Physics-informed Diagnostic and Prognostic Models for Rolling Element Bearings Using Oil Debris Data

By

Hengyangcan (Alex) Gu

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of Master of Applied Science in Aerospace Engineering

Carleton University
Ottawa, Canada

©2023
Hengyangcan Gu
Abstract

Rolling element bearings are susceptible to rolling contact fatigue failure. This poses challenges in industrial applications like wind turbines and aircraft. To address this challenge, the thesis constructs diagnostic and prognostic models due to spalling in the inner raceway. This is achieved by integrating real-time oil debris data and a comprehensive understanding of the underlying bearing degradation mechanisms. The diagnostic model incorporates spall physical information, a heuristic "Kneedle" algorithm and a Random Forest to determine the severity of the spall propagation. For prognostics, the author thoroughly evaluates the effectiveness of the particle filter and its variants. Subsequently, the enhanced version of the Auxiliary Particle Filter with Resample Move is deployed to estimate the remaining useful life. The paramount significance of this developed model lies in its adaptability to bearings of various sizes. This versatility ensures its applicability across various industrial contexts, thus offering an effective solution.
Acknowledgements

First, I want to express my heartfelt gratitude and appreciation for the support and guidance I received from my supervisor Prof. Jie Liu throughout my academic journey. His mentorship and dedication to teaching have been instrumental in shaping my research skills and intellectual growth.

I would also like to express my sincere appreciation to Nicholas Breeuwer and Stephen Swart from Gastops Ltd. Their industry expertise, constructive feedback, and encouragement have not only enriched my research work but also equipped me with practical insights and skills.

In addition, I want to acknowledge the invaluable contributions and cooperation of my research group members: Hassan Mahmoud, Lourd Arun Raj, Shengcai Deng and Xinyi Ma. Working alongside such a talented and committed team has enriched my understanding of the subject matter and made the research process enjoyable.

Most importantly, I would like to extend my deepest gratitude to my parents for their unconditional love, encouragement, and financial support. Their belief in my abilities has been a constant source of motivation, and I am truly grateful for their sacrifices, which have made my education and research pursuits possible.
# Table of Contents

List of Tables \hspace{14.5cm} V

List of Figures \hspace{13.5cm} VI

Acronyms \hspace{13.5cm} VIII

List of Symbols \hspace{13.5cm} IX

1 Bearing Fundamentals \hspace{13.5cm} 1

1.1 Rolling Bearing Geometry \hspace{13.5cm} 1

1.2 Bearing Failure Type \hspace{13.5cm} 2

1.2.1 Fatigue Failure Mode \hspace{13.5cm} 2

1.2.2 Wear Failure Mode \hspace{13.5cm} 2

1.2.3 Corrosion Failure Mode \hspace{13.5cm} 3

1.2.4 Electrical Erosion Failure Mode \hspace{13.5cm} 3

1.2.5 Plastic Deformation Failure Mode \hspace{13.5cm} 4

1.2.6 Fracture and Cracking Failure Mode \hspace{13.5cm} 4

1.3 Condition Monitoring \hspace{13.5cm} 4

1.3.1 Vibration and Noise Monitoring \hspace{13.5cm} 5

1.3.2 Temperature Monitoring \hspace{13.5cm} 5

1.3.3 Lubrication Monitoring \hspace{13.5cm} 6

1.4 PHM Models \hspace{13.5cm} 6

1.4.1 SKF Bearing Life Rating \hspace{13.5cm} 6

1.4.2 Physics-based Model \hspace{13.5cm} 7

1.4.3 Machine Learning \hspace{13.5cm} 7

1.4.4 Data-Model Fusion \hspace{13.5cm} 8

1.5 Thesis Outline \hspace{13.5cm} 9
2 Bearing Spalling and ODM

2.1 Spall Propagation Mechanism .............................................. 10

2.2 In-line Oil Debris Monitor .................................................. 11
   2.2.1 Operation Principles .................................................. 12

2.3 Bearing Spall Progression Test ............................................ 12
   2.3.1 Test Condition and Data Management ................................. 15

2.4 Spalling Behaviour ........................................................... 16
   2.4.1 Total Debris Counts ..................................................... 16
   2.4.2 TDC and Vibration Observation ....................................... 17
   2.4.3 Binned Data Feature ..................................................... 18

3 Spall Estimation Model ......................................................... 21

3.1 Bearing Spall Dynamic Model ............................................... 21

3.2 Model Parameterization Using Ground Truth Data ...................... 23

4 Bearing Diagnosis ............................................................... 25

4.1 Physics-informed Knee Detection .......................................... 25

4.2 "Kneedle" Algorithm .......................................................... 26
   4.2.1 "Kneedle" Algorithm Results And Discussion ....................... 28

4.3 Machine Learning Classifier ................................................ 29
   4.3.1 Feature Selection .......................................................... 29
   4.3.2 Model Selection ............................................................ 30
   4.3.3 Random Forest Classifier Results ..................................... 31

4.4 Algorithm Integration ......................................................... 32

4.5 Hybrid Diagnostic Model Test Results .................................... 33

5 Particle Filter and Its Variants ............................................. 35

5.1 Frequentist vs. Bayesian ..................................................... 35
5.2 Problem Statement ........................................ 35
5.3 Common Problems and Solutions ......................... 38
  5.3.1 Resampling ........................................ 39
  5.3.2 Choosing an appropriate importance density function .... 40
5.4 Evolved Particle Filters .................................. 41
  5.4.1 Particle Filter Family ................................ 41
5.5 Perception by The PHM Community ....................... 43
  5.5.1 Prognostics Framework Using PFs ..................... 43
  5.5.2 Advantages and Limitations .......................... 44
5.6 Benchmarks For Performance Evaluation .................. 45
  5.6.1 Flight Position For State Estimation ................ 45
  5.6.2 Lithium Battery Prognostics ........................ 45
5.7 Sequential Importance Sampling Resampling ............. 46
  5.7.1 SISR Performance and Discussion ................... 47
5.8 Auxiliary Particle Filter ................................ 49
  5.8.1 APF Performance and Discussion .................... 52
5.9 Regularized Particle Filter ................................ 54
  5.9.1 RPF Performance and Discussion .................... 57
5.10 Regularized Auxiliary Particle Filter .................... 59
  5.10.1 RAPF Performance and Discussion ................... 60
5.11 KLD Sampling .......................................... 62
  5.11.1 KLD-sampling Performance and Discussion ........... 65
5.12 Resample-Move .......................................... 68
  5.12.1 Gibbs Sampler - MCMC Move ....................... 69
  5.12.2 Sufficient Statistics ................................ 70
  5.12.3 Resample-Move Performance and Discussion ........... 72
5.13 Stein Particle Filter .................................... 73
5.13.1 Inference Via Deterministic Map . . . . . . . . . . . . . . . . 74
5.13.2 Optimization Via Variational Analysis . . . . . . . . . . . . . 74
5.13.3 Stein Operator . . . . . . . . . . . . . . . . . . . . . . . . . . 74
5.13.4 Kernelized Stein Discrepancy . . . . . . . . . . . . . . . . . . 75
5.13.5 SVGD Performance Superiority . . . . . . . . . . . . . . . . . 75
5.13.6 SVGD in Particle Filter . . . . . . . . . . . . . . . . . . . . . . 76
5.13.7 SPF Performance and Discussion . . . . . . . . . . . . . . . . 78
5.14 APF with Resample-Move . . . . . . . . . . . . . . . . . . . . . . . 79

6 Bearing Prognostics Using ODM Data 81
6.1 Spall Progression Prognostic Model . . . . . . . . . . . . . . . . . . 81
6.2 Prior Knowledge from Generalized Linear Model . . . . . . . . . . . 81
6.3 Results and Discussion . . . . . . . . . . . . . . . . . . . . . . . . . 82

7 Conclusion 86
7.1 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

Appendices 90

A SKF 7205 BECBY OEM Specification 90

Bibliography and references 91

Declaration of Author 100
List of Tables

1  Summary of sensors ...................................................... 14
2  Load Condition for Each Test ........................................ 15
3  Bin size for each ODM sensor .......................................... 16
4  Spall Ground Truth Data. ............................................... 23
5  Latency for Diagnosis. ................................................. 34
6  Particle Filter and Its Variants for PHM .......................... 42
7  Sampling Strategies ...................................................... 42
8  Measurement data for Li-battery degradation ..................... 46
List of Figures

1. Rolling bearing components [6] ........................................ 1
2. Bearing Failure Mode Summary ....................................... 5
3. Spall growth mechanism .................................................. 11
4. ODM Sensor developed by Gastops Ltd. ............................ 11
5. ODM sensor with field coils surrounding sense coil [48] ....... 12
6. ODM sensor output (a) for ferrous debris, (b) for non-ferrous debris [40] .......................................................... 13
7. Test Rig Schematic Drawing ............................................. 13
8. Spalling rate with different load conditions [30] .................. 15
9. Test 5 spall propagation reflected by TDC using sensors with different sizes ......................................................... 17
10. Total debris counts using the 3/8-inch ODM sensor ................ 18
11. Vibration RMS .............................................................. 18
12. Relationship between debris generation and wear process [43] .......................................................... 19
13. Relationship between debris generation and wear process from Test 5 using 3/8-inch ODM ................................................... 19
14. Spall angle in a damaged raceway ..................................... 21
15. Spall length linear correlation with the total debris counts. .... 24
16. "Kneedle" Algorithm Visualization .................................... 27
17. "Kneedle" algorithm intermediate results from Test 5 .......... 29
18. Score for input features .................................................... 30
19. Score by cross-validation .................................................. 31
20. Comparison between random forest prediction and labelled phases .......................................................... 31
21. Prediction noise at the knee-changing points ....................... 32
22. Diagnosis Flowchart ..................................................... 33
23. Diagnosis results .......................................................... 33
24 Hidden Markov model ........................................ 36
25 Particle Filter estimation process. ............................... 36
26 Bayesian methods’ classification ................................. 37
27 Particle degeneracy, the size of the particle represents the weight ............................... 39
28 (a) Optimal importance density (b) Bad importance density proposal .................. 40
29 Flight position benchmark with different process noise using SISR ................... 48
30 RUL prediction and particle impoverishment problem in SISR ....................... 49
31 Flight position benchmark with different process noise using APF .................. 53
32 RUL prediction and particle impoverishment problem in APF ....................... 54
33 Case examples of impoverishment ........................................ 55
34 Kernel functions for kernel density estimation ........................................ 56
35 Flight position benchmark with different process noise using RPF .................. 58
36 RUL prediction and particle impoverishment problem in RPF ....................... 59
37 Flight position benchmark with different process noise using RAPF .................. 61
38 RUL prediction and particle impoverishment problem in RAPF ....................... 62
39 Flight position benchmark using KLD-sampling ........................................ 66
40 Posterior variation during learning in the context of prognosis ....................... 66
41 Li-battery benchmark using KLD-sampling ........................................ 68
42 Resample-move Strategy ........................................ 69
43 RUL prediction and particle diversity in RM ........................................ 73
44 Physical enforcement of each term in SVGD ........................................ 77
45 Flight position benchmark using SPF ........................................ 78
46 Degradation path prediction in (a) SISR (b) APF-RM ............................... 79
47 Test 5 RUL Prognostic Results ........................................ 83
48 Test 6 RUL Prognostic Results ........................................ 84
49 SKF 7205 BECBY Technical Specifications [109] ........................................ 90
Acronyms

**ANN** Artificial Neural Network
**DT** Decision Tree
**EKF** Extended Kalman filter
**EM** Expectation-Maximization
**HMM** Hidden Markov model
**KDE** Kernel Density Estimation
**KF** Kalman filter
**KLD** Kullback-Leibler divergence
**KNN** K-Nearest Neighbors
**MCMC** Monte Carlo Markov Chain
**MH** Metropolis-Hasting
**MISE** Mean Integrated Squared Error
**NB** Naive Bayes
**NDP** Number of Distinct Particles
**ODM** Online Debris Monitor
**OEM** Original Equipment Manufacturer
**OLS** Ordinary Least Squares
**PF** Particle Filter
**PHM** Prognostics and Health Management
**PMCMC** Particle Monte Carlo Markov Chain
**RCF** Rolling Contact Fatigue
**RF** Random Forest
**RKHS** Reproducing Kernel Hilbert Space
**RM** Resample-Move
**RMS** Root Mean Square
RMSE  Root Mean Square Error
RUL  Remaining Useful Life
SMC  Sequential Monte Carlo
SPF  Stein Particle Filter
SVGD  Stein Variational Gradient Descent
SVM  Support Vector Machine
TDC  Total Debris Counts
UKF  Unscented Kalman filter
1 Bearing Fundamentals

A bearing is a mechanical device that supports the rotating components of a machinery or system. It enables smooth rotation by effectively reducing friction between moving parts and distributing loads. The benefits of using bearings include low friction, high efficiency, reduced wear and increased reliability and lifespan of the machinery [1]. Bearings can be in different forms, such as rolling bearings, magnet bearings, fluid bearings and many more. Rolling bearing is the most common type and is widely used in many industries, including the automotive, aerospace and renewable energy sectors. Investigations have shown that damaged bearings are a common cause of machinery failure [2][3]. And bearing failure usually can lead to serious consequences, ranging from reduced efficiency and increased maintenance costs to equipment damage and safety risks [4]. The increasing need for improving bearing prognostic and diagnostic methodologies stems from the desire to minimize engine downtime and enhance the reliability and safe operation of the machinery. It is important to promptly identify bearing failure and predict its Remaining Useful Life (RUL) with proper maintenance, monitoring, and replacement strategies. Prognostics and Health Management (PHM) is an engineering field aiming for high-accuracy health state estimation and future health state prediction [5]. The study of this thesis focuses on developing a diagnostic and predictive model for the rolling bearing using Online Debris Monitor (ODM) data provided by Gastops Ltd.

1.1 Rolling Bearing Geometry

As shown in Figure 1, a rolling bearing consists of four key geometry aspects: inner and outer ring geometry, rolling elements, cage and sealing.

![Rolling bearing components](image)

Figure 1: Rolling bearing components [6]

The outer ring is usually a stationary component mounted on a housing structure. The inner ring is the component that fits tightly around the shaft and rotates with...
it. Both rings come with outer and inner raceways that enclose the rolling elements to provide sliding guidance. The rolling elements move along the raceways, and the geometry of the rolling elements can vary in size, shape, and spacing depending on the bearing’s load capacity and speed requirements. The most common types of rolling elements are spherical balls and cylindrical or tapered rollers. Both rings and rolling element bearings can experience very high stresses from sliding due to concentrated load on a very small contact area [7]. Therefore, they are typically made of steel or other high-strength materials. A bearing cage holds the rolling elements in place and prevents them from rubbing against each other. A cage is not required if the rolling bearing has a maximum number of rolling elements, known as a complement bearing. In this case, all rolling elements are closely spaced and have minimal clearance between them. It is designed to maximize the load-carrying capacity. Bearing sealing is used to protect bearings from contaminants, moisture, and other environmental factors. It also helps retain lubricant within the bearing.

1.2 Bearing Failure Type

SKF classifies the bearing failure modes into six categories based on ISO 15243:2004 [8]. These failure modes include 1) Fatigue, 2) Wear, 3) Corrosion, 4) Electrical Erosion, 5) Plastic Deformation, and 6) Fracture and Cracking.

1.2.1 Fatigue Failure Mode

Fatigue is a common type of failure that can occur in bearings due to periodical stress cycles [9][10][11]. The rolling elements repeatedly pass over the same point on the raceway as they rotate. This creates cyclic stress changes that can eventually lead to the development of cracks at the weakest location under the surface of the raceway. This type of failure is referred to as Subsurface Initiated Fatigue. The other cause is Surface Initiated Fatigue, which is the damage to the rolling contact surface asperities, generally due to inadequate lubrication [8]. Lack of lubrication can lead to metal-to-metal contact in a bearing. The resulting friction can cause the surface asperities to shear over each other. Therefore, micro-cracks can occur at the asperities.

1.2.2 Wear Failure Mode

Wear in bearing consists of Abrasive Wear and Adhesive Wear. Abrasive Wear occurs due to inadequate lubrication or the ingress of dirt particles in the lubricant, causing progressive material removal [12]. Rings and rolling elements are hardened, while cages are not. Therefore, the previous study suggests that the cage in a bearing may be the first component to experience wear when the lubrication condition is
poor [8]. Adhesive Wear occurs when there is metal-to-metal contact or dry contact between the bearing surfaces, causing smearing, a phenomenon of material transfer from one surface to another [13]. This type of wear is also accompanied by frictional heat, producing the local stress concentration [13]. Adhesive wear is often caused by inadequate lubrication or high loads, speeds, or temperatures. Smearing is uncommon under normal operating conditions but can be prominent in severe acceleration [8].

1.2.3 Corrosion Failure Mode

Two types of corrosion can contribute to bearing failure, Moisture Corrosion and Frictional Corrosion. Moisture Corrosion is primarily from the ingress of water and other liquids due to ineffective sealing. Moisture corrosion can cause premature and extended spalling as the material undergoes a structural and surface change [14]. Frictional Corrosion can be further divided into Fretting Corrosion and False Brinelling. Fretting Corrosion occurs when the contact interface between a bearing ring and its seat on a shaft or in a housing under slight relative motion or vibration [15]. Small material debris can become detached from the bearing surface and its seat. This debris can oxidize quickly when exposed to air. Oxidized debris is larger in volume, and the accumulation of debris can lead to uneven load distribution [15]. Fretting corrosion can be secondary damage from heavy spalling of the raceway in a bearing [8]. False brinelling is a type of surface damage due to a combination of repetitive small vibrations and insufficient relative motion [16]. It is characterized by localized wear and roughening of the bearing raceways, typically in the form of shallow depressions or grooves with oxidized debris released when exposed to the air.

1.2.4 Electrical Erosion Failure Mode

Electrical erosion is a type of damage that can occur in bearings subjected to electrical currents. There are two scenarios for this failure mode, Excessive Current Erosion and Current Leakage Erosion. Excessive Current Erosion happens when the bearing experiences high current passing one ring to the other through the rolling elements [17]. The high current can create an electric arc welding effect at the contact area between the rolling element and the ring. The tempering of material can lead to excessive material removal and wear away due to localized heating. Current Leakage Erosion occurs even if the current is comparatively low [17]. The current leakage can lead to surface damage by shallow craters. The damage’s extent for both scenarios depends on current intensity, duration, bearing load, speed and lubricating conditions.
1.2.5 Plastic Deformation Failure Mode

Plastic deformation failure has two sub-modes. One is *Overload Deformation*, and the other one is *Indentations From Debris*. Overload deformation is usually caused by static overloading and shock loads because of improper mounting methods. Incorrect mounting methods can cause plastic deformation in any bearing component. Indentations can come from solid contaminants over-rolled by the rolling elements. The edge of an indentation can be the initiation of a spall [8].

1.2.6 Fracture and Cracking Failure Mode

Fracture and Cracking failure mode leads to the complete end life of a bearing due to the separation of a solid bearing component. There are three causes of fracture, including *Forced Fracture*, *Fatigue Fracture*, and *Thermal Cracking*. Forced fracture is due to overloading, so stress concentration exceeds the material’s tensile strength. Fatigue fracture occurs when the bearing undergoes cyclic bending, exceeding material fatigue strength. And thermal cracking is a consequence of sliding between surfaces. When the sliding is substantial, the heat generated by the motion can lead to cracks.

As demonstrated, bearing failure is a complex phenomenon since many factors can influence it. Figure 2 summarizes the failure modes in bearing and the common failure location. The locations of failure also vary based on the failure mode. All of these bring challenges to bearing prognostics and health management. Except for the last fracture failure mode, the rest of the failure modes are followed by spall propagation as the crack or defect grows. One should note that while the spalling may affect the bearing’s performance and longevity, a bearing can continue functioning with spalling for a limited period before replacement becomes necessary. This is also the main objective of the thesis to provide an RUL prediction for the bearing with the presence of spalling. The next section will provide a detailed discussion of the spalling mechanism and its phenomena.

1.3 Condition Monitoring

When bearings are worn or damaged, they often display specific symptoms that can be observed and identified through condition monitoring. Condition monitoring of bearing can be done online or offline. Online monitoring involves continuously monitoring the performance of the bearing in operation, while offline monitoring involves inspection during a machine shutdown. Indeed, online monitoring provides timely feedback from the system, allowing for early detection of potential issues. This section reviews the important parameters for bearing condition monitoring, including vibration and noise, temperature and debris.
1.3.1 Vibration and Noise Monitoring

Monitoring vibration and noise is the most common method for bearing deterioration and failure analysis [5]. Bearings typically produce a consistent noise and vibration level during regular operation. Trends and changes in the vibration patterns can be identified, indicating the progression of bearing degradation [18]. The industry has much experience using vibration to address bearing failures [19]. Previous studies have shown notable success using vibration statistic features, spectrum analysis and envelope analysis to detect and diagnose various types of faults and RUL prediction [20][21][22]. There have been some public vibration data on bearing condition monitoring, such as the Case Western Reserve University (CWRU) dataset for IEEE PHM 2009 Challenge and FEMTO-ST dataset for IEEE PHM 2012 Challenge.

1.3.2 Temperature Monitoring

Some literature also suggests the importance of temperature monitoring at the bearing position. Blau [23] investigated the effects of temperature on adhesive wear in dry
sliding contacts. The experimental results suggested a "vicious circle" relationship between surface temperature and friction wear. An increase in surface temperature leads to increased friction wear, which, in turn, generates more heat and further elevates the temperature [23]. This continuous cycle worsens the wear process and can result in accelerated damage to the contacting surfaces. A rise in temperature, without any change in operating conditions, can often indicate imminent bearing damage.

1.3.3 Lubrication Monitoring

The lubrication conditions are usually monitored online and offline separately. Online lubrication conditions involve measuring the oil temperature and pressure. Changes in these parameters can indicate problems with lubrication, such as insufficient oil flow or contamination [24]. The lubrication oil is also assessed periodically offline. The assessment includes a detailed examination of the debris, which can provide more insights into the condition of the bearing. The disadvantage is that it is a time-consuming process and cannot provide real-time information on the condition of the bearing. Gastops Ltd. delivers a solution for advanced fluid sensing and analysis. The flagship sensor can quickly and accurately detect and analyze wear debris and contaminants. The information captured by the sensor directly links to the bearing health condition, particularly the severity of spall propagation. This connection will be discussed in the following sections.

1.4 PHM Models

The reliability of the modern system is generally achieved through preventive maintenance, and this prevention-based maintenance usually causes many operational costs [25]. PHM allows a more efficient and cost-effective approach to maintenance and operation by providing real-time monitoring, early fault detection, predictive insights, and even optimized resource allocation. By maximizing equipment uptime, minimizing downtime, and optimizing maintenance activities, PHM contributes to significant cost savings in operations [26]. There have been tremendous studies building PHM models striving for high-accuracy fault detection and prediction. A comparative review of the advantages and limitations of each diagnostic and prognostic method commonly used for the bearing is presented below.

1.4.1 SKF Bearing Life Rating

The SKF Bearing Life Rating is the industry’s most used standard for life prediction. It is a probabilistic model based on a practical calculation. It predicts the number of revolutions or operating hours a bearing can endure before showing evidence of
fatigue [27]. It is based on empirical data from laboratory testing and considers factors such as load, speed, lubrication, and environmental conditions. The most used rating is L10. L10 life represents a statistical threshold, indicating that 90% of the identical bearings would reach the calculated life expectancy, while 10% would fail earlier due to fatigue [27]. It is important to note that the L10 life rating is an approximation rather than an exact prediction. Indeed, the SKF bearing life rating is a valuable tool for engineers to estimate the expected service life of SKF bearings and make informed decisions regarding bearing maintenance practices. However, as aforementioned, L10 only provides statistical evidence with 90% reliability. And bearing life can be extended with the presence of spall, which will be discussed in detail in the next section. The potential life extension has not been addressed in this model.

1.4.2 Physics-based Model

A physics-based model is often considered ideal for PHM applications because these models leverage the fundamental principles and laws of physics to model and predict the behaviour of a system. Physics-based techniques are particularly important in situations where accuracy is a critical factor and testing capabilities are restricted. Developing these models requires extensive domain knowledge and expertise of the underlying physics and mechanisms governing a system. Therefore, the physics-based model usually has good generalization, meaning that it can be applied to different applications without requiring extensive empirical data [28]. In the case of bearing, the deep knowledge of these underlying physics may be incomplete due to the complexity of the degradation mechanisms. This high complexity is the hurdle to developing such an accurate physics-based model with some limited success [29]. There also have been some efforts using a computational tool such as FEA analysis to provide valuable physical insights for bearing prognosis [30].

1.4.3 Machine Learning

Machine Learning, including various Artificial Neural Network (ANN) models, has gained significant popularity in PHM. Machine learning is a purely data-driven method that can learn relevant features directly from high-dimensional data without any requirement for knowledge about the system [31]. They can capture and model nonlinear relationships between input features and the system’s health or remaining useful life, making them suitable for handling complex and dynamic PHM scenarios. These ANN models have demonstrated improved performance in some bearing PHM studies [31][32], offering significant potential. Readers who wish to explore this topic further are encouraged to refer to [33] for more in-depth information and analysis. Nevertheless, one should always note that they also have limitations and usually are
not preferred for the industry. One of the major problems is the limited amount of data, particularly for the application of PHM. In general, the performance of these ANN models relies on massive data. However, collecting sufficient historical data is not feasible for PHM. The first reason is that industries often follow preventive maintenance, replacing a bearing before failure. Run-to-failure data can only be collected from the laboratory. This limits the amount of data production. Second, even some laboratories can produce good-quality data. Companies may be reluctant to share such data due to proprietary concerns. Besides the data availability, the other concern is the generalization problem when applying ANN models for PHM. Lack of generalization means that the learned model cannot perform well on untrained cases [34]. These untrained cases can include bearings with different types and sizes and different operational conditions. The problem can be exacerbated with a limited amount of data. The lack of interpretability is the other reason limiting ANN model applications in the industry. ANN models are considered "black-box," making explaining the underlying decision-making process difficult [35]. It is a long-lasting problem in the ANN community and can hinder transparency and adoption in safety-critical applications such as aviation for the qualification process.

1.4.4 Data-Model Fusion

Complementary to other prognostic approaches, data-model fusion methods improve the measurement-based predictive results through a Bayesian inference framework [36]. The basic idea of data-model fusion is to use data or measurement to update or calibrate the model parameters and, in turn, use models to predict a system’s future behaviour. Monte Carlo Markov Chain (MCMC) [37], Expectation-Maximization (EM) [38] and Sequential Monte Carlo (SMC) (also known as PF) belong to the family of Bayesian inference algorithms, which have been successfully implemented for the prognostic applications. These algorithms are designed to handle stochastic processes, allowing for more accurate estimation with real-world variability. In this thesis, the author investigates PF and its variants, comprehending their impacts and constraints on performance. The aim is to identify the most effective variant for the prognostic application. Further discussion on PF will be provided in Section 5 of the thesis.

It’s important to note that physics-based techniques are not mutually exclusive with data-driven techniques. In many cases, a combination of both data-driven techniques and physics-informed models, known as hybrid modelling, can lead to even more accurate and reliable predictions by leveraging the strengths of each approach. These models rely on historical data, sensor measurements, and statistical analysis to infer the health status and predict the RUL, bypassing the need for a comprehensive understanding of the underlying physics and the lack of interpretability in ANN
models.

1.5 Thesis Outline

The objective of this research is to construct a diagnostic and prognostic framework for the inner raceway spalling in rolling element bearings. This will be achieved by leveraging the ODM data provided by Gastops Ltd. The author implements the physics-informed data-model fusion method by incorporating physical information in the data analysis. The established model provides interpretability to the raw data and helps rationalize the choices made in the development process. The improved version of APF-RM is designed to overcome the limitations of the standard Particle Filter. The developed model can provide more stable estimates and predictions even when the run-to-failure sample size is small. Moreover, this methodology has the potential to be adapted for other rolling bearings in different sizes and shapes. The following includes a detailed breakdown of the thesis.

Section 2 reviews existing empirical studies related to ODM sensors, establishing the research’s theoretical foundation and contextual background. Section 3 presents the spall dynamic model, which helps to infer some physical information from the raw data. Section 4 discusses an online diagnostic model that promptly determines the spalling phases. Section 5 presents the search for improvements in Particle Filter for the prognostic model. After careful assessment and justifications, the author determines that the Auxiliary Particle Filter with Resample-Move algorithm is considered to have the best performance among all variants for prognostic application. Section 6 demonstrates the results achieved through APF-RM for RUL prediction using ODM data. The thesis finishes with a conclusion summarizing the contributions of the research and suggesting future research directions and recommendations.
2 Bearing Spalling and ODM

As discussed in the previous section, SKF identifies various failure modes that can occur in bearings. In recent times, there has been an increasing significance placed on Rolling Contact Fatigue (RCF). This is primarily attributed to the operating conditions of modern machinery, which involve higher speeds, elevated temperatures, and thinner lubricating film thickness [39]. These factors contribute to the growing attention to addressing and understanding the challenges of RCF in order to ensure optimal performance and reliability of machinery. There are two types of damage associated with RCF. One is surface distress, also referred to as micropitting or grey staining, and the other one is spalling from surface pre-damage [39]. This research only focuses on the second failure mode. Bearing spalling refers to the process by which a localized defect on the raceway of a rolling element bearing gradually increases in size and severity over time [40]. As the spall continues to grow, it reaches a point where it significantly affects the geometric tolerances of the bearing. At this stage, the bearing is no longer able to meet its operational requirements due to the compromised condition caused by the spall. During spall progression, the material discharged from the raceway surfaces and released in the lubrication oil. Debris, as a product of spalling, is an excellent indicator of the spall severity. This section provides an overview of the spalling mechanism and its connection with ODM data with some contextual background from the current research progress.

2.1 Spall Propagation Mechanism

As mentioned, spalling is usually dominated by oil-lubricated rolling contact fatigue [39]. It is a degradation phenomenon that occurs on periodic contact surfaces. When a cyclic and highly localized momentary stress makes material fatigue, irregular blocky debris is removed from the material, resulting in spall propagation [41]. Many researches have been conducted to understand the mechanisms of spall propagation. Nagaraj et al. [42] modelled the growth of the spall and found that the material on the leading and trailing edges of the spall tends to yield easily due to concentrated stress induced by the ball elements, as shown in Figure 3.

As the spall widens, the rolling elements can descend into the spall. A change in bearing stiffness is observed as the load supported by the descending balls must be shared by other balls [42]. When the balls try to climb back to the raceway from the spall, a partial impact at the trailing edge results in rising stress, consequently, more severe material fatigue. Thus, the combined effects of local contact stress and impact at the spall trailing edge led to spall propagation along the raceway in the direction of the spall trailing edge [42]. Spalling is a complex process; load, sliding speed, hardness and roughness of the wear surface, lubrication, and so on can also
lead to secondary spalling in nature [43]. Therefore, building a physics-based model to capture all phenomena leaves a significant challenge to the researchers.

2.2 In-line Oil Debris Monitor

The development of debris detection went through three phases in history: 1) offline weighting, 2) offline detection, and 3) in-line monitoring [43]. In the early stage, debris in the lubrication oil was collected and weighed to infer the bearing degradation severity [44]. However, the debris weight is subject to many noises as they may come from multiple sources in the system. Therefore, spectrograph and ferrograph were introduced to the detection methodologies to identify and analyze the debris compositions [45] [46]. Both technologies improve the bearing debris detection accuracy, but the detection had to be performed during overhaul maintenance, so the wear information could not be provided in real-time. Until the late nineties, MetalSCAN, as shown in Figure 4, developed by Gastops Ltd., is an in-line ODM used to detect and measure the size, concentration, and composition of debris in lubricating oil [47].

The monitor is typically installed in line with the oil flow in an engine or other mechanical system and continuously detects debris in real time. It can also charac-
terize debris in both ferrous and non-ferrous [48]. The information reported from ODM can be used to provide early warning of impending system failures, trigger maintenance actions, or optimize oil change intervals.

2.2.1 Operation Principles

The ODM sensor incorporates a magnetic coil assembly, which consists of three coils, two balanced magnetic fields on the outside and a sensing coil in the middle, as shown in Figure 5.

![Figure 5: ODM sensor with field coils surrounding sense coil [48]](image)

Two outer field coils create a quasi-static field within the tube and near the center sense coil, using a high-frequency alternating current source [48]. When the contaminated oil flows through the tube, a specific voltage in the sense coil is induced as a result of debris disrupting the magnetic fields [47].

The amplitude of the inductive voltage is directly proportional to the debris size [47]. This relationship is often expressed in terms of an equivalent spherical diameter of the debris. And the measurement of diameter really depends on the orientation of the debris in the oil. However, most debris due to RCF is blocky and flat with large aspect ratios (∼ 10.1) [43]. Therefore, the assumption of a spherical shape usually leads to an overestimation [47]. Besides the debris size, different material types will cause different phase signatures. Figure 6 shows the opposite phase signatures for ferrous and non-ferrous debris.

2.3 Bearing Spall Progression Test

The spall progression test was carried out at the Gastops research facility. The objective of the experiment is to limit the spalling and capture the nature of spall propagation due to RCF using in-house ODM sensors. The schematic drawing of the test rig is shown in Figure 7.
Figure 6: ODM sensor output (a) for ferrous debris, (b) for non-ferrous debris [40]

Figure 7: Test Rig Schematic Drawing

The test bearing was installed on a shaft that was connected to a 10-kW electric motor. The motor’s torque and speed were carefully regulated and monitored throughout the test. To apply an axial and radial load to the bearing, adjustable pneumatic cylinders are used. Lubrication for the bearing was provided by a pump that delivers jet-graded oil. The scavenged oil passes through the bearing and three
in-line ODM sensors. The intermediate tank combines the scavenged oil with the clean one to maintain the desired flow and pressure levels, helping to overcome pressure losses and ensure efficient operation. Before returning to the main oil tank, the oil was filtered to remove any impurities or debris. Besides the ODM data, the vibration signal is captured on the testing bearing housing and the support structure. Some other sensors are incorporated into the test rig to provide additional information about the system dynamics. The equipped sensors and the measurement description are listed in Table 1.

<table>
<thead>
<tr>
<th>Sensor Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vibration</td>
<td>Testing bearing housing</td>
</tr>
<tr>
<td></td>
<td>Support structure</td>
</tr>
<tr>
<td>ODM</td>
<td>1/4-inch MetalSCAN 4000</td>
</tr>
<tr>
<td></td>
<td>3/8-inch MetalSCAN 4000</td>
</tr>
<tr>
<td></td>
<td>3/4-inch MetalSCAN 4000</td>
</tr>
<tr>
<td>Temperature</td>
<td>Test Lubrication Inlet</td>
</tr>
<tr>
<td></td>
<td>Test Lubrication Outlet</td>
</tr>
<tr>
<td></td>
<td>Test Bearing Outer Race</td>
</tr>
<tr>
<td>Loading</td>
<td>Motor speed</td>
</tr>
<tr>
<td></td>
<td>Axial load</td>
</tr>
<tr>
<td></td>
<td>Radial load</td>
</tr>
<tr>
<td>Flowrate</td>
<td>Lubrication inlet flowrate</td>
</tr>
</tbody>
</table>

The system consists of three ODM sensors with different detection capabilities. The first ODM sensor, with a 1/4-inch diameter, can detect debris in size range from 65 to 350 \(\mu m\) and sense ferrous debris weighing between 1 and 190 \(\mu gm\). The second ODM sensor with a 3/8-inch diameter has a wider detection range from 121 to 700 \(\mu m\) in size and between 11 and 1524 \(\mu gm\) in weight. The last sensor has a diameter of 3/4 inches with a detection range between 250 and 1000 \(\mu m\), and it can sense ferrous debris weighing between 90 and 4400 \(\mu gm\).

Several factors in the test can alter the spalling behaviour. First, the inner race was seeded with multiple indentations using a Rockwell hardness tester to expedite the experiment. These seeded indentations form a cross pattern with a specific size to incubate the spall. However, the manual indentation process introduces variations in the pattern and the spacing among indentations. These indentations can act as a stress concentration, amplifying the stress and causing localized fatigue. Therefore, the initial indentation can significantly affect the spall incubation time and the spall future behaviour in bearings subjected to cyclic loading. The other factor that needs to be considered is the load condition. As aforementioned, RCF is the most frequent failure mode in a bearing if only considering the mechanical failure [30]. Then, after the spall initiation, the main factors affecting the spall progression are the applied
load and the rotational speed [30], as illustrated in Figure 8.

![Figure 8: Spalling rate with different load conditions [30]](image)

Intuitively, an increase in load elevates the stress at the spall edges, and a higher RPM means a higher frequency of loading cycles, increasing the stress and accelerating the degradation. Some other test conditions may also have an effect on the spalling behaviour. For example, the test was conducted sparingly. The pause of the test may affect the ODM reading due to the freeze of lubrication flow. The impulse load that occurred when the motor system was started is caused by the sudden acceleration of the load from rest to its operating speed. This leads to a significant amount of vibrations from the motor. Thus, bearing spalling behaviour exhibits considerable variation, even when subjected to simulated operating conditions.

### 2.3.1 Test Condition and Data Management

There are seven failure tests in total. Three are operated at constant load and speed, and the rest tests have a mix of loading and speed alterations. The first four tests were under a high load, followed by a low load at 6,000 rpm. And the last three tests had a constant load throughout the test at 6,000 rpm. The load change of each case is tabulated in Table 2. As discussed, the load change had a significant impact on the degradation behaviour. However, the author does not comprehensively study

<table>
<thead>
<tr>
<th>Test #</th>
<th>Load 1 (GPa)</th>
<th>Duration (hr)</th>
<th>Load 2 (GPa)</th>
<th>Duration (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.3</td>
<td>13.92</td>
<td>1.9</td>
<td>18.57</td>
</tr>
<tr>
<td>2</td>
<td>2.3</td>
<td>2.96</td>
<td>1.9</td>
<td>8.24</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>13.78</td>
<td>1.9</td>
<td>15.22</td>
</tr>
<tr>
<td>4</td>
<td>2.3</td>
<td>4.2</td>
<td>1.9</td>
<td>9.93</td>
</tr>
<tr>
<td>5</td>
<td>2.3</td>
<td>8.82</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>2.3</td>
<td>9.01</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>2.3</td>
<td>9.85</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
the spalling behaviour with the changing loads due to the delay in the delivery of data. The current work only focuses on the tests at a fixed load condition. To enable a more accurate and complete analysis of the system, more efforts should be put into analyzing the system behaviours with changing loads in future research.

Once the ODM detects debris, the count will be allocated into 16 pre-defined bins based on their size. The size range of each bin is shown in Table 3. It’s worth mentioning that three ODM sensors have different detection ranges, and a larger detection range means a larger bin interval, which may provide a less detailed view of the data distribution.

Table 3: Bin size for each ODM sensor

<table>
<thead>
<tr>
<th>Bin #</th>
<th>1/4-inch MetalSCAN 4000</th>
<th>3/8-inch MetalSCAN 4000</th>
<th>3/4-inch MetalSCAN 4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65</td>
<td>125</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>85</td>
<td>160</td>
<td>300</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>200</td>
<td>350</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
<td>240</td>
<td>400</td>
</tr>
<tr>
<td>5</td>
<td>140</td>
<td>280</td>
<td>450</td>
</tr>
<tr>
<td>6</td>
<td>160</td>
<td>315</td>
<td>500</td>
</tr>
<tr>
<td>7</td>
<td>180</td>
<td>350</td>
<td>550</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>390</td>
<td>600</td>
</tr>
<tr>
<td>9</td>
<td>220</td>
<td>430</td>
<td>650</td>
</tr>
<tr>
<td>10</td>
<td>235</td>
<td>470</td>
<td>700</td>
</tr>
<tr>
<td>11</td>
<td>255</td>
<td>510</td>
<td>750</td>
</tr>
<tr>
<td>12</td>
<td>275</td>
<td>550</td>
<td>800</td>
</tr>
<tr>
<td>13</td>
<td>295</td>
<td>585</td>
<td>850</td>
</tr>
<tr>
<td>14</td>
<td>315</td>
<td>625</td>
<td>900</td>
</tr>
<tr>
<td>15</td>
<td>330</td>
<td>660</td>
<td>950</td>
</tr>
<tr>
<td>16</td>
<td>350</td>
<td>700</td>
<td>1000</td>
</tr>
</tbody>
</table>

2.4 Spalling Behaviour

Bearing debris counts data exhibit certain characteristics to manifest the severity of the bearing spalling. Extracting debris data is crucial for assessing the condition of the bearing and diagnosing abnormalities. In this section, the author would like to discuss these characteristics in terms of Total Debris Counts (TDC) and binned data distribution. The literature results and observations are also compared.

2.4.1 Total Debris Counts

The process of spall propagation at a fixed load condition, as shown in Figure 9, can be characterized into three distinct stages: 1) spall initiation, 2) steady-state propagation and 3) accelerated propagation. The first stage is regarded as the healthy stage. This stage is commonly approximated by the SKF L10 bearing rating life. The debris counts collected during this phase are primarily from the bearing
wear-in, the initial period of operation during which the bearing surfaces undergo a process of mating and conforming to each other [43]. The end of this phase marks the start of spall initiation. Bearing replacement is required in the protocol for some critical applications. And the spall initiation triggers the first inflection point in the TDC, entering the steady-state propagation phase. The inflection point is sometimes referred to as the "knee point" in some literature [49]. A gradual growth rate with increasing total debris counts and vibration characterize this period of spall propagation. The duration of this period varies even among identical bearings with the same operational conditions. Once the second inflection point is detected in the TDC, the spall reaches the final stage of the progression. The spall growth accelerates with a significant increase in the rate of debris release, accompanied by severe vibration.

2.4.2 TDC and Vibration Observation

The TDC data for continuous data are plotted against each other in Figure 10. The figure shows that the degradation trajectories exhibit exponential-like characteristics in all cases. However, disruptions in the trajectories are observed due to the load change. The decrease in the count rate can be observed in those tests with a drop in load. The vibration measurements were resampled at a frequency of 1 Hz, and the Root Mean Square (RMS) vibration values were obtained for each test in Figure 11. The vertical drops observed in the data correspond to instances when the system was either being shut down or started up. The vibration level also dropped immediately after the load was reduced. It is important to note the various strategies
used to acquire vibration data. For the public vibration data mentioned in Section 1.3.1, those laboratories collected vibration data at a very high frequency. A higher sampling frequency allows for capturing more detailed information about the vibration signal. Most importantly, a large amount of data within a small time frame (1s in most cases) can facilitate feature extraction. For this study, the only vibration information provided by Gastops Ltd. is RMS at every second. A low sampling frequency can indeed limit the effectiveness of feature extraction. However, it is also essential to consider the trade-off between the benefits of high-frequency data and the associated data storage and processing requirements.

2.4.3 Binned Data Feature

ODM processes the counts by grouping them into sixteen bins based on the size of the debris. Figure 12 summarizes the relationship between debris generation and the wear process in the previous literature [43]. The figure suggests that the wear rate in the running-in period is primarily from the small debris size and is gradually
Figure 12: Relationship between debris generation and wear process [43]

dominated by the larger debris size in the later stage. However, this relationship may not have practical significance due to the complication in load condition, wear type, wear mechanism, and debris classification. Therefore, characterizing spalling process from binned data is still a significant challenge. Figure 13 presents the observation from debris generation in size and quantity throughout the spalling progression using the ODM sensors. As shown in the figure, debris with smaller sizes dominates in the late damage stage. Debris size distribution tends to be more leptokurtic and skewed. More importantly, the distribution is very similar among all test bearings. Therefore, the ODM binned data exhibits some statistical distribution

Figure 13: Relationship between debris generation and wear process from Test 5 using 3/8-inch ODM.
features identified in the previous study to predict bearing damage levels [40]. These features include the mean mean size, variance, kurtosis, and relative kurtosis listed in equation (1).

\[
Variance = \sum_{j=1}^{N} (d_j - \mathbb{E}(d))^2 P[d_j],
\]

\[
Kurtosis = \sum_{j=1}^{N} (d_j - \mathbb{E}(d))^4 P[d_j],
\]

\[
Skewness = \sum_{j=1}^{N} (d_j - \mathbb{E}(d))^3 P[d_j],
\]

Relative Kurtosis = \frac{Kurtosis}{(Variance)^2},

Relative Skewness = \frac{Skewness}{(Variance)^{3/2}},

where, \mathbb{E}(d) = \sum_{j=1}^{N} d_j P[d_j]

\text{\(\mathbb{E}(d)\) is the mean debris size. \(d_j\) is the average bin size. \(j\) is the number of bins. And \(P[d_j]\) is the number of debris per bin. These figures show that the bin data features seem to be highly nonlinear along the degradation progression due to the complexities of the bearing stiffness and evolving dynamics. The value range of these features varies from case to case. Highly nonlinear features can present challenges in interpreting the PHM model, making it difficult to understand the underlying factors contributing to the diagnosis and prognosis.}
3 Spall Estimation Model

This section introduces a spall estimation model that leverages ODM data. The model considers the spall’s physical characteristics, including its dimension and shape, to approximate its severity. Adopting this model enables a deeper understanding of spalling behaviour, facilitating RUL prediction and real-time health monitoring. To develop this model, the author parameterizes the spall growth model by establishing a correlation between the ODM data and the spall length, enhancing the significance and applicability of the ODM data.

3.1 Bearing Spall Dynamic Model

In previous studies, the spall angle is a common choice for bearing spall damage indicator [40][50]. Spall angle refers to the angle between the surface of a spall and the adjacent surface of a rolling-element bearing, as seen in Figure 14. It is typically measured at the deepest point of the spall and can provide important information about the nature and severity of the spall damage. It grows towards 360 degrees, and a larger spall angle indicates more severe damage. Monitoring spall angle over time can be a valuable tool for predicting bearing failure and scheduling maintenance or replacement. If assuming the spall follows a rectangular prism, the spall length can be an alternative indicator as it is directly related to the spall angle via the diameter of the raceway. Spall length refers to the dimension of a spall in the direction of rotation of a rolling-element bearing, and it is measured as the distance between the edges of the spall. Some research studies have employed ODM data to establish a correlation between TDC or total mass of debris and the spall length [47][50]. To develop such a model, it is essential to understand the behaviour of spall propagation.

The previous study has shown that the spall progression follows two steps. First, it widens axially until it covers the entire width of the raceway [40] [27]. Second, it grows in the circumferential direction [40] [27]. According to Hertzian contact
theory, the depth of maximum shear stress can determine the average depth of the
spall under the contact of the undamaged surface [1]. Yet, the variations in the
spall depth are negligible when compared to the scale of change in the length of the
spall. Thus, the mass loss due to spalls can be expressed using the ODM sensor
measurements as:

\[ M = k \rho V, \] (2)

Where \( k \) is a calibration factor of the ODM to correct the ODM’s over-prediction
tendency [41]. As aforementioned in Section 2.2.1, the ODM presumes the debris is
a sphere rather than a chip. This presumption leads to an overestimation in size.
Although this calibration factor may vary case by case, the author assumes it is a
constant for the same sensor. \( \rho \) is the raceway material density, and \( V \) is the spall’s
volume which can be expressed as:

\[ V = dwL, \] (3)

Where \( d \) is the spall depth, \( w \) is the width of the spall, which is assumed to cover the
entire width of the raceway, \( L \) is the spall length. By combining equation 2 and 3
and isolating the spall length, we can obtain a linear relationship between the mass
loss inferred from the ODM sensor data and the spall length.

\[ L = \frac{M}{kpdw}. \] (4)

The only unknown parameter in this model is the calibration factor, \( k \). As all
variables in the denominator are constants, the entire denominator can be replaced
with the linear coefficient to describe this linear relationship. Since the mass is also
linearly proportional to the total debris counts, the relation between the spall length
and debris counts can be simplified as,

\[ L \propto M, \propto TDC. \] (5)

These assumptions simplify the parameterization of the model because this model
does not depend on any bearing geometries. Thus, this model can be generalized
for all types of bearings. We can obtain the linear coefficient for a specific bearing
as long as the spall length ground truth is collected from the tests. Arguably, this
model may not be one-fit-all. This model has been assessed in the previous research
conducted in the US Air Force Research Laboratories (AFRL). Some test outliers
rendering the results useless are also noted [49]. However, researchers have not given
any explanations for this false estimation so far. But overall, the results in most
tests have shown good accuracy in spall estimation [49]. In the next section, this
model will be parameterized using the ground truth data.

### 3.2 Model Parameterization Using Ground Truth Data

In the previous section, the relationship between TDC and spall length has been elaborated. In this section, the parametric form of a linear relationship is obtained using Ordinary Least Squares (OLS). The ground truth data, listed in Table 4, is the measurement collection of spall dimension at the end of the test. The width remains consistent at approximately 5.9 mm in all tests, while the lengths range from 12 to 19 mm. It’s worth mentioning the challenge of defining the spall length threshold at failure. The spall can spread across the entire race without catastrophic failure at low RPM. This means the bearing can continue operating without immediate and complete failure, even with a large spall covering the entire race. On the other hand, bearings operate under more demanding conditions, including higher rotational speeds and loads. The end life is usually when the spall length is greater than the circumferential ball [51]. However, the exact critical length is uncertain and fully relies on loading conditions.

<table>
<thead>
<tr>
<th>Test #</th>
<th>Length (mm)</th>
<th>Width (mm)</th>
<th>3/4 TDC</th>
<th>1/4 TDC</th>
<th>3/8 TDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.22</td>
<td>5.76</td>
<td>515</td>
<td>4112</td>
<td>2017</td>
</tr>
<tr>
<td>2</td>
<td>17.30</td>
<td>5.89</td>
<td>350</td>
<td>3044</td>
<td>1438</td>
</tr>
<tr>
<td>3</td>
<td>16.01</td>
<td>5.92</td>
<td>417</td>
<td>3573</td>
<td>1715</td>
</tr>
<tr>
<td>4</td>
<td>18.89</td>
<td>5.92</td>
<td>300</td>
<td>4113</td>
<td>2019</td>
</tr>
<tr>
<td>5</td>
<td>12.42</td>
<td>5.85</td>
<td>256</td>
<td>3000</td>
<td>1450</td>
</tr>
<tr>
<td>6</td>
<td>12.83</td>
<td>5.83</td>
<td>329</td>
<td>3001</td>
<td>1414</td>
</tr>
<tr>
<td>7</td>
<td>42.44</td>
<td>7.06</td>
<td>1255</td>
<td>10140</td>
<td>5346</td>
</tr>
</tbody>
</table>

The correlation coefficient can be determined from the TDC \((C)\) and spall length \((L)\). Each sensor provides the TDC data, and it’s unclear which sensor provides the best fitting. The author applies OLS for all sensors. And the regression results with 95% confidence are attached in Figure 15. The correlation coefficient \((R)\) is also obtained and presented in the figures. The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables, ranging from \(-1\) to \(+1\). 1 indicates a strong linear relationship, while values closer to 0 indicate a weaker relationship. The TDC from 1/4" and 3/8" sensors show a strong positive linear relationship. The correlation coefficient of 3/8" ODM data is slightly lower than the 1/4" one. Nevertheless, the 3/8" ODM data is still chosen to correlate the TDC and the spall length based on other test results. These test results are not disclosed due to proprietary reasons.
Figure 15: Spall length linear correlation with the total debris counts.
4 Bearing Diagnosis

The knee points are the time instant that manifest the phase change in spall propagation. Identifying these change points reveals important information about the spall severity. In the previous research work, Raj [52] extracted the bin features demonstrated in the previous section and labelled the data based on eyeballed knee location. An optimized ensemble technique combining K-Nearest Neighbors (KNN) and Random Forest (RF) was trained to classify the phase with the selected features. Although the evaluation metric indicated fairly good performance, the author wants to raise some concerns about this approach. First, data labelling is an important step in supervised machine learning. The previous study classified data based on the eyeballed knee location, introducing errors in data labelling. Second, machine learning requires sufficient data to learn meaningful patterns and make accurate predictions. For this project, the amount of data is limited since degradation data collection can be time-consuming and expensive. Lastly, more data are labelled as the second phase because it usually lasts longer than the first and tertiary phases. This results in a common problem of class imbalance where the number of examples in one or more classes is significantly smaller than in other classes. Dealing with imbalanced class distributions is a challenge because the model is usually biased and inaccurate without proper consideration of this matter. Full reliance on machine learning models is not ideal. In this section, the author proposes a fusion technique, a physics-informed diagnostic model combining the physical degradation information, the knee detection algorithm and machine learning for phase classifications. This fusion model is particularly beneficial in this complex and uncertain scenario where no single source or model can provide a complete and accurate solution.

4.1 Physics-informed Knee Detection

In the previous section, the author determines that the spall length is linearly proportional to the TDC. The TDC data can infer some critical spall length to reinforce knee detection in the region where the knee is most likely to occur. Andrew [51] proposed a limit theory to provide reliable caution and alarm threshold. The limit theory indicates that the last phase usually starts when the spall length is roughly equal to the arc length between two adjacent balls [51]. When the spall length exceeds this limit, the bearing element can be unloaded and result in further increased loads and a small shaft displacement [51]. Then, the spall undergoes rapid growth as a result of elevated stress and the rolling element colliding with the edges of the spall. This process is also accompanied by intensified vibration. Therefore, the previous literature often sets this spall length as the critical failure limit for the bearing. The critical spall length depends on the geometry of the bearing and the
geometric information about the test bearing of SKF 7205 BECBY is attached in Appendix A. The approximated threshold of spall length is calculated as follows:

\[ S_p = \frac{D_p}{2} \times \phi, \]  

where \( S_p \) is the spall length, \( D_p \) is the pitch diameter - the diameter of the center of the rotating balls, and \( \phi \) is the arc length between adjacent balls in radius. The Original Equipment Manufacturer (OEM) specifications do not include the pitch diameter. The author assumes it is the average of the inner bore diameter and the outer diameter, which yields 38.5 mm. The arc length between adjacent balls is given as

\[ \phi = \frac{2 \times \pi}{n}, \]  

where \( n = 13 \) is the total number of balls. Thus, the critical spall length is about 9.3 mm.

Using the developed relationship between the spall length and accumulative counts, the critical accumulative is 1094, with a 95\% confidence interval of [781, 1407]. This is a specific interval in which the occurrence of the second knee is most likely so that the "Kneedle" algorithm will be activated during this interval to detect the possible knee locations.

### 4.2 "Kneedle" Algorithm

The phenomenon of bearing spall propagation in different phases is the increasing debris discharge rate. The "Kneedle" algorithm is employed here to capture the ascending trend in the TDC data. The algorithm was introduced by Satopää et al. [53] using the concept of curvature by measuring the distance between a data point to the diagonal line formed by the points \((x_{min}, y_{min})\) and \((x_{max}, y_{max})\). A knee is defined as a point with the maximum distance in a dataset to the diagonal line [53]. The author uses the toy data to explain how the 'Kneedle' algorithm works and transfers the algorithm to the ODM data for online knee detection. The toy data is generated from the curve \( y = e^{5xx} + 1 \). Knee detection involves three steps depicted in Figure 16.
(a) Find the Euclidean distance from the data point to $y = x$. The label indicates the maximum distance.

(b) Convert the Euclidean distance to the vertical distance by rotation. The label indicates the maximum distance.

(c) The vertical distance is quantified by the difference from $y = x$ to each data point. The knee is the point with the maximum difference.

Figure 16: "Kneedle" Algorithm Visualization
algorithm functions regardless of the magnitude of the given dataset. The normalization step does not change the overall shape of the underlying data. The dataset can be processed as follows:

\[ D_{sn} = (x_{sn}, y_{sn}), \]
\[ x_{sn} = (x_s - \min(x_s)) / (\max(x_s) - \min(x_s)), \]
\[ y_{sn} = (y_s - \min(y_s)) / (\max(y_s) - \min(y_s)). \]

The next step is to find the distance from each data point to the diagonal line, \( y = x \). This can be calculated using Euclidean distances. However, some computational efforts are required to find the coordinates of the intercept points on the diagonal line. The alternative solution for easy computation is to find the vertical distance to the diagonal line. From Figure 16(b), it can be seen that the distances perpendicular to the \( x \)-axis are proportional to the Euclidean distances. This is true because all Euclidean distances are rotated by the same degree and form similar triangles. Since the actual value of the Euclidean distance is irrelevant, the vertical distance can be calculated directly to save computational costs. The vertical distance can be calculated as follows,

\[ D_d = (x_d, y_d), \]
\[ x_d = x_{sn}, \]
\[ y_d = x_{sn} - y_{sn}. \]

The local maxima of the vertical distance are where the knee is located. It indicates the instants of the increasing rate of change in \( y \). A threshold value, \( T_{mx} \), is introduced to allow the user to adjust the sensitivity of the knee detection.

### 4.2.1 "Kneedle" Algorithm Results And Discussion

The algorithm seems simple and effective, but it also comes with some problems. First, it is important to note that this "Kneedle" algorithm is a heuristic algorithm and may not always be meaningful for identifying transitions in a dataset. Second, the latency of knee detection can be a huge concern for real-time processing. For example, the knee occurs at \( t = i \), yet is detected at time \( t = i + n \), where \( n \) is the latency. The study shows that latency depends on the curvature of the dataset and the predefined threshold. Third, since the data is observed in discrete steps, the TDC data follows step-function growth. The algorithm may detect multiple knees representing possible transition points for online applications. And the author does not suggest using a polynomial fit to smooth the data as the prediction may give
different results for the same knee location with an increasing number of receiving data. Therefore, machine learning trained from historical databases is incorporated into the diagnosis to provide more phase change insights. Figure 17 shows the intermediate results from Test 5. The Y-axis is the normalized TDC, and the X-axis is the normalized number of data.

Figure 17: "Kneedle" algorithm intermediate results from Test 5

4.3 Machine Learning Classifier

Machine learning has gained popularity in the PHM community as it can process large amounts of data and identify complex patterns that may not be apparent to humans. Some study has used machine learning classifiers to build diagnostic tool for phase change identification in [52][54].

4.3.1 Feature Selection

Feature selection is deployed before the model selection. The goal of feature selection is to reduce the complexity of feature space and improve its generalization capability by removing irrelevant, redundant, or noisy features. In the preliminary selection, repetitive features are removed. As an example, the author chooses to utilize data solely from the 3/8" ODM sensor instead of data from all sensors. This decision is based on the observation that all the data from different sensors exhibit the same trend and provide redundant information. Additionally, some other features, such as the flow rate and temperature of the lubricant oil, are eliminated as these features have zero or nearly zero variance. In Raj’s work [52], features were selected using a filter-based technique based on the metrics of Information Gain and the Chi-square statistics. In contrast to the filter-based method, the author implements the embedded methods, a technique that performs feature selection as part of the model training process. One popular example of an embedded feature selection method is using RF. RF can measure the importance of each feature based on how much it contributes to the model’s accuracy. Features with higher importance scores are considered more relevant and are retained, while less important features can be
discarded. Figure 18 shows the importance score calculated by the RF model. Since axial load and rate features have the lowest score, both are eliminated for model training.

![Random Forest - Feature Importance](image)

Figure 18: Score for input features.

### 4.3.2 Model Selection

Finding a proper model is crucial as the choice of model can significantly impact the performance and results obtained. There are five types of multiclass classifiers:

1. tree-based models such as Decision Tree (DT) and RF,
2. Naive Bayes (NB),
3. KNN,
4. Support Vector Machine (SVM), and
5. ANN.

The author runs the "Kneedle" algorithm described in Section 4.2 offline and splits the dataset for data labelling. The author does not include ANN for model selection, as discussed, due to the lack of interpretability and a large amount of labelled training data requirement. All models are trained and evaluated using 10-fold cross-validation. 10-fold cross-validation means that each model is trained and evaluated ten times, each time using a different combination of nine folds for training and one fold for evaluation. This approach allows a more robust estimate of the model's performance by reducing the potential bias introduced by using a single train-test split. Repeating the process ten times and averaging the results provides a better representation of the model's performance across different subsets of the data. As shown in Figure 19, RF outperforms the others. Note that DT is equivalent to CART, the acronym for Classification And Regression Trees.
4.3.3 Random Forest Classifier Results

Figure 20 presented below are the results of the fault phase prediction from the two test cases to evaluate the proposed methodology. All evaluation metrics, including precision, recall, and F-score have a weighted average of 0.95.

![Figure 20: Comparison between random forest prediction and labelled phases.]

The machine learning classifier seems to show promising results. Yet, it is worth mentioning that the performance is heavily influenced by the training data. Different training tests may give different results. Also, one should always remember that this technique has some limitations. Most of them have been discussed in the introduction of this section and Section 1.4.3. The other observed issue with this technique is volatile noise at the phase-changing points in some cases, such as Test 7, shown in Figure 21. This phenomenon is also observed in Raj’s study [52]. The model prediction switches between phases back and forth. In some extreme cases, this noisy prediction can last longer.
4.4 Algorithm Integration

The diagnosis model integrates the physical information, the "Kneedle" algorithm, and the RF model as illustrated in the Flowchart in Figure 22. The physical information indicates the period when the knees are most likely to occur. The "Kneedle" algorithm is activated when the debris count is below 150 for the first knee detection or within the range between 781 and 1401 for the second knee detection. The RF model runs in parallel, and the model identifies the first phase change point and ignores the later ones to avoid the prediction noise. Once the "Kneedle" algorithm detects the knee, it cross-checks with the RF model prediction. If both algorithms agree with the results, this diagnostic model determines the spall progression undergoing a phase change. If not, the "Kneedle" algorithm results will have higher precedence than the RF model's, and the last knee detected at the end of the predefined activation periods is regarded as a phase change. Certainly, the major disadvantage of this model is detection latency, which will be evaluated in the next section. The prognosis will be employed after the first knee and prompt the warning once reaching the last phase of degradation.
4.5 Hybrid Diagnostic Model Test Results

The effectiveness of the algorithm is demonstrated using Test 5 and Test 6, shown in Figure 23. The rest of the cases are used to train the RF model.

Figure 23: Diagnosis results.
The red vertical lines indicate the knee points detected by the online "Kneedle" algorithm. The purple vertical lines indicate the phase change identified by the RF model. The green horizontal lines show the activation range between 781 and 1401.

As mentioned before, the biggest concern about this model is detection latency, particularly for the second knee. The second knee’s latency must be small to give operators enough time for action before catastrophic failure. The latency of the knee detection for each case is summarized in Table 5. The latency ranges from the lowest 0.1 hr to the highest 0.73 hr, which is smaller than the average duration of 1.5 hr of the last spall progression phase. The latency is acceptable yet has considerable space to improve for the timely warning.

Table 5: Latency for Diagnosis.

<table>
<thead>
<tr>
<th></th>
<th>1st Knee Latency (hr)</th>
<th>2nd Knee Latency (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 5</td>
<td>0.25</td>
<td>0.68</td>
</tr>
<tr>
<td>Test 6</td>
<td>0.73</td>
<td>0.1</td>
</tr>
</tbody>
</table>
5 Particle Filter and Its Variants

Prognostics is an important aspect of PHM. Prognostics involves assessing the current system's health state and predicting its future behaviour for RUL estimation. The practical issue is that the health state can sometimes be inaccessible or complicated to measure directly under operating conditions [26]. There can be many reasons, such as inaccessible location, harsh environment, or internal monitoring challenges. The general solution is to utilize the secondary measurement that is related to the health state of the system for estimation. However, estimation in the presence of multi-source variability and uncertainty can significantly increase the complexity of the problem [55]. It becomes difficult to accurately model and quantify the uncertainties associated with the estimation process. And the estimation error can accumulate over time if the estimation process does not address the uncertainty. PF is one of the Bayesian paradigms to apply the SMC method, using weighted particles to represent the posterior distribution of hidden system state [56]. This section presents the author's effort in the search for an optimized PF algorithm for PHM applications. First, the author would like to review this state-of-art technique comprehensively.

5.1 Frequentist vs. Bayesian

Some may raise the question of why the Bayesian approach is valuable in PHM even when simpler models suffice. Frequentist and Bayesian approaches are two different philosophies in statistical inference. In frequentist statistics, the emphasis is placed on analyzing data by treating the parameters as fixed, unknown values. It relies heavily on the properties of estimators and tests when a large sample size is available. Bayesian statistics, on the other hand, treats both the data and parameters as random variables. It uses Bayes' theorem to update beliefs about parameters based on both prior knowledge and observed data. Bayesian approach focuses on estimating the posterior distribution of the model parameters rather than producing point estimates and confidence intervals. This approach provides a more complete representation of uncertainty and allows for a richer exploration of possible outcomes. That's why Bayesian is more beneficial when developing prognostic models.

5.2 Problem Statement

The system degradation process is usually regarded as a nonlinear or non-Gaussian online tracking problem. The degradation process can be seen as a state-space model, more precisely, a hidden Markov model (HMM) as illustrated in Figure 24.
The model consists of a state transition function describing the stochastic degradation process and an observation function relating the measurement data \((z)\) with the hidden health state \((x)\). Consider a state-space model

\[
\begin{align*}
x_k &= f(x_{k-1}, u_{k-1}), \\
z_k &= h(x_k, v_k),
\end{align*}
\]  

(10a) (10b)

where \(f(\cdot)\) is a state transition function of the previous state with some process noise \(u_{k-1} \sim i.i.d\), and \(h(\cdot)\) is the observation function, relating the state and measurement with some measurement noise \(v_k \sim i.i.d\).

For a filtering problem, the objective is to obtain the belief, \(p(x_k|z_{1:k})\), in the current state from the given measurements, recursively up to time \(k\) through prediction and update steps as shown in Figure 25.

For the online estimation application, an update of the belief is required whenever a new measurement becomes available. In this case, a recursive filtering approach can be a convenient solution to process data sequentially. To derive the algorithm recursively, assume the posterior, \(p(x_{k-1}|z_{1:k-1})\), from \(t = 1 : k-1\) is available from the previous time instant. The prediction step takes the previous time measurement to obtain the prior of the state at time \(k\).

\[
p(x_k|z_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1})dx_{k-1},
\]

(11)

where the current state probability, \(p(x_k|x_{k-1})\), is obtained by the system model.
function from equation (10a), following the distribution of the process noise. When the new measurement is available, the prior is updated via Bayes’ theorem to obtain the posterior of the current state.

\[
p(x_k|z_{1:k}) = \frac{p(z_k|x_k)p(x_k|z_{1:k-1})}{p(z_k|z_{1:k-1})},
\]

(12)

where the denominator is the evidence which is a normalization constant given by

\[
p(z_k|z_{1:k-1}) = \int p(z_k|x_k)p(x_k|z_{1:k-1})dx_k,
\]

(13)

where \(p(z_k|x_k)\) is the likelihood and \(p(x_k|z_{1:k-1})\) is the prior from the prediction step. The optimal Bayesian estimation can be computed using the recursive propagation from equations (11) and (12). There are several filters available depending on the specific characteristics and requirements of the problem at hand, as illustrated in Figure 26.

Kalman filter (KF) provides a closed-form solution for linear processes, and the uncertainty of KF is restricted to Gaussian distribution [57]. In the case of nonlinear systems, alternative algorithms such as the Extended Kalman filter (EKF), Unscented Kalman filter (UKF), and Particle Filter (PF) can be used [57]. If a system with low nonlinearity and the underlying distribution is assumed to be Gaussian, EKF and UKF are able to address this nonlinearity by approximating the nonlinear functions through linearization or transformation techniques [57]. However, for a highly nonlinear/non-Gaussian system, the computation of the evidence term can be challenging, particularly when it involves integration over a complex and high-dimensional space [58]. PF is a Monte Carlo-based algorithm that approximates the

![Bayesian Inference Diagram](image-url)
posterior by a set of weighted particles, \( \{x^i_k, w^i_k\}_{i=1}^{N_s} \).

\[
p(x_k|z_{1:k}) \approx \sum_{i=1}^{N_s} w^i_k \delta(x_k - x^i_k),
\]

where \( \delta(\cdot) \) is the Dirac function, and the weight is defined to be

\[
w^i_k \propto \frac{p(x^i_{1:k}|z_{1:k})}{q(x^i_{1:k}|z_{1:k})},
\]

The weight assigned to each particle is based on the principle of importance sampling. Importance sampling is a technique to estimate a target distribution by sampling from a proposal importance distribution, \( q(x^i_k|z_{1:k}) \) [59]. The importance density in the denominator can be factorized as

\[
q(x_{1:k}|z_{1:k}) = q(x_k|x_{k-1}, z_{1:k})q(x_{k-1}|z_{1:k-1}).
\]

And the posterior, \( p(x_k|z_{1:k}) \), in the numerator can be simplified as

\[
p(x_{1:k}|z_{1:k}) \propto p(z_k|x_k)p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1}).
\]

By substituting equations (16) and (17) into (15), the weight update equation can be expressed in a recursive propagation form as

\[
w^i_k \propto w^i_{k-1} \frac{p(z_k|x^i_k)p(x^i_k|x^i_{k-1})}{q(x^i_k|x^i_{k-1}, z_k)}.
\]

As \( N_s \rightarrow \infty \), the result from (14) asymptotically approaches the true posterior density.

### 5.3 Common Problems and Solutions

This recursive propagation of particles comes with a common problem - particle degeneracy. As shown in Figure 27, the weights of many particles become small enough to be ignored, and only a few particles with relatively large weights [60] contribute to the estimation.
This phenomenon is accompanied by a reduction in effective particles, which is characterized by effective sample size, \( \hat{N}_{eff} \), a measure of the degeneracy introduced in [61].

\[
\hat{N}_{eff} \approx \frac{1}{\sum_{i=1}^{N_s} (w_i^k)^2},
\]

(19)

where \( w_i^k \) is the normalized weight after each propagation. \( \hat{N}_{eff} \) ranges from 1 to \( N_s \), and a smaller \( \hat{N}_{eff} \) means a more considerable weight variance, indicating a more severe degeneracy. The most direct way to overcome degeneracy is to increase the number of particles, which will inevitably demand more computational effort. Also, the number of particles increases exponentially for higher dimensions to maintain good performance. Therefore, two strategies are generally proposed to alleviate the problem: 1) resampling and 2) choosing an appropriate importance density function.

### 5.3.1 Resampling

The idea of resampling is that since those particles with small weights do not contribute to the estimation, they are no longer needed. New particles are generated to replace those with low weights. The easiest way is to duplicate new particles from the ones with large weights. Computationally, this can be achieved by sampling from an approximated discrete representation of \( p(x_k|z_{1:k}) \) [59]. The weights associated with particles are reset to equal weights after resampling.

However, it is questioned whether it is always correct to duplicate more times from the ones with larger weights. The particles with large weights are statistically duplicated multiple times, which leads to a loss of diversity. The problem is also known
as sample impoverishment [59]. If this problem cannot be properly addressed, all particles may "collapse" to the same point in the state space. Thus, one common adaption is to activate the resampling step when $\hat{N}_{\text{eff}}$ reaches a predefined threshold. Some studies propose to set a fixed fraction, such as $\frac{N_s}{2}$ or $\frac{N_s}{4}$, of the total number of particles as the effective size threshold.

### 5.3.2 Choosing an appropriate importance density function

One of the design aspects of the PF algorithm is the choice of importance distribution. It is an important factor affecting the algorithm’s performance. Figure 28(a) shows the optimal importance density when it is close to the true posterior. This optimal proposal distribution allows most samples to fall into the high-probability region. Figure 28(b) shows an extreme case in which the proposed importance density significantly deviates from the true posterior. An improper proposal can lead to poor estimation because only a few particles are sampled under the posterior with high probability. Increasing the number of particles may mitigate the problem. With a larger number of particles, the filter is more likely to capture the true characteristics of the posterior distribution and prevent the dominance of a few particles. Yet, as mentioned, it is not practical for high-dimension problems because the required particle size can increase substantially to maintain a good performance [58].

![Optional Importance Density](image)

**Figure 28:** (a) Optimal importance density (b) Bad importance density proposal

The HMM indicates that the current state depends on the past trajectory $x_{k-1}$ and the new measurement $z_k$. Therefore, the optimal importance density is the probabil-
ity of the current state $x_k$, given by the past trajectory $x_{k-1}$ and new measurement $z_k$.

$$q(x_k|x_{k-1}, z_k)_{opt} = p(x_k|x_{k-1}, z_k)$$

$$= \frac{p(z_k|x_k, x_{k-1})p(x_k|x_{k-1})}{p(z_k|x_{k-1})}. \quad (20)$$

Substituting into the weight update equation (18) yields

$$w^i_k \propto w^i_{k-1}p(z_k|x_{k-1})$$

$$= w^i_{k-1} \int p(z_k|x_k)p(x_k|x_{k-1})dx_k. \quad (21)$$

This optimal solution is not straightforward to solve for two reasons: 1) the ability to sample from the unknown posterior $p(x_k|x_{k-1}, z_k)$ 2) evaluate the integral over the new state for a non-linear system. For many models, the analytical solution using optimal importance density is impossible. Some suboptimal solution is available for easy computation and will be discussed in the PF variants.

### 5.4 Evolved Particle Filters

A great family of PFs is proposed to address the problems mentioned above. Evolution usually involves improvements in the choice of the importance distribution and the resampling strategies. Having a complete overview of all particle filters can be difficult as many variants do not have a particular name or just a subtle optimization of the existing ones.

#### 5.4.1 Particle Filter Family

Different types of PFs implemented for prognostics are summarized in Tables 6. The variants in *italics* will be discussed in detail and assessed using benchmarks. The author includes the reference of each variant used in the prognostic applications for readers to review.
Besides the PF algorithm itself, the choice of resampling strategy is also an important factor that can significantly impact the performance of a PF. And the most common resampling strategies in the prognostics literature are listed in Table 7. There is a great number of resampling strategies. However, it seems that most studies implement the simplest one, the inverse CDF method. There is a gap in research to assess the performance of different resampling strategies for prognostic applications.

Choosing an optimal PF with different sampling strategies from such diversity can be hard. The author only considers some of the most profound variants in the thesis.
The following sections will thoroughly examine these highlighted variants and assess their performance using benchmarks.

5.5 Perception by The PHM Community

In the context of prognostics, PFs can be used to estimate the RUL thanks to its framework incorporating physical models and data. Each particle contains a vector that represents the system’s current health state, physical model parameters and measurement noise. The trajectories determined by model parameters drive the RUL prediction. The main objective is to obtain a recursive state estimation based on sensory data and prior knowledge about the degradation models. And then draw the particles with extensive space of the unknown parameters \( \theta \) of the physical model as,

\[
p(x_k|z_k, \theta) \Rightarrow p(x_k, \theta|z_k). \tag{22}
\]

The PF algorithm was originally developed as a solution to the problem of state estimation in the field of signal processing and control. It is necessary to devise the PF framework that can estimate the future state or RUL of a system accurately from the current particle population. For the application of prognosis, the algorithm is meant to solve a recursive parameter estimation problem. And the iterative learning process aims to identify and discard parameter combinations that are unlikely or inconsistent with the system’s behaviour. The framework of PF for prognostics is presented in this section.

5.5.1 Prognostics Framework Using PFs

The generic PF framework involves the following steps for the recursive state and model parameter estimation:

1) Initialization: The filter is initialized with a set of particles. Each particle has a vector form with prior knowledge of all unknown parameters, including a possible health state, model parameters, and measurement noise. One thing worth mentioning is that the initialization must cover the effective range.

2) Prediction: Each particle is propagated over a time step based on the system degradation model and the model parameter assigned to each particle to generate a prediction of the system’s health state at the next time instant.

3) Weight Calculation: This step involves calculating the weight associated with each particle based on how well it explains the observed data, using the concept of importance sampling.

4) Resampling: To reduce the particle set’s variance and improve the prediction’s
accuracy, the particles are resampled based on their weights. Particles representing unrealistic or unreasonable degradation systems will be weeded out through resampling. This step can be adaptive, depending on the value of effective sample size $N_{eff}^*$.

6) Repeat: By repeating these steps over time, particle filters can generate a sequence of estimations based on a probabilistic representation of the system’s health state as new measurements become available. This iteration allows the PF to refine the estimate of the system’s state and RUL over time.

7) Prognosis: After learning from all available measurements, the algorithm retains the particles that represent possible trajectories of future system states. These particles can be used to estimate the RUL with a predefined health state threshold. In this step, the health state is progressed without updating the model parameters. And the measurement error is added to the state at each time step to capture the stochastic process. The prognosis step stops when the predicted health states in all particles exceed the threshold. Statistical information such as confidence interval can be obtained from the RUL inferred by each particle.

5.5.2 Advantages and Limitations

PF gains popularity in the PHM community. Before going deeper into the implementation, the author would like to justify the advantages of using particle filters in prognosis.

Advantages:

- It is a recursive Bayesian algorithm suitable for the online estimation problem. The fusion of multiple sources of observation is possible [75].

- Unlike Kalman filtering, it was developed to provide a closed-form solution strictly for a linear/Gaussian system. Particle Filters can deal with nonlinear/non-Gaussian processes [75].

- The algorithm provides an optimal estimate of the state incorporating the stochasticity of the process and the presence of measurement noise [75].

- The algorithm can also infer the model parameters in conjunction with the state estimation [76].

A certain number of drawbacks are also raised in some research papers.

Drawbacks:
• The algorithm is based on the Markov chain. The degradation dynamics and the measurement equation need to be exactly known and accurate, including the model uncertainty [77].

• The future degradation behaviour is assumed to follow the learning phase [36].

• Any errors or approximation in the initial belief can accumulate and propagate over time, distorting the predicted PDF severely [78].

5.6 Benchmarks For Performance Evaluation

To assess and validate the implementation of the PF and its variants, the author does not want to limit it to the context of prognosis. The application of state estimation will also be explored. Therefore, the performance of each PF variant will be assessed using a flight position benchmark model and a case study of lithium battery prognosis.

5.6.1 Flight Position For State Estimation

The following model has been widely used in previous publications to assess the performance of particle filters [56][79].

\[
x_k = \frac{1}{2} x_{k-1} + \frac{25 x_{k-1}}{1 + x_{k-1}^2} + 8 \cos[1.2(k + 1)] + u_{k-1},
\]
\[
z_k = \frac{x_k^2}{20} + v_k, \tag{23}
\]

where \( u_{k-1} \) and \( v_k \) are i.i.d process and measurement noise, respectively. The initial state is set as \( x_0 = 0 \). The variance of measurement noise is set to 1. To test the performance with different levels of uncertainties, the variance of process noise is set to either \( u_{k-1} \sim \mathcal{N}(0, 1) \) or \( u_{k-1} \sim \mathcal{N}(0, 10) \). All particle filters employ 1000 particles by default unless stated. The author uses the Root Mean Square Error (RMSE) metric based on the average over 75 runs to gauge the performance. One thing that should be noted is RMSE may not have a significant meaning for this problem with multiple modes. However, previous studies have frequently utilized it for comparison numerically.

5.6.2 Lithium Battery Prognostics

All PFs use the lithium battery degradation model demonstrated in [80] for assessment. The capacity of the secondary cell over cycles in use quantifies the lithium battery degradation. A threshold of 30% of the rated value is defined as the end life of a battery. The direct observation of the battery capacity is inaccessible. A related capacity, \( C/1 \), can be obtained through the internal battery performance
from electrolyte resistance $R_E$ and the transfer resistance $R_{CT}$. $C/1$ is inversely proportional to the sum of the two resistance, $R_E + R_{CT}$, at a nominal rated current of 1 A [81]. The empirical degradation process is given by an exponential model as follows:

$$x_k = e^{b_k \Delta t} x_{k-1} + \sigma_{k-1},$$

$$z_k = x_k + w_k,$$  \hspace{1cm} (24)

where $k$ is the time step index, $x$ is the damage state, $b$ is the model parameter, $\Delta t$ is the time step. Each particle consists of three unknown parameters, the damage state $x$, the model parameter $b$ and the measurement noise variance $\sigma_w$ from $w_k \sim \mathcal{N}(0, \sigma_w)$. The process noise is not assigned to the particle as it is handled by the uncertainty in the model parameter. The $C/1$ capacity data measured every five weeks ($\Delta t = 5$) is generated based on the true model parameter $b_k = 0.12$ with additive Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.05)$. The generated toy data is given in Table 8.

<table>
<thead>
<tr>
<th>Time Step, k</th>
<th>Initial,0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weeks</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
</tr>
<tr>
<td>$C/1$(Ahr)</td>
<td>1.0000</td>
<td>0.9351</td>
<td>0.8512</td>
<td>0.9028</td>
<td>0.7754</td>
<td>0.7114</td>
<td>0.6830</td>
<td>0.6147</td>
<td>0.5628</td>
<td>0.7090</td>
</tr>
</tbody>
</table>

Particle size is set at 1000, and the prior for the unknown parameters are set to be $x \sim U(0.9, 1.1)$, $b \sim U(0, 0.03)$, and $\sigma_w \sim U(0.01, 0.1)$. Numerous RUL prediction metrics have been developed to measure the accuracy of algorithms. For simplicity, the author uses the difference between the actual RUL and the mean prediction for performance evaluation.

### 5.7 Sequential Importance Sampling Resampling

The SISR can be easily derived from the generic particle filter discussed in Section 5.2. For convenience, SISR uses the prior as the importance density.

$$q(x_k^i|x_{k-1}^i, z_k) = p(x_k^i|x_{k-1}^i).$$  \hspace{1cm} (25)

This is the most common choice as it is intuitive and easy to implement. Furthermore, to avoid the degeneracy problem, SISR applies resampling at every time step, assigning equal weight to the particles. The weight update equation from (18) can be simplified to

$$w_k^i \propto w_{k-1}^i p(z_k|x_k^i) \propto p(z_k|x_k^i).$$  \hspace{1cm} (26)

The main critique of this algorithm is that the proposal importance density ignores the new measurement $z_k$. And the algorithm is very sensitive to outliers [82]. Addi-
tionally, although resampling is applied to reduce degeneracy problems, the samples could suffer a rapid loss of diversity. To tackle these challenges, other PF variants are proposed to enhance the performance. The pseudo-code of SISR is described in Algorithm 1.

Algorithm 1: $[\{x^i_k, w^i_k\}_{i=1}^{N_s}] = \text{SISR } [x^i_{k-1}, w^i_{k-1}, z_k$

- FOR $i = 1 : N_s$
  - Draw $x^i_k \sim p(x_k|x^i_{k-1})$
  - Calculate $w^i_k = p(z_k|x^i_k)$
- END

- Normalize weight: $w^i_k = w^i_k / \sum\{w^i_j\}_{j=1}^{N_s}$
- Resample $\{x^i_k\}$ from $[\{x^i_k, w^i_k\}_{i=1}^{N_s}]$

5.7.1 SISR Performance and Discussion

The performance of SISR is evaluated using the flight position benchmark demonstrated in equation (23) with different magnitudes of noise. The estimated results are presented in Figure 29(a)(b). The RMSE is 3.38 and 5.141 for process noise $u_{k-1} \sim \mathcal{N}(0, 1)$ and $u_{k-1} \sim \mathcal{N}(0, 10)$ respectively. The metric of Number of Distinct Particles (NDP) is used to reflect the problem of particle impoverishment in the filter. As seen in Figure 29(c)(d), at some time instant, the NDP can reach below 100.

(a) Process noise $\mu_{k-1} \sim \mathcal{N}(0, 1)$  
(b) Process noise $\mu_{k-1} \sim \mathcal{N}(0, 10)$
Figure 29: Flight position benchmark with different process noise using SISR.

Figure 30(a) shows the simulation results when the prognosis is activated at the 45th week. For a short period, the SISR can give a reasonable result with an RUL error of 17.4 weeks. However, the major issue is particle impoverishment, as shown in Figure 30(b). The NDP reduces exponentially from the initial set of 1000 to 36. SISR is sometimes considered the generic PF due to its simplicity of implementation. However, the test results reveal some areas for improvement, such as accuracy and particle impoverishment. This is also the main contribution of this thesis in search of optimizations in PF.
5.8 Auxiliary Particle Filter

The APF proposed by Pitt and Shephard is a popular alternative to the SISR algorithm [82]. The algorithm can be regarded as a modification of the sequence of target distributions and the associated proposals in a sensible way. It improves the SISR algorithm by incorporating the preselection of particles from the past trajectories $x_{k-1}$ that are more likely to generate high weights at the current time step $t = k$ when considering the new measurement $z_k$. One can interpret the APF as providing a rough estimation of the optimal importance density that includes the latest observation [83]. As the mathematical interpretation of the APF algorithm differs from the standard SISR algorithm, the author provides a detailed derivation below, which may allow the readers to have some guidance to ensure good performance.

As shown in the problem statement, the generic weight approximation is defined by equation (15). Alternatively, the posterior, $p(x_k|z_{1:k})$, can be factorized using Bayes’ theorem as

$$
p(x_{1:k}|z_{1:k}) = (x_k|x_{k-1}, z_{1:k})p(x_{k-1}|z_{1:k}).
$$

Similarly, we can factorize the importance density in the same way

$$
q(x_{1:k}|z_{1:k}) = q(x_k|x_{k-1}, z_{1:k})q(x_{k-1}|z_{1:k})
$$

$$
\neq q(x_k|x_{k-1}, z_{1:k})q(x_{k-1}|z_{1:k-1}).
$$

The equation (29) is from the original derivation presented in Section 5.2. The second term of the importance density, $q(x_{k-1}|z_{1:k-1})$, only depends on the past
observation. This assumption is made by choice to make efficient recursion for easy computation. However, it prevents the algorithm from making the importance density optimal as $q(x_{k-1}|z_{1:k-1}) \neq p(x_{k-1}|z_{1:k})$. Therefore, the whole idea is to sample from the marginal density, $p(x_{k-1}|z_{1:k})$, such that $q(x_{k-1}|z_{1:k}) \approx p(x_{k-1}|z_{1:k})$, and then work on the rest of the term. This target distribution can be derived by integrating the marginal distribution of the state at time $t = k$.

$$p(x_{k-1}|z_{1:k}) = \int p(x_{k}|z_{1:k})dx_{k}. \quad (30)$$

By substituting the posterior at $t = k - 1$ from equations (14) and (17) into (27), we get

$$p(x_{k-1}|z_{1:k}) \propto \int p(z_{k}|x_{k})p(x_{k}|x_{k-1})p(x_{k-1}|z_{1:k-1})dx_{k}$$

$$= p(x_{k-1}|z_{1:k-1}) \int p(z_{k}|x_{k})p(x_{k}|x_{k-1})dx_{k}$$

$$\approx \left[ \sum_{i=1}^{N_{s}} w_{k-1}^i \delta(x_{k-1} - x_{k-1}^i) \right] \int p(z_{k}|x_{k})p(x_{k}|x_{k-1})dx_{k}$$

$$= \sum_{i=1}^{N_{s}} v_{k-1}^i \delta(x_{k-1} - x_{k-1}^i), \quad (31)$$

where

$$v_{k-1}^i = w_{k-1}^i \int p(z_{k}|x_{k})p(x_{k}|x_{k-1}^i)dx_{k}$$

$$= w_{k-1}^i p(z_{k}|\mu_{k}^i). \quad (32)$$

This integration in equation (32) can be difficult to compute analytically. Therefore, the auxiliary variables, denoted as $\mu_{k}^i$, are generated as a point estimate to represent some characteristics of $p(x_{k}|x_{k-1}^i)$. This characterization could be the mean, mode, median, or sample $\mu_{k}^i \sim p(x_{k}|x_{k-1}^i)$. Then the integration term becomes the likelihood of the new measurement, conditioned on the auxiliary variables at time $t$. This approximation is effective when the process noise is relatively small, and the auxiliary particles can accurately capture the distribution, $p(x_{k}|x_{k-1})$. If the process noise is too large, this single-point approximation cannot characterize $p(x_{k}|x_{k-1})$ [82]. In such scenarios, its performance may be worse than the SISR algorithm.

As aforementioned, we want to approximate $q(x_{0:k-1}|z_{1:k})$ from the particles sampled
from \( p(x_{0:k-1}|z_{1:k}) \). We can relate the importance density to joint density by,

\[
q(x^j_{k-1}|z_{1:k}) = \sum_{i=1}^{N_s} v^i_{k-1} \delta(x^j_{k-1} - x^i_{k-1})
\]

\[
= \sum_{i=1}^{N_s} \frac{v^i_{k-1}}{w^i_{k-1}} w^i_{k-1} \delta(x^j_{k-1} - x^i_{k-1})
\]

\[
= \left( \frac{v^j_{k-1}}{w^j_{k-1}} \right) \sum_{i=1}^{N_s} w^i_{k-1} \delta(x^j_{k-1} - x^i_{k-1})
\]

\[
\approx \frac{v^j_{k-1}}{w^j_{k-1}} p(x^j_{k-1}|z_{1:k-1}).
\]

The superscript "\( i \)" is the particle sets generated from \( p(x_k|z_{1:k}) \) and "\( j \)" is the new particle set resampled from the "\( i \)" particle set to approximate \( q(x^j_{k-1}|z_{1:k}) \). By combining (17), (32), and (33), the weight update equation is given as,

\[
w_k \propto \frac{p(z_k|x_k)p(x_k|x_{k-1})p(x_{k-1}|z_{1:k-1})}{q(x_k|x_{k-1}, z_{1:k})q(x_{k-1}|z_{1:k-1})} = \frac{p(z_k|x_k)p(x_k|x_{k-1})}{q(x_k|x_{k-1}, z_{1:k})} \frac{w_{k-1} p(x_{0:k-1}|z_{1:k-1})}{v_{k-1} p(x_{0:k-1}|z_{1:k-1})} \frac{q(x_k|x_{k-1}, z_{1:k})}{v_k} \frac{w_{k-1} p(z_k|x_k)p(x_k|x_{k-1})}{v_{k-1} q(x_k|x_{k-1}, z_{1:k})}.\]

Furthermore, we can simplify the equation by assuming \( q(x^i_k|x^i_{k-1}, z_{1:k}) = p(x^i_k|x^i_{k-1}) \). This is the second assumption in the APF. Therefore, the APF does not directly apply the definite optimal importance density but rather incorporates a soft constraint to approximate the optimal density. The weight update equation is then

\[
w^j_k \propto \frac{w^j_{k-1}}{v^j_{k-1}} p(z_k|x_k) = \frac{p(z_k|x^j_k)}{p(z_k|\mu^j_k)}.\]

The above derivation can be described by Algorithm 2.

---

**Algorithm 2:** \([\{x^i_k, w^i_k\}_{i=1}^{N_s}] = \text{APF} [x^i_{k-1}, w^i_{k-1}, z_k]\)

- **FOR** \( i = 1 : N_s \)
  - Draw \( \mu^i_k \sim p(x_k|x^i_{k-1}) \)
  - Calculate \( v^i_k \propto p(z_k|\mu^i_k)w^i_{k-1} \)
- **END**

- Normalize weight: \( v^i_k = v^i_k / \sum \{v^i_k\}_{i=1}^{N_s} \)
- Resample \([\{-, -, i^j\}]\) from \([\{x^i_{k-1}, v^i_k\}_{i=1}^{N_s}]\)
Physics-informed Diagnostic and Prognostic Models for Rolling Element Bearings Using Oil Debris Data

- For $j = 1 : N_s$
  - Draw $x^j_k \sim p(x_k|x_{k-1}^j)$
  - Assign $w^i_k = \frac{p(z_k|x^i_k)}{p(z_k|x^j_k)}$

- END

- Normalize weight: $w^i_k = w^i_k / \sum \{w^i_k\}_{i=1}^{N_s}$

5.8.1 APF Performance and Discussion

The APF essentially involves a "looking-ahead" strategy to predict the quality of particles with respect to the most recent observation. Only those with "high-quality" particles will propagate to the next time step. Another interpretation of the APF is that it swaps the sampling and resampling steps. This alternative strategy leads to more diverse particle sets, hence a better estimation. A comparison of strategies between the SISR and APF is shown below.

SISR Standard Strategy:

- Sample $x^i_k \sim p(x_k|x_{k-1}^i)$
- Compute the weight $w^i_k = p(y_k|x^i_k)$
- Resample $\{x^i_k, w^i_k\}$ to obtain a new particle set $\{x^i_k\}$

APF Alternative Strategy:

- Compute the weight $v^i_k = p(y_k|\mu^i_k)$
- Resample $\{x^i_k, v^i_k\}$ to obtain a new particle set $\{x^i_k\}$
- Sample $x^j_k \sim p(x_k|x_{k-1}^j)$

The flight position benchmark test shows significant improvement in accuracy with 2.9061 RMSE for process noise $\mu_{k-1} \sim \mathcal{N}(0, 1)$ and 4.5781 RMSE for process noise $\mu_{k-1} \sim \mathcal{N}(0, 10)$ as shown in Figure 31(a) and (b) respectively. The improvement also shows in diversifying particles. Figure 31(c)(d) plots the NDP in APF against the SISR. In the case of the APF, the diversity of particles is preserved at the initial set of 1000 particles. On the other hand, in SISR, the NDP decreases significantly.
APF enjoys a high degree of particle diversity due to the elimination of the resampling step after particle propagation. However, it should be pointed out that resampling is sometimes still considered a necessary step in the APF to prevent particle degeneracy and ensure accurate estimation of the system state. Resampling is intentionally excluded in this specific implementation to assess the effectiveness of the APF independently. The absence of resampling allows for a focused evaluation of the APF’s performance without relying on traditional resampling techniques and its ability to mitigate the loss of diversity.

Figure 32 is the prognostic results at the 45th week. The estimated RUL with an average of 114.07 weeks with an error of 13.77 weeks. There is still a notable decrease in diversity observed in the distinct particle size, reducing from 1000 to 43 particles. As the variation in the model parameter controls the process noise, the health state in each particle is propagated directly according to its model parameter without the introduction of additional noise.
5.9 Regularized Particle Filter

As aforementioned, albeit resampling can prevent the degeneracy problem, it can introduce the other problem of particle impoverishment. This problem arises when the particles with larger weights result in repeated draws. This happens particularly in cases where the likelihood is highly peaked or multimodal, as presented in Figure 33.
RPF improves the resampling process by adding the regularization steps, drawing from a continuous distribution rather than a discrete one [84]. This continuous distribution can be approximated using the idea of Kernel Density Estimation (KDE) [85]. The regularized posterior is defined as

\[
p(x_k|z_{1:k})_{KDE} = \sum_{i=1}^{N_x} w_k^i K_h(x_k - x_k^i),
\]  

(36)

where

\[
K_h(x) = \frac{1}{h^{n_x}} K\left(\frac{x}{h}\right),
\]  

(37)

\( K_h(\cdot) \) is the scaled Kernel density function, \( h > 0 \) is a smoothing parameter called Kernel bandwidth, and \( n_x \) is the dimension of the state \( x \). An appropriate Kernel density function has to be a symmetric PDF that meets the following properties

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

(38)

The optimal Kernel function and bandwidth \( h \) are chosen to minimize the Mean Integrated Squared Error (MISE) between the true and estimated density.

\[
MISE(\hat{p}) = E\left[ \int [\hat{p}(x_k|z_{1:k}) - p(x_k|z_{1:k})]^2 dx_k \right].
\]  

(39)

There is a wide range of Kernel functions to select. In the case where all particles have the same weight, the optimal Kernel function is Epanechnikov Kernel shown in Figure 34(a) [86].

\[
K_{opt} = \begin{cases} 
\frac{n_x+2}{2cn_x} (1 - ||x||^2) & \text{if } ||x|| < 1, \\
0, & \text{otherwise}
\end{cases}
\]  

(40)

where \( c_{n_x} \) is the volume of the unit hypersphere based on the dimension \( n_x \). The
corresponding optimal bandwidth is

\[ h_{opt} = AN_s^{1/(n_x=4)}, \quad \text{with} \]
\[ A = [8c_{n_x}^{-1}(n_x + 4)(2\sqrt{\pi})^{n_x}]^{1/(n_x+4)}. \]  

(41)

The Gaussian kernel shown in Figure 34(b) is sometimes used as an alternative to reduce the computational cost of generating samples from the Epanechnikov Kernel [87].

\[ K_{subopt} = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}(x)^2\right). \]  

(42)

The associated bandwidth is then

\[ h_{subopt} = AN_s^{1/(n_x=4)}, \quad \text{with} \]
\[ A = [4/(n_x + 2)]^{1/(n_x+4)}. \]  

(43)

The implementation is outlined in Algorithm 3.

Algorithm 3: \([\{x_i^i, w_i^i\}_{i=1}^{N_s}] = \text{RPF } [x_{k-1}^i, w_{k-1}^i, z_k] \]

- FOR \(i = 1 : N_s\)
  - Draw \(x_k^i \sim p(x_k|x_{k-1}^i)\)
  - Calculate \(w_k^i = p(z_k|x_k^i)\)
- END

- Normalize weight: \(w_k^i = w_k^i / \sum \{w_k^i\}_{i=1}^{N_s}\)
- Calculate the empirical covariance matrix \(S_{k-1} = E[(x_k^i - E[x_{k-1}^i])(x_{k-1}^i - E[x_{k-1}^i])^T]\)
• Compute the lower triangular matrix $D_k$ using Cholesky Decomposition such that $D_kD_k^T = S_k$

• Resample $\{x^i_k\}$ from $[\{x^i_k, w^i_k\}_{i=1}^{N_s}]$
  - For $i = 1 : N_s$
    * Draw $\epsilon_i \sim K$ from the Gaussian Kernel
    * Calculate $x^i = x^i_k + h_{opt}D_k\epsilon_i$
  - END FOR

• END IF

5.9.1 RPF Performance and Discussion

As aforementioned, regularization smooths out the weight distribution and improves particle diversity. This optimization adds ergodicity to ensure its long-term behaviour is representative of its underlying distribution. The flight position benchmark test in Figure 35(a)(b) shows an average RMSE of 3.027 for process noise $\mu_{k-1} = \mathcal{N}(0, 1)$ and an average RMSE of 4.31 for process noise $\mu_{k-1} = \mathcal{N}(0, 10)$. The distinct particle size maintains at the initial set of 1000 throughout the entire run.

(a) Process noise $\mu_{k-1} \sim \mathcal{N}(0, 1)$  
(b) Process noise $\mu_{k-1} \sim \mathcal{N}(0, 10)$
The lithium battery test shown in Figure 36(a) gives an average of 115.41 weeks with a deviation of 15.11 weeks. Figure 36(b) showcases the limitation of regularization in the context of model parameter estimation. Regularization adds perturbations to the resampled particles to encourage exploration and diversity only in state space. Yet, the parameter space still experiences the loss of diversity. This problem will be addressed in the optimization step of Resample-Move.

(a) RUL prediction at the 45th week
5.10 Regularized Auxiliary Particle Filter

Jie Liu first proposed a regularized auxiliary particle filtering approach for prognostic estimation [65]. APF shows some advantages over other variants. The auxiliary particles are designed to augment the particle population. However, if the auxiliary particles fail to introduce sufficient diversity, the APF may struggle to accurately estimate the true system state, leading to suboptimal performance. Liu argues that APF has a resampling step to draw past trajectories based on the auxiliary particles. If the likelihood is highly peaked or multimodal, as illustrated in Figure 33, fewer auxiliary particles fall into the high-likelihood region, resulting in repeated draws. As a result, the rationale of the modification is to add the empirical density regularization to the past trajectory, \( x_{k-1} \), in the resampling procedure to diversify the particles. There is one small optimization from the original implementation. Liu used the fixed optimal bandwidth demonstrated in the RPF. However, this optimal kernel bandwidth is only valid for particles with equal weight. The auxiliary particles have different weights, contributing differently to the distribution approximation. Thus, the kernel bandwidth should be adaptively adjusted according to the weight of each particle [88]. The adaptive kernel bandwidth \( h_{ad} \) is calculated as

\[
h_{ad} = h_0 \left( \frac{v_i^k}{\exp \left( \frac{1}{N_s} \sum_{i=1}^{N_s} \log v_i^k \right)} \right)^{-1/2}.
\]

A pseudo-code description is summarized in Algorithm 4.

Algorithm 4: \( \{x_{k}^i, w_{k}^i\}_{i=1}^{N_s} = \text{RAPF}\ [x_{k-1}^i, w_{k-1}^i, z_k] \)
• FOR $i = 1 : N_s$
  - Draw $\mu^i_k \sim p(x_k|x^i_{k-1})$
  - Calculate $v^i_k \propto p(z_k|\mu^i_k)w^i_{k-1}$

• END

• Normalize weight: $v^i_k = v^i_k / \sum\{v^i_k\}_{i=1}^{N_s}$

• Calculate the empirical covariance matrix $S_{k-1} = E[(x^i_{k-1} - E[x^i_{k-1}])(x^i_{k-1} - E[x^i_{k-1}])^T]$}

• Compute the lower triangular matrix $D_k$ using Cholesky Decomposition such that $D_kD_k^T = S_k$

• Resample $\{-,-,i^j\}$ from $\{x^i_{k-1}, w^i_k\}_{i=1}^{N_s}$

• For $j = 1 : N_s$
  - Draw $x^j_k \sim p(x_k|x^j_{k-1})$
  - Draw $\epsilon^i \sim K$ from the Epanechnikov Kernel
  - Calculate the adaptive kernel bandwidth using (44)
  - Calculate $x^*_k = x^i_k + h_{lad}D_k\epsilon^i$
  - Assign $w^i_k = \frac{p(z_k|x^*_k)}{p(z_k|x^i_k)}$

• END

• Normalize weight: $w^i_k = w^i_k / \sum\{w^i_k\}_{i=1}^{N_s}$

---

5.10.1 RAPF Performance and Discussion

The flight position benchmark test shows an improvement in accuracy with an RMSE of 2.871 for process noise $\mu_{k-1} = \mathcal{N}(0,1)$ and an RMSE of 4.65 for process noise $\mu_{k-1} = \mathcal{N}(0,10)$ as shown in Figure 37(a)(b). Observations have indicated that the performance of RAPF can sometimes be inferior to that of SISR with high process noise levels. The justification is that RAPF introduces perturbations twice to every particle throughout the process at each time step. The first perturbation is introduced to the past trajectory $(x_{k-1})$ from the regularization step. And the second perturbation occurs during the propagation step from the process noise. This approach can help maintain diversity in the filter. While introducing perturbations is necessary to capture uncertainties and explore the state space, excessive noise can lead to inaccurate and unreliable estimates.
The performance improvement in prognostics is evident with the RAPF. The algorithm is an extension of APF and RPF. RAPF combines the advantages of both APF and RPF to improve performance by enhancing the weight update process and maintaining diversity in the particle population in state space. However, the problem of impoverishment in the parameter space remains after resampling, resulting in a significant reduction in the NDP from an initial count of 1000 to only 35. The estimated RUL is calculated to be 114.85 weeks, with an error of 14.55 weeks.
5.11 KLD Sampling

In the traditional PF, one of the hyperparameters is the number of particles used for state estimation. Typically, a fixed particle size is employed throughout the entire estimation process. However, the posterior distribution can exhibit significant variations over time. When the target distribution is more statistically dispersed or spread out, a larger number of particles is necessary to represent the distribution effectively. Conversely, if the target distribution is more concentrated or narrow, fewer particles are needed to capture the essential characteristics of the distribution accurately. Fox proposed an adaption of the particle size in the sampling process.
based on Kullback–Leibler divergence (KLD) [89]. The idea is to use the KLD metric to assess the dissimilarity between the current weighted particle distribution and the target distribution [90]. By employing the KLD-sampling, the algorithm adaptively adjusts the number of particles while minimizing the KLD value, ensuring it stays within a predetermined statistical threshold. This improves the accuracy of the particle filter by generating enough samples representative of the target distribution while being computationally efficient.

Suppose a discrete distribution with $k$ different bins represents the posterior $p(x)$. The number of samples in each bin $X = (X_1, X_2, X_3, \ldots, X_k)$ with associated probabilities $p = (p_1, p_2, p_3, \ldots, p_k)$ of each bin that follows a multinomial distribution, $X \sim \text{Multinomial}(n, p)$, such that $\sum p_i = 1$ [91]. Then, the maximum likelihood estimate $\hat{p}$ is given by,

$$\hat{p}(x) = \frac{X}{n}. \quad (45)$$

We can write the KLD between the true distribution $p(x)$ and the estimated distribution $\hat{p}(x)$ in a discrete form as:

$$d_{KL}(\hat{p}||p) = \sum_x \hat{p}(x) \log \left( \frac{\hat{p}(x)}{p(x)} \right). \quad (46)$$

According to Wilks’ theorem[92], when comparing two nested models using the likelihood ratio test, the test statistic denoted as $D$, is asymptotically twice the log-likelihood ratio of the estimated $\hat{p}(x)$ and $p(x)$. If $p(x)$ is indeed the true distribution, then as the sample size becomes very large, the test statistic follows a chi-squared distribution with $k - 1$ degree of freedom, $\chi^2_{k-1}$ [93].

$$D = 2 \log \frac{\hat{p}(x)}{p(x)}$$
$$= 2 \sum_{j=1}^{k} X_j \log \left( \frac{\hat{p}_j}{p_j} \right)$$
$$= 2n \sum_{j=1}^{k} \hat{p}_j \log \left( \frac{\hat{p}_j}{p_j} \right)$$
$$= 2nd_{KL}(\hat{p}||p) \rightarrow \chi^2_{k-1} \text{ as } n \rightarrow \infty. \quad (47)$$

The probability of KLD that is smaller than a given error $\epsilon$ is given as,

$$P(d_{KL}(\hat{p}||p) \leq \epsilon) = P(2nd_{KL}(\hat{p}||p) \leq 2n\epsilon) = P(\chi^2_{k-1} \leq 2n\epsilon). \quad (48)$$
Then, we can get the quantiles of the chi-square distribution,

\[ P(\chi^2_{k-1} \leq \chi^2_{k-1,1-\delta}) = 1 - \delta. \]  

(49)

From equation (48) and (49), the number of samples \( n \) is determined as,

\[ 2n\epsilon = \chi^2_{k-1,1-\delta}, \]
\[ n = \frac{1}{2\epsilon} \chi^2_{k-1,1-\delta}. \]  

(50)

Finding the quantiles of the chi-square distribution can be more challenging compared to the Gaussian distribution, as there isn’t a simple direct lookup table like the Z-scale for the normal distribution. One common approach is using the Wilson-Hilferty transformation [94]. The Wilson-Hilferty transformation is a method to approximate the distribution of a chi-square variable by transforming it to the standard normal (Z-scale) distribution so that the p-values generated by the chi-square variable closely approximate those of the standard normal distribution [94]. Then, the minimum particle size for a good approximation yields

\[ n = \frac{1}{2\epsilon} \chi^2_{k-1,1-\delta} = \frac{k - 1}{2\epsilon} \left\{ 1 - \frac{2}{9(k - 1)} + \sqrt{\frac{2}{9(k - 1)}} z_{1-\delta} \right\}^3, \]  

(51)

where \( z_{1-\delta} \) is the upper \( 1 - \delta \) quantile of the standard normal \( N(0, 1) \) distribution. Based on previous research, it has been reported that setting the value of \( \delta = 0.01 \) or equivalently \( z = 2.326348 \) and \( \epsilon = 0.05 \) yields the optimal performance [89]. Algorithm 5 provides the pseudocode for the KLD-sampling method.

Algorithm 5: \( \{x^i_k\} \) = KLD-Sampling \( [x^i_{k-1}, z_k] \)

- INITIALIZE \( n = 0, k = 0, z, \epsilon, \) bin_list = [ ] and bin size

- WHILE \( n < \frac{1}{2\epsilon} \chi^2_{k-1,1-\delta} \)
  - sample \( x^{(n)}_k \) from \( p(x_k|x_{k-1}) \)
  - find the bin where this sample is allocated
    - IF \( x^{(n)}_k \) falls into a new bin
      - \( k = k + 1 \)
      - update bin_list for the non-empty bins
      - update \( \frac{1}{2\epsilon} \chi^2_{k-1,1-\delta} \) from (51)
      - END
  - END

- END
5.11.1 KLD-sampling Performance and Discussion

The author assesses the KLD-sampling approach using the flight position benchmark with a $\delta$ value of 0.99 and a fixed bin size of 0.1. The state estimation results are compared against the standard SISR method in Figure 39. The evaluation of the KLD-sampling performance can be deliberate. As discussed, increasing the number of particles can enhance the filter’s performance. The author could set a very low particle size for SISR to show the superiority of the KLD-sampling method. However, the author maintains objectivity in the case study and keeps the particle size at 1000 in SISR for performance comparison. Figure 39(b) shows that even with a particle size exceeding 3000 with the KLD-sampling, the improvement in accuracy is limited. The modest performance benefits do not justify the additional computational cost. However, it is noted that in certain applications, particularly robotics, the KLD-sampling approach can be beneficial in improving the accuracy and efficiency of the filtering process, particularly when dealing with complex system state distributions [89]. The conclusion drawn here is that the performance improvement with the KLD-sampling approach can vary depending on the specific problem and dataset being used. It is important to thoroughly evaluate its performance and limitations within the context of the given problem and dataset.
In the context of prognosis, the early iterations may exhibit more dispersed posterior distributions. This phenomenon is more notable when the prior distribution is not informative to constrain the estimates effectively. As the particle filter progresses and learns from more data through the measurement update step, the model estimation is refined by progressively narrowing down the range of plausible parameter values to those that align better with observed data. The remaining particles become more informative to represent the true degradation mode as illustrated in Figure 40.

To showcase the effectiveness of the KLD-sampling approach, the author conducts
the Li-battery case study using a bin size of 0.005. The result, shown in Figure 41, demonstrates that the required particle size decreases through learning. In addition, it is observed that this sampling strategy helps alleviate the particle degeneracy problem to some extent. In the case of the SISR, only 24 NDP remained after 9 learning steps, whereas 75 NDP were retained after the KLD-sampling. The improvement does not directly benefit from the KLD-sampling but from the larger particle size. In other words, increasing particle size in SISR also can lead to the same improvement. The KLD-sampling essentially provides the adaption in the particle size to meet the minimum requirement for effectively representing the posterior distribution. It ensures that the number of particles used strikes an optimal balance between accuracy and computational efficiency. The choice of bin size is a significant hyperparameter in the algorithm, and it can be challenging to determine the most suitable value. A larger bin size is preferred when dealing with a dispersed posterior distribution, which can be beneficial in capturing the overall shape of the distribution without overwhelming computational resources. On the other hand, a smaller bin size should be used when the posterior distribution is narrow, enabling a finer representation of the distribution’s detailed features. Regrettably, the current algorithm does not support adaptive bin sizes.

(a) RUL prediction at the 45th week
5.12 Resample-Move

To recapitulate, RUL prediction can be viewed as a recursive model parameter estimation problem. The objective is to infer the unknown model parameter $\theta$ with a prior $p(\theta)$ based on the sequence of distribution, $p(x_{1:k}, \theta|z_{1:k})$. The particles are drawn from the state space and extended to an extensive model parameter space, $P(x_{1:k}, \theta|z_{1:k})$, which ultimately helps estimate the distribution $p(\theta|z_{1:k})$. PF targeting $p(\theta|z_{1:k})$ provides an asymptotically consistent estimate under this weak assumption [95]. Unfortunately, the standard PF suffers predominant particle impoverishment. The issue can be worse for parameter estimation as successive resampling only samples the particles in the $\theta$ space at time $t = 1$. The lack of ergodicity can propagate the error, making the performance worse over time [95]. Albeit the regularization step introduced earlier helps diversify the particles in the state space, the issue of degeneracy still persists in the parameter space.

An especially appealing approach is the introduction of an MCMC move on the parameters. Berzuini & Gilks first demonstrated this approach named Resample-Move (RM) [96]. As the name indicates, an MCMC move kernel $K_n(x_{1:k}, \theta|z_{1:k})$ of invariant distribution $p(x_{1:k}, \theta|z_{1:k})$ is applied right after the resampling [96]. This strategy can be demonstrated in Figure 42. The particle at $t = k$ is obtained from the PF after resampling. An MCMC move is applied for each particle from time $t = 1 : k$ sequentially.
A fixed number of parameters can be updated to keep the algorithm online. For instance, in the context of RUL prediction, the model parameters can be updated because they play a crucial role in determining the degradation trajectories. In this work, the Gibbs sampler is implemented to update the parameter values such that

$$K_n(x_{1:k}, \theta|z_{1:k}) = \delta_{x_{1:k}} p(x'_{1:k}) p(\theta'|x_{1:k}, z_{1:k}).$$

(52)

It is worth mentioning that if the marginal probability cannot be obtained due to the complexity of the model, the Gibbs sampler can be replaced by the Metropolis-Hasting (MH) sampler, which is more versatile and has more flexibility [97]. Nonetheless, the Gibbs sampler is often preferred over the MH sampler due to its efficiency in accepting new samples. Upon observation of Figure 42, it seems difficult to sample from $p(\theta|x_{1:k}, z_{1:k})$ since it requires storing the paths of $x_{1:k}$. The memory requirements increase as time progresses. However, in many applications, it is possible to find a sufficient statistic, denoted as $T$, from a model and its associated unknown parameters [98]. The sufficient statistic summarizes the essential information from a dataset without losing any information. It allows for reducing the dimension of the data while retaining all the relevant information about the parameter of interest, such that the new parameter can be sampled from $p(\theta|T(x_{1:k}, z_{1:k}))$ [99].

5.12.1 Gibbs Sampler - MCMC Move

Before delving into the derivation, it is essential to explore the relationship between the resample-move algorithm and Gibbs sampling. Gibbs sampling is a commonly used algorithm for sampling from complex multivariate probability distributions. When sampling directly from the joint probability distribution is challenging, the Gibbs sampler offers a solution by reducing the dimensionality and sampling from conditional distributions iteratively [100]. For instance, instead of sampling directly from a joint distribution $p(x, y)$, the sampler reduces the dimension and samples from the conditional distribution $p(x|y)$ and $p(y|x)$. The Gibbs sampling strategy
is shown in Algorithm 6.

Algorithm 6: $[X = \{x_1, x_2, x_3, ..., x_k\}] = \text{Gibbs-Sampler}$

- INITIALIZE $X_0 = \{x_1^0, x_2^0, x_3^0, ..., x_k^0\}$
- FOR iteration $i = 1, 2, 3, ...$
  - sample $x_1^i$ from $p(x_1|x_2^{i-1}, x_3^{i-1}, ..., x_k^{i-1})$
  - sample $x_2^i$ from $p(x_2|x_1^i, x_3^{i-1}, ..., x_k^{i-1})$
  - sample $x_3^i$ from $p(x_3|x_1^i, x_2^i, ..., x_{k-1}^{i-1})$
  - ...  
  - sample $x_k^i$ from $p(x_k|x_1^i, x_2^i, ..., x_{k-1}^i)$
- END

The algorithm is pretty straightforward. The variables are sampled from a sequence of conditional probability. It should be noted that after sampling a new value for $x_i$, its updated value is then used in the subsequent sampling steps for other variables $x_j$. RM essentially is a partial implementation of the Gibbs sampling. After resampling, PF generates samples of the state space, $p(x_1:k|z_1:k, \theta)$. The model parameter is updated by sampling the conditional probability of $p(\theta|x_1:k, z_1:k)$. Working out the conditional probabilities in Gibbs sampling can require some effort, especially in complex models. For instance of a conditional probability of $p(x|y)$,

$$p(x|y) = \frac{p(x, y)}{p(y)},$$

(53)

As equation (53) shows, the conditional probability is the joint distribution $p(x, y)$ up to a constant proportionality. The goal in Gibbs sampling is to remove terms in the joint distribution $p(x, y)$ that does not involve the variable of $x$, which allows us to isolate a functional dependence of the conditional distribution on $x$. By doing so, we can identify a valid probability distribution for the conditional probability of $p(x|y)$ from this functional dependence.

5.12.2 Sufficient Statistics

The other objective is to find a sufficient statistic. Fisher’s factorization theorem is applied to find a characterization of a sufficient statistic [98]. The theorem states that a statistic $T$ is sufficient for a parameter $\theta$ if and only if the likelihood function of the data can be factorized into two functions, one depending only on the data
and the other depending only on the sufficient statistic $T$ and the parameter $\theta$ [98]. That is,

$$f_\theta(x) = g_\theta(T(x))h(x),$$  \hspace{1cm} (54)

where $f_\theta(x)$ is the likelihood function, $x$ is the data, $\theta$ is the parameter of interest, and $T(x)$ is the sufficient statistic. The author would like to use the lithium battery model from equation (24) to demonstrate the theory. The conditional probability of the model parameter $b$ is given by,

$$p(b|x_{1:k}, z_{1:k}) \propto p(x_{1:k}, z_{1:k}|b)p(b).$$  \hspace{1cm} (55)

Since the prior $p(b)$ is a uniform distribution not involving any term associated with $b$, this term can be eliminated. Then, the conditional probability is essentially the likelihood of $x$ and $y$ from $t = 1 : k$ given $b$. To derive the conditional probability, we need to remove all terms that are not involved in model parameter $b$. Meanwhile, it is feasible to identify the sufficient statistics.

$$p(b|x_{1:k}, z_{1:k}) \propto \left(\frac{1}{w_k \sqrt{2\pi}}\right)^k \prod_{n=1}^{k} \exp\left(-\frac{(x_n^i - z_n^i)^2}{2w_k^2}\right) \propto \exp\left(-\sum_{n=2}^{k} \frac{(\exp(-b\Delta t)x_{n-1}^i - x_n^i)^2}{2w_k^2}\right) \propto \exp\left(-\sum_{n=2}^{k} \frac{(\exp(-b\Delta t))2x_{n-1}^i - 2\exp(-b\Delta t)x_n^i + x_n^i + x_{n-1}^i}{2w_k^2}\right)$$

Let $A = \sum_{n=2}^{k} \exp(-b\Delta t) = (k - 1)\exp(-b\Delta t)$

$$\propto \exp\left(-\frac{A^2 - 2A \sum_{n=2}^{k} \frac{x_n^i}{x_{n-1}^i}}{2w_k^2}\right) \propto \exp\left(-\frac{(A - \sum_{n=2}^{k} \frac{x_n^i}{x_{n-1}^i})^2}{2w_k^2}\right)$$

$$A \sim \mathcal{N}\left(\sum_{n=2}^{k} \frac{x_n^i}{x_{n-1}^i}, w_k^2\right)$$

$$b = -\frac{\log(A/(k - 1))}{\Delta t}$$

After converting to quadratic form, the expression resembles the normal distribution. And $\sum_{n=2}^{k} \frac{x_n}{x_{n-1}}$ is the sufficient statistic of $A$ that involves the model parameter of $b$. Instead of storing the entire path of $x$ from $t = 1 : k$, we only need to update the sufficient statistic $\sum_{n=2}^{k} \frac{x_n}{x_{n-1}}$ at each time step.

This is a very elegant algorithm. Nonetheless, as the algorithm does not search the full parameter space and implicitly relies on the approximation of a sequence
of distributions $p(x_{1:n}|z_{1:n}, \theta)$, the particle degeneracy problem creeps in for a considerable period. This drawback has been addressed in [101] [102]. However, the previous study shows that the particle filter with resample-move can perform reasonably well for a fixed time horizon [103]. This algorithm can extend the complete MCMC step to explore the entire parameter space, called Particle Monte Carlo Markov Chain (PMCMC) [104]. This algorithm can fully rejuvenate the particles at subsequent time steps. The only concern is how to keep the algorithm online due to the expensive computational cost.

5.12.3 Resample-Move Performance and Discussion

It’s worth mentioning that the RM algorithm also can be used to diversify the particles in the state space for the state estimation problem. However, the alternative regularization step has already addressed the problem, so the author only presents its efficacy for the model parameter estimation problem. The attached Figure 43(a) shows the RUL prediction results using the Li-battery data.

(a) RUL prediction at the 45th week
As seen in Figure 43(b), although the diversity in state space is affected after resampling, the perturbation in the model parameter allows the rejuvenation of particles in the state space at the next time step. This can effectively help in exploring and refining the estimates. Unlike RPF can only mitigate the impoverishment in the state space, the RM can diversify particles in both state and model parameter space.

5.13 Stein Particle Filter

Stein Particle Filter (SPF) is based on the Stein Variational Gradient Descent (SVGD) framework, which uses a different means for Bayesian Inference. The SVGD algorithm was proposed by Qiang Liu [105]. Unlike Monte Carlo methods relying on random sampling to approximate the distribution, SVGD is a gradient-based optimization method that iteratively updates a set of particles to estimate the target distribution. It has a deterministic form of functional gradient descent to allow particles to move toward the target distribution. By doing so, the algorithm minimizes the KLD between the distribution represented by the particles and the target distribution. SPF incorporates the SVGD algorithm, merging the standard two-step paradigm of the generic PF into one step. Different from calculating the weight associated with each particle, all particles will have equal weight in the SPF algorithm. The name, Stein Variational Gradient Descent, implies this algorithm involves the ideas of the "Stein method", "Variational Inference", and "Gradient Descent". There are also many more, such as the basics of measure theory and discrepancy measures between distributions. This section reviews the mathematics
and methodologies to help understand the principle behind this powerful tool. Instead of showing the detailed derivation in the original paper, the thesis provides a less verbose description to make it easier to understand. Unfortunately, the author fails to extend the algorithm for the prognostic application and will discuss the underlying obstacles. Further research and efforts are required when trying to fit more complicated models.

5.13.1 Inference Via Deterministic Map

SVGD is another inference paradigm via the deterministic map $T(x)$. The strategy is to update the particles iteratively according to the map $T(x)$ at discrete times.

$$x \rightarrow T(x) = x + \epsilon \phi(x),$$

where $\epsilon$ is the step size and $\phi$ is a vector field. Each particle has its own deterministic map at the current position. Infinitesimally, the map pushes the particle a small distance along the direction dictated by the vector field. And eventually, all particles will move closer to the target distribution in a higher probability region.

5.13.2 Optimization Via Variational Analysis

The question left from (57) is to find the vector field $\phi(x)$. The original paper reformulates the problem as an optimization problem in which the vector field at each time step will maximize the rate of decrease of the KLD. Then, the variational problem can be characterized as,

$$\phi^* \in \arg \max_{\phi \in Q} \{-\partial_t KL(T_{x \sim q} \| p(x))_{\epsilon = 0}\}.$$  

This formula tries to find a vector field, $\phi$, in a set of sufficient functions, $Q$, to be chosen, such that the rate of decrease of KLD between the distribution $q$, represented by the particles pushed forward under the transformation map $T(x)$ and the target distribution $p(x)$ is maximized.

5.13.3 Stein Operator

It turns out the infinitesimal rate of change of the KLD has a closed form up to expectation.

$$-\partial_t KL(T_{x \sim q} \| p(x))_{\epsilon = 0} = \mathbb{E}_{x \sim q}[\mathbb{S}_p \phi(x)].$$
The rate of change is the expected value for the current distribution, $q$, of an operator $S_p\phi(x)$. $S_p\phi(x)$ is the Stein operator defined by,

$$S_p\phi(x) = \nabla_x \log p(x) \phi(x) + \nabla_x \phi(x). \quad (60)$$

Therefore, the variational problem becomes,

$$\phi^* \in \arg\max_{\phi \in \mathcal{Q}} \mathbb{E}_{x \sim q}[S_p\phi(x)]. \quad (61)$$

The maximum of $\mathbb{E}_{x \sim q}[S_p\phi(x)]$ is also known as Stein discrepancy introduced by Charles Stein in the 1980s [106]. It measures the difference between the expected value of the empirical distribution, $q$, and the expected value of the theoretical distribution, $p$. Now, all left is to solve this variational problem.

### 5.13.4 Kernelized Stein Discrepancy

The approach in the original paper to solving the optimization problem is to exploit the kernel trick. The kernel trick implicitly maps particles from a random data space into a higher-dimensional feature space. The function set, $\mathcal{Q}$, is chosen to be the unit ball of a d-dimensional vector-valued Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k^d$ with kernel $k_{d \times d}$.

$$\mathcal{Q} = \{\phi \in \mathcal{H}_k^d : ||\phi|| \leq 1\}. \quad (62)$$

Then, substitute (62) into (60) and (61), and the variational problem can be solved exactly to give,

$$\phi^*(\cdot) = \mathbb{E}_{x \sim q}[\nabla_x \log(p(x))k(x, \cdot) + \nabla_x k(x, \cdot)],$$

$$= \frac{1}{n} \sum_{j=1}^{n} [\nabla_x \log(p(x))k(x, x') + \nabla_x k(x, x')]. \quad (63)$$

### 5.13.5 SVGD Performance Superiority

Incorporating with the mapping equation (57), the recursive update form can be written as,

$$x^i_{n+1} = x^i_n + \frac{\epsilon}{N} (\nabla_x \log(p(x))k(x, x') + \nabla_x k(x, x')) \quad (64)$$

Instead of inferring through random samples, the algorithm has a deterministic form to update the particles, which allows consistent and reproducible results.

Let us reiterate the problem setup for the Bayesian inference. The difficulty in solving the Bayesian equation is due to the evidence term in the denominator, a
normalized constant involving integration for nonlinear equations. SVGD is an attractive algorithm for inference because it does not depend on the evidence due to the gradient operator. It can be approved mathematically as,

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)},$$

$$\nabla \log(p(x|z)) = \nabla \log\left(\frac{p(z|x)p(x)}{p(z)}\right),$$

$$= \nabla \left(\log[p(z|x)p(x)] - \log(p(z))\right),$$

$$= \nabla \log[p(z|x)p(x)].$$

Some may raise the question. If the algorithm pushes the particles toward the high-probability region, do all particles collapse to one mode? There are two terms in the vector field expression, and each term has a different physical meaning.

$$\frac{1}{N} \nabla_x \log(p(x)) k(x, x') \rightarrow \text{weighted average of potential gradient}$$

$$\frac{1}{N} \nabla_x k(x, x') \rightarrow \text{repulsion between particles}$$

The $\nabla \log(p(x))$ term provides a smoothed gradient direction, pushing the particles towards the region with a high probability [105]. The second term acts as a repulsive force, preventing particles from collapsing into one local mode [105]. The following experiment demonstrates this phenomenon. The author runs SVGD with 50 particles and 300 iterations targeting a multivariate Gaussian Distribution. The first run shows the results of full SVGD. The second run only enforces gradient descent, and the last run only enables the repulsion term. The results are shown in Figure 44.

### 5.13.6 SVGD in Particle Filter

Compared with traditional Monte-Carlo-based PFs, this is a good algorithm for two simple reasons. First, SVGD avoids issues related to sampling noise and can achieve better convergence with fewer particles. Second, SVGD tends to provide a better exploration of the target distribution. The integration of SVGD with PF has been studied in [107][108].

Based on the framework of PFs, the inference involves two steps: prediction and update. For the prediction step, the particles from the previous time step propagate through the transition function to obtain the predictive distribution at time $t$,

$$p(x_k|z_{k-1}) \approx \frac{1}{N} \sum_{i=1}^{N} p(x_k|x_{k-1}).$$
This prediction equation provides a rough estimation for the prior. For the update step, the filtering distribution from equation (12) can be simplified using the SVGD algorithm as,

\[
p(x_k|z_{1:k}) = \frac{p(z_k|x_k)p(x_k|z_{2:k-1})}{p(z_{1:k-1})},
\]

\[
\log(p(x_k|z_{1:k})) = \log(p(z_k|x_k)) + \log(p(x_k|z_{2:k-1})) - \log(p(z_{1:k-1})),
\]

\[
\nabla_x \log(p(x_k|z_{1:k})) = \nabla_x \log(p(z_k|x_k)) + \nabla_x \log(p(x_k|z_{2:k-1})).
\]

The algorithm runs these two steps through iterations for the filtering problem. The posterior distribution \(p(x_{1:n}|y_{1:n})\) can be approximated by the empirical distribution.
of $n$ particles with equal weight. Then, the SPF algorithm is summarized in Algorithm 7.

Algorithm 7: $[\{x^i_k\}_{i=1}^{N_s}] = \text{SPF} \left[ x^i_{k-1}, z_k \right]

- \text{FOR } i = 1 : N_s
  - \text{Draw } x^i_k \sim p(x_k|x^i_{k-1})

- \text{END}

- \text{Compute } \{\hat{x}^i_k\} = \text{SVGD}(p(x_k|z_k), \{x^i_k\})

5.13.7 SPF Performance and Discussion

Thanks for the courtesy of sharing the original SVGD code from Qiang Liu’s research team. The SPF algorithm is modified based on their work in PyTorch. To run the algorithm, the gradient of the log empirical distribution needs to be defined. Previous research work in the field of robotics has demonstrated successful implementations of SPF. The system’s inherent noise and uncertainties are characterized by a normal distribution, simplifying the implementation process significantly. However, the flight position benchmark exhibits strong nonlinearity and includes multiple modes. Computing the exact log gradient from non-parametric distribution can be challenging and might require additional mathematical manipulations. As an alternative, the author simplifies the problem by assuming a normal distribution, which is a significant deviation from the underlying distribution in this specific benchmark. This assumption imposes considerable constraints on the algorithm’s performance. The flight position benchmark test shown in Figure 45 results in an RMSE of 5.807 for the process noise level of $\mu_{k-1} \sim \mathcal{N}(0, 1)$.

![Flight position benchmark using SPF](image-url)
The error in this assumption can accumulate over time, leading to poor performance. Furthermore, SPF is only applicable in state estimation. One of the main challenges in using SVGD for prognostics is that it requires access to the likelihood function, which may not be available or computationally tractable in this context. Thus, SVGD cannot be directly applied since it relies on the computation of gradients based on the likelihood function.

5.14 APF with Resample-Move

The PF variants presented above were designed to address specific challenges and improve performance. Each variant focuses on different aspects, such as accuracy and resilience to particle degeneracy. Among the variants discussed, the enhanced APF-RM is considered an optimized version that outperforms others. It combines the strengths of APF to correct the particles’ weight based on the latest measurement and the additional RM step to diversify particles in both state and parameter space. The superiority of this improved version can be illustrated in Figure 46, which showcases the contrast in the learning approach between the standard SISR version and APF-RM.

![Figure 46: Degradation path prediction in (a) SISR (b) APF-RM](image)

Each particle represents a potential trajectory, with accepted trajectories depicted in red while those rejected are in grey. The transparency of the colour indicates the frequency of selection during the resampling process. On the left is the generic filter, where the trajectory remains constant throughout learning. In contrast, on the right is the APF-RM, introducing perturbations in the parameter space at each
learning step. This enhances algorithm diversity in both state and parameter spaces, as evident in the zoomed image. With this said, we implement this algorithm for the bearing online prognostics model. The pseudo-code of APF-RM is summarized in Algorithm 8.

Algorithm 8: \[ \{x_k^i, w_k^i\}_{i=1}^{N_s} = \text{APF} \left[ x_{k-1}^i, w_{k-1}^i, z_k \right] \]

- FOR \( i = 1 : N_s \)
  - Draw \( \mu_k^i \sim p(x_k|x_{k-1}^i) \)
  - Calculate \( v_k^i \propto p(z_k|x_k^i)w_{k-1}^i \)
- END

- Normalize weight: \( v_k^i = v_k^i / \sum\{v_k^i\}_{i=1}^{N_s} \)
- Resample \[ \{-, -, i^j\} \] from \( \{x_k^i, v_k^i\}_{i=1}^{N_s} \)
- For \( j = 1 : N_s \)
  - Draw \( x_k^j \sim p(x_k|x_{k-1}^j) \)
  - Assign \( w_k^j = \frac{p(z_k|x_k^j)}{p(z_k|\mu_{ij}^k)} \)
- END

- Normalize weight: \( w_k^j = w_k^j / \sum\{w_k^j\}_{i=1}^{N_s} \)
- Resample \[ \{-, -, i^j\} \] from \( \{x_k^j, w_k^j\}_{j=1}^{N_s} \)
- Kernel Move \( K_k(x_{1:k}, \theta|y_{1:k}) \)
6 Bearing Prognostics Using ODM Data

The objective of this section is to build the physics-informed degradation exponential model that provides an accurate RUL for the second phase of spall propagation. The current prognostic model does not include the tertiary phase prediction for two reasons. First, as aforementioned, the tertiary phase indicates a significant growth in spall length, which can lead to increased vibration levels in the system. And excessive vibration can have detrimental effects on machine performance and reliability. It is recommended to have the faulty bearing replaced before the last phase. Second, in the context of life prediction application, it is generally advisable to adopt a conservative approach. One should also note the challenges associated with load-dependent spall length at failure mentioned in Section 2.4. When the load is small, the spall can spread and cover the entire raceway, while under high load conditions, the bearing is more likely to fail earlier. The uncertainty with this phenomenon has not been explored, and the current model does not have the capability to account for the uncertainties with the critical length at the end of spall failure. This section will include the considerations within the model implementation and discuss the results.

6.1 Spall Progression Prognostic Model

It is observed that exponential models can be highly successful in capturing the spall propagation curves of bearings in experimental data shown in Figure 10. Consider the exponential model,

$$x_k = e^{b_k \Delta t}x_{k-1} + \sigma_k,$$
$$z_k = 0.00838233x_k + w_k,$$

(69)

where $x$ is the spall length, $b$ is the unknown model parameter, 0.00838233 is the correlation coefficient from Figure 15. $\sigma_k$ and $w_k \sim \mathcal{N}(0,\sigma_w)$ are the process noise and measurement noise, respectively. Therefore, similar to the Li-battery case study, each particle has a vector form of $[x; b; \sigma_w]$.

6.2 Prior Knowledge from Generalized Linear Model

Prior knowledge is important in PF because it provides valuable information about the system being modelled and helps improve the efficiency and accuracy of the filtering process. As aforementioned in Figure 40, if the prior is less informative, it may yield more dispersed posterior, meaning 1) more particles are required to represent the distribution and 2) more learning data is needed to capture the model behaviour well. This section provides insights into prior settings for the mode parameter $b$, the spall length $x$, and the measurement noise variance $\sigma_w$. 

81
Fortunately, the exponential model can be easily transformed into a generalized linear model by applying logarithms to the data.

\[ \log(x) = \Delta t b_k + x_0, \]

where \( b_k \) corresponds to the parameter in the state-space model (69), \( x_0 \) reflects the initial spall length, which is not a concern in the predictive model. A linear regression model can be utilized to find the best-fitting line and its associated parameter with a confidence interval. The confidence interval provides a means to inform the initial belief for the model parameter. The prior knowledge of the spall length can be approximated from the correlation from Figure 15, and the measurement noise is about 15% of the TDC given by Gastops Ltd.

All parameters use uniform non-informative prior, reflecting minimal prior knowledge about the parameters, allowing the data to dominate the posterior distribution. The author finds non-informative prior effective in the context of prognostics to avoid introducing biases that might arise from subjective prior beliefs.

### 6.3 Results and Discussion

The author implements the sliding window technique. Instead of learning from \( t = 1 \), the prediction at the current time step only depends on a fixed number of data points. The reason for this implementation has two reasons. First, the assumption of an exponential model may not always hold in real situations; the observed data may deviate from the expected trend predicted by the model. To overcome this issue and minimize the accumulation of errors, the prediction is solely based on the most recent data points. The second reason is that the measurement uncertainty is about 15% of TDC, increasing over time as the count measurement accumulates. The prior knowledge of the measurement noise needs to be updated from time to time. A moving window allows for the inference in a localized manner, reflecting the local behaviours within the dataset rather than considering the entire dataset as a whole, such that the model is less susceptible to accumulative error from the model assumption. Additionally, based on the previous discussion, PF is robust only for a fixed time horizon. As the times of iteration increase, particle degeneracy is inevitable.

Selecting the appropriate window size can be challenging because we don’t want to use an excessive amount of data for training while still ensuring there is sufficient data to facilitate learning effectively. A fixed window size of 20 data points is adopted to balance the computational cost and the accuracy for inference. The particle size is set at 1000. The threshold is set at 6.55 mm, which corresponds to the lower 95%
confidence interval bound from Figure (15)(c). The prognostic results for Test 5 and Test 6 at different time steps are presented in Figure 47 and 48, respectively.

Figure 47: Test 5 RUL Prognostic Results
The results show that PF offers a powerful prognostic tool since it can effectively handle the uncertainties and nonlinearities inherent in degradation processes. Combining observed data with a mathematical model of the degradation process can estimate the RUL with full statistical information rather than a point estimation. Some discrepancies between the actual RUL and the estimated RUL can be observed.
from the plots, particularly at the beginning of the prediction. This is primarily due to the poor assumption of the exponential model. It is perceivable that the degradation trajectory deviates from the exponential trend at some point in both cases. However, with the moving window, the model parameter will correct itself by learning from the most recent observations and giving a more accurate prediction at a late stage.
7 Conclusion

Bearings play a critical role in the industry; their failure or malfunction can have severe consequences for industrial operations. In the failure theory, cycle fatigue is the most common failure mechanism that leads to progressive deterioration in bearings. The repeated stress cycles cause microscopic cracks to initiate within the bearing raceway; over time, these cracks can grow and lead to the development of spalling. Although bearings can experience various failure modes and failure locations, to narrow down the topic, this study focuses explicitly on the spalling on the inner race of angular contact ball bearings due to cycle fatigue.

Currently, the industry adopts the SKF L10 life standard for bearing life calculations. The SKF L10 life is a reliability rating that indicates the estimated number of operating hours or rotations a bearing can achieve before showing signs of fatigue. However, the extensive service life with spalling has not been addressed in L10. The thesis aims to develop a PHM model for life extension using the ODM data collected by Gastops MetalSCAN 4000.

The bearing degradation experiences three spall propagation phases: healthy, steady propagation, and accelerated propagation. The SKF L10 rating covers the healthy phase, and the extensive bearing lifespan for the steady propagation phase is studied in this work. The study does not include the life prediction for the last phase for two reasons. First, as the spalling progresses, the vibration signal shows a considerable increase in the last stage of propagation. This excessive vibration is undesirable and can have negative consequences for a machine. Second, there are significant uncertainties associated with the end-of-life prediction of bearings. Under low loading conditions, it is possible for a bearing to fail when a spall covers the entire raceway. Yet, a bearing can fail at a smaller spall length under high loading conditions. With this said, the developed novel PHM model should have the capabilities to 1) identify the spall propagation phase and 2) provide an accurate estimation of the RUL for the second propagation phase.

One of the significant contributions of this study is to develop a novel diagnostic tool. The developed diagnostic tool incorporates domain-specific knowledge, the knee detection algorithm, and machine learning to identify the spall propagation phrase. The limit theory defines the possible range of knee occurrence based on the geometric information of the bearing. The online "Kneedle" algorithm guides the search for knees by detecting the deflection point in a curve of sequential data. One critique of the "Kneedle" algorithm is that it is a heuristic algorithm, meaning that the detected knee points do not reflect the spall severity. Multiple keen detections are observed during testing. Therefore, the Random Forest model, trained from the
historical run-to-failure data, validates those possible knees. This fusion technique is preferable over individual models or single-source approaches as it enhances accuracy and reduces uncertainty. It also explores the synergies that leverage the strengths of each model.

The second contribution is to explore the particle filter and its variants. One should know that the PF was invented for the state estimation problem. In the context of prognosis, the algorithm is extended to cope with the recursive model parameter estimation problem. PF suffers from two common issues: particle degeneracy and particle impoverishment. Some of its variants or extensions have addressed these specific challenges and improvements in their performance. The author selected the most profound variants to investigate. The performance of these variants is tested in a case study of Li-battery life prediction. From the results, the author determined that the Auxiliary Particle Filter with the Resample-Move step provides the most effective performance improvement. APF provides the soft constraint to the optimal importance distribution to mitigate the particle degeneracy, while the RM offers the means to diversify the particles in both state and parameter space.

The other contribution of this study is that the designed methodology can be generalized for other rolling bearings with different sizes using a small amount of run-to-failure data. The designed methodology can be summarized in the following steps.

- Correlate the TDC to spall length using the ground truth data
- Find the critical spall length from the bearing geometry and its associated statistical interval from the correlation in Step 1

For diagnostic model:

1. Extract features from the binned data
2. Run the "Kneedle" Algorithm offline and label the features based on detected "knees"
3. Train and optimize the RF model
4. Run the fusion prognostic model online, depicted in Figure [22]

For prognostic model:

1. Establish a state-space model based on the Counts-to-Spall correlation.
2. Find sufficient statistics from the state-space model $T(x_{1:k}, y_{1:k})$ and derive the Gibbs sampler $p(\theta|T(x_{1:k}, y_{1:k}))$
3. Run the APF-RM algorithm depicted in Algorithm 8
Nevertheless, the developed methodology comes with some limitations. First, the current model does not consider any loading conditions in the prognosis. This is not trivial for accurate bearing life prediction. The loading condition, including factors such as stress magnitude, speed and variation over time, directly affects the fatigue experienced by the bearing. Different loading conditions can significantly impact the bearing’s lifespan and failure modes. Incorporating the loading condition into the developed model enables a more reliable estimation of the bearing’s remaining useful life.

Second, the accuracy and reliability of the predictions generated by the exponential model are heavily influenced by the quality and relevance of the data used to train the model. If the ODM data is accurate and representative of the exponential trend, the model is most likely to produce reliable and accurate results. On the other hand, if the ODM data is noisy or not representative, the model’s effectiveness may be compromised, leading to less accurate predictions.

Third, the analysis focuses solely on the detection of spalling occurring within the inner raceway of the bearing. However, this narrow scope might not encompass all potential failure modes that a bearing could experience. It’s important to acknowledge that while the study provides valuable insights into detecting inner raceway spalling, a comprehensive understanding of bearing health requires consideration of a broader spectrum of failure modes to ensure a more holistic diagnostic and prognostic approach.

7.1 Future Work

First, although the ODM data shows a direct correlation between wear debris and spall length, the previous study indicates vibration as a critical indicator of bearing degradation and can provide valuable insights into the health condition. Exploring the vibration features and integrating them with the bearing spall level estimation is possible. This integration may leverage the correlation between spall severity and vibration and debris information, allowing for a more accurate estimation of the bearing’s health condition and potential failure.

Second, there has been a lack of extensive research studying the influence of different resampling schemes on the performance of PF in the context of prognostics. In particular, some literature suggests using Block sampling. Block sampling aims to resample particles in high-probability regions, a practical approach to tackling particle degeneracy.

Third, the current PF is trained in situ based on a fixed window of 20 data points for future prediction. This approach may provide a reasonably good prognosis for
the spalling behaviour in a short horizon. However, in some cases, such as Test 6, the future behaviour may deviate from the trained model under the same operating condition. It remains to be explained what leads to such behaviour and whether the information from the historical database can be incorporated into the training. One potential solution is to investigate a fusion framework integrating a pure data-driven technique such as a neural network. Considering Gastops Ltd. has concerns regarding data ownership, exploring and adopting federated learning can be a viable approach for model training. Federated learning is a decentralized machine learning approach that allows multiple parties to train a shared model collaboratively without sharing raw data with a central server.

Furthermore, the current dynamic model uses trend analysis for bearing condition monitoring and prognosis. As fitting the exact dynamic model for the degradation behaviour can be challenging, it is also recommended to explore frequency analysis, such as the Poisson process. Poisson processes are widely used in various fields to model the occurrence of events that happen randomly in time. In the context of bearing prognosis, the Poisson process can model the random process of debris release in a given time interval based on the loading conditions and past events. Even though the author believes this approach can also provide promising results, the effectiveness of this approach needs to be studied and evaluated in future studies.
Appendices

A  SKF 7205 BECBY OEM Specification

The OEM specifications for the testing bearing SKF 7205 BECBY is presented in Figure 49: SKF 7205 BECBY Technical Specifications

<table>
<thead>
<tr>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>d₁</td>
</tr>
<tr>
<td>d₂</td>
</tr>
<tr>
<td>D₁</td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>r₁,₂</td>
</tr>
<tr>
<td>r₃,₄</td>
</tr>
</tbody>
</table>

Figure 49: SKF 7205 BECBY Technical Specifications [109]
Bibliography and References


[93] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.


Declaration

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and to determine whether the report as a whole or parts incorporated in it may be deemed plagiarism. For the comparison of my work with existing sources, I agree that it shall be entered in a database where it shall also remain after examination to enable comparison with future Theses submitted. THE INFORMATION AND TECHNICAL DATA DISCLOSED HEREIN ARE CONFIDENTIAL AND INTENDED SOLEY FOR THE PURPOSE OF THE DEVELOPMENT OF A NOVEL HEALTH MONITORING AND LIFE PROGNOSTICS SYSTEM AS OUTLINED IN THE RESEARCH GRANT AGREEMENT BETWEEN CARLETON UNIVERSITY AND GASTOPS LTD. EFFECTIVE JANUARY 3, 2020. THE INFORMATION AND TECHNICAL DATA IN THIS DOCUMENT AND ASSOCIATED DATA PACKAGES ARE THE EXCLUSIVE PROPERTY OF GASTOPS LTD. AND ARE NOT TO BE USED OR DISCLOSED TO ANY OTHERS WITHOUT THE WRITTEN CONSENT OF GASTOPS LTD. ANY RECIPIENT, BY ITS RETENTION AND USE, AGREES TO HOLD IN CONFIDENCE ANY INFORMATION AND TECHNICAL DATA DISCLOSED HEREIN. THE FOREGOING SHALL NOT APPLY TO PERSONS HAVING PROPRIETARY RIGHTS TO SUCH INFORMATION AND TECHNICAL DATA TO THE EXTENT THAT SUCH RIGHTS EXIST.

Ottawa, 08.2023

Location, date

Hengyangcan Gu

Name