrophic failure for the engine.

Figure 2.1.4-2. Prognosis results for crack growth in a turbine engine blade
Even considering that the 95% confidence interval is accurate enough for the purposes of this particular prognosis problem, and that it has been validated using the feature data beyond the 320th cycle of operation, it can be noticed that the predicted upper and lower bounds offer a precise representation for the trend of future feature data for a limited amount of time. In that sense, it is important to note that those bounds are constructed by using the current state estimate for $x_{c,2}(t)$ and that by no means this model parameter is fixed. In fact the value of the parameter depends on the length of the crack and hence, the current estimate is obsolete after a certain period of time, which is exactly what appears depicted in Figure 2.1.4-2. Thus, there should be a compromise between the desired accuracy in the prediction and the prediction window allowed for early prognosis. There is no clear general solution for the optimal size of the prediction window and it greatly depends on the given customer specifications. For the time being, it is important to note that the prediction window must be large enough to allow corrective actions in the system and avoid catastrophic failures.

2.1.5 Case Study: UH-60 Planetary Gear Plate. Analysis of Axial Crack Growth

As it has been previously mentioned, the use of particle filtering as a primary tool for state estimation in nonlinear non-Gaussian processes allows us to manage the uncertainty inherent to the long term prediction problem. As a consequence, not only it is possible to obtain the expectation (mean value of the prediction) for the growth of a failure mode (e.g. a crack in a rotatory piece of equipment) by using features as input data
(e.g. vibration-based features), but also to establish a consistent methodology for the generation of a probability density function (pdf) associated with the prediction.

The former characteristics are especially relevant when dealing with processes and phenomena that are not entirely understood, as the growth of cracks due to fatigue in the material. Consider, for instance, the case of prognosis for the evolution of an axial crack on the plate of the UH-60 planetary gear plate, see Figure 2.1.5-1.

Figure 2.1.5-1. ANSYS model of the planetary gear plate, showing crack location

Although it is well known that the mentioned fault mode can lead to a critical failure condition in the aircraft, there was no certain way to determine its existence save by a detailed inspection of this piece of equipment; procedure which obviously involves large economical costs. Under this scenario, the use of algorithms capable of estimating the RUL by only analyzing vibration-based features becomes extremely attractive and would help to dramatically decrease operational costs.
With the purpose of testing the feasibility and efficiency of such techniques, a seeded fault test was conducted in order to collect fault data under a fixed known loading profile. In this test, the crack was artificially grown until it reached a total length of 1.34”, after which the gearbox was forced to operate emulating load changes that can vary from 20% to 120% in a 3 (min) ground-air-ground (GAG) cycle (see Figure 2.1.5-2). Given the fact that the initial crack length was perfectly known in this case, a deterministic prognosis approach was considered at first to estimate bounds for the failure time instant.

From material structure theory [29], it is well known that the crack growth evolution may be explained by using an empirical model such as the Paris’ equation (2.1.16), given the proper set of coefficients.

\[
\frac{dL}{dn} = C \cdot (U(n) \cdot \Delta K(n))^n
\]

\[(2.1.16)\]
Where \( L \) is the total crack length, \( C \) and \( m \) are material related coefficients, \( n \) is the cycle index, \( U(n) \) is a parameter that models the effect of crack closure during cycle \( n \) and \( \Delta K(n) \) is the crack tip stress variation during the cycle \( n \), measured in \( \text{MN/m}^{3/2} \).

Although simple, model (2.1.16) requires the computation of two critical parameters in order to be used in any prognosis routine: \( \Delta K(n) \) and \( U(n) \).

The stress \( K(n) \) may be estimated for a constant load (usually 100\%) by using finite element analysis (FEA) tools such as ANSYS, for different crack lengths and crack orientation geometries. Considering a proportional relationship between the stress in the tip of the crack and the load percentage, it is in fact possible to construct a mapping relating both the current crack length and load variation per cycle with \( \Delta K(n) \).

Albeit the former piece of information is extremely helpful, it is insufficient to estimate the evolution of the crack length. On one hand, the closure effect parameter \( U(n) \) cannot be efficiently measured and only empirical approximations exist for certain materials, such as Ti-6Al-4V. Even in the case of this particular material, only upper and lower bounds may be computed and therefore it is impossible to compute expectations and/or determine statistically the validity of confidence intervals.

On the other hand, the crack length has to be first estimated in order to come up with an approximate value for \( \Delta K(n) \) and therefore any estimation error will affect tremendously the accuracy of the long term prediction.
Keeping in mind all previously mentioned limitations, and using both the known initial condition for the crack length as starting point and the deterministic model (2.1.16) in a recursively manner, it is still possible to generate coarse estimates for both the upper and lower bounds for the crack growth evolution, see Figure 2.1.5-3. Long term predictions and bounds generated by means of a deterministic model are reasonably good for regular maintenance scheduling, though insufficient for the on-line determination of confidence intervals for on-flight corrective actions.

![Figure 2.1.5-3. Deterministic bounds for crack length evolution vs. GAG cycles](image)

The inclusion of process data measured and pre-processed in an on-line fashion, however, improves tremendously the prospect of what can be achieved in terms of RUL estimation and prognosis, since it provides with a feedback about the health condition of the process under observation. Indeed, the use of features based on the ratio between the fundamental harmonic and the sidebands inside the spectrum of vibration data gives the
basis for the implementation of any of the particle-filtering-based prognosis methodologies introduced in subsection 2.1.2. Consequently, under this new approach, not only it is possible to estimate the expected growth of the crack, but also the unknown closure parameter in the crack growth model (2.1.16) and the RUL pdf, enabling the computation of any statistics such as expectations, confidence intervals, etc.

As any filter-based technique, all proposed prognosis methods require the definition of a process model in order to incorporate the information present in the feature data. Therefore, the following crack growth state model (based on Paris’ equation) has been implemented for purposes of on-line state and model parameter estimation and RUL pdf estimation by using a particle-filtering-based framework for prognosis:

\[
\begin{align*}
L(t + 1) &= L(t) + C \cdot \alpha(t) \cdot \left\{ \left( \Delta K_{inboard}(t) \right)^m + \left( \Delta K_{outboard}(t) \right)^m \right\} + \omega_1(t) \\
\alpha(t + 1) &= \alpha(t) + \omega_2(t) \\
\Delta K_{inboard}(t) &= f_{inboard} \left( \text{Load}(t), L(t) \right) \\
\Delta K_{outboard}(t) &= f_{outboard} \left( \text{Load}(t), L(t) \right) \\
\text{Feature}(t) &= h(L(t)) + \nu(t)
\end{align*}
\]  

(2.1.17)

Where \( L(t) \) is the total crack length estimation at GAG cycle \( t \), \( \alpha(t) \) is an unknown time-varying model parameter to be estimated (unitary initial condition), \( C \) and \( m \) are model constants related to material properties, \( \Delta K \) is the variation in crack tips stress due to the load profile and the current crack length (estimated through off-line analysis of the system with ANSYS) and \( \omega_1(t) \), \( \omega_2(t) \) and \( \nu(t) \) are non-Gaussian white noises.
Process model (2.1.17) necessitates a noisy estimate of the crack length, based on
the value of the feature data point, to be used in on-line applications. This requirement is
satisfied via a nonlinear mapping \( h(\cdot) \), which is corrected or improved according to the
ground truth crack length data that is acquired (at specific and very limited time instants)
from strain gages sensors allocated on the surface of the planetary gear plate.

As a result, in the proposed scheme, two update loops run in parallel. The first
one, referred to as the *inner loop*, basically uses the feature data and the previous state
pdf estimate to update the crack length and model parameters and thus, the RUL pdf
estimate through any of the prognosis approach discussed in subsection 2.1.2. On the
other hand a second loop, namely the *outer loop*, revises the nonlinear mapping
\( h(\cdot) \) between the vibration-based feature value and the crack length every time strain gage
data is received. It is expected, for future on-line applications, that the nonlinear mapping
\( h(\cdot) \) would be still valid, save for minor adjustments.

In this manner, at every GAG cycle, each particle represents a realization of the
state (namely, both the crack length and the unknown model parameter) that is used as an
initial condition for a predicted trajectory, under some reasonable assumptions about the
future operational regime, such as an expected loading profile. The statistical information
contained in all predicted trajectories (particularly the evolution in time of the estimated
state probability density function) is summarized into a RUL pdf via the definition of
hazardous thresholds (a hazard threshold may be understood as a hazard zone with null
variance) for the system under analysis. Indeed, at any given time instant, each particle
from the current particle population determines both an initial condition for a long term prediction and a probability associated with that prediction, see Figures 2.1.5-4 and 2.1.5-5 where each plausible long term prediction is depicted with a different color.

The time instant when each predicted trajectory reaches a given threshold defines a probable failure time and thus, a realization of the RUL probability density function – also referred to as the TTF pdf – for the system under testing (in this particular case the planetary gear plate). The probability associated with this event is the same as the one linked to the particle that was used as initial condition for the corresponding long term prediction. The collection of all these failure times and their probabilities utterly defines the RUL pdf; and once this pdf is estimated, RUL expectations, 95% confidence interval for long term predictions and ±3 sigma intervals may be also computed. Table 2.1.5-1 shows the results for this particular case study, and compares them with the ground truth data that was supplied from strain gages allocated on the surface of the plate.

Table 2.1.5-1. Prediction results for particle-filtering-based approach for prognosis

<table>
<thead>
<tr>
<th>GAG</th>
<th>Total Crack Length (inches)</th>
<th>Confidence Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>−3 σ</td>
</tr>
<tr>
<td>0</td>
<td>1.34</td>
<td>N/A</td>
</tr>
<tr>
<td>36</td>
<td>2.00</td>
<td>0.74</td>
</tr>
<tr>
<td>100</td>
<td>2.50</td>
<td>1.93</td>
</tr>
<tr>
<td>230</td>
<td>3.02</td>
<td>2.73</td>
</tr>
<tr>
<td>400</td>
<td>3.54</td>
<td>3.41</td>
</tr>
<tr>
<td>550</td>
<td>4.07</td>
<td>3.85</td>
</tr>
<tr>
<td>650</td>
<td>4.52</td>
<td>4.20</td>
</tr>
<tr>
<td>714</td>
<td>6.21</td>
<td>5.27</td>
</tr>
<tr>
<td>750</td>
<td>6.78</td>
<td>6.38</td>
</tr>
</tbody>
</table>
Figure 2.1.5-4. Particle-filtering-based approach for prognosis for the study of crack growth in a planetary gear plate
Ground truth data points, i.e., strain gages crack length measurements, shown in Table 2.1.5-1 were provided incrementally up to 650 GAG cycles in a “blind” test format. Thus, for instance, the prediction result of Table 2.1.5-1 for GAG #36 (1.60”) has been obtained at GAG #0 knowing only the initial crack length. Subsequently, the predicted value for GAG #100 (2.40”) has been obtained at GAG #36 after the ground truth data value of 2.00” was used to adjust the nonlinear mapping \( h(\cdot) \). The prediction for GAG #230 was made at GAG #100, and so on so forth.

Every time a new point of ground truth data is included, a more accurate initial condition for the prediction algorithm is estimated, and hence the overall precision of the algorithm is enhanced. The modularity of the proposed approach allows even modifying the set of thresholds considered in the analysis, every time that it is required to increase the hazard level. Compare, for example, the different thresholds that are shown in both Figures 2.1.5-4 and 2.1.5-5.

In order to illustrate this fact more clearly, consider that the prediction algorithm is launched at GAG cycle 100. Crack length thresholds at 3.0”, 3.5” and 4.5” may be established at that time. Given this scenario, the prediction algorithm provides answers to the question: what are the expected (in a probabilistic sense) times at which the crack will reach the corresponding lengths of 3.0”, 3.5” and 4.5”? By estimating the RUL pdf, the algorithm may also supply the RUL expectation (mean time) and the 95% confidence interval for each case. As the crack length evolves in time, however, the hazard thresholds can be easily modified in order to continue the analysis of its growth,
eventually reaching the condition of Figure 2.1.5-5 where only one remaining hazard threshold is of interest (~6.2”) with a TTF expected value of 713 GAG cycles, or equivalently an expected RUL of 325 GAG cycles, which is extremely close to the value of 714 GAG cycles that was provided in the ground truth data set for the failure time.

At this point, it is essential to note that the accuracy of the algorithm has been validated at every step of the “blind” test, confirming the robustness of the approach with respect to changes in the load profile (depicted in Figure 2.1.5-2) and/or the signal to noise ratio of the feature-based noisy crack length estimation (which steadily improved as the crack length increased, see Figure 2.1.5-5). In that sense, it is interesting to note how the estimate of the model parameter is indicative of changes in the testing operational conditions, as it is shown in Figure 2.1.5-6 where the load profile change at the GAG cycle #320 results evident to the naked eye.
Figure 2.1.5-6. Time-varying model parameter vs. GAG cycles

When only one particular threshold is of interest, it is always possible to use historical failure data or customer specifications to construct a hazard zone, as in Figure 2.1.5-7 where an average length of 6.2” has been considered to generate the RUL pdf estimate, and hence, its expectation.

The 95% confidence interval generated under these conditions has been successfully validated (see vertical blue line across the non-Gaussian RUL pdf in Figure 2.1.5-7, representing the ground truth failure data point) during the “blind” test, and thus the prognosis approach in general has proven to meet all necessary requirements to be considered a satisfactory solution for the crack growth problem.
Figure 2.1.5-7. Prediction results for a unique hazard zone at 6.2”

It must be also noticed that the prognosis technique will lose accuracy when analyzing the growth of crack lengths exceeding 6.75” (and thus, when one of the tips of the crack is about to exit from the plate surface), since the FEA model developed for the planetary gear plate does not intend to describe the stresses that affect the tips of the crack under those conditions. Regardless of the latter, the proposed approach has proven to be efficient, accurate and precise to solve the prognosis problem for any crack length within the range of interest.

Finally, it is important to mention that the proposed methodology has been thoroughly compared with an EKF-based approach for long term predictions. Results were always favorable for the proposed particle-filtering-based prognosis scheme in
terms of accuracy and precision of the RUL pdf estimate, just as was already shown in
the results of subsection 2.1.3.

Considering all of the above, it is possible to affirm that the proposed
methodology offers, in this case, a complete and modular solution to the prognosis
problem, which has been tested with excellent results and validated at several stages
within the progression of the seeded fault.

Furthermore, the particle filtering framework for the prediction of the RUL may
be easily implemented in real time on-board a HUMS or other health monitoring platform
for on-line applications; in fact, an integrated architecture that combines vibration data
processing, feature extraction, fault diagnosis and failure prognosis based on this concept
is described in [27].

The excellent quality of the obtained results not only validates the proposed
methodology, but also gives support for the implementation of more sophisticated
techniques such as APF or RPF and noise structure adaptation techniques to improve in
both the on-line state and RUL pdf estimates.

2.1.6 Discussion & Conclusion

This section has introduced an architecture for the development, implementation,
testing and assessment of a particle-filtering-based framework for failure FDI and
prognosis. Several approaches to solve these problems were presented hereby, all of them
based on the fact that the current state pdf estimate may be used to determine the operational condition of the system and/or to predict the progression of a fault indicator, given a dynamic state model and process measurements. In that sense, the task of estimating the current value of the fault indicator, as well as other important changing parameters in the environment, involves two basic steps: the prediction step, based on the process model and an update step, which incorporates the new measurement into the a priori state estimate.

Concerning the general framework for FDI proposed in this section, it must be noticed that it has been successful and very efficient in pinpointing abnormal conditions in a variety of cases (change in growth dynamics, reaching of a pre-determined threshold, etc.). In that sense, given the excellent preliminary results obtained, it would be interesting to analyze the case of several failure modes and how other techniques such as VRPF [34] can help in the solution of the main challenges faced in that type of fault scenarios.

Regarding prognosis, all methods were successfully tested in an illustrative example. Furthermore, it was shown that a prediction method based on a combination of a resampling scheme and Epanechnikov kernels (for pdf approximation in long term predictions) is able to simultaneously reduce the impact of model errors and provide a balanced result in terms of accuracy and precision in the RUL estimates. On the other hand, it was shown that an approach simply based on the expectation of the long term prediction also provides acceptable results and it is suitable for on-line implementation.
In addition, two successful case studies were presented to illustrate the performance of a simple implementation (SIR particle filter and an expectation-based long term prediction generation) with real failure data. Both of them provided an excellent insight about the impact that model inaccuracies and/or customer specifications (hazard zone definition or desired prediction window) may have in the algorithm’s performance.

A particular application of the concept of an *outer correction loop* has also been implemented in the case of the analysis of the crack growth on a planetary gear plate. Specifically, ground truth data has been used to modify a nonlinear mapping $h(\cdot)$ that related the feature value with a noisy estimate of the crack length. The on-line mapping update concept illustrated how the algorithm’s accuracy may be significantly improved by the use of several learning loops — combining model-based (PF) and data driven techniques — working in parallel within the prognosis framework. More work has to be done in the direction of an improved parameter estimation technique, as well in the possibility of modifying the noise structure parameters given the measurement data collected so far. These topics are part of the discussion that follows next.
Section 2.2  Remaining Research

Preliminary research work has been oriented to demonstrate the feasibility of a particle-filtering-based framework for on-line FDI and prognosis. In that sense, most of the effort spent so far has been invested in the development of a solid theoretical foundation, as well as performance assessment on several case studies where, as a general result, the proposed approach has proven to be efficient. There are, though, some particular research topics that will need special attention throughout next research stages, given the significance of the improvements that they may offer in the performance of the overall methodology.

2.2.1  The Inner and Outer Loops for Diagnosis and Prognosis

Given a set of measurements, it is intended for the *inner* and *outer loops* to process the data, capture the available information and implement a learning scheme where the newly acquired knowledge is transformed into a set of corrections for the parameters that characterize the performance of FDI and/or prognostic algorithms.

During the preliminary part of the research, most of the effort has been focused in the implementation of the *inner loop*, both in the case of FDI and prognosis. Actually, the whole particle filtering approach represents, per se, an *inner correction loop* where the state pdf estimate, including unknown time-varying model parameters, is updated with each new piece of observation data.
However, little work has been done with respect to the outer correction loops for FDI and prognosis, except for the correction algorithm used to improve the accuracy of the expectation for the RUL of components (subsection 2.1.2.2), the inclusion of exogenous variables in the prognosis scheme, such as the loading profile in section 2.1.5 (illustrated in Figure 2.2.1-1) and the update of the nonlinear mapping $h(\cdot)$, which related the feature value and the preliminary crack length estimation.

In that sense, and in order to accomplish the objectives stated in this proposal, two different guidelines are proposed for the implementation of data-driven outer corrective loops for the remaining part of the research work.

**Figure 2.2.1-1.** Application example for inner and outer loops in failure prognosis
Regarding the implementation of an outer loop for FDI, research will be focused in the development of a methodology combining RSPF [31] and ASIR [28]. In addition, and only if the number of fault modes considered in the case studies is sufficiently large, a VRPF [34] may be used to group similar fault modes into a single category of anomalies. Since ASIR does not need additional memory to store sigma points and covariance matrices, it is expected that this combined approach will be more suitable than the VUF algorithm [33] (which uses the UPF instead of ASIR, see subsection 1.2.5.1) for the particular case of real-time applications.

With respect to the implementation of an outer correction loop for prognosis, a second and different guideline will be followed. Instead of only adjusting the size of the particle population, feature data collected within a sliding window will be used to update noise parameters inside the dynamic state model. This approach, where “parameters of model parameters” are adjusted, allows dealing with sudden changes in the signal to noise ratio of the feature data, inclusion of new exogenous variables and other time-varying phenomena. The ability of learning and adjusting will translate in an increased accuracy and precision for the RUL pdf estimates and eventually will lead to an improved performance when compared with the preliminary results shown hereby.
Section 2.3  Facilities Needed

2.3.1 Computational Requirements

In terms of computational requirements, most of routines proposed here only need a personal computer in order to be implemented. In particular, it is expected to perform the remaining research work by using an Intel® Pentium® D CPU Processor, 2.80 GHz and 2.0 GB of RAM in order to deal with the large amounts of data that need to be processed in real time.

Algorithm development will be performed entirely in MATLAB® environment, since it provides with a large and complete set of tools that may used for coding (MATLAB® Editor, Statistics Toolbox, and MATLAB® Compiler), testing, debugging and display of obtained results (MATLAB® General User Interface Tools).

Thus, considering all of the above, the list of required software packages for the remaining research work includes:

- Microsoft Windows XP Professional, Service Pack 2
- Microsoft® Office 2003
- MATLAB® R2006a
- Microsoft® Visual C++ 6.0
2.3.2 Failure Ground Truth Data

The need of failure ground truth data, as well as no-fault baseline data, is imperative in order to properly validate the algorithms developed, especially if it is intended to apply this framework in an on-line monitoring system.

Regarding FDI applications, baseline data is required to characterize the no-fault state pdf and to determine the minimum detection threshold for which the required detection confidence is met. On the other hand, prognostic routines need vibration-based data to complete the inner loop, which basically consist in the update of the crack length and model parameter estimates.

In that sense, data that will be provided from a second planetary gear plate seeded fault test will be used to complete the current database for insipient and middle grown crack lengths. Although the plates are obviously not the same and the test conditions may differ with respect to the loading profile or location of the seeded fault, there is confidence in the fact that the vibration-based features used in both the FDI and prognostics algorithms are always normalized (ratio in the frequency domain) and hence, it will be possible to compare results with respect to the first seeded fault test and even complement the nonlinear mapping between the feature value and the preliminary crack length estimate. The latter is essential for the on-line implementation of the proposed algorithms since the fact that the algorithm is capable to detect and prognoses cracks with
comparable measures of accuracy and precision, regardless of the small differences that may exist between two different gearboxes, still has to be validated.

The second seeded fault test for the planetary gear plate is expected to be run in November 2006. Data from strain gages will be only used for validation purposes and it is expected to generate algorithms executables, based on the analysis performed on those data sets, by the first half of 2007.
References


[22] Liu, J. S., “Metropolized independent sampling with comparison to rejection sampling and importance sampling,” *Statistics and Computing*, vol. 6, pp. 113 – 119, 1996.


