Path-based splitting methods for SDEs and machine learning for battery lifetime prognostics

Calum David Strange

Doctor of Philosophy

The University of Edinburgh

2023

To my loving partner Chi,

You made me the man that I am today.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Prof. Gonçalo dos Reis and Dr. James Foster, for their support and guidance. They have been of great help through my intellectual journey and I have always enjoyed our discussions. Thank you for the responses to my many questions and messages. Gonçalo has given me many opportunities for growth and he has always trusted me to make the most of them. Through his supervision, my goals have been made reality. James has shown tremendous patience in sharing his great breadth and depth of knowledge.

I would like to thank Prof. Gareth Peters and Prof. Andrea Macrina for our discussions around mathematical finance. And I am grateful to my mentor David Semmens for his continued interest in my professional development and growth.

I am extremely grateful to my partner, Chi, for her love and company through these many years. And finally, I would like to thank my family and friends for their belief and support.

Abstract

In the first half of this Thesis, we present the numerical analysis of splitting methods for stochastic differential equations (SDEs) using a novel path-based approach. The application of splitting methods to SDEs can be viewed as replacing the driving Brownian-time path with a piecewise linear path, producing a 'controlled-differential-equation' (CDE). By Taylor expansion of the SDE and resulting CDE, we show that the global strong and weak errors of splitting schemes can be obtained by comparison of the iterated integrals in each. Matching all integrals up to order p+1 in expectation will produce a weak order p+0.5 scheme, and in addition matching the integrals up to order p+0.5 strongly will produce a strong order p scheme. In addition, we present new splitting methods utilising the 'space-time' Lévy area of Brownian motion which obtain global strong $O(h^{1.5})$ and $O(h^2)$ weak errors for a class of SDEs satisfying a commutativity condition. We then present several numerical examples including Multilevel Monte Carlo.

In the second half of this Thesis, we present a series of papers focusing on lifetime prognostics for lithium-ion batteries. Lithium-ion batteries are fuelling the advancing renewable-energy based world. At the core of transformational developments in battery design, modelling and management is data. We start with a comprehensive review of publicly available datasets. This is followed by a study which explores the evolution of internal resistance (IR) in cells, introducing the original concept of 'elbows' for IR. The IR of cells increases as a cell degrades and this often happens in a non-linear fashion: where early degradation is linear until an inflection point (the elbow) is reached followed by increased rapid degradation. As a follow up to the exploration of IR, we present a model able to predict the full IR and capacity evolution of a cell from one charge/discharge cycle. At the time of publication, this represented a significant reduction (100x) in the number of cycles required for prediction. The published paper was the first to show that such results were possible. In the final paper, we consider experimental design for battery testing. Where we focus on the important question of how many cells are required to accurately capture statistical variation.

Lay Summary

The first half of this thesis is dedicated to the study of numerical schemes for stochastic differential equations. Stochastic differential equations are used to model many real world phenomena and numerical schemes allow us to simulate them. We focus on a special class of numerical schemes called 'splitting schemes' and present a new framework for their understanding and analysis.

In the second half of this thesis, we investigate the ageing of lithium-ion batteries. As anyone with a mobile phone will know, batteries do not last forever and their performance degrades with time. This is a problem of great importance as governments look towards electric vehicles and many parts of the economy move towards electrification. We look at the available public data for batteries and propose several state of the art models for the prediction of battery lifetime. As a key finding, we provide a model which is able to test used batteries to aid their reuse, recycling and replacement.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Calum David Strange

Contents

Acknowledgements **iii**

Abstract **iv**

Lay Summary **v**

Declaration **vi**

Figures and Tables **xi**

I Path-based splitting methods for SDEs **1**

1 Introduction **2**

1.1 Convergence of numerical schemes 4

1.2 Numerical schemes 5

1.2.1 Basic schemes for Itô SDEs 5

1.2.2 Splitting schemes 7

1.3 Multilevel Monte Carlo 8

1.4 A brief introduction to tensors 11

1.5 Notation and spaces 13

2 Splitting methods as piecewise linear paths **14**

2.1 Introduction 14

2.1.1 Overview of part I 15

2.2 The path perspective 16

2.3 Assumptions on the splitting path and a preliminary result 19

2.4 Stochastic Taylor expansions 20

2.4.1 Stratonovich Taylor expansion 21

2.4.2 Controlled Taylor expansion 22

3 Iterated integrals of Brownian motion and Piecewise linear paths **24**

3.1 Algebraic properties of iterated integrals 24

3.2 Lévy areas of Brownian motion 28

3.2.1 Relational properties between Lévy areas 31

3.3 Computation of Brownian integrals 32

3.3.1 Expected value of Brownian integrals 34

3.4 Algorithmic computation of path integrals 35
3.4.1 Generating all words of a given order 35
3.4.2 Expansion of iterated integrals 36
3.4.3 Computing path integrals 38
3.5 L2 estimates of iterated path integrals 40
3.6 Commutativity simplifies Taylor expansions 42

4 Error analysis of path-based splitting schemes 44
4.1 Local estimates 44
4.2 Global strong error 46
4.2.1 Global strong error for commutative SDEs 47
4.3 Global weak error 49
4.3.1 Global weak error for commutative SDEs 51
4.4 A brief comparison with cubature on Wiener space 53

5 Examples of path-based splitting schemes 55
5.1 Piecewise linear splitting paths 55
5.2 Theoretical convergence rates 59
5.2.1 Strang splitting path 59
5.2.2 HS1 path 61
5.2.3 SO2 path 63
5.3 Derivation and optimality of splittings 69

6 Numerical examples 82
6.1 FitzHugh-Nagumo model 83
6.1.1 Derivation of FitzHugh-Nagumo splitting 84
6.1.2 Numerical results 85
6.2 Lotka-Volterra model 86
6.2.1 Derivation of Lotka-Volterra splitting 88
6.2.2 Numerical results 89
6.3 Multilevel Monte Carlo 91
6.3.1 Example: Interacting stock model 91
6.4 Testing Long-time integration 95

7 Conclusions and future work 97

II Machine learning for battery lifetime prognostics 104

Nomenclature 105

8 A brief outline of part II 106

9 Review of publicly available data for Li-ion batteries. 107

9.1 Introduction . . . 107

9.2 Where is the Data? . . . 109

9.2.1 Cycle ageing data . . . 110

9.2.2 Drive cycle data . . . 119

9.2.3 Characterisation data for cell modelling . . . 122

9.2.4 Calendar ageing . . . 125

9.2.5 Aeroplanes, satellites and energy storage . . . 126

9.2.6 Synthetic data . . . 127

9.2.7 Data by request . . . 128

9.3 Data governance, repositories, tools and future outlook . . . 129

9.3.1 Data repositories and platforms . . . 129

9.3.2 Community maintained reviews and standards . . . 130

9.3.3 Data processing and validation tools . . . 131

9.3.4 Current limitations . . . 131

9.4 Conclusions . . . 133

10 Elbows of Internal resistance rise curves 134

10.1 Introduction . . . 134

10.2 Battery Data Framework and Data Pre-Processing Procedures . . . 136

10.2.1 Data Pre-Processing via a Machine Learning Approach: Completing the Missing IR Data . . . 136

10.3 Identification of Elbows, Knees and Their Relations . . . 142

10.3.1 Methodology . . . 142

10.3.2 Linear Relations . . . 146

10.4 Early Prediction of Elbows . . . 148

10.5 Conclusions and Future Work . . . 150

11 One cycle prediction 152

11.1 Introduction . . . 152

11.2 Predicting future capacity and internal resistance . . . 155

11.2.1 Modelling approach . . . 157

11.2.2 Prediction intervals via the forward-dropout method . . . 159

11.3 Model performance . . . 160

11.3.1 Performance metrics . . . 160

11.3.2 Full curve prediction 161
11.3.3 Comparison with prior art 163
11.4 Conclusions 166

12 Sample sizes required to understand cell-to-cell variability 167
12.1 Introduction 168
12.2 Design of a sequential analysis 169
12.3 Datasets overview 171
12.4 Methodology and estimation 173
12.5 Results 177
12.6 Two applied examples 179
12.6.1 Using prior knowledge to inform new testing 179
12.6.2 Generating test cases in an online format to inform larger and longer experiments 180
12.7 Conclusions and outlook 183

13 Conclusions and future work 186

Figures and Tables

Figures

2.1 In the Monte Carlo paradigm, information about the Brownian motion is generated and then mapped to a numerical solution of the SDE. Typically, only Brownian increments are sampled. 15

3.1 Space-time Lévy area gives the area between a Brownian path and its linear approximant. 29

3.2 Space-time Lévy swing gives the side where the path has greater space-time Lévy area. 30

3.3 Space-time-time Lévy area corresponds to a cubic approximation of the Brownian arch (which is a Brownian motion conditioned on having zero increment and space-time Lévy area [17]). 31

3.4 Subdividing integration over a triangle. 36

3.5 All non-increasing routes. 37

3.6 Specific routes. 37

5.1 Illustration of piecewise linear paths associated with various splitting methods for SDEs. 58

6.1 S_N estimated for (6.2) using 1,000 sample paths as a function of step size $h = \frac{T}{N}$. 85

6.2 Example dynamics for the LV model, with parameter values $a_{11} = a_{22} = 0.0005$, $a_{12} = -a_{21} = 0.003$, $b_1 = -b_2 = 1$ and $G_1 = G_2 = 0.2$. The left plot displays the dynamics in phase space, and the right plot shows how the populations evolve through time. 87

6.3 Strong errors for (6.11), estimated from 1000 sample paths. 90

6.4 Example price path for three interacting assets. 93

6.5 Plots comparing MLMC performance for Strang splitting and HS2 splitting. . . 94

6.6 Comparison of splitting (6.17) with the non-Markovian Euler scheme, based on 10^6 samples. 96

9.1 Hierarchical architecture of the existing battery datasets from an historical point of view. 109

9.2 The typical plots of a high-throughput cycling dataset encompassing measured terminal current, voltage and temperature variations. Capacity, IR, voltage and temperature can then be used for the ageing analysis. 110

9.3 A typical Nyquist plot: Battery characterisation using EIS measurements 124

9.4 Lithium battery sample applications. 132

10.1 Schematic of machine learning model for internal resistance (IR) prediction. . . 137

10.2 The predicted IR data for cell b8c4 are given by the black continuous line and is formed from the average of 20 predictions. We display 80% and 95% prediction intervals. Beyond the intuition of extrapolation, these intervals show that predictions past the EOL (capacity) should not be trusted. 140

10.3 Graphical abstract for the proposed algorithmic framework. 141

10.4 Steps of Algorithm 1 applied to the internal resistance degradation curve of cell b1c29 (non-predicted data). (**a**), step 1. (**b**), step 3. (**c**), step 4. (**d**), step 5. Step 2 is omitted as it has no impact here: n^* is chosen as the final cycle number. The width of the 95% confidence interval (computed by the non-parametric bootstrapping procedure) for the elbow-point of this curve is 23 cycles, and for the elbow-onset it is 38 cycles. 145

10.5 (**a**), Comparison of elbow-points obtained with Algorithm 1, [78]'s Bacon–Watts, maximum curvature and slope changing ratio methods on a sample of cells from the A123 dataset (from left to right b2c34, b1c30, b3c15, b3c1, b1c3). (**b**), Comparison of elbow-points for all cells in the A123 dataset. One expects to see a linear relationship between EOL and elbow-point; of the methods compared only Algorithm 1 and the algorithm of Satopaa et al. [2011] recover a linear relationship reliably, however, by examining plot (**a**), we see that Satopaa's algorithm selects the end point as the elbow. 145

10.6 (**a**), Linear regression model linking the knee-point to end of life. (**b**), elbow-point to end of life. (**c**), knee-point to elbow-point. (**d**), knee-onset to elbow-onset. Every linear model is presented with a 95% confidence band on the plotted regression line; all linear relations here are calculated from the A123 dataset enriched with the predicted IR data for batch 8. Elbow points derived from the predicted IR data are highlighted as open black circles; the reader will appreciate that their inclusion did not significantly influence the linear regression results obtained. 147

11.1 Empirical model fitted to the capacity and IR curves of cell b3c12. 156

11.2 Representation of the CNN architecture. The '$\times 3$' notation denotes three repeated blocks with the displayed configuration. 158

11.3 Plots of model predictions for nine randomly chosen test cells at randomly selected input cycles. Model predictions are produced from data of a single cycle, given the measured capacity/IR at that cycle. The prediction intervals (95% and 80%) are calculated under a normality assumption from 100 predictions with forward dropout applied. Moving from left to right, and then down a row repeating, the 'test cell' displayed is b2c30, b8c20, b3c9, b2c14, b1c7, b3c34, b2c45, b3c26 and b1c4. Here, we present the IR and capacity values relative to the values measured at the input cycle. 162

11.4 **a.** Confusion matrix for prediction of EOL classes 'short life' (EOL < 550), 'medium life' ($550 <$ EOL < 1200) and 'long life' (EOL > 1200), based on quantitative prediction for test set. **b.** Actual EOL vs predicted EOL with distribution of errors, showing a slight bias to under-predict EOL especially for longer lived cells. **c.** Distribution of training data for EOL showing a significantly lower amount of training data for long lived cells. 162

11.5 Plot of model's performance to predict RUL in the full setting, showing a mostly linear improvement in performance as the prediction point approaches the EOL. 163

12.1 Sketch of high-throughput test example using 1000 channels and different phases of the design of experiment and continuous test enhancement. 170

12.2 Example for an ageing test plan for 2 years with 24 channels available. Each row represents a channel and a) and b) describe two scenarios with a change of the test plan after one year depending of having tested "enough" to capture variability. Yellow denotes channels used to capture the cell-to-cell variability, green shows the initial run of other stress factors, and blue and red show sequential tests on the same channels. In a) after 1 year enough cells have been tested under the same conditions, so additional test conditions can be tested with the available channels. In b) the amount of data collected is not yet enough and additional tests are started with the same ageing conditions. 171

12.3 Capacity fade trajectories (y-axis) over time (x-axis) for the six datasets mentioned in Section 12.3 172

12.4 Relative standard error (RSE) of sample standard deviation as a function of sample size. Black continuous line given by deterministic RSE asymptotic approximation of (12.4). 177

12.5 Probability of a random sample estimate with relative error not more than 25%. 178

12.6 Q-Q plots for the standardised distribution of cell slopes β. The inset histogram plots show the true (non-standardised) distribution of slopes for each dataset. . 179

12.7 Q-Q plots for the standardised distribution of cell slopes β across the percent input of time data as according to Table 12.4. Datasets are left-to-right and percent input from top-to-bottom starting at 20% until 100%. The inset histogram plots show the true (non-standardised) distribution of slopes for each dataset and percentage of input data. 182

Tables

5.1 Theoretical convergence rates of splitting paths. Without the commutativity condition the higher order schemes fail to match the word $\omega\omega$ a.s. and the word $\omega\omega\omega\omega$ in expectation. Strang splitting is a high order weak scheme as it matches the integral $\omega\omega\tau$ in expectation, which both $SO3$ and $SO4$ fail to do. However, while the paths $SO3$ and $SO4$ only obtain a global strong error of $O(h)$, we note that they have a local strong error of $O(h^2)$ compared with a local strong error of $O(h^{1.5})$ for the Strang splitting (see Table 5.2). Green text marks high order convergence. 59

5.2 Checking the conditions of Theorem 4.2.3 for example paths. Highlighted in red are where the paths fail to satisfy the conditions in the theorem. D_d^2, as defined in (3.14), represents the $d \times d$ identity matrix. 68

6.1 Compute time to simulate 1000 sample paths of (6.2) with 100 steps (seconds) 85

6.2 Estimated compute time to produce 1000 sample paths of (6.2) with an error of $S_N = 10^{-3}$ (seconds) 86

6.3 Compute time to simulate 1000 sample paths of (6.11) with 100 steps (seconds) 90

6.4 Estimated compute time to produce 1000 sample paths of (6.11) with an error of $S_N = 10^{-3}$ (seconds). $(*)$ based on Figure 6.3 the Euler method would require $\sim 2^{24}$ steps to achieve an accuracy of $S_N = 10^{-3}$, in combination with Table 6.3 this gives an approximate time to run of 40.5 hours. 90

9.1 NASA dataset repository: Related papers and the corresponding research conducted. (See additionally Supplementary material Table 3 for full details.) . 111

9.2 Overview of cycle ageing datasets. 'gr' stands for 'graphite', 'Cal' denotes calendar ageing, 'Chrg' charge protocol and 'Dhrg' discharge, 'E' denotes 'energy'. Here, we use 'IR' to denote both internal resistance and impedance. No 'test variables' indicates that all cells in the experiment were cycled in the same way. 118

9.3 Panasonic 18650PF Li-ion Battery Data: Related paper and the corresponding research conducted. 120

9.4 LG 18650HG2 Li-ion Battery Data: Related paper and the corresponding research conducted. 120

9.5 Overview of Driving cycle data. ‘cycle’ here denotes the use of different drive cycle profiles, ‘E’ denotes ‘energy’. No ‘test variables’ indicates that all cells in the experiment were cycled in the same way. 122

9.6 Overview of Calendar Ageing degradation data. Here, ‘time’ denotes the frequency at which ageing was interrupted to take measurements. 126

9.7 Non-publicly available Battery Data: Related paper and the corresponding research conducted. (See additionally Supplementary material Section 3.) . . . 129

10.1 Proposed architecture of convolutional neural network (CNN) model for prediction of IR. Hyper-parameters are given in the format: filters, kernel size, activation for conv1d layers; pool size for max_pooling; dropout rate for dropout; nodes, activation for dense layers. 138

10.2 Average performance of model to predict IR with 95% prediction intervals. . . 139

10.3 Average performance of capacity model trained on batches 1–3 tested on batches 4–8, with 95% prediction intervals. 139

10.4 Coefficients of four linear regression models relating the knee-point (**a**) and the elbow-point (**b**) to the end of life, the knee-point to the elbow-point (**c**) and the knee-onset to elbow-onset (**d**), respectively. The p-values for β_1 were computed using the Wald test, and the small values allow the rejection of the null hypothesis that a linear relationship does not exist. The 95% confidence intervals for the estimated coefficients are calculated via bootstrapping. The coefficient of determination, R^2, of these linear regression models is (**a**) 0.9822, (**b**) 0.9896, (**c**) 0.9818 and (**d**) 0.9520; all close to 1, showing that the fitted models explain the observed data well. 148

10.5 Result of RVM regressor for elbow-onset (**a**) and elbow-point (**b**) when predictions are made from the first 50 cycles. The 90% confidence intervals (CI) were calculated via bootstrapping. The entry ‘With b8?’ refers to results computed with (‘Yes’) and without (‘No’) the inclusion of the artificially predicted IR data of batch 8. 149

11.1 Proposed architecture of the CNN model for the prediction of ttk-o, Q@k-o, ttk-p, Q@k-p, and RUL. The model can be trained to predict multiple points at once (joint prediction) or separately. Hyper-parameters are given in the format: filters, kernel size, activation for conv1d layers; pool size for max_pooling; dropout for dropout; nodes, activation for dense layers. 158

11.2 Performance of proposed model to predict Capacity’s: ttk-o, ttk-p and RUL; and IR’s: tte-o and tte-p. 160

11.3 Performance of model to predict future capacity and IR, when current capacity/IR is known. 160

11.4 Comparison of results from works using the Severson dataset. For comparison purposes, reported results exclude data from batch 8. Ordered by number of cycles used and reported MAPE. Inputs listed are in-cycle measurements of voltage (V), current (I), temperature (T), capacity as a function of voltage (Q(V)); and cycle-to-cycle measurements of capacity (SOH), internal resistance (IR) and time to charge (ttc). 165

11.5 Performance of proposed model to predict EOL and RUL when restricted to batches 1, 2 and 3. We report performance both in the initial setting (input cycle < 100 cycles) and the full setting. 165

12.1 Theoretical number of samples required to estimate the population standard deviation for a range of maximum acceptable deviations (relative to σ_β) and confidence levels that this maximum will not be exceeded. The 68% confidence row corresponds to the asymptotic RSE result of Equation (12.4) with the boxed $n = 9$ stands for $\texttt{RSE} = 0.25$. 176

12.2 Sample sizes for the datasets of this study (at 25% maximum acceptable deviation and 68% confidence) per dataset, and theoretical sample size estimate (see Table 12.1). N is the total number of cells tested. 178

12.3 Sample size for Attia-predicted (at 25% maximum acceptable deviation and 68% confidence) per dataset, and sample size theoretical estimate (see Table 12.1). Information on Severson-2019 and Attia-2020 kept for comparison. . . . 180

12.4 Required sample size given the first $20, 40, \ldots, 100\%$ of input data (at 25% maximum acceptable deviation and 68% confidence). It should be noted that the theoretical number of required cells is 9 regardless of input size (see Table 12.1). Severson-2019* denotes the results after an outlier cell is removed from the Severson-2019 dataset (as justified below). 181

PART I

Path-based splitting methods for SDEs

Chapter 1

Introduction

In this chapter, we give a brief introduction to the numerical analysis of SDEs, focusing on the necessary background material for this thesis.

A wide range of processes (such as in Finance, Biology and Physics [56]) can be modelled by the following class of Itô SDEs:

$$dy_t = f(y_t)dt + g(y_t)dW_t \ , \quad y_0 = y \ , \tag{1.1}$$

with initial value $y \in \mathbb{R}^e$, where $W = (W^1, \cdots, W^d) = \{W_t\}_{t\in[0,T]}$ denotes a d-dimensional Brownian motion, and $f : \mathbb{R}^e \to \mathbb{R}^e$ and $g : \mathbb{R}^e \to \mathbb{R}^{e\times d}$. For many choices of f and g there is no explicit solution available. It is thus necessary to develop and analyse numerical schemes for approximating the solution. There exists a vast body of research dedicated to this topic, for a wide range of applications and assumptions (see e.g. [57] and [59]).

In many applications, one is interested in calculating the expected value of some functional ψ of the SDE (1.1)

$$\mathbb{E}\Big[\psi(y_t)\Big|y_0 = y\Big] \ .$$

For example, in the pricing of financial derivatives, ψ would be the payoff function. The usual approach to estimating this expected value is the Monte Carlo estimator:

Definition 1.0.1 (Monte Carlo estimator)**.** *To approximate the expected value of some functional $\psi : \mathbb{R}^e \to \mathbb{R}^l$, as is usual in numerics, we adopt the standard Monte Carlo estimator. Given M independent samples $(y^i)_{i=1}^M$ of a random variable y, define the estimator*

$$\psi^M := \frac{1}{M}\sum_{i=1}^{M} \psi(y^i) \ . \tag{1.2}$$

Remark 1.0.2. *If $\psi(y^i)$ has finite second moment then by the central limit theorem*

$$\sqrt{M}(\psi^M - \mathbb{E}[\psi(y)]) \xrightarrow{d} \mathcal{N}\big(0, \mathit{Var}(\psi(y))\big) \ ,$$

that is, ψ^M converges to its expectation with a rate of $O\left(\frac{1}{\sqrt{M}}\right)$. Where $\xrightarrow{d}$ denotes convergence in distribution (see e.g. [59]). Where by $\mathcal{N}(\mu, \sigma^2)$ we denote the normal distribution with mean μ and variance σ^2.

Of course, we do not always have the ability to draw independent samples of a random variable. For us, the random variable in question is the value of the SDE (2.1) at some time $T > 0$. Our approach (when an explicit solution is not available) is then to construct a numerical approximation $(Y_k)_{k=1}^N$ of the solution. Loosely speaking, we construct this numerical scheme iteratively as a function of its previous value. Over a grid $\pi_N = \{t_0, t_1, \ldots, t_N\}$ with timestep $h = t_{k+1} - t_k$ and some random variables ω_{k+1} (usually approximating the Brownian motion). That is

$$Y_{k+1} := F(Y_k, h, \omega_{k+1}) \ , \qquad \text{with} \qquad Y_0 := y_0 \ .$$

Each step Y_k provides our approximation for y_{t_k}. If ψ depends only on the final value y_T, we then replace ψ^M by

$$\bar{\psi}^M := \frac{1}{M} \sum_{i=1}^{M} \psi(Y_N^i) \ ,$$

where each Y_N^i is an independent realisation of our numerical approximation. We are interested in the error of this estimation, and in particular the speed at which this error decreases. This error can be broken into multiple components:

$$\left\| \mathbb{E}[\psi(y_T)] - \bar{\psi}^M \right\| = \underbrace{\left\| \mathbb{E}[\psi(y_T)] - \mathbb{E}[\psi(Y_N)] \right\|}_{\text{`Weak error'}} + \underbrace{\left\| \mathbb{E}[\psi(Y_N)] - \bar{\psi}^M \right\|}_{\text{Monte Carlo error}} \ .$$

As in Remark 1.0.2, the Monte Carlo error is controlled by the sample size M and is of size $O(1/\sqrt{M})$. This error can be improved by taking more samples, or by using a different estimator than the Monte Carlo estimator. Examples of such alternative approaches are various variance reduction techniques (see e.g. [34] for an overview) and Multi-level Monte Carlo (see [24] and Section 1.3 below). The main focus of this thesis will be on the error which we incur by our choice of numerical scheme (in the above example we would interested in the 'weak error'), and in constructing numerical schemes which minimise this error.

1.1 Convergence of numerical schemes

We concern ourselves with two types of convergence: 'strong' convergence and 'weak' convergence. 'Strong' means convergence in an L^2 sense and 'weak' means convergence with respect to certain test statistics. Concretely, for an approximation Y and a true solution y, we define

$$\text{'Strong error'} := \mathbb{E}\Big[\big\|y - Y\big\|^2\Big]^{\frac{1}{2}} ,$$

$$\text{'Weak error'} := \big\|\mathbb{E}\big[\psi(y) - \psi(Y)\big]\big\| ,$$

where $\psi : \mathbb{R}^e \to \mathbb{R}^l$ is some test function. Our interest is then how quickly these errors converge as a function of the step size h with which we run our numerical schemes.

Definition 1.1.1. *A numerical scheme is said to 'converge with global strong order p' if*

$$\mathbb{E}\Big[\big\|y_{k\times h} - Y_k\big\|^2\Big]^{\frac{1}{2}} = O(h^p) , \tag{1.3}$$

for every $k \in \{1, \dots, N\}$.

Definition 1.1.2. *A numerical scheme is said to 'converge with global weak order p' if, for some smooth test function $\psi \in C^\infty(\mathbb{R}^e, \mathbb{R}^l)$,*

$$\big\|\mathbb{E}\big[\psi(y_{k\times h}) - \psi(Y_k)\big]\big\| = O(h^p) ,$$

for every $k \in \{1, \dots, N\}$.

The above error rates are defined globally; However, it is usually much easier to prove 'local' error rates. Local errors are the error made over a single step of the numerical scheme and are smaller than the global error. The global error rates can then be inferred from the local rates. To obtain the global strong error rate we appeal to the following classical result of Milstein and Tretyakov [48]:

Theorem 1.1.3 (Fundamental Theorem on the mean-square order of convergence [48, Theorem 1.1.1])**.** *Let Y_1 denote a one-step approximation of y_h, both from initial value $y \in \mathbb{R}^e$ with time step h. Suppose that the following inequalities hold*

$$\mathbb{E}\big[\|y_h - Y_1\|^2\big]^{\frac{1}{2}} \leqslant K\sqrt{1 + \|y\|^2} \times h^p \qquad \textit{and} \qquad \big\|\mathbb{E}\big[y_h - Y_1\big]\big\| \leqslant K\sqrt{1 + \|y\|^2} \times h^q ,$$

for

$$p \geqslant \frac{1}{2} \qquad \textit{and} \qquad q \geqslant p + \frac{1}{2} .$$

Then, for any N and for any $k \in \{1, \ldots, N\}$

$$\mathbb{E}\big[\|y_{kh} - Y_k\|^2\big]^{\frac{1}{2}} \leqslant K\sqrt{1 + \mathbb{E}\|y_0\|^2} \times h^{p-\frac{1}{2}} \ .$$

Remark 1.1.4. *As we will show later in the text (see Section 4.3) the global weak error can easily be obtained from the local weak error, by a telescoping sum argument. With a local weak error of $O(h^p)$ implying a global weak error of $O(h^{p-1})$.*

1.2 Numerical schemes

The number and variety of numerical schemes designed for SDEs is vast, and a full review of such methods is beyond the scope of this thesis. We instead point the reader to the books [48, 57, 59] and mention a few basic schemes to provide motivation or for later use in numerical examples. We will also introduce 'splitting schemes' which will be the main focus of this work.

1.2.1 Basic schemes for Itô SDEs

Perhaps the most basic and widely used scheme is the Euler-Maruyama (or 'Euler') method. This method achieves a strong order of $O(h^{1/2})$ and a weak order of $O(h)$. While it is often the first choice of numerical scheme, in many applications it is a poor one, performing far worse than more informed schemes (See e.g. our numerical examples in Chapter 6).

Definition 1.2.1 (The Euler-Maruyama method)**.** *We can construct a numerical approximation of* (1.1) *by setting $Y_k = y_0$ and for time step h iterating*

$$Y_{k+1} := Y_k + f(Y_k)h + g(Y_k)W_{t_k,t_{k+1}} \ , \tag{1.4}$$

where $W_{s,t} := W_t - W_s$ denotes the Brownian increment.

The Milstein scheme is an improvement on the Euler scheme. It achieves this improvement by including an additional term from the Taylor expansion of the stochastic integral. This method achieves a strong order of $O(h)$ and a weak order of $O(h)$. In fact, no 'increment only' scheme (one using only the Brownian increment $W_{s,t}$) can achieve a strong rate of convergence higher than $O(h^{1/2})$ [7].

Definition 1.2.2 (The Milstein scheme)**.** *We can construct a numerical approximation of* (1.1) *by setting $Y_k = y_0$ and for time step h iterating*

$$Y_{k+1} := Y_k + f(Y_k)h + g(Y_k)W_{t_k,t_{k+1}} + \frac{1}{2}g'(Y_k)g(Y_k)\Big(W^{\otimes 2}_{t_k,t_{k+1}} + A_{t_k,t_{k+1}} - hD^2_d\Big) \ ,$$

where D_d^2 (defined in (3.14)*) denotes the $d \times d$ identity matrix, $\otimes$ denotes the tensor product (see Definition 1.4.7) which for two vectors $x, y \in \mathbb{R}^d$ is given by $x \otimes y = xy^\top = \{x_i y_j\}_{1 \leqslant i,j \leqslant d}$, and $A_{s,t}$ denotes the* Lévy area *of Brownian motion (here defined with Itô integrals, see Definition 3.2.1 for the equivalent definition with Stratonovich integrals), which is defined as*

$$A_{s,t} = \int_s^t W_{s,r} \otimes dW_r - \left(\int_s^t W_{s,r} \otimes dW_r \right)^\top .$$

When $d = 1$ the Lévy area is zero and so the Milstein method becomes

$$Y_{k+1} := Y_k + f(Y_k)h + g(Y_k)W_{t_k,t_{k+1}} + \frac{1}{2}g'(Y_k)g(Y_k)\Big(W_{t_k,t_{k+1}}^2 - h\Big) . \tag{1.5}$$

The Euler-Maruyama scheme (1.4) and the 1D Milstein method (1.5) are both easy to implement. Since Brownian motion has independent increments and is independent in its coordinate processes, at each step we simply need to generate

$$W_{t_k,t_{k+1}} \sim \mathcal{N}\big(0, h\mathbf{1}_d\big) ,$$

(where by $\mathbf{1}_d$ we denote the vector of length d containing only 1's) and to propagate forward the schemes as defined.

However, in multidimensions the Milstein method also requires the simulation of the Lévy area $A_{s,t}$. Exact generation of $A_{s,t}$ has only been proposed for $d = 2$ [22] and this is known to be a difficult problem in general [9]. However, as we will show in Section 3.6, when the diffusion matrix g of the SDE satisfies the commutativity condition (2.2) the Lévy area does not need to be generated due to cancellations in the Taylor expansion.

The explicit schemes above are designed for Itô SDEs; However, it is actually easier to develop high order schemes for Stratonovich SDEs. So, we instead write the Itô SDE (1.1) in Stratonovich form:

$$dy_t = \tilde{f}(y_t)dt + g(y_t) \circ dW_t, \quad y_0 = y \tag{1.6}$$

where $\tilde{f}(y) := f(y) - \frac{1}{2}g'(y)g(y)$ is the drift after applying the Itô-Stratonovich correction. The main benefit to working with Stratonovich SDEs is algebraic in nature and will be explained throughout the text.

1.2.2 Splitting schemes

In this thesis, we will focus on 'splitting schemes' as an approach to discretize the Stratonovich SDE (1.6). Spitting schemes for SDEs are inspired by an idea originating in ODE numerics (often referred to as 'operator splitting' see e.g. [45]). Consider the following ODE

$$dx_t = (A + B)x_t dt \ , \quad x_0 = x$$

where $x \in \mathbb{R}^d$, and A and B are time independent $d \times d$ matrices. Which has exact solution

$$x_t = \exp\Big((A + B)t\Big)x_0 \ .$$

The observation of operator splitting is that, if A and B commute, we can instead write $x_t = \exp\big(At\big)\exp\big(Bt\big)x_0$. And we may then solve the sub problems

$$\exp\Big(At\Big) \quad \text{and} \quad \exp\Big(Bt\Big) \ ,$$

sequentially. In fact, this splitting can be justified (using the Baker-Campbell-Hausdorff formula) in general and has proven to be an effective approach to numerically solving ODEs. For example, a Strang splitting scheme of the above ODE over the interval $h := t_{k+1} - t_k$ is

$$X_{k+1} = \exp\Big(\frac{h}{2}A\Big)\exp\Big(hB\Big)\exp\Big(\frac{h}{2}A\Big)X_k \ ,$$

where $\exp(V)x$ denotes the solution z_1 at time $u = 1$ of the ODE $z' = V(z), \ z(0) = x$.

Splitting schemes for SDEs are also defined through composition of subsystems. For example, we could separate the drift and diffusion part of the SDE. And in this way, higher order convergence can be obtained. A Lie-Trotter splitting solves either the drift or diffusion first, followed by the diffusion of drift respectively. At each stage, the solution to the previous step is used as the initial value for the next. This gives the following splitting

Definition 1.2.3 (The Lie-Trotter splitting scheme)**.** *We may construct a numerical approximation for the SDE* (1.6) *by setting* $Y_0 = y_0$ *and for timestep* h *iterating*

$$Y_{k+1} := \exp\Big(g(\cdot)W_{t_k,t_{k+1}}\Big)\exp\Big(f(\cdot)h\Big)Y_k \ .$$

An improvement upon the Lie-Trotter splitting is a Strang splitting which symmetrises the approach. The Strang splitting solves the drift part up to time $h/2$, then solves the diffusion part up to time h, before again solving the drift part up to time $h/2$. Which gives the following splitting:

Definition 1.2.4 (The Strang splitting scheme). *We may construct a numerical approximation for the SDE* (1.6) *by setting $Y_0 = y_0$ and for timestep h iterating*

$$Y_{k+1} := \exp\left(f(\cdot)\frac{h}{2}\right)\exp\left(g(\cdot)W_{t_k,t_{k+1}}\right)\exp\left(f(\cdot)\frac{h}{2}\right)Y_k \; .$$

In the commutative case, both the Lie-Trotter and the Strang splitting achieve a strong order of $O(h)$. But the Strang splitting achieves a weak order of $O(h^2)$ compared with the Lie-Trotter's weak order of $O(h)$ (see Table 5.1).

An extension of the SDE Strang splitting is the Ninomiya-Victoir splitting, which solves each column of the diffusion matrix in turn. Concretely:

Definition 1.2.5 (The Ninomiya-Victoir scheme [55]). *We may construct a numerical approximation for the SDE* (1.6) *by setting $Y_0 = y_0$ and for timestep h iterating*

$$Y_{k+1} := \begin{cases} \widehat{Y}_{k+1}, & \textit{if } n_k = 1, \\ \widecheck{Y}_{k+1}, & \textit{if } n_k = -1, \end{cases}$$

where n_k is an independent Rademacher random variable and $\widehat{Y}_{k+1}, \widecheck{Y}_{k+1}$ are given by

$$\begin{aligned}\widehat{Y}_{k+1} &:= \exp\left(f(\cdot)\frac{h}{2}\right)\exp\left(g_1(\cdot)W^{(1)}_{t_k,t_{k+1}}\right)\cdots\exp\left(g_d(\cdot)W^{(d)}_{t_k,t_{k+1}}\right)\exp\left(f(\cdot)\frac{h}{2}\right)Y_k \; , \\ \widecheck{Y}_{k+1} &:= \exp\left(f(\cdot)\frac{h}{2}\right)\exp\left(g_d(\cdot)W^{(d)}_{t_k,t_{k+1}}\right)\cdots\exp\left(g_1(\cdot)W^{(1)}_{t_k,t_{k+1}}\right)\exp\left(f(\cdot)\frac{h}{2}\right)Y_k \; ,\end{aligned}$$

where g_i denotes the i'th column of g.

1.3 Multilevel Monte Carlo

As noted in Remark 1.0.2, the error of the standard Monte Carlo estimator is governed by the variance. One well established and popular approach to improving upon this and reducing the variance of the resulting estimator is Multilevel Monte Carlo (MLMC). We point the reader to [24] for a general overview of the topic, and this section is largely based on the description therein. In Section 6.3 we will show how some of the high order splitting paths presented in this thesis can be incorporated with MLMC. The idea works as follows. Let P_ℓ denote an estimate of $\mathbb{E}[\psi(y_T)]$ generated with timestep

$$h_\ell = \frac{T}{2^\ell}, \quad \ell = 0, 1, \cdots, L \; .$$

By a telescoping sum argument we have that

$$\mathbb{E}[P_L] = \mathbb{E}[P_0] + \sum_{\ell=1}^{L} \mathbb{E}[P_\ell - P_{\ell-1}] \ ,$$

and each expected value on the RHS can then be replaced by its own coupled Monte Carlo estimate. Introducing the following independent Monte Carlo estimators

$$Z_0 := \frac{1}{N_0} \sum_{i=1}^{N_0} P_0^{(0,i)} \qquad \text{and} \qquad Z_\ell := \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \left(P_\ell^{(\ell,i)} - P_{\ell-1}^{(\ell,i)} \right) \ , \tag{1.7}$$

where, for each i, $P_\ell^{(\ell,i)}$ is an independent sample of P_ℓ at level ℓ. To calculate the difference $P_\ell^{(\ell,i)} - P_{\ell-1}^{(\ell,i)}$ it is typical to simulate both $P_\ell^{(\ell,i)}$ and $P_{\ell-1}^{(\ell,i)}$ with the same Brownian path. The superscript ℓ is included to indicate that each correction term uses a different Brownian path. The MLMC estimator is then defined by

$$\bar{\psi}^{\text{MLMC}}(N_0, \ldots, N_L) := \sum_{\ell=0}^{L} Z_\ell \ . \tag{1.8}$$

The key freedom here is the choice of values for $N_0 > N_1 > \cdots > N_L$. A smaller step size means taking more steps to reach the final time T and thus an increase in computation cost, but this comes with the benefit of improved accuracy. A larger step size will take less time to compute, meaning many more (less accurate) samples can be generated in the same length of time. MLMC exploits this fact by calculating a large number of less accurate estimates and a smaller number of more accurate estimates. By a careful balancing of the different levels involved, a higher accuracy can be achieved for a fixed computational cost than could have been achieved with a standard Monte Carlo estimator.

Theorem 1.3.1 (MLMC complexity (Theorem 1 of [24])). *Let P denote a random variable, and let P_ℓ denote the corresponding ℓ level approximation.*

If there exists independent estimators Y_ℓ based on N_ℓ Monte Carlo samples (as defined in (1.7)*), and positive constants $\alpha, \beta, \gamma, c_1, c_2, c_3$ such that $\alpha \geqslant \frac{1}{2}\min(\beta, \gamma)$ and*

1. $|\mathbb{E}[P_\ell - P]| \leqslant c_1 2^{\alpha\ell}$,
2. $\mathbb{E}[Z_\ell] = \begin{cases} \mathbb{E}[P_0] \ , & \ell = 0 \\ \mathbb{E}[P_\ell - P_{\ell-1}] \ , & \ell > 0 \end{cases}$
3. *Var*$(Z_\ell) \leqslant c_2 N_\ell^{-1} 2^{-\beta\ell}$
4. $\mathbb{E}[C_\ell] \leqslant c_3 N_\ell 2^{\gamma^\ell}$

where C_ℓ is the computational cost of Y_ℓ. Then there exists a positive constant c_4 such that for any $\varepsilon < e^{-1}$ there are values L and N_ℓ for which the MLMC estimator (1.8) *has a mean square error with bound*

$$\mathbb{E}\Big[\Big(\bar{\psi}^{MLMC}(N_0,\dots,N_L)-\mathbb{E}[P]\Big)^2\Big] < \epsilon^2 \ ,$$

with a computational complexity C with bound

$$\mathbb{E}[C] \leqslant \begin{cases} c_4\varepsilon^{-2} \ , & \beta > \gamma \ , \\ c_4\varepsilon^{-2}(\log(\varepsilon))^2 & \beta = \gamma \ , \\ c_4\varepsilon^{-2-(\gamma-\beta)/\alpha} \ , & \beta < \gamma \ . \end{cases}$$

Let V_ℓ denote the variance of $P_\ell - P_{\ell-1}$ (which we estimate) and C_ℓ the cost to produce a single sample of P_ℓ To achieve a total variance of $\varepsilon^2/2$, the optimal choice of L_ℓ is given by

$$N_\ell = \left\lceil 2\epsilon^{-2}\sqrt{V_\ell/C_\ell}\left(\sum_{\ell=0}^{L}\sqrt{V_\ell C_\ell}\right)\right\rceil \ , \tag{1.9}$$

and the total computational cost is therefore

$$C = \varepsilon^{-2}\left(\sum_{\ell=0}^{L}\sqrt{V_\ell C_\ell}\right)^2 \ . \tag{1.10}$$

The constants α and β appearing in the MLMC complexity theorem (Theorem 1.3.1) are related to the 'weak' and 'strong' error rates of the numerical scheme used to generate the samples $P_\ell^{(\ell,i)}$. For $P_\ell^{(\ell,i)} := \psi(Y_N^{(\ell,i)})$, where $Y_N^{(\ell,i)}$ is a sample generated by our chosen numerical scheme with time step h_ℓ, α is the weak error rate of the numerical scheme. And when the functional ψ is assumed to be Lipschitz with Lipschitz constant L we have that

$$\text{Var}(Z_\ell) \leqslant N_\ell^{-1}\mathbb{E}\left[\|\psi(Y_N^{(\ell,i)}) - \psi(Y_N^{(\ell-1,i)})\|^2\right] \leqslant LN_\ell^{-1}\ \mathbb{E}\left[\|Y_N^{(\ell,i)} - Y_N^{(\ell-1,i)}\|^2\right] \ ,$$

by the triangular inequality (adding and subtracting y) we can then link the right hand side of this to the strong error as defined in (1.3). A numerical scheme with a high strong order will thus reduce the variance at each level in the MLMC. This will reduce the number of samples needed at each level and reduce the number of levels required. We demonstrate this in Section 6.3 using one of the high order splitting schemes presented in this thesis. The parameter γ is related to how the computational cost C_ℓ increases with ℓ.

1.4 A brief introduction to tensors

The central object of our study will be the iterated integrals of the time-Brownian path $\{(t, W_t)\}_t$. These integrals are 'tensors'. For our purposes (working in $\mathbb{R}^e$) it is sufficient (and perhaps more enlightening) to adopt the following interpretation of 'tensor' common in the context of machine learning:

Definition 1.4.1. *A tensor is a multidimensional array (or matrix).*

With this perspective, we can easily isolate scalar components of the tensor by indexing. We refer to a tensor which requires m indices to isolate a scalar component as an 'm-tensor'.

Example 1.4.2. *As an example, for the* 3*-tensor*

$$A := \left(\begin{pmatrix} a_{111} & a_{121} \\ a_{211} & a_{221} \\ a_{311} & a_{321} \end{pmatrix} , \quad \begin{pmatrix} a_{112} & a_{122} \\ a_{212} & a_{222} \\ a_{312} & a_{322} \end{pmatrix} \right) ,$$

$A_{111} = a_{111}$*,* $A_{112} = a_{112}$ *,* $A_{321} = a_{321}$ *and* $A_{122} = a_{122}$*.* A*, here, is in* $\mathbb{R}^3 \otimes \mathbb{R}^2 \otimes \mathbb{R}^2$*.*

At times it will also useful to refer to the columns of a tensor, in which case we drop the first index. So, when indexing an m-tensor by $m-1$ indices we are referring to the columns of the tensor (one could extend this convention to refer to matrices, lists of matrices etc.). For the above example, we would write

$$A_{11} = \begin{pmatrix} a_{111} \\ a_{211} \\ a_{311} \end{pmatrix} , \quad A_{21} = \begin{pmatrix} a_{121} \\ a_{221} \\ a_{321} \end{pmatrix} , \quad A_{12} = \begin{pmatrix} a_{112} \\ a_{212} \\ a_{312} \end{pmatrix} \quad \text{and} \quad A_{12} = \begin{pmatrix} a_{122} \\ a_{222} \\ a_{322} \end{pmatrix} .$$

Remark 1.4.3. *An* m*-tensor* A *is* $\mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \otimes \cdots \otimes \mathbb{R}^{d_m}$*-valued. As the spaces are isomorphic, we will often instead write that* $A \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_m}$*.*

We will now define two different notions of multiplication for tensors. The first of these is a generalisation of matrix multiplication:

Definition 1.4.4 (Matrix multiplication of tensors)**.** *Let* $A \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_m}$ *be an* m*-tensor and* $B \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_n}$ *be an* n*-tensor. If* $(d_m, \ldots, d_{m-k+1}) = (p_1, \ldots, p_k)$ *for* $k < m$ *and* $k \leqslant n$*, then* AB *is an* $(m+n-2k)$*-tensor and*

$$\big(AB\big)_{\alpha\beta} = \sum_{i_{m-k+1}=1}^{d_{m-k+1}} \cdots \sum_{i_m=1}^{d_m} A_{\alpha i_{m-k+1} \cdots i_m} B_{i_m \cdots i_{m-k+1} \beta} ,$$

where $\alpha := i_1 \cdots i_{m-k+2}$ *and* $\beta := j_{k+1} \cdots j_n$*. If* $n = k$*, then* $\beta := \{\}$ *(the empty index).*

In the above definition, it should be emphasised that the matrix multiplication of tensors is only defined for *compatible* tensors. So, for $A \in \mathbb{R}^{2\times 3}$, $B \in \mathbb{R}^{3\times 2}$ and $C \in \mathbb{R}^{d\times 2\times 3}$, we have that $AB \in \mathbb{R}^{2\times 2}$, $BA \in \mathbb{R}^{3\times 3}$, $CB \in \mathbb{R}^{d}$ and CA is not defined. The multiplications AC and BC are defined if $d = 3$ or $d = 2$, respectively. One can also encode 'vector' multiplication in this way by 're-indexing': where we understand that column vectors are $\mathbb{R}^{d\times 1}$-valued and row vectors are $\mathbb{R}^{1\times d}$ valued. To demonstrate the matrix multiplication of compatible tensors we give the following example:

Example 1.4.5. *Taking A as defined in Example 1.4.2 and*

$$B := \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \;, \quad \textit{we have} \quad AB = \begin{pmatrix} a_{111}b_{11} + a_{121}b_{12} + a_{112}b_{21} + a_{122}b_{22} \\ a_{211}b_{11} + a_{221}b_{12} + a_{212}b_{21} + a_{222}b_{22} \\ a_{311}b_{11} + a_{321}b_{12} + a_{312}b_{21} + a_{322}b_{22} \end{pmatrix} .$$

Remark 1.4.6. *With the notion of multiplication as defined in Definition 1.4.4, we may view a tensor $A \in \mathbb{R}^{d_1\times\cdots\times d_m}$ as a mulilinear map from $\mathbb{R}^{d_k\times\cdots\times d_m}$ to $\mathbb{R}^{d_1\times\cdots\times d_{k-1}}$ (for any $2 \leqslant k \leqslant m$). When this understanding is useful, we write $A \in L\big(\mathbb{R}^{d_k\times\cdots\times d_m}, \mathbb{R}^{d_1\times\cdots\times d_{k-1}}\big)$, where $L(E, F)$ is the space of multilinear maps from E to F.*

The second notion of multiplication we are interested in is the 'tensor product' between two tensors, which can be defined as follows:

Definition 1.4.7. *For two tensors $A \in \mathbb{R}^{d_1\times\cdots\times d_m}$ and $B \in \mathbb{R}^{p_1\times\cdots\times p_n}$ we define the tensor product $\otimes : \mathbb{R}^{d_1\times\cdots\times d_m} \times \mathbb{R}^{p_1\times\cdots\times p_n} \to \mathbb{R}^{d_1\times\cdots\times d_m\times p_1\times\cdots\times p_n}$ component wise by*

$$\big(A \otimes B\big)_{i_1\cdots i_m j_1\cdots j_n} := A_{i_1\cdots i_m} B_{j_1\cdots j_n} \;.$$

For an m-tensor A and an n-tensor B, $A \otimes B$ is an $(m + n)$-tensor. Multiplication by scalars is understood component wise.

Writing explicit examples for the tensors product could quickly become cumbersome, so we will only provide the following two simple illustrations:

Example 1.4.8. *Let*

$$x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;, \qquad y := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \qquad \textit{and} \qquad z := \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \;,$$

then we have that

$$x \otimes y = \begin{pmatrix} x_1y_1 & x_1y_2 \\ x_2y_1 & x_2y_2 \end{pmatrix} \qquad \textit{and}$$

$$x \otimes y \otimes z = \left(\begin{pmatrix} x_1y_1z_1 & x_1y_2z_1 \\ x_2y_1z_1 & x_2y_2z_1 \end{pmatrix} , \begin{pmatrix} x_1y_1z_2 & x_1y_2z_2 \\ x_2y_1z_2 & x_2y_2z_2 \end{pmatrix} \right) .$$

For two tensors of the same dimension (living in the same space) we understand addition in the direct element wise sense. One can see that these two operations endow the space of tensors with an associative algebra.

Proposition 1.4.9. *The tensor product $\otimes$ is right and left distributive, and associative. That is, for tensors $A, B, C \in (\mathbb{R}^e)^{\otimes m}$,*

$$(A+B)\otimes C = A\otimes C + B\otimes C \ , \tag{1.11}$$

$$A\otimes(B+C) = A\otimes B + A\otimes C \ , \tag{1.12}$$

$$(A\otimes B)\otimes C = A\otimes(B\otimes C) \ . \tag{1.13}$$

The tensor product is not commutative.

Proof. These properties are clear from the definition of the tensor sum as component wise scalar addition, and the tensor product as a component wise scalar multiplication. That the tensor product is not commutative should be clear from Example 1.4.8. □

Properties (1.11) and (1.12) hold separately so long as the addition is defined (i.e. when $A, B \in (\mathbb{R}^e)^{\otimes m}$ or $B, C \in (\mathbb{R}^e)^{\otimes m}$). The associative property (1.13) holds in general.

Remark 1.4.10. *When we write iterated integrals in tensor form it should be understood in an element wise sense. So, formally, for*

$$dX := \begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix} \quad \text{and} \quad dY := \begin{pmatrix} dy_1 \\ dy_2 \end{pmatrix} \ ,$$

$$\int\!\!\int dX\otimes dY = \int\!\!\int \big(dX\otimes dY\big) = \begin{pmatrix} \int\!\int dx_1dy_1 & \int\!\int dx_1dy_2 \\ \int\!\int dx_2dy_1 & \int\!\int dx_2dy_2 \end{pmatrix} \ .$$

1.5 Notation and spaces

Here we introduce some notation and spaces that will be used throughout the text.

Let $\mathcal{C}^k(\mathbb{R}^e, \mathbb{R}^d)$ denote the space of k-times continuously differentiable functions from R^e to $\mathbb{R}^d$, and let $\mathcal{C}^k_{Lip}(\mathbb{R}^e, \mathbb{R}^d) \subseteq \mathcal{C}^k(\mathbb{R}^e, \mathbb{R}^d)$ denote its subspace of Lipschitz functions. Throughout, $\|\cdot\|$ will denote the standard Euclidean norm on $\mathbb{R}^n$, and by $L^p(\mathbb{R}^e)$ we denote the space of $\mathbb{R}^e$-valued p-integrable random variables (i.e. $\mathbb{E}\big[\|X\|^p\big] < \infty$). Given a continuous path $\gamma : [0,1] \to \mathbb{R}^n$, we will denote its length using the notation,

$$\|\gamma\|_{\text{1-var},[0,1]} = \int_0^1 |d\gamma(r)| := \sup_{\substack{0=r_0<r_1<\cdots<r_N=1,\\ N\geqslant 0.}} \bigg(\sum_{i=0}^{N-1} \|\gamma(r_{i+1}) - \gamma(r_i)\|\bigg).$$

Chapter 2

Splitting methods as piecewise linear paths

2.1 Introduction

We present a study of high order path-based splitting methods for Stratonovich SDEs of the form

$$dy_t = f(y_t)dt + g(y_t) \circ dW_t, \quad y_0 \in L^2(\mathbb{R}^e), \tag{2.1}$$

where $W = (W^1, \cdots, W^d) = \{W_t\}_{t\in[0,T]}$ denotes a d-dimensional Brownian motion, and the vector fields are given by $f \in \mathcal{C}^2(\mathbb{R}^e, \mathbb{R}^e)$ and $g = (g_1, \cdots, g_d) \in \mathcal{C}^3(\mathbb{R}^e, \mathbb{R}^{e\times d})$ where we understand $g(y_t) \circ dW_t = \sum_{i=1}^d g_i(y_t) \circ dW_t^i$. We focus on Stratonovich SDEs for algebraic reasons: as we will make clear later the text, Stratonovich integrals yield the 'same' chain rule and integration by parts formula as Riemann-Stieltjes integrals, aiding our analysis. The schemes we propose are of course still applicable to Itô SDEs, after the appropriate correction term is added to convert the SDE to Stratonovich form.

We are particularly interested in SDEs satisfying the following commutativity condition, as this will help us obtain higher order convergence,

$$g_i'(y)g_j(y) = g_j'(y)g_i(y) \ , \quad \forall y \in \mathbb{R}^e \ . \tag{2.2}$$

Where the columns $\{g_i\}_{1\leqslant i\leqslant d}$ of g can each be viewed as a vector field on $\mathbb{R}^e$. We also assume g_i are globally Lipschitz continuous with globally Lipschitz derivatives.

Without the condition (2.2), high order numerical methods for SDEs require the use, or approximation, of second iterated integrals of the Brownian motion [61]. Generating both the increments and iterated integrals, or equivalently Lévy areas, of Brownian motion is a difficult problem [7, 9] and beyond the scope of this study. We refer the reader to [8, 14, 15, 22, 52, 68] for studies on Lévy area approximation. Nevertheless, there is a large variety of

SDEs used in applications that satisfy (2.2), such as SDEs with scalar, diagonal or additive noise. While we focus on schemes for SDEs satisfying the commutativity condition (2.2), the error analysis that we introduce for establishing convergence is generic and does not rely on this condition.

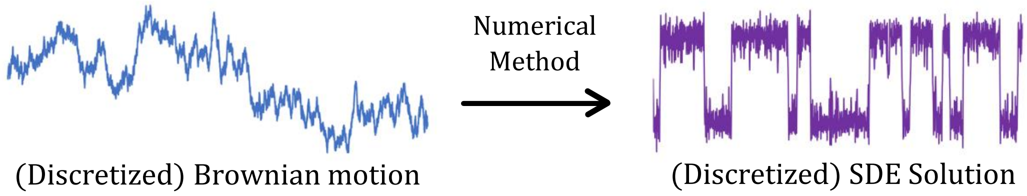


Fig. 2.1: In the Monte Carlo paradigm, information about the Brownian motion is generated and then mapped to a numerical solution of the SDE. Typically, only Brownian increments are sampled.

2.1.1 Overview of part I

Our aim (in the first part of this thesis) is to present a unified framework for the design and error analysis of splitting schemes. This analysis is built on a novel perspective which views splitting schemes as replacing the driving time-Brownian path $\{(t, W_t)\}_{t\in[0,T]}$ of the SDE (2.1) with a piecewise linear path with random coefficients (see Section 2.2). With this perspective we can compare the solutions of the SDE and the differential equation driven by this 'splitting path' by Taylor expanding both. As we show in Section 2.4, these Taylor expansions are remarkably similar. And, it turns out that the key difference between these two Taylor expansions is the iterated integrals which appear in them. By constructing a 'splitting path' whose iterated integrals (*path* integrals) 'match' the iterated time-Brownian integrals we can obtain a high order of numerical convergence. In Section 2.3 we introduce assumptions on the splitting paths needed to enable our analysis.

In Chapter 3 we focus our attention on these iterated integrals. In Section 3.1 we look at the algebraic properties of the iterated integrals, where we see that both types of iterated integrals share the same integration by parts formula. This enables us to isolate the symmetric and anti-symmetric components of the iterated integrals which will ultimately allow us to prove the importance of the commutativity condition (2.2) in Theorem 3.6.2. In Section 3.2 we introduce Lévy Area's of the Brownian motion, these can be viewed as components of the iterated integrals and will be used to construct the high order splitting schemes we propose. In Section 3.3 we calculate the values of the iterated Brownian integrals, showing us the values which the path integrals need to match. In Section 3.4 we present an algorithmic approach to automate the calculation of the iterated path integrals, this will speed up our comparisons. In Section 3.5 we obtain L^2 estimates for the iterated path integrals, these will be used for our strong error analysis. We conclude the chapter in Section 3.6 where we show that the commutativity condition (2.2) simplifies the Taylor expansions making it easier to obtain high order convergence.

In Chapter 4 we present our error analysis of path-based splitting schemes. Section 4.1 presents our local error analysis which will be used in Sections 4.2 and 4.3 to obtain global strong and weak error rates. Our main convergence results are given in Theorems 4.2.1 and 4.3.1.

In Chapter 5 we present new high order path-based splitting using the Lévy areas of Section 3.2. We apply our error analysis to these paths in Section 5.2 and in Section 5.3 we show how these paths were derived.

We conclude our study in Chapter 6, applying the path-based splitting schemes to several numerical examples.

2.2 The path perspective

Inspired by rough path theory [20], which views SDEs as functions that map Brownian motion to continuous paths (see Figure 2.1), we propose an approximation $y^\gamma = \{y_r^\gamma\}_{r\in[0,1]}$ for (2.1) that comes from the controlled differential equation (CDE),

$$dy_r^\gamma = f(y_r^\gamma)\, d\gamma^\tau(r) + g(y_r^\gamma)\, d\gamma^\omega(r), \qquad y_0^\gamma = y_0, \tag{2.3}$$

or equivalently

$$y_r^\gamma = y_0^\gamma + \int_0^r f(y_u^\gamma)\, d\gamma^\tau(u) + \int_0^r g(y_u^\gamma)\, d\gamma^\omega(u),$$

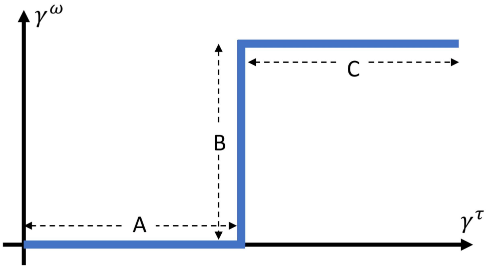
where $\gamma = (\gamma^\tau, \gamma^\omega)^\top : [0,1] \to \mathbb{R}^{1+d}$ is a parameterised (continuous) piecewise linear path designed to match certain iterated integrals of the 'space-time' Brownian motion $\{(t, W_t)\}_{t\in[0,T]}$. Since the path γ is piecewise linear, it immediately follows that

$$d\gamma(r) = \frac{1}{r_{i+1} - r_i}\gamma_{r_i,r_{i+1}}\, dr \ ,$$

for $r \in [r_i, r_{i+1}]$. Where $r_i \in [0,1]$ is the parameter value at the start of the i'th piece of γ, and $\gamma_{r_i,r_{i+1}}$ is the increment of the linear piece. Therefore the CDE (2.3) reduces to a sequence of ODEs, corresponding to each piece of γ. These ODEs can then be discretized by a suitable ODE solver, such a Runge-Kutta method. Furthermore, we will see that this approach can be interpreted as a splitting method. We thus refer to these paths as 'splitting paths'. We refer the reader to [4, Section 3] for an overview of splitting methods for SDEs.

Example 2.2.1. *Let $\gamma = (\gamma^\tau, \gamma^\omega)^\top : [0,1] \to \mathbb{R}^{1+d}$ denote a piecewise linear path where the vertices between pieces are at $r_i := \frac{i}{3}$ for $0 \leqslant i \leqslant 3$ and the increments are*

$$\gamma_{r_i, r_{i+1}} = \begin{cases} (A, 0), & \textit{if } i = 0 \\ (0, B), & \textit{if } i = 1 \\ (C, 0), & \textit{if } i = 2\ . \end{cases}$$



Replacing the driving Brownian path $t \to (t, W_t)$ in the SDE (2.1) *with the parametrisation $r \to \gamma_r$, the approximating CDE* (2.3) *reduces to the sequence of ODEs*

$$\begin{aligned} dy_r^\gamma &= f(y_r^\gamma) \times 3A dr\ , \quad \textit{for} \quad r \in \left[0, \frac{1}{3}\right] \\ dy_r^\gamma &= g(y_r^\gamma) \times 3B dr\ , \quad \textit{for} \quad r \in \left[\frac{1}{3}, \frac{2}{3}\right]\ , \qquad \textit{with} \quad y_0^\gamma = y_0\ . \\ dy_r^\gamma &= f(y_r^\gamma) \times 3C dr\ , \quad \textit{for} \quad r \in \left[\frac{2}{3}, 1\right] \end{aligned}$$

Practically, it makes more sense to solve the sequence of ODEs each over the interval $[0,1]$. Thus removing the artificial scaling factor of 3 (corresponding to the number pieces in the path). In which case, we may formulate the CDE driven by path γ as the splitting

$$y_1^\gamma = \exp\Big(f(\cdot)A\Big) \exp\Big(g(\cdot)B\Big) \exp\Big(f(\cdot)C\Big) y_0^\gamma\ ,$$

where $\exp(V)x$ denotes the solution z_1 at time $u = 1$ of the ODE $z' = V(z),\ z(0) = x$.

More generally, CDEs are one of the key objects in rough path theory [20, 21, 43] (often referred to as 'rough' differential equations). However, we emphasise that while our approach takes inspiration from rough paths we are not working with rough paths – and no p-variation or lift maps are used. Instead, we will heavily draw upon ideas and interpretations from rough path theory. Similarly, we point towards [5, 17, 30, 32, 50, 51, 58] as works presenting results for stochastic processes or continuous data streams, without 'rough path' statements, but making use of the machinery and insights that are provided by rough path theory.

In terms of proof methodologies for the error analysis, it is worth noting that ours differs from previous works on splitting methods for SDEs [4, 16, 29, 38, 49, 53, 55, 65], which are either extensions of the Strang splitting [64] or use the Baker-Campbell-Hausdorff formula for expanding the compositions of ODEs (see [49] for the latter). For some perspective, we present an informal version of our main result, Theorem 4.2.1, which describes our approach to high order splitting methods for commutative SDEs.

Theorem 2.2.2 (Convergence of path-based splitting for SDEs (informal version))**.** *Given a fixed number of steps N, we will define a numerical solution $Y = \{Y_k\}_{0\leqslant k\leqslant N}$ for the SDE (2.1) over the finite time horizon $[0,T]$ as follows,*

$$Y_{k+1} := \big(\textit{Solution at time } r = 1 \textit{ of CDE (2.3) driven by } \gamma_k : [0,1] \to \mathbb{R}^{1+d}\big)\big(Y_k\big),$$

where each piecewise linear path γ_k is constructed from $\big\{W_t \,:\, t \in \big[\frac{kT}{N}, \frac{(k+1)T}{N}\big]\big\}$, is sufficiently regular (see Assumption 2.3.1), and for some fixed $p \in \{\frac{m}{2}\}_{m\in\mathbb{N}}$ satisfies

1. *the iterated integrals of γ_k and (t, W_t) with order[1] less than $p - \frac{1}{2}$ coincide,*
2. *the iterated integrals of γ_k and (t, W_t) with order p match in expectation.*

Then, there exists a constant $C > 0$, such that for sufficiently small $h = \frac{T}{N}$, we have

$$\mathbb{E}\Big[\|Y_k - y_{kh}\|^2\Big]^{\frac{1}{2}} \leqslant Ch^{p-\frac{1}{2}}. \tag{2.4}$$

for $k \in \{1, \cdots, N\}$. If $p = 2$, and the SDE satisfies the commutativity condition (2.2), then the estimate (2.4) holds under the assumption that each γ_k is sufficiently regular and has coordinate processes $\{\gamma_k^{\omega,i}\}_{1\leqslant i\leqslant d}$ that are independent, symmetric and satisfy

$$\gamma_k^{\omega,i}(1) - \gamma_k^{\omega,i}(0) = W^i_{kh,(k+1)h}, \qquad \gamma_k^{\tau}(1) - \gamma_k^{\tau}(0) = h, \tag{2.5}$$

$$\int_0^1 \big(\gamma_k^{\omega,i}(r) - \gamma_k^{\omega,i}(0)\big)\, d\gamma_k^{\tau}(r) = \int_{kh}^{(k+1)h} W^i_{kh,u}\, du, \tag{2.6}$$

$$\mathbb{E}\bigg[\int_0^1 \big(\gamma_k^{\omega,i}(r) - \gamma_k^{\omega,i}(0)\big)^2 d\gamma_k^{\tau}(r)\bigg] = \frac{1}{2}h^2. \tag{2.7}$$

[1] For iterated integrals of the Brownian motion or path γ we denote by 'order' the size of the integral with respect to the interval $t - s$. Each dW_t adds $1/2$, and each dt adds 1 to the order. This is made precise in Section 2.4.

2.3 Assumptions on the splitting path and a preliminary result

Here we introduce our main assumption on the splitting path, namely that it scales like Brownian motion. This assumption is easily justified as the splitting path is replacing and estimating a Brownian motion.

Assumption 2.3.1 (Brownian-like scaling)**.** *Let $\gamma = (\gamma^\tau, \gamma^\omega)^\top : [0,1] \to \mathbb{R}^{1+d}$ be a piecewise linear path with $m \in \mathbb{N}$ components that have, almost surely, finite length. For $i \geqslant 0$, we denote the increment of the i'th piece of γ by $\gamma_{r_i,r_{i+1}}$ and assume that*

1. *$\gamma^\tau_{r_i,r_{i+1}}$, the increment in the time component of γ, is deterministic.*
2. *$\gamma^\tau_{r_i,r_{i+1}}$ scales with the step size h and the increment in the space component, $\gamma^\omega_{r_i,r_{i+1}}$, has finite even moments scaling with h. Concretely, we have*

$$\gamma^\tau_{r_i,r_{i+1}} = O(h), \quad \textit{and} \quad \mathbb{E}\big[|(\gamma^\omega_{r_i,r_{i+1}})_j|^{2k}\big] = O(h^k),$$

for every $j \in \{1, \dots, d\}$.

Remark 2.3.2 (Comment on Assumption 2.3.1)**.** *We impose that γ^τ is deterministic for convenience and, inspecting the proof, one may be able to lift this constraint. Moreover, we expect our methodology can accommodate for randomised algorithms (see [3, 27, 36, 37, 62] for examples of SDE solvers with a randomised time component).*

We now present a moment bound for the CDE, which will be used to control remainder terms of the Taylor expansion discussed later. Following the approach of [20, Theorem 3.7], we obtain the following result, Theorem 2.3.3.

Theorem 2.3.3 (Fourth moment bound for CDEs)**.** *Let γ satisfy Assumption 2.3.1 and let y^γ denote the solution to* (2.3) *with $y_0^\gamma \in L^4(\mathbb{R}^e)$. Suppose that f and g satisfy*

$$\|f(y)\| \leqslant C(1 + \|y\|), \quad \textit{and} \quad \|g(y)\| \leqslant C(1 + \|y\|), \tag{2.8}$$

with $\mathbb{E}\big[\exp\big(16C\int_0^1 |d\gamma(u)|\big)\big] < \infty$. Then there exists a positive constant $\widetilde{C} > 0$, depending only the path γ and the growth constant C in (2.8), such that for $r \in [0,1]$,

$$\mathbb{E}\big[\|y_r^\gamma - y_0^\gamma\|^4\big] \leqslant \widetilde{C}h^2\big(1 + \mathbb{E}\big[\|y_0\|^4\big]\big). \tag{2.9}$$

Proof. Let $G : \mathbb{R}^e \to \mathbb{R}^{e\times(d+1)}$ have first column given by $f : \mathbb{R}^e \to \mathbb{R}^e$ and the rest of the matrix given by $g : \mathbb{R}^e \to \mathbb{R}^{e\times d}$. Then the growth assumption (2.8) implies that $\|G(y)\| \leqslant C(1 + \|y\|)$. Thus, by direct application of [20, Theorem 3.7], we have

$$\|y_r^\gamma - y_0^\gamma\| \leqslant C(1 + \|y_0^\gamma\|)\exp\left(2C\int_0^r |d\gamma(u)|\right)\int_0^r |d\gamma(u)|,$$

for $r \in [0,1]$. Consequently,

$$\begin{aligned}\mathbb{E}\big[\|y_r^\gamma - y_0^\gamma\|^4\big] &\leqslant C^4\mathbb{E}\Bigg[\big(1+\|y_0^\gamma\|\big)^4\exp\bigg(8C\int_0^r |d\gamma(u)|\bigg)\bigg(\int_0^r |d\gamma(u)|\bigg)^4\Bigg]\\ &\leqslant C^4\big(1+\mathbb{E}\big[\|y_0^\gamma\|^4\big]\big)\mathbb{E}\bigg[\exp\bigg(16C\int_0^r |d\gamma(u)|\bigg)\bigg]^{\frac{1}{2}}\mathbb{E}\bigg[\bigg(\int_0^r |d\gamma(u)|\bigg)^8\bigg]^{\frac{1}{2}}\\ &\leqslant \widetilde{C}h^2\big(1+\mathbb{E}\big[\|y_0\|^4\big]\big),\end{aligned}$$

where we used the independence of y_0 and γ, the Cauchy-Schwarz inequality and

$$\begin{aligned}\mathbb{E}\bigg[\bigg(\int_0^r |d\gamma(u)|\bigg)^8\bigg] \leqslant \mathbb{E}\bigg[\bigg(\int_0^1 |d\gamma(u)|\bigg)^8\bigg] &= \mathbb{E}\bigg[\bigg(\sum_{i=0}^{m-1}\|\gamma_{r_i,r_{i+1}}\|\bigg)^8\bigg]\\ &\leqslant m^7\sum_{i=0}^{m-1}\mathbb{E}\Big[\|\gamma_{r_i,r_{i+1}}\|^8\Big] = O(h^4),\end{aligned}$$

which follows by Jensen's inequality and Assumption 2.3.1. □

Remark 2.3.4. *The assumptions we make on the functions f and g are more relaxed than the assumptions of [17] (which assumes $\mathcal{C}^\infty$ and bounded), where the optimal estimates of Brownian integrals used in this work were originally derived.*

2.4 Stochastic Taylor expansions

Our approach to the error analysis of splitting schemes is to Taylor expand both the SDE (2.1) and the CDE (2.3) (driven by a splitting path γ), and to compare the corresponding iterated integrals from each. In this section we present these Taylor expansions. By matching the lower order terms in the Taylor expansions and showing that the remainder terms are higher order, we can bound local errors for our splitting schemes. We then apply Milstein and Tretyakov's framework for mean-square error analysis [48] to obtain a global strong convergence rate – which is our main result in Theorem 4.2.1. Similarly, by a telescoping sum argument, we obtain the global weak error – presented in Theorem 4.3.1.

First, let us introduce some short hand notation to aid presentation. We encode the order of integration (dW vs dt) by the multi-index (word) $\alpha = (\alpha_1, \cdots, \alpha_n) \in \{\tau, \omega\}^n$, and define the iterated integrals of W and $\gamma : [0,1] \to \mathbb{R}^{1+d}$ as

$$I_\alpha(F) := \int_0^h\int_0^{r_1}\cdots\int_0^{r_{n-1}} F(y_{r_n})\,dB_{r_n}^{\alpha_1}dB_{r_{n-1}}^{\alpha_2}\cdots dB_{r_2}^{\alpha_{n-1}}dB_{r_1}^{\alpha_n}, \tag{2.10}$$

$$I_\alpha^\gamma(F) := \int_0^1\int_0^{r_1}\cdots\int_0^{r_{n-1}} F(y_{r_n}^\gamma)\,d\gamma^{\alpha_1}(r_n)\cdots d\gamma^{\alpha_n}(r_1), \tag{2.11}$$

where y and y^γ are the solutions of the SDE (2.1) and its CDE approximation (2.3), $F : \mathbb{R}^e \to \mathbb{R}^e$, $dB_r^\tau = dr$, $dB_r^\omega = \otimes \circ dW_r$ and, if $\gamma(r) = (\gamma^\tau(r), \gamma^\omega(r))^\top$, then we write $d\gamma(r)^\tau = d\gamma^\tau(r)$ and $d\gamma(r)^\omega = \otimes\, d\gamma^\omega(r)$. We denote the set of words by $\mathcal{A} = \cup_{n\geqslant 0}\{\tau,\omega\}^n$. We also define the integrals,

$$J_\alpha(F) := I_\alpha(F) - F(y_0)I_\alpha(1), \qquad J_\alpha^\gamma(F) := I_\alpha^\gamma(F) - F(y_0^\gamma)I_\alpha^\gamma(1), \tag{2.12}$$

where we understand $I_\alpha(1)$ and $I_\alpha^\gamma(1)$ as defined in (2.10) and (2.11), but with $F(y)$ replaced by the scalar 1. For a given word $\alpha = (\alpha_i)_{1\leqslant i\leqslant n}$, we will define its order by $\text{ord}(\alpha) := |\alpha|_\tau + \frac{1}{2}|\alpha|_\omega$, where $|\alpha|_\tau := \sum_{i=1}^n \mathbf{1}_{\alpha_i=\tau}$ and $|\alpha|_\omega := \sum_{i=1}^n \mathbf{1}_{\alpha_i=\omega}$. We denote the length of a word by $|\alpha|$.

2.4.1 Stratonovich Taylor expansion

Letting y denote the solution to (2.1), we have the following chain rule (see [20, Exercise 3.17]) for any function $F \in \mathcal{C}^1(\mathbb{R}^e)$,

$$F(y_r) = F(y_0) + \int_0^r F'(y_s) \circ dy_s, \quad r \geqslant 0\ . \tag{2.13}$$

By expanding 'dy_s' and iteratively applying (2.13), we obtain the Taylor expansion.

Proposition 2.4.1 (Stochastic Taylor expansion of the Stratonovich SDE (2.1) [2, Proposition 1.1], [35, Theorem 5.6.1]). *Let $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, $f \in \mathcal{C}_{\text{Lip}}^{[p-1]}(\mathbb{R}^e, \mathbb{R}^e)$, $g \in \mathcal{C}_{\text{Lip}}^{2p-1}(\mathbb{R}^e, \mathbb{R}^{e\times d})$ and $\psi \in \mathcal{C}_{\text{Lip}}^{2p}(\mathbb{R}^e, \mathbb{R}^l)$. The Stratonovich Taylor expansion of (2.1), up to order p, is*

$$\psi(y_h) = \psi(y_0) + \sum_{\substack{\alpha\in\mathcal{A}, \\ ord(\alpha)\leqslant p}} V_\psi(\alpha)(y_0)I_\alpha(1) + R_{p,\psi}(h, y_0), \tag{2.14}$$

where, we recall the definition of $\text{ord}(\alpha) := |\alpha|_\tau + \frac{1}{2}|\alpha|_\omega$ after equation (2.12), and

$$R_{p,\psi}(h, y_0) := \sum_{\substack{\alpha\in\mathcal{A}, \\ ord(\alpha)=p}} J_\alpha(V_\psi(\alpha)), \tag{2.15}$$

with the vector field derivatives $V_\psi(\alpha) : \mathbb{R}^e \to L((\mathbb{R}^d)^{\otimes|\alpha|_\omega}, \mathbb{R}^l)$ defined for multi-indices recursively by $V_\psi(\tau)(y) := \psi'(y)f(y)$, $V_\psi(\omega)(y) := \psi'(y)g(y)$ and

$$V_\psi(l\beta)(y) = V_\psi(\beta)'V(l)(y), \tag{2.16}$$

where $l \in \{\tau,\omega\}$, $V(\tau)(y) := f(y)$, $V(\omega)(y) := g(y)$ and $l\beta := (l, \beta_1, \cdots, \beta_n)$ denotes concatenation. Moreover, we have

$$\mathbb{E}\big[\|R_{p,\psi}(h, y_0)\|^2\big]^{1/2} = O(h^{p+\frac{1}{2}}). \tag{2.17}$$

Informal derivation of Proposition 2.4.1. Applying the chain rule (2.13) to ψ and substituting in the SDE (2.1), we have that

$$\psi(y_h) = \psi(y_0) + \int_0^r \psi'(y_{r_1}) f(y_{r_1}) dr_1 + \int_0^r \psi'(y_{r_1}) g(y_{r_1}) \circ dW_{r_1} \ ,$$

which we may equivalently write as

$$\begin{aligned}\psi(y_h) = \psi(y_0) + \psi'(y_0) f(y_0) I_\tau(1) + \int_0^r \Big(\psi'(y_{r_1}) f(y_{r_1}) - \psi'(y_0) f(y_0)\Big) dr_1 \\ + \psi'(y_0) g(y_0) I_\omega(1) + \int_0^r \Big(\psi'(y_{r_1}) g(y_{r_1}) - \psi'(y_0) g(y_0)\Big) \circ dW_{r_1} \ .\end{aligned}$$

Applying the chain rule (2.13) to the differences contained in the integrals, we obtain

$$\begin{aligned}\psi(y_h) &= \psi(y_0) + \psi'(y_0) f(y_0) I_\tau(1) + \psi'(y_0) g(y_0) I_\omega(1) \\ &+ \int_0^r \int_0^{r_1} \big(\psi'(y_{r_2}) f(y_{r_2})\big)' \circ dy_{r_2} dr_1 + \int_0^r \int_0^{r_1} \big(\psi'(y_{r_2}) g(y_{r_2})\big)' \circ dy_{r_2} dW_{r_1} \ .\end{aligned}$$

Proceeding in this way we can obtain the Taylor expansion (2.14) up to arbitrary order p. As $\operatorname{ord}(\tau) = 1$ and $\operatorname{ord}(\omega) = 1/2$, we will not need to expand every integral at each stage. We thus require more differentiability in the diffusion matrix g than the drift vector field f. □

2.4.2 Controlled Taylor expansion

We now present a CDE Taylor expansion. Just as with the Stratonovich SDE, we have the following chain rule for $F \in \mathcal{C}^1(R^e)$,

$$F(y^\gamma_r) = F(y^\gamma_0) + \int_0^r F'(y^\gamma_s)\, dy^\gamma_s, \qquad r \in [0,1], \tag{2.18}$$

where y^γ denotes the solution to the CDE (2.3). Again, just as in the SDE setting, by expanding 'dy^γ_s' and iteratively applying (2.18), we can obtain a Taylor expansion.

Proposition 2.4.2. *Let $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, $f \in \mathcal{C}^{[p-1]}_{\mathrm{Lip}}(\mathbb{R}^e, \mathbb{R}^e)$, $g \in \mathcal{C}^{2p-1}_{\mathrm{Lip}}(\mathbb{R}^e, \mathbb{R}^{e\times d})$ and $\psi \in \mathcal{C}^{2p}_{\mathrm{Lip}}(\mathbb{R}^e, \mathbb{R}^l)$. The Taylor expansion of (2.3), up to order p, is*

$$\psi(y^\gamma_1) = \psi(y^\gamma_0) + \sum_{\substack{\alpha\in\mathcal{A}, \\ ord(\alpha)\leqslant p}} V_\psi(\alpha)(y^\gamma_0) I^\gamma_\alpha(1) + R^\gamma_{p,\psi}(h, y^\gamma_0), \tag{2.19}$$

where

$$R^\gamma_{p,\psi}(h, y^\gamma_0) := \sum_{\substack{\alpha\in\mathcal{A}, \\ ord(\alpha)=p}} J^\gamma_\alpha(V_\psi(\alpha)), \tag{2.20}$$

with V_ψ defined as in Proposition 2.4.1

Remark 2.4.3. *We note the proof of Proposition 2.4.2 is essentially identical to that of the Stratonovich Taylor expansion (but with $t \mapsto (t, W_t)$ replaced by $r \mapsto (\gamma_r^\tau, \gamma_r^\omega)$). Moreover, the Taylor expansions are of the same form. The only difference between the Taylor expansions is that the Stratonovich Taylor expansion involves the integrals $I_\alpha(1)$ and the CDE Taylor expansion contains the integrals $I_\alpha^\gamma(1)$. Thus, we are able to focus largely on comparing these iterated integrals. By matching all of the iterated integrals up to a given level, we will in turn match the Taylor expansion up to that level.*

Chapter 3

Iterated integrals of Brownian motion and Piecewise linear paths

Central to our analysis of splitting methods are the iterated integrals with respect to the Brownian path driving the SDE (2.1) and the splitting path driving the CDE (2.3). These objects appear in the Taylor expansions of the SDE and CDE (Section 2.4) which, as discussed previously, will be key to our error analysis in Chapter 4. We thus dedicate this chapter to their study. We begin by considering the algebraic properties of the iterated integrals. These properties will allow us to prove several representations for iterated Brownian integrals in terms of the Lévy areas introduced in Section 3.2 (see Section 3.3 below), and to show how the commutativity condition (2.2) simplifies certain terms in the Taylor expansions (see Section 3.6), unlocking higher order convergence (see Theorem 4.2.3).

3.1 Algebraic properties of iterated integrals

When considering the algebraic properties of iterated integrals, it is more convenient to work with 1-D integrals than the tensor integrals we use for the Taylor expansion. So we introduce here an element-wise notation for the iterated integrals. For an element-wise representation, we must know against which component of the Brownian motion we are integrating. We will encode this by denoting $dW_t^\tau := dt$ and letting dW_t^i denote integration against the i'th dimension of the Brownian motion $W = (W^1, \cdots, W^d)$. Similarly, we write $d\gamma_r^\tau$ to denote integration with respect to the time component of γ, and $d\gamma_r^i$, the i'th dimension of the space component of γ. Let us denote the set of letters for this encoding by

$$\mathcal{A}_d = \{\tau, 1, \dots, d\} ,$$

and let $\mathcal{A}_d^*$ denote the space of words with letters from $\mathcal{A}_d$, where e is the empty word. This should not be confused with the notation introduced above in Section 2.4 where $\mathcal{A}$ denotes the space of words with letters from $\{\tau, \omega\}$. We then define the following short hand notation:

Definition 3.1.1. *For $m \geqslant 0$ and $i_1, i_2, \cdots, i_m \in \mathcal{A}_d$, we denote the (1-D) iterated Brownian integrals and iterated integrals against the components of the path γ by*

$$I_{i_1 \cdots i_m} := \int_s^t \int_s^{r_1} \cdots \int_s^{r_{m-1}} \circ\, dW_{r_m}^{i_1} \circ dW_{r_{m-1}}^{i_2} \cdots \circ dW_{r_2}^{i_{m-1}} \circ dW_{r_1}^{i_m} ,$$

$$I_{i_1 \cdots i_m}^{\gamma} := \int_s^t \int_s^{r_1} \cdots \int_s^{r_{m-1}} d\gamma_{r_m}^{i_1} d\gamma_{r_{m-1}}^{i_2} \cdots d\gamma_{r_2}^{i_{m-1}} d\gamma_{r_1}^{i_m} , \quad \textit{for} \quad 0 \leqslant s \leqslant t < \infty$$

where we understand $I_e = I_e^{\gamma} = 1$. To encode the additive property of integrals, for words $u, v \in \mathcal{A}_d^*$ and scalars $\lambda_1, \lambda_2 \in \mathbb{R}$, let

$$I_{\lambda_1 u + \lambda_2 v} := \lambda_1 I_u + \lambda_2 I_v \ .$$

We now introduce the shuffle product, which is commonly used in rough path theory [43].

Definition 3.1.2. *Let $\mathbb{R}\langle \mathcal{A}_d \rangle$ be the space of non-commutative polynomials in $\mathcal{A}_d$ with real coefficients (note that $\mathbb{R}\langle \mathcal{A}_d \rangle$ contains $\mathcal{A}_d^*$). Then the* **shuffle product** $ш : \mathbb{R}\langle \mathcal{A}_d \rangle \times \mathbb{R}\langle \mathcal{A}_d \rangle \to \mathbb{R}\langle \mathcal{A}_d \rangle$ *is the unique bilinear map such that*

$$ua \sqcup\!\sqcup\, vb = (u \sqcup\!\sqcup\, vb)a + (ua \sqcup\!\sqcup\, v)b \ ,$$

$$u \sqcup\!\sqcup\, e = e \sqcup\!\sqcup\, u = u \ ,$$

for all $u, v \in \mathbb{R}\langle \mathcal{A}_d \rangle$ and $a, b \in \mathcal{A}_d$, where we understand ua to denote concatenation.

The shuffle product of two words u and v can be viewed as the sum over all possible words constructed from the letters of u and v such that the ordering of u and v remains intact. That is, for any word w found in the summation $u \sqcup\!\sqcup\, v$, removing the letters which came from v leaves us with u and vice versa. Or symbolically $w \backslash v = u$. For two words with length m and n, the shuffle produces a sum of $\frac{(m+n)!}{m!n!}$ terms. To further demonstrate our interpretation of the shuffle product we give the following examples:

Example 3.1.3. *For a letter $a \in \mathcal{A}_d$ we have that*

$$aaa \sqcup\!\sqcup\, a = 4aaaa = a \sqcup\!\sqcup\, aaa \qquad \textit{and} \qquad aa \sqcup\!\sqcup\, aa = 6aaaa \ .$$

Example 3.1.4. *For letters $a, b, c, d \in \mathcal{A}_d$ we have that*

$$\begin{aligned} ab \sqcup\!\sqcup\, cd &= (a \sqcup\!\sqcup\, cd)b + (ab \sqcup\!\sqcup\, c)d \\ &= \big((e \sqcup\!\sqcup\, cd)a + (a \sqcup\!\sqcup\, c)d\big)b + \big((a \sqcup\!\sqcup\, c)b + (ab \sqcup\!\sqcup\, e)c\big)d \\ &= cdab + (ac + ca)db + (ac + ca)bd + abcd \\ &= cdab + acdb + cadb + acbd + cabd + abcd \ . \end{aligned}$$

With this notation, we can link the shuffle product to the integration by parts formula. As a result, the shuffle product will allow us to expand products of iterated integrals. Similarly to [14, Theorem 3.2.30] we obtain Theorem 3.1.5, but crucially we observe here that the result holds for both the Brownian integrals **and** the path integrals.

Theorem 3.1.5 (Integration by parts formula for integrals). *For all words $u, v \in \mathcal{A}_d^*$, we have*

$$I_u \cdot I_v = I_{u \sqcup\!\sqcup v} \tag{3.1}$$

Proof. We prove this result by induction over the length of the words in $\mathcal{A}_d^*$.

Base case. It is clear that the identity (3.1) holds when $u = e$ or $v = e$ since $I_e = 1$.

Induction step. Suppose that (3.1) holds for all words $u, v \in \mathcal{A}_d^*$ with a combined length less than m. Then for words $u, v \in \mathcal{A}_d^*$ and letters $a, b \in \mathcal{A}_d$ such that $|ua| + |vb| = m$, we have

$$\begin{aligned}
I_{ua} \cdot I_{vb} &= \int_s^t I_u(r) \circ dW_r^a \int_s^t I_v(r) \circ dW_r^b \\
&= \int_s^t \left(\int_s^{r_1} I_u(r_2) \circ dW_{r_2}^a \right) \circ d\left(\int_s^{r_1} I_v(r_2) \circ dW_{r_2}^b \right) \\
&\quad + \int_s^t \left(\int_s^{r_1} I_v(r_2) \circ dW_{r_2}^b \right) \circ d\left(\int_s^{r_1} I_u(r_2) \circ dW_{r_2}^a \right) \\
&= \int_s^t I_{ua}(r_1) I_v(r_1) \circ dW_{r_1}^b + \int_s^t I_{vb}(r_1) I_u(r_1) \circ dW_{r_1}^a \\
&= I_{(ua \sqcup\!\sqcup v)b + (u \sqcup\!\sqcup vb)a},
\end{aligned}$$

where the second line uses integration by parts (which holds for Stratonovich integrals) and the last line uses the induction hypothesis. The result now follows by linearity and induction. The same argument gives (3.1) for iterated integrals with respect to the path γ. □

Using Theorem 3.1.5, it will be straightforward to rewrite products of integrals as linear combinations of (high order) integrals. In addition, it shall enable us to establish decompositions of iterated integrals into symmetric (where swapping two indices gives the same value) and antisymmetric (where swapping two indices gives negative the same value) components. The following result presents these decompositions and is key to proving that the commutativity condition (2.2) simplifies certain terms in the Taylor expansions (see Section 3.6).

Theorem 3.1.6 (Symmetric and antisymmetric components of iterated integrals)**.** *Let the Lie bracket* $[\,\cdot,\cdot\,] : \mathbb{R}\langle\mathcal{A}_d\rangle \times \mathbb{R}\langle\mathcal{A}_d\rangle \to \mathbb{R}\langle\mathcal{A}_d\rangle$ *be the unique bilinear map satisfying*

$$[u, v] = uv - vu, \tag{3.2}$$

for words $u, v \in \mathcal{A}_d^*$*. Then, adopting the notation of Definition 3.1.1 and Theorem 3.1.5, we have the following identities*

$$\begin{aligned}
I_{ij} &= \frac{1}{2}I_i \cdot I_j + \frac{1}{2}I_{[i,j]}, && (3.3)\\
I_{ijk} &= \frac{1}{6}I_i \cdot I_j \cdot I_k + \frac{1}{4}I_i \cdot I_{[j,k]} + \frac{1}{4}I_{[i,j]} \cdot I_k + \frac{1}{6}I_{[[i,j],k]} + \frac{1}{6}I_{[i,[j,k]]}, && (3.4)\\
I_{ijkl} &= \frac{1}{24}I_i \cdot I_j \cdot I_k \cdot I_l + \frac{1}{12}I_i \cdot I_{[j,[k,l]]} + \frac{1}{12}I_i \cdot I_{[[j,k],l]} + \frac{1}{12}I_{[i,[j,k]]} \cdot I_l && (3.5)\\
&\quad + \frac{1}{12}I_{[[i,j],k]} \cdot I_l + \frac{1}{12}I_i \cdot I_j \cdot I_{[k,l]} + \frac{1}{12}I_i \cdot I_{[j,k]} \cdot I_l + \frac{1}{12}I_{[i,j]} \cdot I_k \cdot I_l\\
&\quad + \frac{1}{8}I_{[i,j]} \cdot I_{[k,l]} + \frac{1}{12}I_{[i,[j,[k,l]]]} + \frac{1}{12}I_{[[i,[j,k]],l]} + \frac{1}{12}I_{[[[i,j],k],l]}\\
&\quad + \frac{1}{12}\big(I_{kjli} - I_{kjil} + I_{lijk} - I_{iljk}\big) + \frac{1}{12}\big(I_{jilk} - I_{kilj} + I_{jlik} - I_{klij}\big),
\end{aligned}$$

for $i, j, k, l \in \mathcal{A}_d$ *and where we understand (for any* $i, j \in \mathcal{A}_d$*)*

$$I_{[i,j]} := I_{ij} - I_{ji}\ .$$

Proof. The results follow by expanding the Lie brackets $[\,\cdot,\cdot\,]$ in the right hand sides using (3.2), applying the integration by parts formula (3.1) and the linear operation of Definition 3.1.6. For example,

$$\frac{1}{2}I_i \cdot I_j + \frac{1}{2}I_{[i,j]} = \frac{1}{2}I_{i \sqcup\!\sqcup j} + \frac{1}{2}(I_{ij} - I_{ji}) = \frac{1}{2}(I_{ij} + I_{ji}) + \frac{1}{2}(I_{ij} - I_{ji}) = I_{ij}\ ,$$

proving (3.3). Equations (3.4) and (3.5) are proven in the same manner. □

Remark 3.1.7. *Results for higher order integrals in terms of symmetric and anti-symmetric components will also exist; However, to prove a high order convergence strong of* $O(h^{1.5})$ *(and weak order* $O(h^2)$*) we only need to consider words of at most length 4 (which includes the integral* $I_{\omega\omega\omega\omega}(1)$*). It is thus sufficient for our purposes to only have explicit formula for these first three levels.*

We do, however, have use of the following result (Theorem 3.1.8) coming from rough path theory. As the focus of this Thesis is not rough paths (and to avoid introducing a large number of definitions) we will not provide a full proof of this result and point the reader to the citations contained in the proof for further clarification.

Theorem 3.1.8. *Let $I_\alpha(1)$ and $I^\gamma_\alpha(1)$ be defined as in* (2.10) *and* (2.10)*, for a path γ satisfying Assumption 2.3.1. Let the word $\alpha \in \mathcal{A}$ consist of only ω's (i.e. $|\alpha| = |\alpha|^\omega$) with length $|\alpha| = m$. The symmetric components of the integrals $I_\alpha(1)$ and $I^\gamma_\alpha(1)$ are given, respectively, by*

$$\frac{1}{m!}\left(I_\omega(1)\right)^{\otimes m} \qquad \textit{and} \qquad \frac{1}{m!}\left(I^\gamma_\omega(1)\right)^{\otimes m} .$$

Proof. This result follows by [43, Exercise 3.15] which shows that the above result holds for all 'weakly-geometric rough paths' [43, Definition 3.14]. The 'Stratonovich enhanced Brownian path' is a geometric rough path (see e.g. [21, Chapter 3]) and thus also a weakly-geometric rough path. Since γ has finite length (i.e. finite 1-variation), it can be directly viewed as a geometric p-rough path and thus the result holds. □

Remark 3.1.9. *The reader may note that these symmetric components look like they come from the expansion of an exponential function. And indeed the 'Signature' of a path (which is the infinite series of iterated integrals) can be described as its full non-commutative exponential [43, Chapter 2].*

3.2 Lévy areas of Brownian motion

Here we define several different Lévy areas, describing the shape of the time-Brownian path (t, W_t), which will be used in our construction of splitting paths in Chapter 5. As we will show in Section 3.3, these Lévy areas can be thought of as components of the Brownian integrals, and indeed we can write many iterated integrals in terms of $W_{s,t}$ and the Lévy areas introduced here.

Firstly we introduce the Brownian Lévy area, which encodes the distance between the Brownian motion in its different dimensions.

Definition 3.2.1. *The* **space-space (or Brownian) Lévy** *area of Brownian motion over the interval $[s,t]$ is defined as*

$$A_{s,t} := I_{\omega\omega}(1) - \left(I_{\omega\omega}(1)\right)^\top ,$$

or equivalently, for $i, j \in \{1, \ldots, d\}$, component wise by

$$(A_{s,t})_{ij} := I_{ij} - I_{ji} = \int_s^t \int_s^{r_1} \circ dW^i_{r_2} \circ dW^j_{r_1} - \int_s^t \int_s^{r_1} \circ dW^j_{r_2} \circ dW^i_{r_1} .$$

As we will show in Section 3.6 the commutativity condition (2.2) means that $A_{s,t}$ will not appear in the Taylor expansion of the SDE and CDE. Thus, we do not need to simulate the Brownian Lévy area for our numerical schemes. This is a benefit as the simulation of $A_{s,t}$ is computationally expensive [9].

The most important of the Lévy areas for our construction of high order splitting paths is the 'space-time' Lévy area. And in fact its inclusion is already enough to construct a high order scheme (see path HS1 in Section 5.1 and the paper [61]). In Theorem 3.3.1 we see that the space-time Lévy area appears as a component of the integrals $I_{\tau\omega}(1)$ and $I_{\omega\tau}(1)$, which we must match to achieve a high order of convergence.

Definition 3.2.2. *The rescaled* **space-time Lévy** *area of a Brownian motion W over an interval $[s,t]$ corresponds to the signed area of the associated bridge process.*

$$H_{s,t} := \frac{1}{h}\int_s^t \Big(W_{s,u} - \frac{u-s}{h}W_{s,t}\Big)\,du,$$

where $h := t - s$ and $W_{s,u} := W_u - W_s$ for $u \in [s,t]$.

The space-time Lévy area encodes how far on average the Brownian path was from the shortest path between W_s and W_t (a straight line). In Figure 3.1 the Brownian motion is above this shortest path for most of the period $[s,t]$ and thus has $H_{s,t} > 0$.

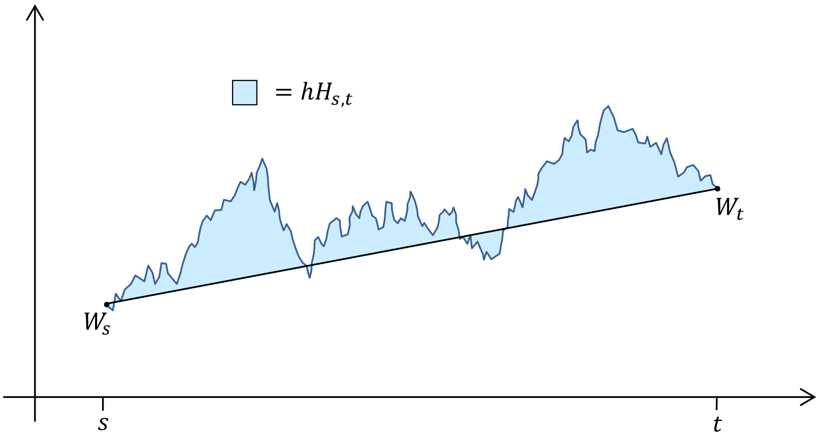


Fig. 3.1: Space-time Lévy area gives the area between a Brownian path and its linear approximant.

While we can construct a high order splitting scheme with $W_{s,t}$ and $H_{s,t}$ alone, it is possible to produce more accurate schemes (of the same order) through the inclusion or estimation of the following additional quantities (see Section 5.3 for details). The first of these is the 'space-time Lévy swing' which gives finer detail on the distribution $H_{s,t}$ (see Figure 3.2).

Definition 3.2.3. *The* **space-time Lévy swing**[1] *of Brownian motion over $[s,t]$ is defined as*

$$n_{s,t} := \operatorname{sgn}\big(H_{s,u} - H_{u,t}\big),$$

where $u := \frac{1}{2}(s+t)$ is the interval's midpoint.

[1] **S**ide **W**ith **IN**tegral **G**reater

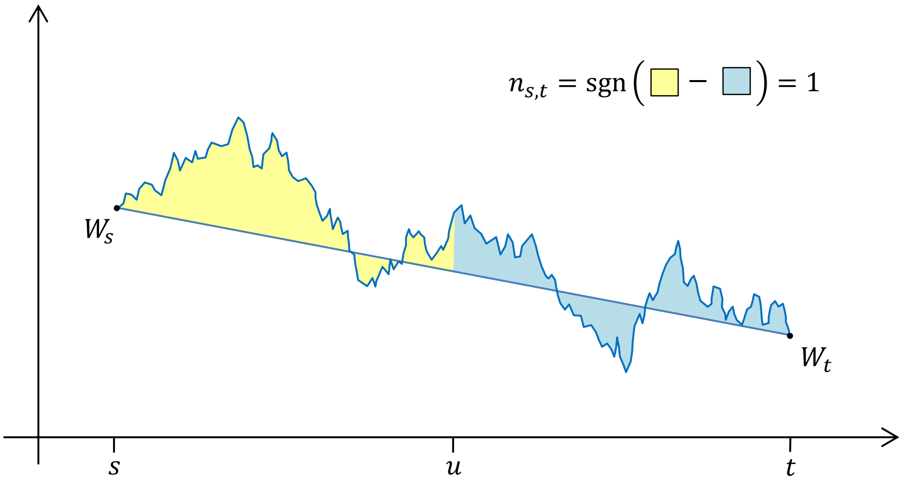


Fig. 3.2: Space-time Lévy swing gives the side where the path has greater space-time Lévy area.

In Section 5.3 the 'optimal' splitting schemes are derived in such a way as to approximate the 'space-space-time' Lévy area. Which is defined as follows.

Definition 3.2.4. *Over an interval $[s,t]$, the* **space-space-time Lévy** *area $L_{s,t}$ of a standard Brownian motion is defined, for $i, j \in \{1, \ldots, d\}$, component wise by*

$$\begin{aligned}(L_{s,t})_{ij} &:= \frac{1}{6}\left(I_{[[i,j],\tau]} + I_{[i,[j,\tau]]}\right)\\ &= \frac{1}{6}\Bigg(2\int_s^t\int_s^{r_1}\int_s^{r_2} \circ dW^i_{r_3} \circ dW^j_{r_2} dr_1 - \int_s^t\int_s^{r_1}\int_s^{r_2} \circ dW^j_{r_3} \circ dW^i_{r_2} dr_1\\ &\qquad - \int_s^t\int_s^{r_1}\int_s^{r_2} \circ dW^i_{r_3} dr_2 \circ dW^j_{r_1} - \int_s^t\int_s^{r_1}\int_s^{r_2} \circ dW^j_{r_3} dr_2 \circ dW^i_{r_1}\\ &\qquad + 2\int_s^t\int_s^{r_1}\int_s^{r_2} dr_3 \circ dW^j_{r_2} \circ dW^i_{r_1} - \int_s^t\int_s^{r_1}\int_s^{r_2} dr_3 \circ dW^i_{r_2} \circ dW^j_{r_1}\Bigg) .\end{aligned}$$

While this area is more difficult to visualise, by inspecting the definition the reader should note that it corresponds to the last two terms of the expansion (3.4). And, in Theorem 3.3.3 we see that it is a component of the integrals corresponding to the words $\omega\omega\tau$, $\omega\tau\omega$ and $\tau\omega\omega$. As these integrals are of order 2, we do not need to match $L_{s,t}$ in an almost sure sense to achieve high order strong convergence of $O(h^{1.5})$.

Remark 3.2.5. *Along with the path increment $W_{s,t}$, the Lévy areas $H_{s,t}$ and $L_{s,t}$ are sufficient to construct the iterated integrals appearing in the stochastic Taylor expansion (2.14), up to order 2 for SDEs satisfying the commutativity condition (2.2).*

The final Lévy area used in the definition and derivation of the splitting paths we present is the 'space-time-time' Lévy area. Which is defined as follows and described in Figure 3.3.

Definition 3.2.6. *The rescaled* **space-time-time Lévy area** *of Brownian motion over an interval $[s,t]$ is defined as*

$$K_{s,t} := \frac{1}{h^2}\int_s^t \left(W_{s,u} - \frac{u-s}{h}W_{s,t}\right)\left(\frac{1}{2}h - (u-s)\right)du.$$

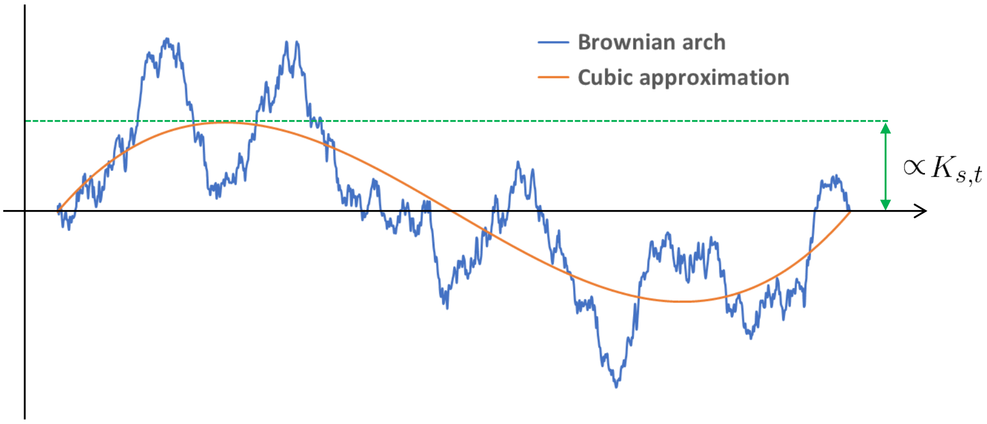


Fig. 3.3: Space-time-time Lévy area corresponds to a cubic approximation of the Brownian arch (which is a Brownian motion conditioned on having zero increment and space-time Lévy area [17]).

3.2.1 Relational properties between Lévy areas

In Section 5.1 we will introduce piecewise linear splitting paths defined using the Lévy areas defined above. We note their distributions and relationships in the following few results. As these results were originally shown in [14] we point the reader to the relevant proofs therein. Further details are also contained in Section 5.3 of this thesis.

Lemma 3.2.7. *For $H_{s,t}$ and $K_{s,t}$ defined as above*

$$H_{s,t} \sim \mathcal{N}\left(0, \frac{h}{12}\mathbf{1}_d\right) \qquad \text{and} \qquad K_{s,t} \sim \mathcal{N}\left(0, \frac{h}{720}\mathbf{1}_d\right) .$$

Proof. See the remarks in [14, Definitions 4.2.1 and 4.2.3]. □

Lemma 3.2.8. *For $n_{s,t}$ defined as in Definition 3.2.3, for every $i \in \{1,\dots,d\}$*

$$\mathbb{P}\big((n_{s,t})_i = 1\big) = \mathbb{P}\big((n_{s,t})_i = -1\big) = 0.5 .$$

Proof. This follows by the symmetry of the Brownian motion. □

With lemmas 3.2.7 and 3.2.8 it is clear how to generate each Lévy area separately. In the following results we note the independence between terms.

Lemma 3.2.9. *The space-time Lévy area $H_{s,t}$ is independent of $W_{s,t}$.*

Proof. It was shown in [17, Remark 3.6] that $H_{s,t} \sim \mathcal{N}(0, \frac{1}{12}h)$ is independent of $W_{s,t}$ when $d = 1$. Since the coordinate processes of a Brownian motion are independent, it therefore follows that $W_{s,t} \sim \mathcal{N}(0, h\mathbf{1}_d)$ and $H_{s,t} \sim \mathcal{N}(0, \frac{1}{12}h\mathbf{1}_d)$ are independent. □

Lemma 3.2.10. *The space-time Lévy swing $n_{s,t}$ is independent of $W_{s,t}$ and $H_{s,t}$.*

Proof. This is a direct result of [14, Theorem 5.0.6] again noting that the coordinate processes of Brownian motion are independent. □

Lemma 3.2.11. *The space-time-time Lévy area $K_{s,t}$ is independent of $W_{s,t}$ and $H_{s,t}$.*

Proof. See [14, Definition 4.2.3]. □

3.3 Computation of Brownian integrals

In this section, we use Theorems 3.1.5 and 3.1.6 to prove several representations of iterated Brownian integrals in terms of the Lévy areas introduced in Section 3.2. These representations will be useful later for checking the convergence rates of the proposed splitting schemes. Throughout, we set $0 \leqslant s < t$ and $h = t - s > 0$.

The results presented here are largely drawn from or inspired by results in [14, Chapter 4.2] (there presented in the 1-D case); Theorem 3.3.1 and Lemma 3.3.2 both follow directly from results therein (we include here proofs to aid understanding), whereas Theorem 3.3.3 is an extension of a previous result and the proof is new: based on our Theorem 3.1.6.

Theorem 3.3.1. *Let $H_{s,t}$ denote the space-time Lévy area of Brownian motion as defined in Definition 3.2.2, then*

$$I_{\omega\tau}(1) = \frac{1}{2}hW_{s,t} + hH_{s,t} \ , \tag{3.6}$$

$$I_{\tau\omega}(1) = \frac{1}{2}hW_{s,t} - hH_{s,t} \ . \tag{3.7}$$

Proof. Recalling Definition 3.2.2, we have

$$\begin{aligned} H_{s,t} &= \frac{1}{h}\int_s^t \left(W_{s,u} - \frac{u-s}{h}W_{s,t}\right) du \\ &= \frac{1}{h}I_{\omega\tau}(1) - \frac{1}{2}W_{s,t} \ , \end{aligned}$$

which proves equation (3.6).

To prove (3.7) we note that by Theorem 3.1.5 (we have one Brownian motion here, so we may view this as performing integration by parts component wise)

$$hW_{s,t} = I_\tau(1)I_\omega(1) = I_{\tau \sqcup\!\sqcup \omega}(1) = I_{\tau\omega}(1) + I_{\omega\tau}(1) \ ,$$

which, in combination with (3.6), provides (3.7). □

We present next a version of [14, Theorem 4.2.7].

Lemma 3.3.2. *Let $H_{s,t}$, $L_{s,t}$ and $K_{s,t}$ denote the Lévy areas of Brownian motion as defined in Definitions 3.2.2, 3.2.4 and 3.2.6 respectively. Then*

$$hH_{s,t} = \frac{1}{2}I_{[\omega,\tau]}(1), \tag{3.8}$$

$$h^2K_{s,t} = \frac{1}{6}I_{[[\omega,\tau],\tau]}(1) \ . \tag{3.9}$$

Proof. Equation (3.8) follows as a direct result of Theorem 3.3.1. To prove (3.9) we start from the definition of $K_{s,t}$ (Definition 3.2.6), expand and apply integration by parts, which (with some abuse of notation) gives the following:

$$\begin{aligned} h^2K_{s,t} &= \frac{1}{2}h\int_s^t W_{s,u}du - \int_s^t W_{s,u}(u-s)du \\ &\quad - \frac{1}{2}\int_s^t (u-s)ds \times W_{s,t} + \int_s^t \frac{(u-s)^2}{h}du \times W_{s,t} \\ &= \frac{1}{2}I_\tau \cdot I_{\omega\tau} - (I_{\omega\tau\tau} + I_{\tau\omega\tau}) - \frac{1}{2}I_{\tau\tau}\cdot I_\omega + \frac{1}{3}I_\tau \cdot I_\tau \cdot I_\omega \\ &= \frac{1}{6}I_{\omega\tau\tau} - \frac{1}{3}I_{\tau\omega\tau} + \frac{1}{6}I_{\tau\tau\omega} \ , \end{aligned}$$

which, applying (3.2), gives us (3.9). □

Based on Theorem 3.1.6 and Lemma 3.3.2, we obtain the following extension of [14, Lemma 4.2.5] to multi-dimensions. In contrast with the result presented there, we see how the cross terms $(A_{s,t})_{ij}$, $(L_{s,t})_{ij}$ and $(L_{s,t})_{ji}$ appear in the multidimensional case.

Theorem 3.3.3. *Let $A_{s,t}$, $H_{s,t}$ and $L_{s,t}$ denote the Lévy areas of Brownian motion as defined in Definitions 3.2.1, 3.2.2 and 3.2.4 respectively for $0 < s < t$, with $h := t - s$. Then we have the following identities for all $i, j \in \{1, \dots, d\}$*

$$(I_{\omega\omega\tau}(1))_{ij} = \frac{1}{6}h(W_{s,t})_i(W_{s,t})_j + \frac{1}{2}h(W_{s,t})_i(H_{s,t})_j + \frac{1}{4}h(A_{s,t})_{ij} + (L_{s,t})_{ij} \ , \tag{3.10}$$

$$\begin{aligned}(I_{\omega\tau\omega}(1))_{ij} &= \frac{1}{6}h(W_{s,t})_i(W_{s,t})_j - \frac{1}{2}h(W_{s,t})_i(H_{s,t})_j + \frac{1}{2}h(W_{s,t})_j(H_{s,t})_i \\ &\quad - (L_{s,t})_{ij} - (L_{s,t})_{ji} \ , \end{aligned}\tag{3.11}$$

$$(I_{\tau\omega\omega}(1))_{ij} = \frac{1}{6}h(W_{s,t})_i(W_{s,t})_j - \frac{1}{2}h(W_{s,t})_j(H_{s,t})_i + \frac{1}{4}h(A_{s,t})_{ij} + (L_{s,t})_{ji} \ . \tag{3.12}$$

Proof. For all $i, j \in \{1, \dots, d\}$, by Theorem 3.1.6 and applying Lemma 3.3.2 we have that

$$\begin{aligned} I_{ij\tau} &= \frac{1}{6}I_i \cdot I_j \cdot I_\tau + \frac{1}{4}I_i \cdot I_{[j,\tau]} + \frac{1}{4}I_{[i,j]}\cdot I_\tau + \frac{1}{6}I_{[[i,j],\tau]} + \frac{1}{6}I_{[i,[j,\tau]]} \\ &= \frac{1}{6}h(W_{s,t})_i(W_{s,t})_j + \frac{1}{2}h(W_{s,t})_i(H_{s,t})_j + \frac{1}{4}h(I_{ij} - I_{ji}) \\ &\quad + \frac{1}{6}I_{[[i,j],\tau]} + \frac{1}{6}I_{[i,[j,\tau]]} \ , \end{aligned}$$

$$\begin{aligned} I_{i\tau j} &= \frac{1}{6}I_i \cdot I_\tau \cdot I_j + \frac{1}{4}I_i \cdot I_{[\tau,j]} + \frac{1}{4}I_{[i,\tau]} \cdot I_j + \frac{1}{6}I_{[[i,\tau],j]} + \frac{1}{6}I_{[i,[\tau,j]]} \\ &= \frac{1}{6}h(W_{s,t})_i(W_{s,t})_j - \frac{1}{2}h(W_{s,t})_i(H_{s,t})_j + \frac{1}{2}h(H_{s,t})_i(W_{s,t})_j \\ &\quad + \frac{1}{6}I_{[[i,\tau],j]} + \frac{1}{6}I_{[i,[\tau,j]]} \ , \end{aligned}$$

$$\begin{aligned} I_{\tau ij} &= \frac{1}{6}I_\tau \cdot I_i \cdot I_j + \frac{1}{4}I_\tau \cdot I_{[i,j]} + \frac{1}{4}I_{[\tau,i]} \cdot I_j + \frac{1}{6}I_{[[\tau,i],j]} + \frac{1}{6}I_{[\tau,[i,j]]} \\ &= \frac{1}{6}h(W_{s,t})_i(W_{s,t})_j - \frac{1}{2}h(W_{s,t})_i(H_{s,t})_j + \frac{1}{4}h(I_{ij} - I_{ji}) \\ &\quad + \frac{1}{6}I_{[[\tau,i],j]} + \frac{1}{6}I_{[\tau,[i,j]]} \ , \end{aligned}$$

comparing with the definition of $L_{s,t}$ (Definition 3.2.4), we conclude by noting that

$$I_{[[i,j],\tau]} + I_{[i,[j,\tau]]} = I_{[[\tau,j],i]} + I_{[\tau,[j,i]]} \ , \qquad \text{and}$$

$$I_{[[i,j],\tau]} + I_{[i,[j,\tau]]} + I_{[[\tau,i],j]} + I_{[\tau,[i,j]]} = -(I_{[[i,\tau],j]} + I_{[i,[\tau,j]]}) \ .$$

□

3.3.1 Expected value of Brownian integrals

As we will prove in Chapter 4, to obtain a certain global error we must match higher order terms from the Taylor expansion (2.14) in expectation. For the Stratonovich integrals $I_\alpha(1)$ (here written in tensor form) the expected value is easily obtained using the following result.

Theorem 3.3.4. *Let the word $\alpha \in \mathcal{A}$ have order $ord(\alpha) = p \in \{\frac{k}{2}\}_{k \in \mathbb{N}}$ (recall $ord(\alpha) = |\alpha|^\tau + |\alpha|^\omega/2$). Let $S^\omega(\alpha)$ denote the set of subwords of α formed by all consecutive ω's (e.g. $S^\omega(\omega\tau\omega\omega\omega) = \{\omega, \omega\omega\omega\}$). If there is at least one word $\beta \in S^\omega(\alpha)$ with odd length (i.e. $|\beta| = 2k+1$ for $k \in \mathbb{N}_0$) then $\mathbb{E}[I_\alpha(1)] = 0$, otherwise*

$$\mathbb{E}[I_\alpha(1)] = \frac{h^p}{\sqrt{2^{|\alpha|^\omega}}p!} D_d^{|\alpha|^\omega} \ . \tag{3.13}$$

Where, for $i = 2m$ with $m, d \in \mathbb{N}^+$, D_d^i denotes the i-tensor defined component wise by

$$(D_d^i)_\alpha = \begin{cases} 1 & \textit{if } \alpha \in \mathcal{A}_2^i \ , \\ 0 & \textit{otherwise} \end{cases} \tag{3.14}$$

and $\mathcal{A}_2^i$ denotes the set of pair-wise equal multi-indices $\alpha = \alpha_1\alpha_2 \ldots \alpha_{i/2-1}\alpha_{i/2}$, where each $\alpha_j = (k,k)$ for some $k \in \{1, \ldots, d\}$. D_d^2 corresponds to the $d \times d$ identity matrix.

Proof. This follows as a direct result of [23, Lemma 3]. Alternatively, one can obtain (3.13) from the expected signature of the Brownian path see [44, Proposition 4.10]. □

3.4 Algorithmic computation of path integrals

Our error analysis hinges on the computation of the path integrals $I_\alpha^\gamma(1)$ ($\alpha \in \mathcal{A} = \cup_{n\geqslant 0}\{\tau,\omega\}^n$) appearing in the CDE Taylor expansion (2.19). However, this quickly becomes tedious: involving many terms and multiplications. We thus propose an algorithmic approach to computing these integrals that is applicable to strong and weak error analysis. This algorithmic approach is made possible by the piecewise nature of the paths and the additive property of the integrals. A Python implementation of the methodology presented next is available at github.com/calum-strange/auto_splitting_integrals.

3.4.1 Generating all words of a given order

Before dealing with the path integrals, we first consider the generation of all words $\alpha \in \mathcal{A}$ with $\mathrm{ord}(\alpha) = p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, which we denote by $\mathcal{A}(p)$ – see notation introduced in Section 2.4.

Lemma 3.4.1. *Set ord*$(\tau) = 1$ *and ord*$(\omega) = 1/2$*. Then the following simple recursion holds for* $p \geqslant 1$

$$\mathcal{A}(p) = \tau\mathcal{A}(p-1) \cup \omega\mathcal{A}(p-0.5) \ ,$$

$$\textit{with} \quad \mathcal{A}(0.5) = \{\omega\} \quad \textit{and} \quad \mathcal{A}(0) = \{\} \ ,$$

where "$\{\}$*" is the empty set, and we understand* $l\{\alpha_1,\ldots,\alpha_j\} := \{l\alpha_1,\ldots,l\alpha_j\}$ *to denote elementwise concatenation for* $l \in \{\tau,\omega\}$ *with* $l\{\} := \{l\}$*.*

Proof. This result is proven by induction over the order of the words $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$.

Base case. The list of all words with order 1 is given by $\{\tau,\omega\omega\}$ and we have that

$$\mathcal{A}(1) = \{\tau,\omega\omega\} = \tau\{\} \cup \omega\{\omega\} = \tau\mathcal{A}(0) \cup \omega\mathcal{A}(0.5) \ .$$

In addition $\mathcal{A}(1.5) = \{\tau\omega,\omega\tau,\omega\omega\omega\}$, so we have

$$\mathcal{A}(1.5) = \{\tau\omega,\omega\tau,\omega\omega\omega\} = \tau\{\omega\} \cup \omega\{\tau,\omega\omega\} = \tau\mathcal{A}(0.5) \cup \omega\mathcal{A}(1)\,,$$

proving the base case.

Induction step: $p = k+1$. Assume the recursion holds for $p = k$ and $p = k+0.5$. Note that all words in $\mathcal{A}(k+1)$ must start with τ or ω, thus splitting the set based on this and removing the first letter from each word we are left with: a set of words of order $p = k$ (for the words that started with τ), and a set of words of order $p = k+0.5$ (for those starting with ω). These two sets are $\mathcal{A}(k)$ and $\mathcal{A}(k+0.5)$ respectively.

We conclude by induction. □

3.4.2 Expansion of iterated integrals

In order to calculate the integral $I^{\gamma}_{\alpha}(1)$ we first expand it into a finite sum of iterated integrals over which the derivatives of γ are constant. We then use the fact that $d\gamma^{\tau}(r) = \frac{1}{r_{i+1}-r_i}\gamma^{\tau}_{r_i,r_{i+1}}dr$ and $d\gamma^{\omega}(r) = \frac{1}{r_{i+1}-r_i}\gamma^{\omega}_{r_i,r_{i+1}}dr$ for $r \in [r_i, r_{i+1}]$. These derivatives of γ are independent of the integration and thus for each iterated integral we obtain the tensor product of the γ derivatives multiplied by a constant coming from the integrals. The exact form of this expansion depends on $|\alpha|$ (the number of integrals) and m, the number of pieces in the path γ. For example, integrating a double integral against a three-piece path (e.g. γ of Example 2.2.1) we first expand the integral as follows (with some abuse of notation)

$$\begin{aligned}\int_0^1\int_0^{r_1} &= \int_{\frac{2}{3}}^1\int_0^{r_1} + \int_{\frac{1}{3}}^{\frac{2}{3}}\int_0^{r_1} + \int_0^{\frac{1}{3}}\int_0^{r_1}\\ &= \int_{\frac{2}{3}}^1\int_{\frac{2}{3}}^{r_1} + \int_{\frac{2}{3}}^1\int_{\frac{1}{3}}^{\frac{2}{3}} + \int_{\frac{2}{3}}^1\int_0^{\frac{1}{3}} + \int_{\frac{1}{3}}^{\frac{2}{3}}\int_{\frac{1}{3}}^{r_1} + \int_{\frac{1}{3}}^{\frac{2}{3}}\int_0^{\frac{1}{3}} + \int_0^{\frac{1}{3}}\int_0^{r_1} \quad . \qquad (3.15)\end{aligned}$$

As we display in Figure 3.4, we can view (3.15) as subdividing the area of a triangle. Where we see (corresponding sequentially to the terms in (3.15)) that (a) = (b) = (c) . In higher dimensions we can imagine subdividing the volume of a simplex.

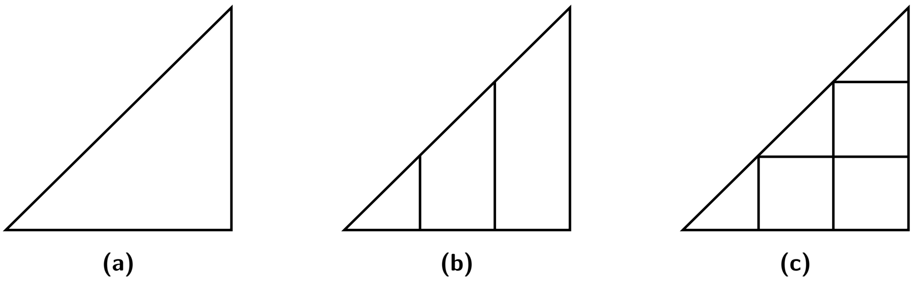


Fig. 3.4: Subdividing integration over a triangle.

For our algorithmic approach, we instead adopt the following perspective: In Fig 3.5 (for a triple integral against a three piece path) we view the expansion of integrals against piecewise linear paths as finding all non-increasing routes through a series of connected layers. Where the layers contain nodes equal to the number of pieces in the path, and the number of layers is equal to the number of integrals. A route passing through the ith node (numbered bottom up) in the jth layer (numbered left to right) then represents an iterated integral where the jth integral (outside to inside) is against the ith piece of the path. The integrals in (3.15) would correspond to the routes in the first two layers of Figure 3.5.

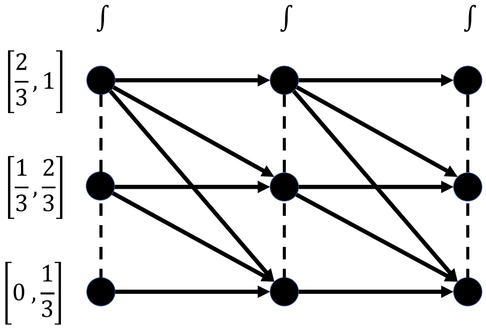

∫
∫
∫
$\left[\frac{2}{3},1\right]$
$\left[\frac{1}{3},\frac{2}{3}\right]$
$\left[0,\frac{1}{3}\right]$

Fig. 3.5: All non-increasing routes.

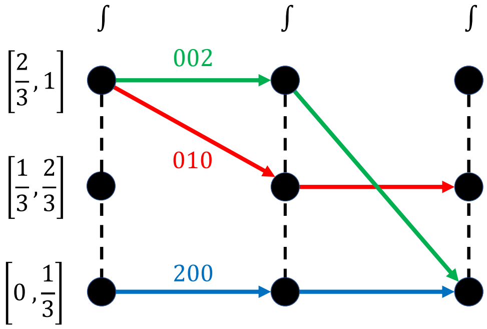

∫
∫
∫
$\left[\frac{2}{3},1\right]$
$\left[\frac{1}{3},\frac{2}{3}\right]$
$\left[0,\frac{1}{3}\right]$
002
010
200

Fig. 3.6: Specific routes.

We can encode such a route by adding an additional ‘dummy’ node (connected to each node in the first layer and at the height of the top node) and then considering the size of the downward movement between each layer: a *drop*. For example, in Fig 3.6

$$002 = \int_{\frac{2}{3}}^{1}\int_{\frac{2}{3}}^{r_1}\int_{0}^{\frac{1}{3}} \quad , \quad 010 = \int_{\frac{2}{3}}^{1}\int_{\frac{1}{3}}^{\frac{2}{3}}\int_{\frac{1}{3}}^{r_2} \quad \text{and} \quad 200 = \int_{0}^{\frac{1}{3}}\int_{0}^{r_1}\int_{0}^{r_2} \ , \tag{3.16}$$

and we can label the integrals in (3.15) sequentially as $00, 01, 02, 10, 11, 20$. For a given number of pieces $m \in \mathbb{N}$ and number of layers (integrals) $l \in \mathbb{N}$, we denote the set of all such non-increasing routes by $\mathcal{R}(m, l)$.

Lemma 3.4.2. *For a fixed path with $m \in \mathbb{N}$ pieces and an integral with $l \in \mathbb{N}$ layers, the set of all non-increasing routes $\mathcal{R}(m, l)$ can be generated by the following recursive formula*

$$\mathcal{R}(m,l) = \bigcup_{i=0}^{m-1} i\mathcal{R}(m-i, l-1) \quad \textit{with} \quad \mathcal{R}(m,0) = \{\} \ ,$$

where $i\mathcal{R}(m-i, l-1)$ denotes element wise concatenation as in Lemma 3.4.1.

Proof. This proof is an induction on $l \in \mathbb{N}$.

First, note that no route can go below the bottom (1st) node, and thus the drops on a route can sum to at most $m - 1$.

Base case. $l = 1$. The possible routes with a single layer are to drop to any node or stay at the top node, i.e. $\mathcal{R}(m, 1) = \{0, 1, \ldots, m-1\}$. And, we have

$$\mathcal{R}(m,1) = \{0,1,\ldots,m-1\} = \bigcup_{i=0}^{m-1} i\{\} = \bigcup_{i=0}^{m-1} i\mathcal{R}(m-i,0) \ ,$$

so, the base case is proven.

Induction step. Assume that the recursion holds for $l = k$. All routes in $\mathcal{R}(m, k+1)$ start with $i \in \{0, 1, \ldots, m-1\}$. A drop of i places us on the $(m-i)$'th node, and we can view removing i from the start of the route as removing the first layer of nodes and setting our 'dummy' node at height $m - i$. As the route is non-increasing, the set of all possible routes from this position is $\mathcal{R}(m-i, k)$.

We conclude by induction. □

Remark 3.4.3. *We can easily obtain the constant produced by the integrals (without evaluating them), by counting the number of consecutive flat parts in the corresponding route. Ignoring the first number in any route, each* 0 *adds one to a factorial (starting at* $\frac{1}{2!}$*), and each non-zero element restarts a factorial count. For example*

$$|1000| = |0000| = \frac{1}{4!} \quad \textit{and} \quad |001001000| = \frac{1}{2!}\frac{1}{3!}\frac{1}{4!} = \frac{1}{288}\,,$$

where for $|001001000|$ *we obtain* $\frac{1}{2!}$ *from the first* 00*,* $\frac{1}{3!}$ *from the* 100 *and* $\frac{1}{4!}$ *from the* 1000*. Again, considering Figure 3.6 and comparing with* (3.16) *we have*

$$|002| = \frac{1}{2!}\,, \quad |010| = \frac{1}{2!} \quad \textit{and} \quad |200| = \frac{1}{3!}\,.$$

3.4.3 Computing path integrals

To calculate a given iterated integral I^{γ}_{α} for an m-piece path γ we first proceed by calculating $\mathcal{R}(m, |\alpha|)$. This, in combination with the word α, tells us against which piece of the path to evaluate the integral. We combine each route with the word as follows: matching left-to-right on the route with right-to-left on the word. For example, for route 011 and word $\tau\omega\omega$ (when $m = 3$) we visit:

i) the space component of the 3rd piece of γ,
ii) the space component of the 2nd piece, and
iii) the time component of the 1st piece.

If any visited (time or space) components are equal to zero (e.g. the space component in the 1st and 3rd piece of the Strang splitting, and time component in the 2nd) then we call the route a 'zero route' for that combination of γ and α.

Comparing all routes with the word α, we collect the non-zero routes, multiply the components of γ along these routes and calculate the constant coming from the integral. If we consider the Strang splitting path $\gamma = \gamma^{S}$ (see (5.3) below), we obtain (see (3.15) corresponding sequentially to $00, 01, 02, 10, 11, 20$)

$$\begin{aligned}
I^{\gamma}_{\omega\omega}(1) &= 3^2 \int_{\frac{1}{3}}^{\frac{2}{3}} \int_{\frac{1}{3}}^{r_1} dr_2 dr_1 \times W_{s,t} \otimes W_{s,t} \\
&= |10| \times W_{s,t} \otimes W_{s,t} = \frac{1}{2} W_{s,t}^{\otimes 2} \ , \\
I^{\gamma}_{\tau\omega}(1) &= |11| \times \frac{1}{2} h W_{s,t} = \frac{1}{2} h W_{s,t} \ , \\
I^{\gamma}_{\omega\tau}(1) &= |01| \times \frac{1}{2} h W_{s,t} = \frac{1}{2} h W_{s,t} \ , \\
I^{\gamma}_{\tau\tau}(1) &= (|00| + |02| + |20|) \times \frac{1}{4} h^2 = \frac{1}{2} h^2 \ .
\end{aligned}$$

Paths with more (and more complex) pieces will result in slightly more involved computations. For example taking the high order Strang path $\gamma = \gamma^{\mathrm{HS1}}$ (see (5.4) below) we obtain

$$\begin{aligned}
I^{\gamma}_{\omega\omega}(1) = & \ |10| \times (W_{s,t}/2 - \sqrt{3} H_{s,t}) \otimes (W_{s,t}/2 - \sqrt{3} H_{s,t}) \\
& + |30| \times (W_{s,t}/2 + \sqrt{3} H_{s,t}) \otimes (W_{s,t}/2 + \sqrt{3} H_{s,t}) \\
& + |12| \times (W_{s,t}/2 + \sqrt{3} H_{s,t}) \otimes (W_{s,t}/2 - \sqrt{3} H_{s,t}) \\
= & \frac{1}{2} W_{s,t}^{\otimes 2} - \frac{\sqrt{3}}{2} W_{s,t} \otimes H_{s,t} + \frac{\sqrt{3}}{2} H_{s,t} \otimes W_{s,t} \ .
\end{aligned}$$

Expanding brackets and collecting terms in this fashion can quickly become burdensome. In order to automate this task we may encode each derivative of γ as a list of tuples: (‘constant’, ‘random variable’). So that $W_{s,t}/2 - \sqrt{3} H_{s,t}$ becomes $\{(\frac{1}{2}, \text{‘W’}), (-\sqrt{3}, \text{‘H’})\}$. We then understand the multiplication of two tuples

$$(a, \text{‘X’}) \times (b, \text{‘Y’}) = (a \times b, \text{‘XY’}) \ ,$$

where we append ‘Y’ to the end of string ‘X’. We multiply two lists of tuples by multiplying each tuple in the first list by all tuples in the second and writing the resultant tuples to a new list. This can be performed sequentially as we traverse each route. To collect like terms, at the end of all non-zero routes we may then update a dictionary with keys given by the ‘random variable’ (e.g. ‘hhWH’ $= h^2 W_{s,t} \otimes H_{s,t}$) adding the ‘constant’ to a running total. In our Python implementation, this addition of constants was performed using the package *SymPy* [47] and the built-in module *fractions*. If the time components of γ are assumed scalar, then their multiplication may be performed separately from the space components.

Remark 3.4.4. *An alternative approach, inspired by rough paths, would be to instead consider the signature of the splitting paths. By Chen's identity [6, Theorem 2] the signature of our piece-wise linear paths can be written as the tensor product of the signature of each linear piece. This is the approach taken in the 'signatory' python package, which can be found at https://github.com/patrick-kidger/signatory. Our approach, being symbolic in nature, allows us to compare the exact form of the Brownian and path integrals.*

3.5 L^2 estimates of iterated path integrals

As described previously, our error analysis involves Taylor expanding both the SDE and CDE, and then comparing the resulting terms. The Stratonovich Taylor expansion (2.14) and the Controlled Taylor expansion (2.19) differ by involving $I_\alpha(1)$ and $J_\alpha(V_\psi(\alpha))$ or $I^\gamma_\alpha(1)$ and $J^\gamma_\alpha(V_\psi(\alpha))$. For the strong error analysis (the L^2 difference) it is useful to have L^2 estimates on the iterated path integrals $I^\gamma_\alpha(1)$ and $J^\gamma_\alpha(V_\psi(\alpha))$. Under Assumption 2.3.1 and the assumptions of Theorem 2.3.3 on the path γ we obtain the following two results.

Lemma 3.5.1. *Suppose the path γ satisfies Assumption 2.3.1 and let $\alpha \in \mathcal{A}$. Then*

$$\mathbb{E}\big[\|I^\gamma_\alpha(1)\|^2\big] = O\big(h^{2ord(\alpha)}\big).$$

Proof. Since γ is piecewise linear, we may split the iterated integral of γ into a finite sum of iterated integrals as described in Section 3.4.2. This directly follows by the standard additive property of Riemann-Stieltjes integrals. We may convert these path integrals into regular (deterministic) integrals over these routes as $d\gamma^\tau(r) = \frac{1}{r_{i+1}-r_i}\gamma^\tau_{r_i,r_{i+1}}dr$ and $d\gamma^\omega(r) = \frac{1}{r_{i+1}-r_i}\gamma^\omega_{r_i,r_{i+1}}dr$ for $r \in [r_i, r_{i+1}]$. By Assumption 2.3.1, $\gamma^\tau_{r_i,r_{i+1}} = O(h)$ is a deterministic constant and we therefore have

$$\begin{aligned}\mathbb{E}\big[\|I^\gamma_\alpha(1)\|^2\big] &\leqslant Ch^{2|\alpha|_\tau} \sum_{\mathcal{R}(m,|\alpha|)} \mathbb{E}\Bigg[\Bigg\|\bigotimes_{j=1}^{|\alpha|_\omega} \gamma^\omega_{r^j_i,r^j_{i+1}}\Bigg\|^2\Bigg],\\ &= Ch^{2|\alpha|_\tau} \sum_{\mathcal{R}(m,|\alpha|)} \mathbb{E}\Bigg[\sum_{i_1=1}^{d} \cdots \sum_{i_{|\alpha|_\omega}=1}^{d} \big(\gamma^\omega_{r^1_i,r^1_{i+1}}\big)^2_{i_1} \times \cdots \times \big(\gamma^\omega_{r^{|\alpha|_\omega}_i,r^{|\alpha|_\omega}_{i+1}}\big)^2_{i_{|\alpha|_\omega}}\Bigg],\end{aligned}$$

where $\mathcal{R}(m,|\alpha|)$ denotes the set of routes as defined in lemma 3.4.2. We can then estimate the $\gamma^\omega_{r_i,r_{i+1}}$ terms by iteratively applying Hölder's inequality to the expectation and applying the assumption that $\mathbb{E}\big[|(\gamma^\omega_{r_i,r_{i+1}})_j|^{2k}\big] = O(h^k)$ for $k \in \mathbb{N}$. This implies that

$$\mathbb{E}\big[\|I^\gamma_\alpha(1)\|^2\big] \leqslant C_{d,m,|\alpha|}h^{2\mathrm{ord}(\alpha)}.$$

□

We now consider the J^γ_α terms, which will follow in much the same way as for I^γ_α.

Lemma 3.5.2. *Suppose that the assumptions of Theorem 2.3.3 hold. Let $\alpha \in \mathcal{A}$ and $F : \mathbb{R}^e \to L\big((\mathbb{R}^d)^{\otimes b}, \mathbb{R}^e\big)$ be a globally Lipschitz continuous map for some $b \in \mathbb{N}$, then*

$$\mathbb{E}\big[\|J^\gamma_\alpha(F)\|^2\big] = O\big(h^{2\mathrm{ord}(\alpha)+1}\big).$$

Proof. Just as in the previous proof, we may split the iterated integral of the piecewise linear path γ into a finite sum of iterated integrals over the intervals where both $d\gamma^\tau(r) = \frac{1}{r_{i+1}-r_i}\gamma^\tau_{r_i,r_{i+1}}dr$ and $d\gamma^\omega(r) = \frac{1}{r_{i+1}-r_i}\gamma^\omega_{r_i,r_{i+1}}dr$ for all $r \in [r_i, r_{i+1}]$. Applying Jensen's and Hölder's inequalities to the finite sum in $\mathbb{E}\big[\|J^\gamma_\alpha(F)\|^2\big]$ yields

$$\mathbb{E}\big[\|J^\gamma_\alpha(F)\|^2\big] \leqslant Ch^{2|\alpha|_\tau} \sum_{\mathcal{R}(m,|\alpha|)} \mathbb{E}\left[\left\| \int\cdots\int_{0<r_{|\alpha|}<\cdots<r_1<1} F(y^\gamma_{r_{|\alpha|}}) - F(y^\gamma_0)\,dr_{|\alpha|}\cdots dr_1\right\|^4\right]^{\frac{1}{2}}$$
$$\times\, \mathbb{E}\left[\left\|\bigotimes_{j=1}^{|\alpha|_\omega} \gamma^\omega_{r_i^j, r_{i+1}^j}\right\|^4\right]^{\frac{1}{2}}.$$

By applying Jensen's inequality to the uniform distribution on $[s,t]$, we have that

$$\left\|\int_s^t z_r\,dr\right\|^4 = (t-s)^4\left\|\int_s^t \frac{z_r}{t-s}\,dr\right\|^4 \leqslant (t-s)^4\int_s^t \frac{\|z_r\|^4}{t-s}\,dr = (t-s)^3\int_s^t \|z_r\|^4\,dr\ ,$$

for any continuous integrable process z_r. Therefore,

$$\mathbb{E}\big[\|J^\gamma_\alpha(F)\|^2\big] \leqslant C_1h^{2|\alpha|_\tau} \sum_{\mathcal{R}(m,|\alpha|)} \mathbb{E}\left[\int\cdots\int_{0<r_{|\alpha|}<\cdots<r_1<1} \|F(y^\gamma_{r_{|\alpha|}}) - F(y^\gamma_0)\|^4 dr_{|\alpha|}\cdots dr_1\right]^{\frac{1}{2}}$$
$$\times\, \mathbb{E}\left[\left\|\bigotimes_{j=1}^{|\alpha|_\omega} \gamma^\omega_{r_i^j, r_{i+1}^j}\right\|^4\right]^{\frac{1}{2}}.$$

By repeatedly applying Hölder's inequality, we can estimate the last term as $O(h^{|\alpha|_\omega})$. Since $\mathrm{ord}(\alpha) = |\alpha|_\tau + \frac{1}{2}|\alpha|_\omega$, it follows from the global Lipschitz continuity of F that

$$\mathbb{E}\big[\|J^\gamma_\alpha(F)\|^2\big] \leqslant C_2\|F\|^2_{\text{Lip-}1}h^{2\mathrm{ord}(\alpha)}$$
$$\times \sum_{\mathcal{R}(m,|\alpha|)} \left(\int\cdots\int_{0<r_{|\alpha|}<\cdots<r_1<1} \mathbb{E}\Big[\|y^\gamma_{r_{|\alpha|}} - y^\gamma_0\|^4\Big] dr_{|\alpha|}\cdots dr_1\right)^{\frac{1}{2}}.$$

By Theorem 2.3.3, we have $\mathbb{E}\Big[\|y^\gamma_{r_{|\alpha|}} - y^\gamma_0\|^4\Big] = O(h^2)$ and thus the result follows. □

3.6 Commutativity simplifies Taylor expansions

In this section we show how the commutativity condition (2.2) simplifies certain terms in the Taylor expansions of the SDE (2.1) and CDE (2.3). In particular, for the integrals corresponding to words $\alpha \in \mathcal{A}$ containing only ω's the commutativity condition implies that only the symmetric part of the integrals appear in the Taylor expansion. This is made clear in Theorem 3.6.2. In the following lemma we show why this simplification occurs.

Lemma 3.6.1. *Suppose that the following commutativity condition holds*

$$g_i'(y)g_j(y) = g_j'(y)g_i(y), \quad \forall y \in \mathbb{R}^e, \tag{3.17}$$

for $i, j \in \{1, \cdots, d\}$*. Then for all words* $\alpha \in \mathcal{A}$ *containing only* ω*'s (i.e.* $|\alpha| = |\alpha|^{\omega}$*)* $V_\psi(\alpha) : \mathbb{R}^e \to L((\mathbb{R}^d)^{\otimes|\alpha|_\omega}, \mathbb{R}^l)$ *is symmetric in all but its first index. Where* $V_\psi(\alpha)$*, as defined in* (2.16)*, are the tensors multiplying the iterated integrals in the Taylor expansion.*

Proof. We will argue by induction on α.

We consider the case where $\psi(y) := y$, and so the recursion (2.16) becomes

$$V(l\beta)(y) = V(\beta)'V(l)(y)\ , \tag{3.18}$$

where $l \in \{\tau, \omega\}$, $V(\tau)(y) := f(y)$, $V(\omega)(y) := g(y)$ and $\beta \in \mathcal{A}$. Here we have that $V(\alpha) : \mathbb{R}^e \to L((\mathbb{R}^d)^{\otimes|\alpha|_\omega}, \mathbb{R}^e)$. The reader should note that the derivatives of ψ will always be multiplying on the left hand side in $V_\psi(\alpha)(y)$ and so its inclusion would not change the method of this proof.

Base case. $\alpha = \omega$ *and* $\alpha = \omega\omega$*.* $V(\omega)(y) = g(y)$ is a 2-tensor, it is thus trivially symmetric in its last index. For $\alpha = \omega\omega$, we have that

$$V(\omega\omega)(y) = g'(y)g(y)\ ,$$

thus (3.17) implies that $(V(\omega\omega))_{ij} = (V(\omega\omega))_{ji}$.

Induction step: $\alpha = \omega\omega\beta$. Let $\beta \in \mathcal{A}$ consist of only ω's and assume that $V(\beta)(y)$ is symmetric in all but its first index. By recursion on (3.18) and using the product rule

$$\begin{aligned} V(\omega\omega\beta)(y) &= V(\omega\beta)'V(\omega)(y) \\ &= \big(V(\beta)'V(\omega)\big)'V(\omega)(y) \\ &= V(\beta)''\big(g(y), g(y)\big) + V(\beta)'g'(y)g(y)\ . \end{aligned} \tag{3.19}$$

By our induction hypothesis, $V(\beta)(y)$ is symmetric and, therefore, its derivatives are symmetric in the same indices. We see that $V(\beta)''$ is symmetric and bilinear. Thus the left term in (3.19) is symmetric in all indices but the first. And, by our commutativity assumption (3.17), the right term is the multiplication of two symmetric tensors which is symmetric. We conclude by induction. □

This symmetry then tells us that the Taylor expansions simplify: leaving us with only the symmetric path of the integrals for all words α containing only ω's.

Theorem 3.6.2. *Suppose that the commutativity condition* (3.17) *holds, then for all words $\alpha \in \mathcal{A}$ containing only ω's (i.e. $|\alpha| = |\alpha|^{\omega}$) the following relations hold for $y \in \mathbb{R}^e$*

$$\begin{aligned} V_{\psi}(\alpha)(y)I_{\alpha}(1) &= \frac{1}{|\alpha|!}V_{\psi}(\alpha)(y)\big(I_{\omega}(1)\big)^{\otimes|\alpha|} , \\ V_{\psi}(\alpha)(y)I_{\alpha}^{\gamma}(1) &= \frac{1}{|\alpha|!}V_{\psi}(\alpha)(y)\big(I_{\omega}^{\gamma}(1)\big)^{\otimes|\alpha|} , \end{aligned}$$

where $V_{\psi}(\alpha)$ is defined as in Proposition 2.4.1.

Proof. The result follows by Theorem 3.1.8 as, in Lemma 3.6.1, we have proven that $V(\alpha)(y)$ is symmetric for all α containing only ω's. For example, by the symmetric-antisymmetric decomposition of the iterated integral I_{ij} (3.3), we have that

$$V(\omega\omega)I_{\omega\omega}(1) = \sum_{i,j=1}^{d} \big(V(\omega\omega)\big)_{ij} I_{ji} = \frac{1}{2}\sum_{i,j=1}^{d} \big(V(\omega\omega)\big)_{ij}\big(I_j \cdot I_i + I_{[j,i]}\big) .$$

And since, by Lemma 3.6.1, $\big(V(\omega\omega)\big)_{ij} = \big(V(\omega\omega)\big)_{ji}$ the antisymmetric component $I_{[j,i]} = -I_{[i,j]}$ will cancel out in the above sum. Thus, we see that

$$V(\omega\omega)I_{\omega\omega}(1) = \frac{1}{2}\sum_{i,j=1}^{d} \big(V(\omega\omega)\big)_{ij}\big(I_i \cdot I_j\big) = \frac{1}{2}V(\omega\omega)\big(I_{\omega}(1)\big)^{\otimes 2} .$$

In a similar way, for the higher order 'ω only' integrals, we are left with the symmetric component as given by Theorem 3.1.8. □

Remark 3.6.3. *Theorem 3.6.2 tells us that, when the commutativity condition is satisfied, the Taylor expansions of the SDE and CDE both simplify for words α containing only ω's. Thus, in order to match these terms in the Taylor expansions, we require only that $I_{\omega}^{\gamma}(1) = I_{\omega}(1)$. This will be made explicit in Sections 4.2.1 and 4.3.1.*

Chapter 4

Error analysis of path-based splitting schemes

Now that we have Taylor expansions for both the CDE and the Stratonovich SDE, along with a control over the size of the remainder terms in each, we can establish the strong and weak convergence properties of path-based splitting schemes. We first obtain local strong and weak error estimates using a direct application of Lemmas 3.5.1 and 3.5.2 before applying the framework of Milstein and Tretyakov (recall Theorem 1.1.3), which allows us to prove a global strong convergence rate for the approximating CDE. Similarly, we obtain the global weak error by a telescoping sum argument.

4.1 Local estimates

Theorem 4.1.1 (Local error estimates). *Suppose that the path $\gamma : [0,1] \to \mathbb{R}^{1+d}$ satisfies Assumption 2.3.1 and for a fixed $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, let $f \in \mathcal{C}_{\mathrm{Lip}}^{[p-1]}(\mathbb{R}^e, \mathbb{R}^e)$ and $g \in \mathcal{C}_{\mathrm{Lip}}^{2p-1}(\mathbb{R}^e, \mathbb{R}^{e\times d})$. Suppose also that the assumptions of Theorem 2.3.3 hold and the integrals $I_\alpha^\gamma(1)$ and $I_\alpha(1)$ agree almost surely for $\alpha \in \mathcal{A}$ with $\mathrm{ord}(\alpha) \leqslant p - \frac{1}{2}$ and in expectation for all $\alpha \in \mathcal{A}$ with $\mathrm{ord}(\alpha) = p$. Let Y_1 denote an approximation (e.g. using an ODE solver) of the CDE solution $\{y_r^\gamma\}_{r\in[0,1]}$ driven by γ, such that $y_0^\gamma = y_0$ and*

$$\mathbb{E}\big[\|y_1^\gamma - Y_1\|^2\big]^{\frac{1}{2}} = O(h^p), \quad \text{and} \quad \big\|\mathbb{E}[y_1^\gamma] - \mathbb{E}[Y_1]\big\| = O\big(h^{p+\frac{1}{2}}\big),$$

where $y = \{y_t\}_{t\in[0,h]}$ is the solution of the SDE (2.1) and $h > 0$ is the step size. Then

$$\mathbb{E}\big[\|y_h - Y_1\|^2\big]^{\frac{1}{2}} = O(h^p)\ , \quad \text{and} \quad \big\|\mathbb{E}[y_h] - \mathbb{E}[Y_1]\big\| = O\big(h^{p+\frac{1}{2}}\big)\ .$$

Where we remind the reader that, due to our parametrization, y_1^γ approximates y_h.

Proof. We start by proving the local strong error. By the triangle inequality,

$$\mathbb{E}\big[\|y_h - Y_1\|^2\big]^{\frac{1}{2}} \leqslant \mathbb{E}\big[\|y_h - y_1^\gamma\|^2\big]^{\frac{1}{2}} + \mathbb{E}\big[\|y_1^\gamma - Y_1\|^2\big]^{\frac{1}{2}} = \mathbb{E}\big[\|y_h - y_1^\gamma\|^2\big]^{\frac{1}{2}} + O(h^p),$$

as the second term is the difference between the CDE solution and its approximation.

Recall the remainder terms $R_p(h, y_0)$ and $R_p^\gamma(h, y_0)$ in Propositions 2.4.1 and 2.4.2. Then, by another two applications of the triangle inequality, it directly follows that

$$\begin{aligned}\mathbb{E}\big[\|y_h - Y_1\|^2\big]^{\frac{1}{2}} \leqslant \mathbb{E}\big[\|(y_h - R_p(h, y_0)) - (y_1^\gamma - R_p^\gamma(h, y_0))\|^2\big]^{\frac{1}{2}}\\ + \mathbb{E}\big[\|R_p(h, y_0)\|^2\big]^{\frac{1}{2}} + \mathbb{E}\big[\|R_p^\gamma(h, y_0)\|^2\big]^{\frac{1}{2}} + O(h^p),\end{aligned}$$

where the first term is simply the difference in the Taylor expansions, up to order p, of the SDE solution y_h and the CDE solution y_1^γ. Therefore, by the assumption that all integrals of the form $I_\alpha^\gamma(1)$ are matched almost surely for $\mathrm{ord}(\alpha) \leqslant p - \frac{1}{2}$, we have

$$\mathbb{E}\big[\|y_h - Y_1\|^2\big]^{\frac{1}{2}} \leqslant \mathbb{E}\big[\|R_p(h, y_0)\|^2\big]^{\frac{1}{2}} + \mathbb{E}\big[\|R_p^\gamma(h, y_0)\|^2\big]^{\frac{1}{2}} + O(h^p).$$

By Proposition 2.4.1, the SDE remainder term will satisfy $\mathbb{E}\big[\|R_p(h, y_0)\|^2\big]^{\frac{1}{2}} = O(h^{p+\frac{1}{2}})$. On the other hand, $R_p^\gamma(h, y_0)$ is given by (2.20) and therefore, by Lemma 3.5.2, we have

$$\mathbb{E}\big[\|R_p^\gamma(h, y_0)\|^2\big]^{\frac{1}{2}} = O\big(h^{p+\frac{1}{2}}\big).$$

This gives the desired result for the local strong error, that $\mathbb{E}\big[\|y_h - Y_1\|^2\big]^{\frac{1}{2}} = O(h^p)$.

We now turn our attention to the local weak error. Using the triangle inequality and the same Taylor expansions as in the proof of local strong error, it follows that

$$\begin{aligned}\big\|\mathbb{E}[y_h] - \mathbb{E}[Y_1]\big\| \leqslant \big\|\mathbb{E}\big[y_h - R_p(h, y_0)\big] - \mathbb{E}\big[y_1^\gamma - R_p^\gamma(h, y_0)\big]\big\|\\ + \big\|\mathbb{E}[y_1^\gamma] - \mathbb{E}[Y_1]\big\| + \big\|\mathbb{E}[R_p(h, y_0)]\big\| + \big\|\mathbb{E}[R_p^\gamma(h, y_0)]\big\|.\end{aligned}$$

From our assumption, the $I_\alpha^\gamma(1)$ terms in the SDE and CDE Taylor expansions are matched in expectation for $\mathrm{ord}(\alpha) \leqslant p$ and, therefore, the first term disappears. Moreover, we assume $\|\mathbb{E}[y_h] - \mathbb{E}[Y_1]\| = O\big(h^{p+\frac{1}{2}}\big)$ and, by Jensen's inequality, we have

$$\|\mathbb{E}[R_p(h, y_0)]\| \leqslant \mathbb{E}\big[\|R_p(h, y_0)\|^2\big]^{\frac{1}{2}}, \quad \text{and} \quad \|\mathbb{E}[R_p^\gamma(h, y_0)]\| \leqslant \mathbb{E}\big[\|R_p^\gamma(h, y_0)\|^2\big]^{\frac{1}{2}}.$$

Since the above terms were previously shown to be $O(h^{p+\frac{1}{2}})$, the result follows. □

As a simple extension of the above result we have the following estimate on the local weak error when including a test function $\psi : \mathbb{R}^e \to \mathbb{R}^l$.

Theorem 4.1.2 (Local weak error estimate.). *Suppose that the path $\gamma : [0, 1] \to \mathbb{R}^{1+d}$ satisfies Assumption 2.3.1 and for a fixed $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, let $f \in \mathcal{C}_{\mathrm{Lip}}^{[p-1]}(\mathbb{R}^e, \mathbb{R}^e)$, $g \in \mathcal{C}_{\mathrm{Lip}}^{2p-1}(\mathbb{R}^e, \mathbb{R}^{e\times d})$ and $\psi \in \mathcal{C}_b^{2p}(\mathbb{R}^e, \mathbb{R})$. Suppose also that the assumptions of Theorem 2.3.3 hold and the integrals $I_\alpha^\gamma(1)$ and $I_\alpha(1)$ agree in expectation for all $\alpha \in \mathcal{A}$ with* *ord$(\alpha) \leqslant p$. Let Y_1 denote an approximation (e.g. using an ODE solver) of the CDE solution $\{y_r^\gamma\}_{r\in[0,1]}$ driven by γ, such that $y_0^\gamma = y_0$ and*

$$\|\mathbb{E}[\psi(y_1^\gamma)] - \mathbb{E}[\psi(Y_1)]\| = O\big(h^{p+\frac{1}{2}}\big) ,$$

where $y = \{y_t\}_{t\in[0,h]}$ is the solution of the SDE (2.1) and $h > 0$ is the step size. Then

$$\|\mathbb{E}[\psi(y_h)] - \mathbb{E}[\psi(Y_1)]\| = O\big(h^{p+\frac{1}{2}}\big) . \tag{4.1}$$

4.2 Global strong error

As a consequence of the local estimates given by Theorem 4.1.1 and following the framework of Milstein and Tretyakov (recall Theorem 1.1.3), we obtain the global strong error rate:

Theorem 4.2.1 (Global strong error estimate)**.** *Given a fixed number of steps N, we define a numerical solution $\{Y_k\}_{0\leqslant k\leqslant N}$ for the SDE (2.1) over $[0,T]$ as follows,*

$$Y_{k+1} := \big(\textit{Solution at } r = 1 \textit{ of CDE (2.3) driven by } \gamma_k : [0,1] \to \mathbb{R}^{1+d}\big)\big(Y_k\big) + E_k,$$

where $Y_0 := y_0$. We assume that a suitable numerical scheme for the CDE (2.3) exists, and denote by $\{E_k\}$ the numerical errors ('CDE errors') made approximating its solution. Which we assume, for a fixed $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, uniformly satisfy

$$\mathbb{E}\big[\|E_k\|^2\big]^{\frac{1}{2}} = O(h^p) ,$$
$$\big\|\mathbb{E}[E_k]\big\| = O\big(h^{p+\frac{1}{2}}\big) .$$

In addition, we assume that each path $\gamma_k : [0,1] \to \mathbb{R}^{1+d}$ is expressible as $\gamma_k = \varphi\big(\big\{(t, W_t) : t \in \big[\frac{kT}{N}, \frac{(k+1)T}{N}\big]\big\}\big)$ for some fixed path-valued function φ. We will assume that the paths $\{\gamma_k\}$ uniformly satisfy Assumption 2.3.1 and that $f \in C_{\mathrm{Lip}}^{[p-1]}(\mathbb{R}^e, \mathbb{R}^e)$ and $g \in C_{\mathrm{Lip}}^{2p-1}(\mathbb{R}^e, \mathbb{R}^{e\times d})$. Suppose also that the assumptions of Theorem 2.3.3 hold and that the integrals $I_\alpha^{\gamma_k}(1)$ and $I_\alpha(1)$ agree almost surely for all $\alpha \in \mathcal{A}$ with ord$(\alpha) \leqslant p - \frac{1}{2}$ and in expectation for all $\alpha \in \mathcal{A}$ with ord$(\alpha) = p$. Then over the finite interval $[0,T]$, for $k \in \{1, 2, \cdots, N\}$, we have

$$\mathbb{E}\big[\|y_{kh} - Y_k\|^2\big]^{1/2} = O\big(h^{p-\frac{1}{2}}\big).$$

Remark 4.2.2 (Infinite time horizon)**.** *The convergence results in this paper are established over a finite time horizon T. However, our framework could be employed to deal with the infinite time horizon setting under suitable conditions on the SDE. For instance, if the SDE is ergodic with an exponential contraction property then contributions of local* *errors to the global error are reduced (exponentially in time). An extension of the classical Milstein-Tretyakov mean-square error analysis [48] to the infinite time horizon case for such contractive SDEs is given by [41, Theorem 3.3.]. See also [12, Section 3] for similar, but employing Multilevel Monte Carlo (MLMC).*

4.2.1 Global strong error for commutative SDEs

Although Theorem 4.2.1 identifies conditions on the splitting path γ to achieve a given strong convergence rate, it can be difficult to generate the required integrals (as discussed in the introduction). Fortunately, as shown in Section 3.6, the commutativity condition (2.2) leads to certain simplifications in the Taylor expansions of the Stratonovich SDE (2.1) and its CDE approximation (2.3). As a consequence, we have the following theorem.

Theorem 4.2.3 (Global strong error estimate for commutative SDEs)**.** *Assume that the SDE* (2.1) *satisfies the commutativity condition* (2.2), *and suppose that the assumptions of Theorem 4.2.1 hold for $p = 2$, but with the exception that each path γ_k is only assumed to satisfy the following equalities:*

$$I_{\omega}^{\gamma_k}(1) = W_{kh,(k+1)h} \ , \qquad I_{\tau}^{\gamma_k}(1) = h \ ,$$

$$I_{\omega\tau}^{\gamma_k}(1) = \frac{1}{2}hW_{kh,(k+1)h} + hH_{kh,(k+1)h}$$

$$\textit{and} \qquad \mathbb{E}\left[I_{\omega\omega\tau}^{\gamma_k}(1)\right] = \frac{1}{4}h^2 D_d^2 \ .$$

And for $i, j \in \{1, \cdots, d\}$ with $i \neq j$, we have that

$$\mathbb{E}\bigg[\int_0^1\int_0^{r_1}\int_0^{r_2} d\big(\gamma_k^{\omega}\big)^i(r_3)\, d\gamma_k^{\tau}(r_2)\, d\big(\gamma_k^{\omega}\big)^j(r_1)\bigg] = 0,$$
$$\mathbb{E}\bigg[\int_0^1\int_0^{r_1}\int_0^{r_2} d\gamma_k^{\tau}(r_3)\, d\big(\gamma_k^{\omega}\big)^i(r_2)\, d\big(\gamma_k^{\omega}\big)^j(r_1)\bigg] = 0.$$

Then on the interval $[0, T]$, for $k \in \{1, 2, \cdots, N\}$, the numerical solution $\{Y_k\}$ satisfies

$$\mathbb{E}\big[\, \|y_{kh} - Y_k\|^2\big]^{1/2} = O\big(h^{\frac{3}{2}}\big) \ .$$

Proof. By Theorem 3.6.2, we see that the CDE Taylor expansion can match all the 'noise only' terms simply by the path γ_k having the increment $\gamma_k^\omega(1) - \gamma_k^\omega(0) = W_{kh,(k+1)h}$. Adopting the element wise notation introduced in Definition 3.1.1, we recall Theorem 3.1.6, which gives the following decompositions of iterated integrals:

$$I_{ij} = \frac{1}{2} I_i \cdot I_j + \frac{1}{2} I_{[i,j]} \ , \tag{4.2}$$

$$I_{ijk} = \frac{1}{6} I_i \cdot I_j \cdot I_k + \frac{1}{4} I_i \cdot I_{[j,k]} + \frac{1}{4} I_{[i,j]} \cdot I_k + \frac{1}{6}\big(I_{[[i,j],k]} + I_{[i,[j,k]]}\big) \ , \tag{4.3}$$

for indices $i, j, k \in \{1, \cdots, d\}$. However, by identifying a coordinate of the Brownian motion with time, we see that the above would still hold when $i, j, k \in \{\tau, 1, \cdots, d\}$. We note also that these expansions hold for both the Stratonovich integrals $I_\cdot$ and the path integrals $I_\cdot^{\gamma_k}$. (In the following paragraph, the verb 'match' refers to when $r \mapsto (\gamma_k^\tau, \gamma_k^\omega)(r)$ and $t \mapsto (t, W_t)$ give the same iterated integral – either almost surely or in expectation).

Rearranging (4.2) we see that

$$I_{\tau\omega} = I_\tau \cdot I_\omega - I_{\omega\tau} \ ,$$

thus, by virtue of matching I_ω, I_τ and $I_{\omega\tau}$ (recall Theorem 3.3.1), γ_k matches $I_{\tau\omega}$.

By assumption, γ_k matches $I_{\omega\omega\tau}$ in expectation (recall (3.13)) and the lower order terms exactly, by (4.3) and the fact that $I_{[\tau,[i,i]]} = 0$, we see that

$$\mathbb{E}\Big[I_{ii\tau} - I_{ii\tau}^{\gamma_k}\Big] = \mathbb{E}\Big[I_{[i,[i,\tau]]} - I_{[i,[i,\tau]]}^{\gamma_k}\Big] = 0 \ .$$

By the antisymmetry of $[\,\cdot,\cdot\,]$ (recall $[i,j] = -[j,i]$), this implies that

$$\mathbb{E}\Big[I_{[[i,\tau],i]} - I_{[[i,\tau],i]}^{\gamma_k}\Big] = \mathbb{E}\Big[I_{[[\tau,i],i]} - I_{[[\tau,i],i]}^{\gamma_k}\Big] = \mathbb{E}\Big[I_{[i,[\tau,i]]} - I_{[i,[\tau,i]]}^{\gamma_k}\Big] = 0 \ . \tag{4.4}$$

Consulting (4.3) and again noting that lower order terms are matched, we thus see that γ matches the diagonals of $I_{\omega\tau\omega}$ and $I_{\tau\omega\omega}$ in expectation. By assumption, the off diagonals $\{I_{i\tau j}, I_{\tau ij}\}$ are matched in expectation and thus $I_{\omega\tau\omega}$ and $I_{\tau\omega\omega}$ are matched in expectation.

From the above, we see the Taylor expansions of the SDE (2.1) and CDE (2.3) coincide up to order $p = 2$, as required by Theorem 4.2.1. The result now follows. □

Remark 4.2.4. *In Theorem 4.2.1, we have accounted for the fact that the CDE (2.3), or rather the resulting sequence of ODEs, may be approximated using an ODE solver. However, obtaining the required estimates for these additional 'CDE errors' $\{E_k\}$ may be non-trivial and thus, we leave such an error analysis as a topic of future work. That said, to achieve strong order $3/2$ convergence, we expect that a single step of a second order ODE solver* *would be sufficient to discretize ODEs depending on just f and a single step of a fourth order solver (such as RK4) to suffice for the other ODEs. The intuition is that γ has Brownian-like scaling and so vector fields are either $O(h)$ or $O\big(h^{\frac{1}{2}}\big)$. Hence, we expect the local errors to be $O(h^3)$ or $O\big(h^{\frac{5}{2}}\big)$ in these two cases.*

Remark 4.2.5. *When the commutativity condition* (2.2) *is satisfied, to achieve a global strong convergence of $O(h)$ the splitting path need only match the integrals*

$$I^{\gamma}_{\omega}(1) = W_{s,t} \ , \qquad I^{\gamma}_{\tau}(1) = h \qquad \textit{and} \qquad \mathbb{E}[I^{\gamma}_{\omega\tau}(1)] = 0 \ .$$

4.3 Global weak error

Inspired by the approach in the Thesis [23], we see that the global weak error can be obtained from the local weak error.

Given some functional $\psi : \mathbb{R}^e \to \mathbb{R}^l$, we introduce the notation

$$P_t\psi(y) := \mathbb{E}[\psi(y_t)|y_0 = y] \qquad \text{and} \qquad Q_r\psi(y) := \mathbb{E}[\psi(y^{\gamma}_r)|y^{\gamma}_0 = y] \ , \tag{4.5}$$

where y_t is the solution to the SDE (2.1) at time $t > 0$ and y^{γ}_r is the solution to the CDE (2.3) at $r \in [0,1]$, both with initial value $y \in \mathbb{R}^d$.

Over an interval of time $T = N \times h$, for $h > 0$ and $N \in \mathbb{N}$, we can view our CDE approximation of $P_T\psi(y)$ as an iterative application of Q. That is, starting at $y^{\gamma}_0 = y_0$, we solve for y^{γ}_1 N times, consecutively, each time starting from the previous solution. As $(y^{\gamma}_r)_{r\in[0,1]}$ approximates y_t over the interval $[0,h]$, this process produces an approximation for y_T. We are then interested in the order of the weak error

$$\epsilon_N := \big\| P_T\psi(y) - \underbrace{Q_1 \ldots Q_1}_{N \text{ times}} \psi(y) \big\| \ . \tag{4.6}$$

Theorem 4.3.1. *Suppose that the path $\gamma : [0,1] \to \mathbb{R}^{1+d}$ satisfies Assumption 2.3.1 and for a fixed $p \in \{\frac{k}{2}\}_{k\in\mathbb{N}}$, let $f \in \mathcal{C}^{[p-1]}_{\mathrm{Lip}}(\mathbb{R}^e, \mathbb{R}^e)$, $g \in \mathcal{C}^{2p-1}_{\mathrm{Lip}}(\mathbb{R}^e, \mathbb{R}^{e\times d})$ and $\psi \in \mathcal{C}^{2p}_b(\mathbb{R}^e, \mathbb{R})$. Suppose also that the assumptions of Theorem 2.3.3 hold and the integrals $I^{\gamma}_{\alpha}(1)$ and $I_{\alpha}(1)$ agree in expectation for all $\alpha \in \mathcal{A}$ with ord$(\alpha) \leqslant p$, then*

$$\epsilon_N = O(h^{p-\frac{1}{2}}) \ .$$

Proof. We start by writing the left hand side as a telescoping sum

$$\begin{aligned}\|P_T\psi(y) - \underbrace{Q_1\dots Q_1}_{N \text{ times}}\psi(y)\| \leqslant& \|P_{h_N}(P_{T-h_N}\psi) - Q_1(P_{T-h_N}\psi)\| \\ &+ \sum_{j=1}^{N-1} \|\underbrace{Q_1\dots Q_1}_{N-j \text{ times}}(P_{h_j}(P_{t_{j-1}}\psi) - Q_1(P_{t_{j-1}}\psi))\| \ .\end{aligned}$$

As Q is a markov operator and P is a semigroup, we have that [10, Theorem 13.2]

$$\|\underbrace{Q_1\dots Q_1}_{N-j \text{ times}}(P_{h_j}(P_{t_{j-1}}\psi) - Q_1(P_{t_{j-1}}\psi))\| \leqslant \|P_{h_j}(P_{t_{j-1}}\psi) - Q_1(P_{t_{j-1}}\psi)\| \ ,$$

and

$$\|\underbrace{Q_1\dots Q_1}_{N-1 \text{ times}}(P_h\psi - Q_1\psi)\| \leqslant \|P_h\psi - Q_1\psi\| \ .$$

Thus, we see that

$$\epsilon_N \leqslant \|P_{h_1}\psi(y) - Q_1\psi(y)\| + \sum_{j=2}^{N} \|P_{h_j}(P_{t_{j-1}}\psi(y)) - Q_1(P_{t_{j-1}}\psi(y))\| \ ,$$

which is a sum of local errors. By Theorem 4.1.2, we know that the local errors are of order $O(h^{p+\frac{1}{2}})$. Thus we have that

$$\epsilon_N = \frac{T}{h}O(h^{p+\frac{1}{2}}) = O(h^{p-\frac{1}{2}}) \ .$$

□

Remark 4.3.2. *As for with the Strong error result, if we can discretize the CDE approximation with enough accuracy then the discretization will converge to the true solution with the rate of the path γ. Concretely, Let Y_1 denote an approximation (e.g. using an ODE solver) of the CDE solution $\{y_r^\gamma\}_{r\in[0,1]}$ driven by γ, such that $y_0^\gamma = y_0$ and for $p \geqslant \frac{1}{2}$*

$$\|\mathbb{E}[\psi(y_1^\gamma)] - \mathbb{E}[\psi(Y_1)]\| = O\big(h^{p+\frac{1}{2}}\big) \ ,$$

where $h > 0$ is the step size. If γ satisfies the assumptions of Theorem 4.3.1 for the same p and $q \geqslant p + \frac{1}{2}$, then

$$\|\mathbb{E}[\psi(y_T)] - \mathbb{E}[\psi(Y_N)]\| = O(h^{p-\frac{1}{2}}) \ .$$

4.3.1 Global weak error for commutative SDEs

Again, as a result of Theorem 3.6.2, the commutativity condition (2.2) makes it easier to obtain high order weak convergence. Here we present a version of Theorem 4.3.1, simplified slightly by the commutativity condition. Unfortunately, as we do not match the term $I_{\omega\omega\tau}$ in an a.s. sense, we must check the expected value for all permutations of the word $\omega\omega\omega\tau$ (see Remark 4.3.4). Given our path meets the conditions of Theorem 4.2.3 we must check that six integrals have expectation zero in order to prove a global weak order of 2.

Theorem 4.3.3 (Global weak error estimate for commutative SDEs)**.** *Suppose that the assumptions of Theorem 4.3.1 hold for $p = 2.5$, but with the exception that each path γ_k will now match only the following iterated integrals of the Brownian motion:*

$$I^{\gamma_k}_{\omega}(1) = W_{kh,(k+1)h} \ , \qquad I^{\gamma_k}_{\tau}(1) = h \ ,$$

$$I^{\gamma_k}_{\omega\tau}(1) = \frac{1}{2} hW_{kh,(k+1)h} + hH_{kh,(k+1)h} \ ,$$

$$\mathbb{E}\left[I^{\gamma_k}_{\omega\omega\tau}(1)\right] = \frac{1}{4}h^2 D_d^2 \ , \qquad \mathbb{E}\left[I^{\gamma_k}_{\omega\tau\tau}(1)\right] = 0 \ , \qquad \textit{and} \, ,$$

$$\mathbb{E}\left[I^{\gamma_k}_{\tau\omega\omega\omega}(1)\right] = \mathbb{E}\left[I^{\gamma_k}_{\omega\tau\omega\omega}(1)\right] = \mathbb{E}\left[I^{\gamma_k}_{\omega\omega\tau\omega}(1)\right] = \mathbb{E}\left[I^{\gamma_k}_{\omega\omega\omega\tau}(1)\right] = 0 \ ,$$

where 0 denotes the appropriate tensor of zeros. And that the 'off-diagonal' terms have expectation zero, that is (recalling the element wise notation of Definition 3.1.1) for $i, j \in \{1, \cdots, d\}$ with $i \neq j$

$$\mathbb{E}\Big[I^{\gamma_k}_{i\tau j}\Big] = \mathbb{E}\Big[I^{\gamma_k}_{\tau ij}\Big] = 0 \ .$$

Then on the interval $[0, T]$, for $k \in \{1, 2, \cdots, N\}$, the numerical solution $\{Y_k\}$ satisfies

$$\epsilon_k = O\big(h^2\big) \ .$$

Proof. As with the proof of Theorem 4.2.3, we note that matching the integrals I_τ and I_ω implies that we match all of the 'time only' integrals, and by Theorem 3.6.2 we match all of the 'noise only' integrals. We also recall that, by Theorem 3.3.4 for a word $\alpha \in \mathcal{A}$ with $ord(\alpha) = p \notin \mathbb{N}$

$$\mathbb{E}[I_\alpha(1)] = 0 \ .$$

That we match the integrals $I_{\tau\omega}$ almost surely, and $I_{\omega\tau\omega}$ and $I_{\tau\omega\omega}$ in expectation was already shown in the proof of Theorem 4.2.3. In a similar way as to how we obtained (4.4) in the strong order proof, we see that matching $I_{\omega\tau\tau}$ in expectation implies that

$$\mathbb{E}\Big[I_{[[i,\tau],\tau]} - I^{\gamma k}_{[[i,\tau],\tau]}\Big] = \mathbb{E}\Big[I_{[[\tau,i],\tau]} - I^{\gamma k}_{[[\tau,i],\tau]}\Big] = \mathbb{E}\Big[I_{[\tau,[\tau,i]]} - I^{\gamma k}_{[\tau,[\tau,i]]}\Big] = 0 \;,$$

which (again consulting the expansion (4.3)) implies that we match the integrals $I_{\tau\omega\tau}$ and $I_{\tau\tau\omega}$ in expectation. The other order 2.5 terms are matched in expectation by assumption. We thus match all terms up to order 2.5 in expectation and Theorem 4.3.1 then gives the claimed convergence rate. □

Remark 4.3.4. *In Theorem 4.3.3 we must check the expected value of the integrals for all permutations of the word $\omega\omega\omega\tau$. This is in contrast to Theorem 4.2.3 where we saw that it was enough to check that we match a single permutation of $\omega\omega\tau$ in expectation; This was possible as we assumed that $I_{\omega\tau}(1)$ is matched almost surely. As we do not aim to match the term $I_{\omega\omega\tau}(1)$ in an a.s. sense, this symmetry is not present in Theorem 4.3.3. Here we demonstrate this fact: Recalling* (3.5) *and expanding for $ii i\tau$, we see that*

$$\begin{aligned}\mathbb{E}[I_{iii\tau} - I^{\gamma}_{iii\tau}] = \frac{1}{12}\mathbb{E}\Big[&\big(I_i \cdot I_{[i,[i,\tau]]} - I^{\gamma}_i \cdot I^{\gamma}_{[i,[i,\tau]]}\big) + \big(I_{[i,[i,[i,\tau]]]} - I^{\gamma}_{[i,[i,[i,\tau]]]}\big) \\ &+ \Big(\big(I_{ii\tau i} - I_{iii\tau} + I_{\tau iii} - I_{i\tau ii}\big) - \big(I^{\gamma}_{ii\tau i} - I^{\gamma}_{iii\tau} + I^{\gamma}_{\tau iii} - I^{\gamma}_{i\tau ii}\big)\Big)\Big] \;.\end{aligned} \tag{4.7}$$

Expanding instead for τiii we find

$$\begin{aligned}\mathbb{E}[I_{\tau iii} - I^{\gamma}_{\tau iii}] = \frac{1}{12}\mathbb{E}\Big[&\big(I_{[[\tau,i],i]} \cdot I_i - I^{\gamma}_{[[\tau,i],i]} \cdot I^{\gamma}_i\big) + \big(I_{[[[\tau,i],i],i]} - I^{\gamma}_{[[[\tau,i],i],i]}\big) \\ &- \Big(\big(I_{ii\tau i} - I_{iii\tau} + I_{\tau iii} - I_{i\tau ii}\big) - \big(I^{\gamma}_{ii\tau i} - I^{\gamma}_{iii\tau} + I^{\gamma}_{\tau iii} - I^{\gamma}_{i\tau ii}\big)\Big)\Big] \;.\end{aligned}$$

Let $B := \mathbb{E}\big[I_i \cdot I_{[i,[i,\tau]]} - I^{\gamma}_i \cdot I^{\gamma}_{[i,[i,\tau]]}\big]$ and C denote the expected value of the remaining two differences in (4.7)*. Inspecting the terms in each of the above equations we see that*

$$\mathbb{E}[I_{iii\tau} - I^{\gamma}_{iii\tau}] = B + C \qquad \textit{and} \qquad \mathbb{E}[I_{\tau iii} - I^{\gamma}_{\tau iii}] = B - C \;.$$

It should thus be clear that, unless $B = 0$ (e.g. if $I_{\omega\omega\tau}(1) = I^{\gamma}_{\omega\omega\tau}(1)$), matching $I_{\omega\omega\omega\tau}$ does not imply that we match $I_{\tau\omega\omega\omega}$. Similar arguments can be made for $I_{\omega\tau\omega\omega}$ and $I_{\omega\tau\omega\omega}$.

4.4 A brief comparison with cubature on Wiener space

As they were an inspiration for this work, it is worth briefly pausing to discuss Cubature methods for SDEs [25, 26, 44]. Cubature methods provide weak approximations to SDEs, where stochastic integrals are replaced with collections of deterministic points and weights. For example, a Brownian increment $W_{s,t}$ may be replaced by the points $\pm\sqrt{(t-s)}$ with weights 0.5 (which we may interpret as replacing the Brownian motion with two paths with increments $\pm\sqrt{(t-s)}$). This approximation matches the first three moments of the Brownian increment. In practice, this replaces the SDE with a number of parallel ODEs which represent the distribution of the SDE solution. However, this approximation is only accurate for small time steps (and small noise). This means that the scheme must be iterated, and this takes the form of branching, which results in an exponential increase in the number of particles required to represent the SDE's solution (see for example Figure 1 in [26]). Thus, in order for Cubature methods to be practical, either the number of steps must be small or the number of particles must eventually be reduced [42]. This exponential growth limitation is not present with splitting methods as explored in this thesis: instead we are limited by Monte Carlo error (which is of $O(1/\sqrt{M})$ for M samples).

This thesis also takes the perspective of replacing the driving Brownian path with other paths; However, the paths we choose are still allowed to be random, and our numerical approximations thus fall within the Monte Carlo paradigm. Moreover, Cubature is a weak approximation scheme and the splitting schemes we focus on provide strong approximations. Therefore, our analysis differs in this regard.

For comparison with our methodology we recall the definition of Cubature on Wiener space [44, Definition 2.2], adopting our notation of Definition 3.1.1:

Definition 4.4.1. *Let m be a natural number. The n continuous paths with initial value 0 and bounded variation*

$$\omega_1,\ \dots,\ \omega_n \in C^0_{0,bv}([0,t],\mathbb{R}^d)\ ,$$

and the positive weights $\lambda_1,\ \dots,\ \lambda_n$ define a cubature formula on Wiener space of degree m at time t, if and only if, for all $(i_1,\dots,i_k)\in\mathcal{A}_m$,

$$\mathbb{E}\left[I_{i_1,\dots,i_k}\right] = \sum_{j=1}^{n}\lambda_j I^{\omega_j}_{i_1,\dots,i_k}\ . \tag{4.8}$$

We emphasise the involvement of the iterated integrals in (4.8) and its similarity to our requirement that integrals are matched in expectation in Theorem 4.1.2. Indeed, much of the approach and algebraic structure exploited in the analysis of Cubature schemes is of a similar flavour to that employed in this thesis. We may also rewrite (4.8) as

$$\mathbb{E}\left[I_{i_1,\ldots,i_k}\right] = \mathbb{E}_{\mathbb{Q}_t}\left[I_{i_1,\ldots,i_k}\right] \ ,$$

where the measure $\mathbb{Q}_t$ is associated to the paths $\omega_1, \ \ldots, \ \omega_n$. We point the reader to [44, Section 3] for details, but the above should make the comparison with Theorem 4.1.2 clear.

More generally, Cubature can be viewed as replacing random variables with a collection of random points and weights. And, with this perspective, it is worth noting that splitting methods and Cubature can be combined. For example, in the Strang splitting (see (5.3)), we may replace W_k with a degree 5 Cubature formula such as

$$\left\{\omega_1 = \sqrt{3h}, \lambda_1 = \frac{1}{6}\right\} \cup \left\{\omega_2 = -\sqrt{3h}, \lambda_2 = \frac{1}{6}\right\} \cup \left\{\omega_3 = 0, \lambda_3 = \frac{2}{3}\right\} \ .$$

A "Strang splitting + Cubature" methodology was considered in Theorem 2 of [54].

Chapter 5

Examples of path-based splitting schemes

5.1 Piecewise linear splitting paths

In this section, we present a variety of piecewise linear paths which fall into the proposed framework for developing SDE splitting methods (Theorem 4.2.1). These 'splitting paths' correspond to both well-known numerical methods (such as Lie-Trotter and Strang splitting) as well as the new high order splitting methods, which can exploit the optimal integral estimators that are derived in Section 5.3. Furthermore, we illustrate both the Strang and high order splitting paths in Figure 5.1. Throughout, we use the notation in Example 2.2.1 and define paths by their increments.

Example 5.1.1 (Lie-Trotter). *A Lie-Trotter splitting can be defined by one of two possible two-piece paths* $\gamma^{LT1}, \gamma^{LT2} : [0,1] \to \mathbb{R}^{1+d}$ *given by* $\gamma^{LT}(z) = (\gamma^{\tau}, \gamma^{\omega})(z)$ *with*

$$\gamma^{LT1}_{r_i, r_{i+1}} := \begin{cases} (h, 0), & \text{if } i = 0 \\ (0, W_{s,t}), & \text{if } i = 1, \end{cases} \tag{5.1}$$

$$\gamma^{LT2}_{r_i, r_{i+1}} := \begin{cases} (0, W_{s,t})\,, & \text{if } i = 0 \\ (h, 0)\,, & \text{if } i = 1. \end{cases} \tag{5.2}$$

Example 5.1.2 (Strang splitting). *The Strang splitting, see Figure 5.1, can be defined as a three-piece path* $\gamma^{S} : [0,1] \to \mathbb{R}^{1+d}$ *given by* $\gamma^{S}(z) = (\gamma^{\tau}, \gamma^{\omega})(z)$ *with the pieces:*

$$\gamma^{S}_{r_i, r_{i+1}} := \begin{cases} \left(\frac{1}{2}h, 0\right), & \text{if } i = 0 \\ (0, W_{s,t}), & \text{if } i = 1 \\ \left(\frac{1}{2}h, 0\right), & \text{if } i = 2. \end{cases} \tag{5.3}$$

We now proceed to 'higher order' piecewise linear paths that are constructed to match the increment $W_{s,t}$ and space-time Lévy area $H_{s,t}$ of the Brownian motion. For these paths to match higher order iterated integrals in expectation (for example, to achieve 3/2 strong convergence), then we necessarily require at least three pieces. The inability of paths with two pieces to match the conditions (2.6) and (2.7) required for high order strong convergence was explicitly shown in [14, p97 and Appendix A]. We begin by presenting paths (5.4) and (5.5), which each have a total of five pieces (vertical and horizontal), and can thus be seen as extensions of the Strang splitting.

Example 5.1.3 (High order Strang splitting (linear version)). *A high order Strang splitting, see Figure 5.1, can be defined using a five-piece path $\gamma^{HS1} : [0,1] \to \mathbb{R}^{1+d}$, which is linear in the Brownian motion and has the pieces:*

$$\gamma^{HS1}_{r_i,r_{i+1}} := \begin{cases} \big(\frac{3-\sqrt{3}}{6}h, 0\big), & \textit{if } i = 0 \\ \big(0, \frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\big), & \textit{if } i = 1 \\ \big(\frac{\sqrt{3}}{3}h, 0\big), & \textit{if } i = 2 \\ \big(0, \frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\big), & \textit{if } i = 3 \\ \big(\frac{3-\sqrt{3}}{6}h, 0\big), & \textit{if } i = 4. \end{cases} \tag{5.4}$$

Example 5.1.4 (High order Strang splitting (non-linear version)). *A high order Strang splitting, see Figure 5.1, can be defined as a five-piece path $\gamma^{HS2} : [0,1] \to \mathbb{R}^{1+d}$, which is based on an optimal estimator for a certain Brownian integral and has pieces:*

$$\gamma^{HS2}_{r_i,r_{i+1}} := \begin{cases} \big(0, \frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\big), & \textit{if } i = 0 \\ \big(\frac{1}{2}h, 0\big), & \textit{if } i = 1 \\ \big(0, C_{s,t}\big), & \textit{if } i = 2 \\ \big(\frac{1}{2}h, 0\big), & \textit{if } i = 3 \\ \big(0, \frac{1}{2}W_{s,t} - H_{s,t} - \frac{1}{2}C_{s,t}\big), & \textit{if } i = 4. \end{cases} \tag{5.5}$$

where the random vector $C_{s,t}$ is defined component-wise by

$$C^j_{s,t} := \epsilon^j_{s,t}\bigg(\frac{1}{3}\big(W^j_{s,t}\big)^2 + \frac{4}{5}\big(H^j_{s,t}\big)^2 + \frac{4}{15}h - \frac{1}{\sqrt{6\pi}}h^{\frac{1}{2}}n^j_{s,t}W^j_{s,t}\bigg)^{\frac{1}{2}}, \tag{5.6}$$

$$\epsilon^j_{s,t} := \operatorname{sgn}\bigg(W^j_{s,t} - \frac{3}{\sqrt{24\pi}}h^{\frac{1}{2}}n^j_{s,t}\bigg), \tag{5.7}$$

where $n^j_{s,t} := \operatorname{sgn}\big(H^j_{s,s+\frac{1}{2}h} - H^j_{s+\frac{1}{2}h,t}\big)$ are independent Rademacher random variables.

Remark 5.1.5. *The formula for $C_{s,t}$ is derived such that $\int_0^1 \big((\gamma_r^\omega)^j - (\gamma_0^\omega)^j\big)^2 d\gamma_r^\tau$ matches the optimal estimator $\mathbb{E}\big[\int_s^t \big(W_{s,u}^j\big)^2\, du \,\big|\, W_{s,t}, H_{s,t}, n_{s,t}\big]$ (see Theorem 5.3.3).*

For paths with three pieces, two of which are vertical and only relate to diffusion vector field, we refer to the resulting approximation as the 'Shifted ODE' approach. We use this terminology as, in the additive noise setting, the vertical pieces correspond to additive shifts for the numerical solution and so there is only one non-trivial ODE. As before, paths can be linear or non-linear functions of the input random variables.

Example 5.1.6 (Shifted ODE splitting (high order and linear))**.** *We can define a high order splitting by a three-piece path $\gamma^{SO1} : [0,1] \to \mathbb{R}^{1+d}$ with the following pieces:*

$$\gamma^{SO1}_{r_i, r_{i+1}} := \begin{cases} \big(0, H_{s,t} + \frac{1}{2}\sqrt{h}n_{s,t}\big), & \textit{if } i = 0 \\ \big(h, W_{s,t} - \sqrt{h}n_{s,t}\big), & \textit{if } i = 1 \\ \big(0, -H_{s,t} + \frac{1}{2}\sqrt{h}n_{s,t}\big), & \textit{if } i = 2. \end{cases} \tag{5.8}$$

Example 5.1.7 (Shifted ODE splitting (high order and non-linear))**.** *We can define a high order splitting, see Figure 5.1, using a three-piece path $\gamma^{SO2} : [0,1] \to \mathbb{R}^{1+d}$, which is based on an optimal estimator for a certain Brownian integral and has pieces:*

$$\gamma^{SO2}_{r_i, r_{i+1}} := \begin{cases} \big(0, \frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\big), & \textit{if } i = 0 \\ \big(h, C_{s,t}\big), & \textit{if } i = 1 \\ \big(0, \frac{1}{2}W_{s,t} - H_{s,t} - \frac{1}{2}C_{s,t}\big), & \textit{if } i = 2. \end{cases} \tag{5.9}$$

where the random vector $C_{s,t}$ is defined component-wise by

$$\begin{aligned} C_{s,t}^j &:= \epsilon_{s,t}^j\bigg(\big(W_{s,t}^j\big)^2 + \frac{12}{5}\big(H_{s,t}^j\big)^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}h^{\frac{1}{2}}n_{s,t}^j W_{s,t}^j\bigg)^{\frac{1}{2}}, \\ \epsilon_{s,t}^j &:= \operatorname{sgn}\bigg(W_{s,t}^j - \frac{3}{\sqrt{24\pi}}h^{\frac{1}{2}}n_{s,t}^j\bigg), \end{aligned}$$

where $n_{s,t}^j := \operatorname{sgn}\big(H^j_{s,s+\frac{1}{2}h} - H^j_{s+\frac{1}{2}h,t}\big)$ are independent Rademacher random variables.

Remark 5.1.8. *Just like Example 5.1.4, $C_{s,t}$ is derived so that $\int_0^1 \big((\gamma_r^\omega)^j - (\gamma_0^\omega)^j\big)^2 d\gamma_r^\tau$ matches the optimal estimator $\mathbb{E}\big[\int_s^t \big(W_{s,u}^j\big)^2\, du \,\big|\, W_{s,t}, H_{s,t}, n_{s,t}\big]$ (see Theorem 5.3.3).*

The following paths do not generally result in high order approximations for SDEs satisfying the commutativity condition (2.2). However, the piecewise linear path given by (5.10) results in the 'Shifted Euler' method for SDEs with additive noise, which we demonstrate can outperform the standard Euler-Maruyama method in [19, Section 5.2].

Example 5.1.9 (Shifted ODE splitting (low order; suitable for Euler's method))**.** *We can define a low order splitting by a three-piece path* $\gamma^{SO3} : [0,1] \to \mathbb{R}^{1+d}$ *with*

$$\gamma^{SO3}_{r_i, r_{i+1}} := \begin{cases} \big(0, \frac{1}{2}W_{s,t} + H_{s,t}\big), & \textit{if } i = 0 \\ \big(h, 0\big), & \textit{if } i = 1 \\ \big(0, \frac{1}{2}W_{s,t} - H_{s,t}\big), & \textit{if } i = 2. \end{cases} \tag{5.10}$$

The path (5.11) is also not usually high order, but gives a third order approximation when applied to underdamped Langevin dynamics (see [18] for details).

Example 5.1.10 (Shifted ODE splitting for the underdamped Langevin diffusion [18])**.** *We can define a splitting by a three-piece path* $\gamma^{SO4} : [0,1] \to \mathbb{R}^{1+d}$ *with pieces:*

$$\gamma^{SO4}_{r_i, r_{i+1}} := \begin{cases} \big(0, H_{s,t} + 6K_{s,t}\big), & \textit{if } i = 0 \\ \big(h, W_{s,t} - 12K_{s,t}\big), & \textit{if } i = 1 \\ \big(0, -H_{s,t} + 6K_{s,t}\big), & \textit{if } i = 2. \end{cases} \tag{5.11}$$

where $K_{s,t} \sim \mathcal{N}\big(0, \frac{1}{720}h\mathbf{1}_d\big)$ *(defined in Definition 3.2.6) is independent of* $\big(W_{s,t}, H_{s,t}\big)$.

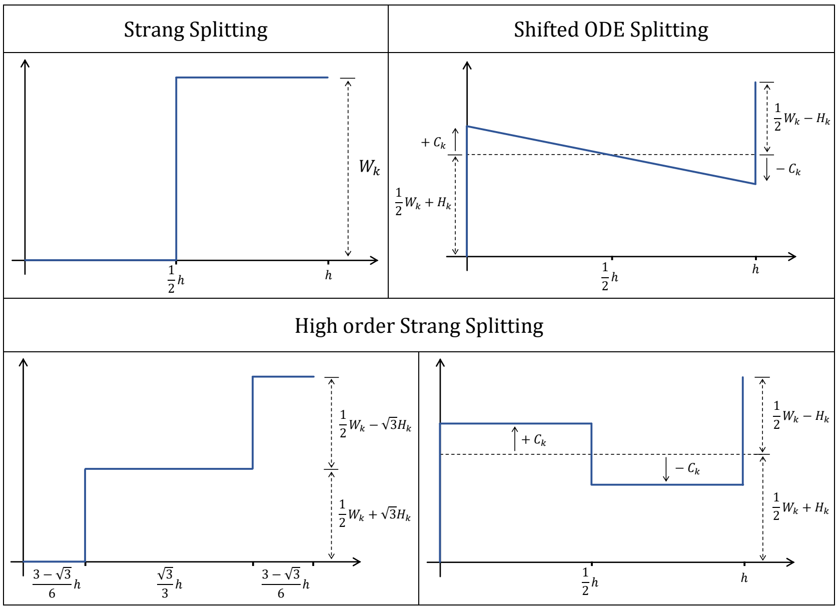


Fig. 5.1: Illustration of piecewise linear paths associated with various splitting methods for SDEs.

5.2 Theoretical convergence rates

We present here the application of the error analysis developed in Chapter 4 to the piecewise linear paths proposed in Section 5.1. We start by showing how our analysis applies to the Strang splitting path, recovering the expected error rates. We then consider the paths proposed in this thesis. For the sake of brevity, we will only explicit the results for our paths γ^{HS1} (5.4) and γ^{SO2} (5.9). From the working for these two paths it should be clear how to go about checking the terms for the remaining paths. The theoretical error rates for all the proposed paths can be found in Table 5.1 and a comparison with the conditions of Theorem 4.2.3 (global strong error for SDEs under the commutativity condition (2.2)) in Table 5.2.

	Strong error		Weak error	
	Commutative	Non-Commutative	Commutative	Non-Commutative
γ^{LT1}	$O(h)$	$O(h^{\frac{1}{2}})$	$O(h)$	$O(h)$
γ^{LT2}	$O(h)$	$O(h^{\frac{1}{2}})$	$O(h)$	$O(h)$
γ^{S}	$O(h)$	$O(h^{\frac{1}{2}})$	$O(h^2)$	$O(h)$
γ^{HS1}	$O(h^{\frac{3}{2}})$	$O(h^{\frac{1}{2}})$	$O(h^2)$	$O(h)$
γ^{HS2}	$O(h^{\frac{3}{2}})$	$O(h^{\frac{1}{2}})$	$O(h^2)$	$O(h)$
γ^{SO1}	$O(h^{\frac{3}{2}})$	$O(h^{\frac{1}{2}})$	$O(h^2)$	$O(h)$
γ^{SO2}	$O(h^{\frac{3}{2}})$	$O(h^{\frac{1}{2}})$	$O(h^2)$	$O(h)$
γ^{SO3}	$O(h)$	$O(h^{\frac{1}{2}})$	$O(h)$	$O(h)$
γ^{SO4}	$O(h)$	$O(h^{\frac{1}{2}})$	$O(h)$	$O(h)$

Table 5.1: Theoretical convergence rates of splitting paths. Without the commutativity condition the higher order schemes fail to match the word $\omega\omega$ a.s. and the word $\omega\omega\omega\omega$ in expectation. Strang splitting is a high order weak scheme as it matches the integral $\omega\omega\tau$ in expectation, which both $SO3$ and $SO4$ fail to do. However, while the paths $SO3$ and $SO4$ only obtain a global strong error of $O(h)$, we note that they have a local strong error of $O(h^2)$ compared with a local strong error of $O(h^{1.5})$ for the Strang splitting (see Table 5.2). Green text marks high order convergence.

5.2.1 Strang splitting path

As a first example, we apply our analysis to the Strang splitting path (5.3). We show that our analysis recovers the expected global *strong* convergence rates of $O(h)$ for commutative SDEs and $O(h^{\frac{1}{2}})$ for non-commutative SDEs, and the expected global *weak* convergence rates of $O(h^2)$ for commutative SDEs and $O(h)$ for non-commutative SDEs (see e.g. [13] and [4]). As our methodology describes, to obtain the error rates we calculate the iterated integrals for the Strang splitting path. We emphasise that applying the implemented algorithm of Section 3.4 automatically returns the final results of these calculations, but we will present here some intermediary steps for clarity. Proceeding as described in Section 3.4 and taking

$\gamma = \gamma^S$ (5.4), we obtain

$$I^S_\omega(1) = 3\int_{\frac{1}{3}}^{\frac{2}{3}} dr \times W_{s,t} = W_{s,t} \qquad \text{and} \qquad I^S_\tau(1) = 3\int_{\frac{2}{3}}^{1} dr \times \frac{h}{2} + 3\int_0^{\frac{1}{3}} dr \times \frac{h}{2} = h\ .$$

The Strang splitting matches $I_\omega(1)$, and thus, by Theorem 3.6.2, for commutative SDEs it matches all $I_\alpha(1)$ with 'ω only' words α. In general, we see that

$$I^S_{\omega\omega}(1) = 3^2\int_{\frac{1}{3}}^{\frac{2}{3}}\int_{\frac{1}{3}}^{r_1} dr_2 dr_1 \times W_{s,t}\otimes W_{s,t} = \frac{1}{2}W_{s,t}^{\otimes 2} \qquad \text{and} \qquad \mathbb{E}\Big[I^S_{\omega\omega}(1)\Big] = \frac{1}{2}hD_d^2\ ,$$

so, the Strang spitting does not match $I_{\omega\omega}(1) = \frac{1}{2}W_{s,t}^{\otimes 2} + A_{s,t}$ (3.3) (which is expected as we do not include $A_{s,t}$), but it does match the integral in expectation. Now, checking the $O(h^{1.5})$ terms, we have that

$$I^S_{\omega\tau}(1) = 3^2\int_{\frac{1}{3}}^{\frac{2}{3}}\int_0^{\frac{1}{3}} dr_2 dr_1 \times W_{s,t} \times \frac{h}{2} = |11| \times \frac{1}{2}hW_{s,t} = \frac{1}{2}hW_{s,t}\ ,$$

which (comparing with (3.6)) does not match the required value of $I_{\omega\tau}(1) = \frac{1}{2}hW_{s,t} + hH_{s,t}$. Thus, we see that the Strang splitting has a *local* strong error of $O(h^{1.5})$ for SDEs satisfying the commutativity condition, and a local strong error of $O(h)$ in general.

Local weak error

As the splitting matches $I_\tau(1)$, it matches all integrals $I_\alpha(1)$ for 'τ only' words α. We thus only need to check the local weak error for words containing at least one ω. In fact, as the Strang path is constructed from $W_{s,t} \sim \mathcal{N}(0, h\mathbf{1}_d)$ all integrals containing and odd number of ω's have expectation zero. We thus check the $O(h^2)$ terms

$$I^S_{\omega\omega\tau}(1) = 3^3\int_{\frac{2}{3}}^{1}\int_{\frac{1}{3}}^{\frac{2}{3}}\int_{\frac{1}{3}}^{r_2} dr_3 dr_2 dr_1 \times W_{s,t}\otimes W_{s,t} \times \frac{1}{2}h = |010| \times \frac{1}{2}hW_{s,t}^{\otimes 2} = \frac{1}{4}hW_{s,t}^{\otimes 2}\ ,$$

$$I^S_{\omega\tau\omega}(1) = 0 \qquad \text{and} \qquad I^S_{\tau\omega\omega}(1) = |101| \times \frac{1}{2}hW_{s,t}^{\otimes 2} = \frac{1}{4}hW_{s,t}^{\otimes 2}\ ,$$

which, comparing with Theorem 3.3.4, all have the correct expected values. We now come to the term which causes issues for the weak error of the Strang splitting (and all other splittings we present) when commutativity is not assumed. For the Strang splitting, we have

$$I^S_{\omega\omega\omega\omega}(1) = |1000| \times W^{\otimes 4} = \frac{1}{12}W^{\otimes 4}\ ,$$

and (as $\mathbb{E}\big[(W^i_{s,t})^4\big] = 3h^2$ and $\mathbb{E}\big[(W^i_{s,t})^2\big] = h$) we see that

$$\mathbb{E}\Big[\big(I^S_{\omega\omega\omega\omega}(1)\big)_{iiii}\Big] = \frac{1}{4}h^2 \qquad \text{and} \qquad \mathbb{E}\Big[\big(I^S_{\omega\omega\omega\omega}(1)\big)_{iijj}\Big] = \frac{1}{12}h^2 \ ,$$

for $i, j \in \{1, \dots, d\}$ with $i \neq j$. For similar reasons the Strang splitting fails to match the integral $\omega\omega\omega\omega\tau$. Thus we see that the Strang splitting achieves a local weak error of $O(h^3)$ when the commutativity condition is satisfied, and $O(h^2)$ in general. By Theorems 4.2.1 and 4.3.1, these local weak and strong error rates imply the global error rates of Table 5.1.

5.2.2 HS1 path

Strong error for HS1 path

For comparison with Theorem 4.2.3 we must calculate the integrals with words ω, τ, $\omega\tau$ and $\omega\omega\tau$. Proceeding as described in Section 3.4 and taking $\gamma = \gamma^{HS1}$ (5.4), we obtain

$$\begin{aligned} I^{HS1}_{\omega}(1) &= 5\int_{\frac{3}{5}}^{\frac{4}{5}} dr \times \left(\frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\right) + 5\int_{\frac{1}{5}}^{\frac{2}{5}} dr \times \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \\ &= |1| \times \left(\frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\right) + |3| \times \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) = W_{s,t} \ , \end{aligned}$$

where we recalled Remark 3.4.3 to calculate $|1| = |3| = 1$, and (now omitting the first step)

$$I^{HS1}_{\tau}(1) = |0| \times \frac{3-\sqrt{3}}{6}h + |2| \times \frac{\sqrt{3}}{3}h + |4| \times \frac{3-\sqrt{3}}{6}h = h \ .$$

After this level the benefit of the algorithm should become clear. Likewise we see that,

$$\begin{aligned} I^{HS1}_{\omega\tau}(1) &= |01| \times \left(\frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\right) \times \frac{3-\sqrt{3}}{6}h \\ &\quad + |03| \times \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \times \frac{3-\sqrt{3}}{6}h \\ &\quad + |21| \times \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \times \frac{\sqrt{3}}{3}h \\ &= \frac{1}{2}hW_{s,t} + hH_{s,t} \ , \end{aligned}$$

and, leaving the most involved calculation to last,

$$\begin{aligned}
I^{HS1}_{\omega\omega\tau}(1) &= |010| \times \left(\frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\right) \otimes \left(\frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\right) \times \frac{3-\sqrt{3}}{6}h \\
&\quad + |012| \times \left(\frac{1}{2}W_{s,t} - \sqrt{3}H_{s,t}\right) \otimes \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \times \frac{3-\sqrt{3}}{6}h \\
&\quad + |030| \times \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \otimes \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \times \frac{3-\sqrt{3}}{6}h \\
&\quad + |210| \times \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \otimes \left(\frac{1}{2}W_{s,t} + \sqrt{3}H_{s,t}\right) \times \frac{\sqrt{3}}{3}h \\
&= \frac{6-\sqrt{3}}{24}hW^{\otimes 2}_{s,t} + \frac{3-2\sqrt{3}}{4}hW_{s,t}\otimes H_{s,t} + \frac{\sqrt{3}}{4}hH_{s,t}\otimes W_{s,t} + \frac{\sqrt{3}}{2}hH^{\otimes 2}_{s,t} \; .
\end{aligned}$$

As $W_{s,t} \sim \mathcal{N}\big(0, h\mathbf{1}_d\big)$ and $H_{s,t} \sim \mathcal{N}\big(0, \frac{1}{12}h\mathbf{1}_d\big)$ are independent, we easily see that

$$\mathbb{E}[I^{HS1}_{\omega\omega\tau}(1)] = \left(\frac{6-\sqrt{3}}{24} + \frac{\sqrt{3}}{12\times 2}\right) h^2 D_d^2 = \frac{1}{4}h^2 D_d^2 \; . \tag{5.12}$$

It can also be reasoned from the independence of $W_{s,t}$ and $H_{s,t}$, and the construction of the path that the off-diagonal terms will have expectation zero. Alternatively, we can check explicitly as follows

$$\begin{aligned}
I^{HS1}_{\tau\omega\omega}(1) &= \frac{6-\sqrt{3}}{24}hW^{\otimes 2}_{s,t} - \frac{\sqrt{3}}{4}hW_{s,t}\otimes H_{s,t} + \frac{\sqrt{3}-2}{4}hH_{s,t}\otimes W_{s,t} + \frac{\sqrt{3}}{2}hH^{\otimes 2}_{s,t} \; , \\
I^{HS1}_{\omega\tau\omega}(1) &= \frac{\sqrt{3}}{12}hW^{\otimes 2}_{s,t} - \frac{1}{2}hW_{s,t}\otimes H_{s,t} + \frac{1}{2}hH_{s,t}\otimes W_{s,t} - \sqrt{3}hH^{\otimes 2}_{s,t} \; ,
\end{aligned}$$

clearly these terms have zero expectation off the diagonal. And, comparing with Theorem 3.3.4 we obtain the correct expected value (as predicted by Theorem 4.2.3 and (5.12)). Therefore by Theorem 4.2.3, the $HS1$ splitting path obtains a theoretical global strong convergence rate of $O(h^{\frac{3}{2}})$ for SDEs satisfying the commutativity condition (2.2).

Weak error for HS1 path

From the above strong error analysis, we can see that the path γ^{HS1} already satisfies the first three condition of Theorem 4.3.3, and clearly $\mathbb{E}\big[I^{HS1}_{\omega\tau}(1)\big] = \mathbb{E}\big[I^{HS1}_{\tau\omega}(1)\big] = 0$. Thus all we require to obtain global weak order 2 is that

$$\mathbb{E}\Big[I^{HS1}_{\alpha}(1)\Big] = 0 \; ,$$

for all words α with $ord(\alpha) = 2.5$. We note that a word α with $ord(\alpha) = 2.5$ must contain an odd number of ω. That is $|\alpha|^{\omega} \in \{1, 3, 5\}$, and, since γ^{HS1} is constructed from $W_{s,t} \sim \mathcal{N}(0, h\mathbf{1}_d)$ and $H_{s,t} \sim \mathcal{N}(0, \frac{1}{12}h\mathbf{1}_d)$ which are independent, it is easy to conclude that every component in the expansion of I_{α}^{HS1} will thus have expectation zero. We make this conclusion by observing that each term in this expansion must contain an odd number of either $W_{s,t}$ or $H_{s,t}$. Therefore, the path γ^{HS1} achieves a global weak error rate of 2, for SDEs satisfying the commutativity condition (2.2).

Remark 5.2.1. *Extending this analysis to terms of $O(h^3)$ we would see that the path fails to match the expected value of all the terms, and thus the path does not (in general) achieve a weak order of 3. In particular, the path γ^{HS1} fails to match the integrals with words $\omega\omega\omega\omega\tau$, $\omega\omega\omega\tau\omega$, $\omega\omega\tau\omega\omega$, $\omega\tau\omega\omega\omega$ and $\tau\omega\omega\omega\omega$ in expectation.*

5.2.3 SO2 path

Strong error for SO2 path

Firstly, recall that the path SO2 involves the r.v. $C_{s,t}$ defined component-wise by

$$\begin{aligned} C^j_{s,t} &:= \epsilon^j_{s,t}\bigg(\big(W^j_{s,t}\big)^2 + \frac{12}{5}\big(H^j_{s,t}\big)^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}h^{\frac{1}{2}}n^j_{s,t}W^j_{s,t}\bigg)^{\frac{1}{2}}, \\ \epsilon^j_{s,t} &:= \operatorname{sgn}\bigg(W^j_{s,t} - \frac{3}{\sqrt{24\pi}}h^{\frac{1}{2}}n^j_{s,t}\bigg), \end{aligned} \tag{5.13}$$

where $n^j_{s,t} := \operatorname{sgn}\big(H^j_{s,s+\frac{1}{2}h} - H^j_{s+\frac{1}{2}h,t}\big)$ are independent Rademacher random variables. Observe that by flipping the sign of $W_{s,t}$ we in turn flip the sign of $n_{s,t}$ and thus also the sign of $\epsilon_{s,t}$. As $W_{s,t}$ and $-W_{s,t}$ have the same distribution, we thus find that $\mathbb{E}[C^j_{s,t}] = -\mathbb{E}[C^j_{s,t}]$. This then implies that

$$\mathbb{E}[C^j_{s,t}] = 0 \ . \tag{5.14}$$

Now we will calculate the required terms for comparison with Theorem 4.2.3. In particular, the integrals corresponding to the words $\omega, \tau, \omega\tau$ and $\omega\omega\tau$. Taking $\gamma = \gamma^{SO2}$ (5.9) we have

$$\begin{aligned} I^{SO2}_{\omega}(1) &= |0| \times \left(\frac{1}{2}W_{s,t} - H_{s,t} - \frac{1}{2}C_{s,t}\right) + |1| \times C_{s,t} + |2| \times \left(\frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\right) \\ &= W_{s,t} \ , \end{aligned}$$

$$I^{S02}_{\tau}(1) = |1| \times h = h \ ,$$

$$\begin{aligned} I_{\omega\tau}^{SO2} &= |10| \times C_{s,t} \times h + |11| \times \left(\frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\right) \times h \\ &= \frac{1}{2}W_{s,t} + H_{s,t} \ , \end{aligned}$$

recalling that $|10| = \frac{1}{2}$. And we find that

$$\begin{aligned} I_{\omega\omega\tau}^{SO2}(1) = & \ |100| \times C_{s,t} \otimes C_{s,t} \times h \\ & + |101| \times \left(\frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\right) \otimes C_{s,t} \times h \\ & + |110| \times \left(\frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\right) \otimes \left(\frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\right) \times h \\ = & \ \frac{1}{8}hW_{s,t}^{\otimes 2} + \frac{1}{2}hH_{s,t}^{\otimes 2} + \frac{1}{24}hC_{s,t}^{\otimes 2} + \frac{1}{4}h\Big(W_{s,t} \otimes H_{s,t} + H_{s,t} \otimes W_{s,t}\Big) \\ & + \frac{1}{8}h\Big(W_{s,t} \otimes C_{s,t} - C_{s,t} \otimes W_{s,t}\Big) + \frac{1}{4}h\Big(H_{s,t} \otimes C_{s,t} - C_{s,t} \otimes H_{s,t}\Big) \ . \end{aligned} \tag{5.15}$$

As shown in (5.14) $\mathbb{E}[C_{s,t}^j] = 0$. Thus, consulting the definition of $C_{s,t}$ (5.13), it is clear from (5.15) that off the diagonal every term is the multiplication of two independent terms with expectation zero. Hence, for $i, j \in [1, \dots, d]$ with $i \neq j$

$$\mathbb{E}\Big[\big(I_{\omega\omega\tau}^{SO2}\big)_{ij}\Big] = 0 \ .$$

On the diagonal, the last two grouped terms in (5.15) cancel out and by the independence of $W_{s,t}$ and $H_{s,t}$ we see that

$$\begin{aligned} \mathbb{E}\Big[\big(I_{\omega\omega\tau}^{SO2}\big)_{ii}\Big] &= \mathbb{E}\Bigg[\frac{1}{8}h(W_{s,t}^i)^2 + \frac{1}{2}h(H_{s,t}^i)^2 \\ &\qquad + \frac{1}{24}h\bigg(\big(W_{s,t}^i\big)^2 + \frac{12}{5}\big(H_{s,t}^i\big)^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}h^{\frac{1}{2}}n_{s,t}^i W_{s,t}^i\bigg)\Bigg] \\ &= h^2\left(\frac{1}{8} + \frac{1}{24} + \frac{1}{24} + \frac{1}{120} + \frac{1}{30}\right) = \frac{1}{4}h^2 \end{aligned}$$

where we used that $n_{s,t}$ and $W_{s,t}$ are independent (see Lemma 3.2.10). It is also clear from (5.15) and from the independence of the terms that off the diagonal the expected value will be zero. It is also obvious that off the diagonal will have expectation zero for $I_{\omega\tau\omega}^{SO2}$ and $I_{\tau\omega\omega}^{SO2}$.

Weak error for SO2 path

The weak error for the SO2 path is slightly more involved than for the HS1 path: as here we cannot simply appeal to the independence of $W_{s,t}$ and $H_{s,t}$. We must thus check the conditions of Theorem 4.3.3 more carefully. In this section, we will skip straight to the integral calculations obtained by applying the algorithm of Section 3.4. Taking $\gamma = \gamma^{S02}$ we have that

$$I^{SO2}_{\omega\tau\tau}(1) = \frac{1}{4}h^2 W_{s,t} + \frac{1}{2}h^2 H_{s,t} - \frac{1}{12}h^2 C_{s,t}\ ,$$

and, as noted in (5.14), $W_{s,t},\ H_{s,t}$ and $C_{s,t}$ all have expectation zero. Thus we see that $\mathbb{E}[I^{SO2}_{\omega\tau\tau}(1)] = 0$. And, now onto the terms involving more ω's. Which are more involved:

$$\begin{aligned}
I^{SO2}_{\tau\omega\omega\omega} = {} & \frac{1}{48}hW^{\otimes 3}_{s,t} - \frac{1}{6}H^{\otimes 3}_{s,t} && (5.16)\\
& + \frac{1}{12}h\big(H^{\otimes 2}_{s,t}\otimes W_{s,t} + H_{s,t}\otimes W_{s,t}\otimes H_{s,t} + W_{s,t}\otimes H^{\otimes 2}_{s,t}\big) && (5.17)\\
& - \frac{1}{24}h\big(H_{s,t}\otimes W^{\otimes 2}_{s,t} + W_{s,t}\otimes H_{s,t}\otimes W_{s,t} + W^{\otimes 2}_{s,t}\otimes H_{s,t}\big) && (5.18)\\
& + \frac{1}{48}h\big(2C_{s,t}\otimes W^{\otimes 2}_{s,t} - W_{s,t}\otimes C_{s,t}\otimes W_{s,t} - W^{\otimes 2}_{s,t}\otimes C_{s,t}\big) && (5.19)\\
& + \frac{1}{12}h\big(2C_{s,t}\otimes H^{\otimes 2}_{s,t} - H_{s,t}\otimes C_{s,t}\otimes H_{s,t} - H^{\otimes 2}_{s,t}\otimes C_{s,t}\big) && (5.20)\\
& + \frac{1}{24}h\big(C^{\otimes 2}_{s,t}\otimes W_{s,t} - C_{s,t}\otimes W_{s,t}\otimes C_{s,t} + \frac{1}{2}W_{s,t}\otimes C^{\otimes 2}_{s,t}\big) && (5.21)\\
& - \frac{1}{12}h\big(C^{\otimes 2}_{s,t}\otimes H_{s,t} - C_{s,t}\otimes H_{s,t}\otimes C_{s,t} + \frac{1}{2}H_{s,t}\otimes C^{\otimes 2}_{s,t}\big) && (5.22)\\
& + \frac{1}{24}h\Big(H_{s,t}\otimes C_{s,t}\otimes W_{s,t} + H_{s,t}\otimes W_{s,t}\otimes C_{s,t} + W_{s,t}\otimes C_{s,t}\otimes H_{s,t} && (5.23)\\
& \qquad + W_{s,t}\otimes H_{s,t}\otimes C_{s,t} - 2C_{s,t}\otimes W_{s,t}\otimes H_{s,t} - 2C_{s,t}\otimes H_{s,t}\otimes W_{s,t}\Big)\ ,
\end{aligned}$$

lines (5.16), (5.17) and (5.18) have expectation zero as $W_{s,t} \sim \mathcal{N}(0,h)$ and $H_{s,t} \sim \mathcal{N}(0,h/12)$ are independent. For the other lines 'off the diagonal' we have expectation zero by independence and as $\mathbb{E}[C^j_{s,t}] = 0$. For lines (5.19), (5.20) and (5.23) 'on the diagonal' we have zero a.s. as terms cancel out. For lines (5.21) and (5.22) 'on the diagonal' note that

$$W^i_{s,t}(C^i_{s,t})^2 = W^i_{s,t} \times \left(\left(W^i_{s,t}\right)^2 + \frac{12}{5}\left(H^i_{s,t}\right)^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}h^{\frac{1}{2}}n^i_{s,t}W^i_{s,t}\right)\ ,$$

which has expectation zero as $n^i_{s,t}$ and $W^i_{s,t}$ are independent and $\mathbb{E}[n_{s,t}] = 0$, and similarly we observe that $\mathbb{E}[H^i_{s,t}(C^i_{s,t})^2] = 0$. Thus we see that (5.21) and (5.22) have expectation zero, and so

$$\mathbb{E}\big[I^{SO2}_{\tau\omega\omega\omega}(1)\big] = 0\ .$$

The same reasoning can be used to show that

$$\mathbb{E}\big[I^{SO2}_{\omega\tau\omega\omega}(1)\big] = \mathbb{E}\big[I^{SO2}_{\omega\omega\tau\omega}(1)\big] = \mathbb{E}\big[I^{SO2}_{\omega\omega\omega\tau}(1)\big] = 0 \ ,$$

where $I^{SO2}_{\omega\tau\omega\omega}(1)$, $I^{SO2}_{\omega\omega\tau\omega}(1)$ and $I^{SO2}_{\omega\omega\omega\tau}(1)$ are given as follows:

$$\begin{aligned}
I^{SO2}_{\omega\tau\omega\omega} = {} & \frac{1}{16}hW^{\otimes 3}_{s,t} + \frac{1}{2}H^{\otimes 3}_{s,t} \\
& + \frac{1}{4}h\big(W_{s,t}\otimes H^{\otimes 2}_{s,t} + H_{s,t}\otimes W_{s,t}\otimes H_{s,t} - H^{\otimes 2}_{s,t}\otimes W_{s,t}\big) \\
& + \frac{1}{8}h\big(H_{s,t}\otimes W^{\otimes 2}_{s,t} - W_{s,t}\otimes H_{s,t}\otimes W_{s,t} - W^{\otimes 2}_{s,t}\otimes H_{s,t}\big) \\
& + \frac{1}{48}h\big(W_{s,t}\otimes C^{\otimes 2}_{s,t} - 2C^{\otimes 2}_{s,t}\otimes W_{s,t}\big) + \frac{1}{24}h\big(H_{s,t}\otimes C^{\otimes 2}_{s,t} + 2C^{\otimes 2}_{s,t}\otimes H_{s,t}\big) \\
& + \frac{1}{16}h\big(W_{s,t}\otimes C_{s,t}\otimes W_{s,t} - W^{\otimes 2}_{s,t}\otimes C_{s,t}\big) \\
& + \frac{1}{4}h\big(H^{\otimes 2}_{s,t}\otimes C_{s,t} - H_{s,t}\otimes C_{s,t}\otimes H_{s,t}\big) \\
& + \frac{1}{8}h\Big(H_{s,t}\otimes C_{s,t}\otimes W_{s,t} + W_{s,t}\otimes H_{s,t}\otimes C_{s,t} \\
& \qquad - W_{s,t}\otimes C_{s,t}\otimes H_{s,t} - H_{s,t}\otimes W_{s,t}\otimes C_{s,t}\Big) \ ,
\end{aligned}$$

$$\begin{aligned}
I^{SO2}_{\omega\omega\tau\omega} = {} & \frac{1}{16}hW^{\otimes 3}_{s,t} - \frac{1}{2}H^{\otimes 3}_{s,t} \\
& + \frac{1}{4}h\big(H^{\otimes 2}_{s,t}\otimes W_{s,t} - H_{s,t}\otimes W_{s,t}\otimes H_{s,t} - W_{s,t}\otimes H^{\otimes 2}_{s,t}\big) \\
& + \frac{1}{8}h\big(H_{s,t}\otimes W^{\otimes 2}_{s,t} + W_{s,t}\otimes H_{s,t}\otimes W_{s,t} - W^{\otimes 2}_{s,t}\otimes H_{s,t}\big) \\
& + \frac{1}{48}h\big(C^{\otimes 2}_{s,t}\otimes W_{s,t} - 2W_{s,t}\otimes C^{\otimes 2}_{s,t}\big) - \frac{1}{24}h\big(2H_{s,t}\otimes C^{\otimes 2}_{s,t} + C^{\otimes 2}_{s,t}\otimes H_{s,t}\big) \\
& + \frac{1}{16}h\big(W_{s,t}\otimes C_{s,t}\otimes W_{s,t} - C_{s,t}\otimes W^{\otimes 2}_{s,t}\big) \\
& + \frac{1}{4}h\big(C_{s,t}\otimes H^{\otimes 2}_{s,t} - H_{s,t}\otimes C_{s,t}\otimes H_{s,t}\big) \\
& + \frac{1}{8}h\Big(H_{s,t}\otimes C_{s,t}\otimes W_{s,t} + C_{s,t}\otimes W_{s,t}\otimes H_{s,t} \\
& \qquad - W_{s,t}\otimes C_{s,t}\otimes H_{s,t} - C_{s,t}\otimes H_{s,t}\otimes W_{s,t}\Big) \ ,
\end{aligned}$$

$$
\begin{aligned}
I^{SO2}_{\omega\omega\omega\tau} = {} & \frac{1}{48}hW^{\otimes 3}_{s,t} + \frac{1}{6}H^{\otimes 3}_{s,t} \\
& + \frac{1}{12}h\big(H^{\otimes 2}_{s,t} \otimes W_{s,t} + H_{s,t} \otimes W_{s,t} \otimes H_{s,t} + W_{s,t} \otimes H^{\otimes 2}_{s,t}\big) \\
& + \frac{1}{24}h\big(H_{s,t} \otimes W^{\otimes 2}_{s,t} + W_{s,t} \otimes H_{s,t} \otimes W_{s,t} + W^{\otimes 2}_{s,t} \otimes H_{s,t}\big) \\
& + \frac{1}{24}h\big(W_{s,t} \otimes C^{\otimes 2}_{s,t} - C_{s,t} \otimes W_{s,t} \otimes C_{s,t} + \frac{1}{2}C^{\otimes 2}_{s,t} \otimes W_{s,t}\big) \\
& + \frac{1}{12}h\big(H_{s,t} \otimes C^{\otimes 2}_{s,t} - C_{s,t} \otimes H_{s,t} \otimes C_{s,t} + \frac{1}{2}C^{\otimes 2}_{s,t} \otimes H_{s,t}\big) \\
& + \frac{1}{48}h\big(2W^{\otimes 2}_{s,t} \otimes C_{s,t} - W_{s,t} \otimes C_{s,t} \otimes W_{s,t} - C_{s,t} \otimes W^{\otimes 2}_{s,t}\big) \\
& + \frac{1}{12}h\big(2H^{\otimes 2}_{s,t} \otimes C_{s,t} - H_{s,t} \otimes C_{s,t} \otimes H_{s,t} - C_{s,t} \otimes H^{\otimes 2}_{s,t}\big) \\
& + \frac{1}{24}h\Big(2W_{s,t} \otimes H_{s,t} \otimes C_{s,t} + 2H_{s,t} \otimes W_{s,t} \otimes C_{s,t} - W_{s,t} \otimes C_{s,t} \otimes H_{s,t} \\
& \qquad - H_{s,t} \otimes C_{s,t} \otimes W_{s,t} - C_{s,t} \otimes W_{s,t} \otimes H_{s,t} - C_{s,t} \otimes H_{s,t} \otimes W_{s,t}\Big) \,.
\end{aligned}
$$

By Theorem 4.3.3, we thus conclude that the path γ^{SO2} achieves a global weak rate of convergence of $O(h^2)$ for SDEs satifying the commutativity condition (2.2).

	$\alpha =$	ω	τ	$\omega\tau$	$\omega\omega\tau$
γ^{LT1}	$I_\alpha(1)$	$W_{s,t}$	h	0	0
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	0
γ^{LT2}	$I_\alpha(1)$	$W_{s,t}$	h	$hW_{s,t}$	$\frac{1}{2}hW_{s,t}^{\otimes 2}$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{2}h^2 D_d^2$
γ^{S}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t}$	$\frac{1}{4}hW_{s,t}^{\otimes 2}$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{4}h^2 D_d^2$
γ^{HS1}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t} + hH_{s,t}$	$\frac{6-\sqrt{3}}{24}hW_{s,t}^{\otimes 2} + \frac{3-2\sqrt{3}}{4}hW_{s,t} \otimes H_{s,t}$ $+\frac{\sqrt{3}}{4}hH_{s,t} \otimes W_{s,t} + \frac{\sqrt{3}}{2}hH_{s,t}^{\otimes 2}$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{4}h^2 D_d^2$
γ^{HS2}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t} + hH_{s,t}$	$\frac{1}{8}hW_{s,t}^{\otimes 2} + \frac{1}{2}hH_{s,t}^{\otimes 2} + \frac{1}{8}hC_{s,t}^{\otimes 2}$ $+\frac{1}{4}h(W_{s,t} \otimes H_{s,t} + H_{s,t} \otimes W_{s,t})$ $+\frac{1}{8}h(W_{s,t} \otimes C_{s,t} - C_{s,t} \otimes W_{s,t})$ $+\frac{1}{4}h(H_{s,t} \otimes C_{s,t} - C_{s,t} \otimes H_{s,t})$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{4}h^2 D_d^2$
γ^{SO1}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t} + hH_{s,t}$	$\frac{1}{6}hW_{s,t}^{\otimes 2} + \frac{1}{2}hH_{s,t}^{\otimes 2} + \frac{1}{24}h^2 n_{s,t}^{\otimes 2}$ $+\frac{1}{12}h^{3/2}(n_{s,t} \otimes W_{s,t} - 2W_{s,t} \otimes n_{s,t})$ $+\frac{1}{4}h^{3/2}(n_{s,t} \otimes H_{s,t} - H_{s,t} \otimes n_{s,t})$ $+\frac{1}{2}hH_{s,t} \otimes W_{s,t}$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{4}h^2 D_d^2$
γ^{SO2}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t} + hH_{s,t}$	$\frac{1}{8}hW_{s,t}^{\otimes 2} + \frac{1}{2}hH_{s,t}^{\otimes 2} + \frac{1}{24}hC_{s,t}^{\otimes 2}$ $+\frac{1}{4}h(W_{s,t} \otimes H_{s,t} + H_{s,t} \otimes W_{s,t})$ $+\frac{1}{8}h(W_{s,t} \otimes C_{s,t} - C_{s,t} \otimes W_{s,t})$ $+\frac{1}{4}h(H_{s,t} \otimes C_{s,t} - C_{s,t} \otimes H_{s,t})$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{4}h^2 D_d^2$
γ^{SO3}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t} + hH_{s,t}$	$\frac{1}{8}hW_{s,t}^{\otimes 2} + \frac{1}{2}hH_{s,t}^{\otimes 2}$ $+\frac{1}{4}h(W_{s,t} \otimes H_{s,t} + H_{s,t} \otimes W_{s,t})$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{1}{6}h^2 D_d^2$
γ^{SO4}	$I_\alpha(1)$	$W_{s,t}$	h	$\frac{1}{2}hW_{s,t} + hH_{s,t}$	$\frac{1}{6}hW_{s,t}^{\otimes 2} + \frac{1}{2}hH_{s,t}^{\otimes 2} + 6hK_{s,t}^{\otimes 2}$ $+h(K_{s,t} \otimes W_{s,t} - 2W_{s,t} \otimes K_{s,t})$ $+3h(K_{s,t} \otimes H_{s,t} - H_{s,t} \otimes K_{s,t})$ $+\frac{1}{2}hH_{s,t} \otimes W_{s,t}$
	$\mathbb{E}[I_\alpha(1)]$	0	h	0	$\frac{13}{60}h^2 D_d^2$

Table 5.2: Checking the conditions of Theorem 4.2.3 for example paths. Highlighted in red are where the paths fail to satisfy the conditions in the theorem. D_d^2, as defined in (3.14), represents the $d \times d$ identity matrix.

5.3 Derivation and optimality of splittings

Declaration of authorship The following section should be credited to Dr. James Foster[1]. We include it here for completeness.

In this section, we derive estimators for certain iterated stochastic integrals using a polynomial expansion of Brownian motion [17]. We use this expansion since its first two coefficients give the path's increment and space-time Lévy area (Definition 3.2.2). Just as in [17], the integral that we would primarily like to approximate is the so-called 'space-space-time' Lévy area, which we defined above (Definition 3.2.4). We note that a preliminary version of the results in this section were first presented in the doctoral thesis [14].

The key difference between the integral estimators defined in this section and those derived in [17], is that we shall additionally make use of the 'space-time Lévy swing' (Definition 3.2.3). Similar to [17], we propose approximating $L_{s,t}$ using its conditional expectation. That is, we would like to derive a closed-form expression for $\mathbb{E}\big[L_{s,t}\,|\,W_{s,t}, H_{s,t}, n_{s,t}\big]$. In addition, we shall derive the conditional variance of $L_{s,t}$ as this gives the $L^2(\mathbb{P})$ error.

In this section, we focus on the case where Brownian motion is one-dimensional and leave the general case, a matrix of space-space-time Lévy areas, as future work. However, the off-diagonal terms in this matrix will have zero expectation due to the independence and symmetry of the d coordinate processes of the Brownian motion. Therefore, we may construct a high order multidimensional splitting path simply by taking independent copies of the paths detailed in Section 5.1.

Theorem 5.3.1 (An optimal unbiased estimator of space-space-time Lévy area). *Let $H_{s,t}$ and $L_{s,t}$ be the previously defined Lévy areas of Brownian motion and time. Let $n_{s,t} := \operatorname{sgn}(H_{s,u} - H_{u,t})$ denote the space-time Lévy swing given by definition 3.2.3. Then the conditional mean and variance of $L_{s,t}$ given the information $(W, H, n)_{s,t}$ is*

$$\mathbb{E}\big[L_{s,t}\,|\,W_{s,t}, H_{s,t}, n_{s,t}\big] = \frac{1}{30}h^2 + \frac{3}{5}hH_{s,t}^2 - \frac{1}{8\sqrt{6\pi}}n_{s,t}h^{\frac{3}{2}}W_{s,t}, \tag{5.24}$$

$$\begin{aligned}\operatorname{Var}\big(L_{s,t}\,|\,W_{s,t}, H_{s,t}, n_{s,t}\big) = \frac{11}{25200}h^4 + \Big(\frac{1}{720} - \frac{1}{384\pi}\Big)h^3W_{s,t}^2 + \frac{1}{700}h^3H_{s,t}^2 \\ - \frac{1}{320\sqrt{6\pi}}n_{s,t}h^{\frac{7}{2}}W_{s,t}.\end{aligned} \tag{5.25}$$

[1] University of Bath, Department of Mathematical Sciences. jmf68@bath.ac.uk

Proof. We first note by applying [17, Theorem 3.10] on $[s,u]$ and $[u,t]$, we have

$$\begin{aligned}
\mathbb{E}\big[L_{s,u}\,|\,W_{s,u},H_{s,u}\big] &= \frac{1}{120}h^2 + \frac{3}{10}hH_{s,u}^2,\\
\mathrm{Var}\big(L_{s,u}\,|\,W_{s,u},H_{s,u}\big) &= \frac{11}{403200}h^4 + h^3\Big(\frac{1}{5760}W_{s,u}^2 + \frac{1}{5600}H_{s,u}^2\Big),\\
\mathbb{E}\big[L_{u,t}\,|\,W_{u,t},H_{u,t}\big] &= \frac{1}{120}h^2 + \frac{3}{10}hH_{u,t}^2,\\
\mathrm{Var}\big(L_{u,t}\,|\,W_{u,t},H_{u,t}\big) &= \frac{11}{403200}h^4 + h^3\Big(\frac{1}{5760}W_{u,t}^2 + \frac{1}{5600}H_{u,t}^2\Big).
\end{aligned}$$

To utilise the above expectations, we will 'expand' the following integrals over $[s,t]$.

$$\begin{aligned}
\int_s^t W_{s,r}\,dr &= \int_s^u W_{s,r}\,dr + \int_u^t W_{s,r}\,dr \qquad (5.26)\\
&= \int_s^u W_{s,r}\,dr + \frac{1}{2}hW_{s,u} + \int_u^t W_{u,r}\,dr,\\
\int_s^t W_{s,r}^2\,dr &= \int_s^u W_{s,r}^2\,dr + \int_u^t W_{s,r}^2\,dr \qquad (5.27)\\
&= \int_s^u W_{s,r}^2\,dr + \frac{1}{2}hW_{s,u}^2 + 2W_{s,u}\int_u^t W_{u,r}\,dr + \int_u^t W_{u,r}^2\,dr.
\end{aligned}$$

By [17, Theorem 3.9], which follows from integration by parts, we have that, for $u \leqslant v$,

$$\begin{aligned}
\int_u^v W_{v,r}\,dr &= \frac{1}{2}(v-u)W_{u,v} + (v-u)H_{u,v}, \qquad (5.28)\\
\int_u^v W_{u,r}^2\,dr &= \frac{1}{3}(v-u)W_{u,v}^2 + (v-u)W_{u,v}H_{u,v} + 2L_{u,v}. \qquad (5.29)
\end{aligned}$$

From the decomposition (5.26) and identity (5.28) on $[s,u]$ and $[u,t]$, it follows that

$$H_{s,t} = \frac{1}{4}\big(W_{s,u} - W_{u,t}\big) + \frac{1}{2}\big(H_{s,u} + H_{u,t}\big). \qquad (5.30)$$

We now define the following random variables:

$$Z_{s,u} := \frac{1}{8}\big(W_{s,u} - W_{u,t}\big) - \frac{3}{4}\big(H_{s,u} + H_{u,t}\big), \qquad (5.31)$$

$$N_{s,t} := H_{s,u} - H_{u,t}. \qquad (5.32)$$

Since $W_{a,b} \sim \mathcal{N}(0,(b-a))$ and $H_{a,b} \sim \mathcal{N}\big(0,\frac{1}{12}(b-a)\big)$ are independent, we see that $W_{s,t}, H_{s,t}, Z_{s,u}, N_{s,t}$ are jointly normal, uncorrelated and therefore also independent. From (5.31) and (5.32), it directly follows that $Z_{s,u} \sim \mathcal{N}\big(0,\frac{1}{16}h\big)$ and $N_{s,t} \sim \mathcal{N}\big(0,\frac{1}{12}h\big)$. In addition, by rearranging the above expressions for these random variables, we have

$$W_{s,u} = \frac{1}{2}W_{s,t} + \frac{3}{2}H_{s,t} + Z_{s,u}, \tag{5.33}$$

$$W_{u,t} = \frac{1}{2}W_{s,t} - \frac{3}{2}H_{s,t} - Z_{s,u}, \tag{5.34}$$

$$H_{s,u} = \frac{1}{4}H_{s,t} - \frac{1}{2}Z_{s,u} + \frac{1}{2}N_{s,t}, \tag{5.35}$$

$$H_{u,t} = \frac{1}{4}H_{s,t} - \frac{1}{2}Z_{s,u} - \frac{1}{2}N_{s,t}. \tag{5.36}$$

Putting all of this together, and using the independence of Brownian increments, gives

$$\begin{aligned}
&\mathbb{E}\bigg[\int_s^t W_{s,r}^2\,dr\,\bigg|\,W_{s,u}, H_{s,u}, W_{u,t}, H_{u,t}\bigg] \\
&= \mathbb{E}\bigg[\int_s^u W_{s,r}^2\,dr\,\bigg|\,W_{s,u}, H_{s,u}\bigg] + \frac{1}{2}hW_{s,u}^2 + 2W_{s,u}\int_u^t W_{u,r}\,dr \\
&\qquad + \mathbb{E}\bigg[\int_u^t W_{u,r}^2\,dr\,\bigg|\,W_{u,t}, H_{u,t}\bigg] \\
&= \frac{1}{6}hW_{s,u}^2 + \frac{1}{2}hW_{s,u}H_{s,u} + 2\mathbb{E}\big[L_{s,u}\,|\,W_{s,u}, H_{s,u}\big] + \frac{1}{2}hW_{s,u}^2 + \frac{1}{2}hW_{s,u}W_{u,t} \\
&\qquad + hW_{s,u}H_{u,t} + \frac{1}{6}hW_{u,t}^2 + \frac{1}{2}hW_{u,t}H_{u,t} + 2\mathbb{E}\big[L_{u,t}\,|\,W_{u,t}, H_{u,t}\big] \\
&= \frac{1}{6}hW_{s,u}^2 + \frac{1}{2}hW_{s,u}H_{s,u} + \frac{3}{5}hH_{s,u}^2 + \frac{1}{60}h^2 + \frac{1}{2}hW_{s,u}^2 + \frac{1}{2}hW_{s,u}W_{u,t} \\
&\qquad + hW_{s,u}H_{u,t} + \frac{1}{6}hW_{u,t}^2 + \frac{1}{2}hW_{u,t}H_{u,t} + \frac{3}{5}hH_{u,t}^2 + \frac{1}{60}h^2 \\
&= \frac{1}{3}hW_{s,t}^2 + hW_{s,t}H_{s,t} + \frac{6}{5}hH_{s,t}^2 + \frac{1}{30}h^2 \\
&\qquad + \frac{1}{5}hH_{s,t}Z_{s,u} - \frac{1}{4}hW_{s,t}N_{s,t} + \frac{2}{15}hZ_{s,u}^2 + \frac{3}{10}hN_{s,t}^2,
\end{aligned}$$

where the last line was obtained by substituting (5.33)–(5.36) into the previous line. Since $n_{s,t} := \operatorname{sgn}(N_{s,t})$ and $N_{s,t} \sim \mathcal{N}\big(0,\frac{1}{12}h\big)$, it follows that $|N_{s,t}|$ has a half-normal distribution and is independent of $n_{s,t}$. Moreover, this implies that its moments are

$$\mathbb{E}\big[N_{s,t}\,|\,n_{s,t}\big] = \frac{1}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}, \qquad \mathbb{E}\big[N_{s,t}^3\,|\,n_{s,t}\big] = \frac{1}{6\sqrt{6\pi}}n_{s,t}h^{\frac{3}{2}}, \tag{5.37}$$

$$\mathbb{E}\big[N_{s,t}^2\,|\,n_{s,t}\big] = \frac{1}{12}h, \qquad \mathbb{E}\big[N_{s,t}^4\,|\,n_{s,t}\big] = \frac{1}{48}h^2. \tag{5.38}$$

Explicit formulae for the first four central moments of the half-normal distribution are given in [11, Equation (16)]. Since $W_{s,t}, H_{s,t}, Z_{s,u}, N_{s,t}$ are independent, we have

$$\begin{aligned}
&\mathbb{E}\bigg[\int_s^t W_{s,r}^2\, dr \,\bigg|\, W_{s,t}, H_{s,t}, n_{s,t}\bigg] \\
&\qquad = \mathbb{E}\bigg[\mathbb{E}\bigg[\int_s^t W_{s,r}^2\, dr \,\bigg|\, W_{s,t}, H_{s,t}, Z_{s,u}, N_{s,t}\bigg] \,\bigg|\, W_{s,t}, H_{s,t}, n_{s,t}\bigg].
\end{aligned}$$

As $\big(W_{s,t}, H_{s,t}, Z_{s,u}, N_{s,t}\big)$ and $\big(W_{s,u}, H_{s,u}, W_{u,t}, H_{u,t}\big)$ encode the same information, we have

$$\begin{aligned}
&\mathbb{E}\bigg[\int_s^t W_{s,r}^2\, dr \,\bigg|\, W_{s,t}, H_{s,t}, n_{s,t}\bigg] \\
&= \frac{1}{3}hW_{s,t}^2 + hW_{s,t}H_{s,t} + \frac{6}{5}hH_{s,t}^2 + \frac{1}{30}h^2 \\
&\qquad + \frac{1}{5}hH_{s,t}\mathbb{E}\big[Z_{s,u}\big] - \frac{1}{4}hW_{s,t}\mathbb{E}\big[N_{s,t}|n_{s,t}\big] + \frac{2}{15}h\mathbb{E}\big[Z_{s,u}^2\big] + \frac{3}{10}h\mathbb{E}\big[N_{s,t}^2|n_{s,t}\big] \\
&= \frac{1}{3}hW_{s,t}^2 + hW_{s,t}H_{s,t} + \frac{1}{15}h^2 + \frac{6}{5}hH_{s,t}^2 - \frac{1}{4\sqrt{6\pi}}n_{s,t}h^{\frac{3}{2}}W_{s,t},
\end{aligned}$$

where we used the moments $\mathbb{E}\big[Z_{s,u}\big] = 0$, $\mathbb{E}\big[Z_{s,u}^2\big] = \frac{1}{16}h$ as well as (5.37) and (5.38). The condition expectation (5.24) now follows by applying equation (5.29) to the above.

We employ a similar strategy to compute the conditional variance (5.25) of $L_{s,t}$. Using the decomposition (5.27) and independence of $\big(W_{s,u}, H_{s,u}, W_{u,t}, H_{u,t}\big)$, we have

$$\begin{aligned}
&\mathrm{Var}\bigg(\int_s^t W_{s,r}^2 dr \,\bigg|\, W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg) \\
&\quad = \mathrm{Var}\bigg(\int_s^u W_{s,r}^2\, dr + \frac{1}{2}hW_{s,u}^2 \\
&\qquad\qquad + 2W_{s,u}\int_u^t W_{u,r}\, dr + \int_u^t W_{u,r}^2\, dr \,\bigg|\, W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg) \\
&\quad = \mathrm{Var}\bigg(\int_s^u W_{s,r}^2 dr \,\bigg|\, W_{s,u}, H_{s,u}\bigg) + \mathrm{Var}\bigg(\int_u^t W_{u,r}^2 dr \,\bigg|\, W_{u,t}, H_{u,t}\bigg).
\end{aligned}$$

Therefore, by (5.29) and the formulae for the condition variances of $L_{s,u}$ and $L_{u,t}$,

$$\begin{aligned}
&\mathrm{Var}\bigg(\int_s^t W_{s,r}^2 dr \,\bigg|\, W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg) \\
&\quad = \frac{11}{50400}h^4 + h^3\Big(\frac{1}{1440}W_{s,u}^2 + \frac{1}{1440}W_{u,t}^2 + \frac{1}{1400}H_{s,u}^2 + \frac{1}{1400}H_{u,t}^2\Big).
\end{aligned}$$

By plugging in (5.33) – (5.36), we can rewrite this in terms of $W_{s,t}, H_{s,t}, Z_{s,u}, N_{s,t}$.

$$\begin{aligned}
&\mathrm{Var}\bigg(\int_s^t W_{s,r}^2\,dr\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg)\\
&= \frac{11}{50400}h^4 + h^3\Big(\frac{1}{1440}W_{s,u}^2 + \frac{1}{1440}W_{u,t}^2 + \frac{1}{1400}H_{s,u}^2 + \frac{1}{1400}H_{u,t}^2\Big)\\
&= \frac{11}{50400}h^4 + h^3\bigg(\frac{1}{2880}W_{s,t}^2 + \frac{9}{2800}H_{s,t}^2 + \frac{2}{525}H_{s,t}Z_{s,u} + \frac{11}{6300}Z_{s,u}^2 + \frac{1}{2800}N_{s,t}^2\bigg).
\end{aligned}$$

The second conditional moment of the iterated integral can be directly calculated as

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\int_s^t W_{s,r}^2\,dr\bigg)^2\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg]\\
&= \mathbb{E}\bigg[\int_s^t W_{s,r}^2\,dr\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg]^2 + \mathrm{Var}\bigg(\int_s^t W_{s,r}^2 dr\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg).
\end{aligned}$$

Therefore, by substituting the expressions for the above conditional moments, we have

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\int_s^t W_{s,r}^2\,dr\bigg)^2\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg]\\
&\quad= \Big(\frac{1}{3}hW_{s,t}^2 + hW_{s,t}H_{s,t} + \frac{6}{5}hH_{s,t}^2 + \frac{1}{30}h^2\\
&\qquad + \frac{1}{5}hH_{s,t}Z_{s,u} - \frac{1}{4}hW_{s,t}N_{s,t} + \frac{2}{15}hZ_{s,u}^2 + \frac{3}{10}hN_{s,t}^2\Big)^2 + \frac{11}{50400}h^4\\
&\qquad + h^3\bigg(\frac{1}{2880}W_{s,t}^2 + \frac{9}{2800}H_{s,t}^2 + \frac{2}{525}H_{s,t}Z_{s,u} + \frac{11}{6300}Z_{s,u}^2 + \frac{1}{2800}N_{s,t}^2\bigg).
\end{aligned}$$

Expanding the bracket and collecting terms yields

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\int_s^t W_{s,r}^2\,dr\bigg)^2\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg]\\
&= \frac{67}{50400}h^4 + \frac{1}{9}h^2W_{s,t}^4 + \frac{36}{25}h^2H_{s,t}^4 + \frac{4}{225}h^2Z_{s,u}^4 + \frac{9}{100}h^2N_{s,t}^4 + \frac{9}{5}h^2W_{s,t}^2H_{s,t}^2\\
&+ \frac{4}{45}h^2W_{s,t}^2Z_{s,u}^2 + \frac{21}{80}h^2W_{s,t}^2N_{s,t}^2 + \frac{9}{25}h^2H_{s,t}^2Z_{s,u}^2 + \frac{18}{25}h^2H_{s,t}^2N_{s,t}^2 + \frac{2}{25}h^2Z_{s,u}^2N_{s,t}^2\\
&+ \frac{13}{576}h^3W_{s,t}^2 + \frac{233}{2800}h^3H_{s,t}^2 + \frac{67}{6300}h^3Z_{s,u}^2 + \frac{57}{2800}h^3N_{s,t}^2 + \frac{1}{15}h^3W_{s,t}H_{s,t}\\
&- \frac{1}{60}h^3W_{s,t}N_{s,t} + \frac{3}{175}h^3H_{s,t}Z_{s,u} + \frac{2}{3}h^2W_{s,t}^3H_{s,t} - \frac{1}{10}h^2W_{s,t}H_{s,t}Z_{s,u}N_{s,t}\\
&- \frac{1}{6}h^2W_{s,t}^3N_{s,t} + \frac{12}{5}h^2W_{s,t}H_{s,t}^3 - \frac{3}{20}h^2W_{s,t}N_{s,t}^3 + \frac{12}{25}h^2H_{s,t}^3Z_{s,u} + \frac{4}{75}h^2H_{s,t}Z_{s,u}^3\\
&+ \frac{2}{15}h^2W_{s,t}^2H_{s,t}Z_{s,u} - \frac{1}{2}h^2W_{s,t}^2H_{s,t}N_{s,t} + \frac{2}{5}h^2W_{s,t}H_{s,t}^2Z_{s,u} + \frac{4}{15}h^2W_{s,t}H_{s,t}Z_{s,u}^2\\
&+ \frac{3}{5}h^2W_{s,t}H_{s,t}N_{s,t}^2 - \frac{3}{5}h^2W_{s,t}H_{s,t}^2N_{s,t} - \frac{1}{15}h^2W_{s,t}Z_{s,u}^2N_{s,t} + \frac{3}{25}h^2H_{s,t}Z_{s,u}N_{s,t}^2.
\end{aligned}$$

By taking the expectation of the above terms conditional on $\big(W_{s,t}, H_{s,t}, n_{s,t}\big)$ and substituting in the moments of $N_{s,t}\,|\,n_{s,t}$ given by (5.37) and (5.38), it follows that

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\int_s^t W_{s,r}^2\,dr\bigg)^2\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\\
&\quad= \mathbb{E}\bigg[\mathbb{E}\bigg[\bigg(\int_s^t W_{s,r}^2\,dr\bigg)^2\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg]\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\\
&\quad= \frac{13}{2100}h^4 + \frac{1}{9}h^2W_{s,t}^4 + \frac{36}{25}h^2H_{s,t}^4 + \frac{9}{5}h^2W_{s,t}^2H_{s,t}^2 + \frac{1}{20}h^3W_{s,t}^2 + \frac{29}{175}h^3H_{s,t}^2\\
&\qquad + \frac{2}{15}h^3W_{s,t}H_{s,t} + \frac{12}{5}h^2W_{s,t}H_{s,t}^3 + \frac{2}{3}h^2W_{s,t}^3H_{s,t}\\
&\qquad - \frac{1}{\sqrt{6\pi}}n_{s,t}h^{\frac{5}{2}}\bigg(\frac{1}{6}W_{s,t}^3 + \frac{11}{240}hW_{s,t} + \frac{1}{2}W_{s,t}^2H_{s,t} + \frac{3}{5}W_{s,t}H_{s,t}^2\bigg).
\end{aligned}$$

Thus, we can compute the required conditional variance using the following identity:

$$\begin{aligned}
&\mathrm{Var}\bigg(\int_s^t W_{s,r}^2\,dr\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg)\\
&\quad= \mathbb{E}\bigg[\bigg(\int_s^t W_{s,r}^2\,dr\bigg)^2\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg] - \bigg(\mathbb{E}\bigg[\int_s^t W_{s,r}^2\,dr\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\bigg)^2.
\end{aligned}$$

Plugging in the expressions for these conditional moments and simplifying terms gives

$$\begin{aligned}
&\mathrm{Var}\bigg(\int_s^t W_{s,r}^2\,dr\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg)\\
&\quad= \frac{13}{2100}h^4 + \frac{1}{9}h^2W_{s,t}^4 + \frac{36}{25}h^2H_{s,t}^4 + \frac{9}{5}h^2W_{s,t}^2H_{s,t}^2 + \frac{1}{20}h^3W_{s,t}^2 + \frac{29}{175}h^3H_{s,t}^2\\
&\qquad + \frac{2}{15}h^3W_{s,t}H_{s,t} + \frac{12}{5}h^2W_{s,t}H_{s,t}^3 + \frac{2}{3}h^2W_{s,t}^3H_{s,t}\\
&\qquad - \frac{1}{\sqrt{6\pi}}n_{s,t}h^{\frac{5}{2}}\bigg(\frac{1}{6}W_{s,t}^3 + \frac{11}{240}hW_{s,t} + \frac{1}{2}W_{s,t}^2H_{s,t} + \frac{3}{5}W_{s,t}H_{s,t}^2\bigg)\\
&\qquad - \bigg(\frac{1}{3}hW_{s,t} + hW_{s,t}H_{s,t} + \frac{1}{15}h^2 + \frac{6}{5}hH_{s,t}^2 - \frac{1}{4\sqrt{6\pi}}n_{s,t}h^{\frac{3}{2}}W_{s,t}\bigg)^2\\
&\quad= \frac{11}{6300}h^4 + \Big(\frac{1}{180} - \frac{1}{96\pi}\Big)h^3W_{s,t}^2 + \frac{1}{175}h^3H_{s,t}^2 - \frac{1}{80\sqrt{6\pi}}n_{s,t}h^{\frac{7}{2}}W_{s,t}.
\end{aligned}$$

The result now follows as, by (5.29), the above is the conditional variance of $2L_{s,t}$. □

In the construction of the piecewise linear paths defined by (5.5) and (5.9), there are two distinct solutions which result in paths with the required iterated integrals. To decide on the solution, we consider the 'space-time-time' Lévy area of the path. Whilst this quantity is Gaussian for Brownian motion and can be exactly generated, it is asymptotically smaller than space-space-time Lévy area, and so less impactful. Therefore, we propose using the expectation of space-time-time Lévy area conditional on $\big(W_{s,t}, H_{s,t}, n_{s,t}\big)$ and choosing the path γ which best matches this approximation.

Since $W_{s,t}, H_{s,t}$ and $K_{s,t}$ can be identified with coefficients from a polynomial expansion of Brownian motion, it is straightforward to establish their independence. However, $K_{s,t}$ is not independent of $n_{s,t}$ and we can compute the following moments:

Theorem 5.3.2. *The space-time-time Lévy area $K_{s,t}$ is independent of $(W_{s,t}, H_{s,t})$ and has the following distribution and conditional moments,*

$$K_{s,t} \sim \mathcal{N}\Big(0, \frac{1}{720}h\Big), \tag{5.39}$$

$$\mathbb{E}\big[K_{s,t}\,|\,n_{s,t}\big] = \frac{1}{8\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}, \tag{5.40}$$

$$\mathbb{E}\big[K_{s,t}^2\,|\,n_{s,t}\big] = \frac{1}{720}h. \tag{5.41}$$

Proof. It was shown in [17, Theorem 2.2], that for a Brownian bridge B on $[0,1]$ and certain orthogonal polynomials e_1 and e_2, we have

$$I_1 := \int_0^1 B_t \cdot \frac{e_1(t)}{t(1-t)}dt \quad\text{and}\quad I_2 := \int_0^1 B_t \cdot \frac{e_2(t)}{t(1-t)}dt$$

are independent random variables with $I_1 \sim \mathcal{N}\big(0, \frac{1}{2}\big)$ and $I_1 \sim \mathcal{N}\big(0, \frac{1}{6}\big)$. Moreover, by Theorems 2.7 and 2.8 in [17], the orthogonal polynomials e_1 and e_2 are given by

$$\begin{aligned} e_1(t) &= \sqrt{6}t(t-1), \\ e_2(t) &= \sqrt{30}t(t-1)(2t-1). \end{aligned}$$

Thus $I_1 = \sqrt{6}\int_0^1 B_t dt$ and $I_2 = 2\sqrt{30}\int_0^1 B_t(t-\frac{1}{2})dt$. It therefore follows that

$$\int_0^1 B_t dt \sim \mathcal{N}\Big(0, \frac{1}{12}\Big) \quad\text{and}\quad \int_0^1 B_t\Big(\frac{1}{2}-t\Big)dt \sim \mathcal{N}\Big(0, \frac{1}{720}\Big)$$

are independent. By the standard Brownian scaling, this implies $H_{s,t} \sim \mathcal{N}\big(0, \frac{1}{12}h\big)$ and $K_{s,t} \sim \mathcal{N}\big(0, \frac{1}{720}h\big)$ are independent. Moreover, since $H_{s,t}$ and $K_{s,t}$ are functions of the Brownian bridge $\big\{W_{s,u} - \frac{u-s}{h}W_{s,t}\big\}_{u\in[s,t]}$, they are also independent of $W_{s,t}$. We will now compute the expectation of $K_{s,t}$ conditional on $(W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t})$.

$$\begin{aligned}
&h^2\mathbb{E}\Big[K_{s,t}\,\big|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\Big]\\
&\quad= \mathbb{E}\bigg[\int_s^t\bigg(W_{s,r} - \frac{r-s}{h}W_{s,t}\bigg)\bigg(\frac{1}{2}h - (r-s)\bigg)dr\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg]\\
&\quad= \frac{1}{2}h\int_s^t W_{s,r}\,dr - \mathbb{E}\bigg[\int_s^t W_{s,r}(r-s)\,dr\,\bigg|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\bigg] + \frac{1}{12}h^2 W_{s,t}\\
&\quad= \frac{1}{3}h^2W_{s,t} + \frac{1}{2}h^2H_{s,t} - \mathbb{E}\bigg[\int_s^u W_{s,r}(r-s)\,dr\,\bigg|\,W_{s,u}, H_{s,u}\bigg] - W_{s,u}\int_u^t (r-s)\,dr\\
&\qquad - \mathbb{E}\bigg[\int_u^t W_{u,r}(r-s)\,dr\,\bigg|\,W_{u,t}, H_{u,t}\bigg]\\
&\quad= \frac{1}{3}h^2W_{s,t} + \frac{1}{2}h^2H_{s,t} - \int_s^u \mathbb{E}\big[W_{s,r}\,\big|\,W_{s,u}, H_{s,u}\big](r-s)\,dr - \frac{3}{8}h^2W_{s,u}\\
&\qquad - \int_u^t \mathbb{E}\big[W_{u,r}\,\big|\,W_{u,t}, H_{u,t}\big](r-u)\,dr - \int_u^t W_{u,r}(u-s)\,dr.
\end{aligned}$$

In [17], it was shown that $\mathbb{E}\big[W_{s,r}\,\big|\,W_{s,t}, H_{s,t}\big] = \frac{r-s}{t-s}W_{s,t} + \frac{6(r-s)(t-r)}{(t-s)^2}H_{s,t}$ for $r \in [s,t]$. Therefore, plugging this into the previous equation gives

$$\begin{aligned}
&h^2\mathbb{E}\Big[K_{s,t}\,\big|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\Big]\\
&\quad= \frac{1}{3}h^2W_{s,t} + \frac{1}{2}h^2H_{s,t} - \frac{1}{12}h^2W_{s,u} - \frac{1}{8}h^2H_{s,u} - \frac{3}{8}h^2W_{s,u}\\
&\qquad - \frac{1}{12}h^2W_{u,t} - \frac{1}{8}h^2H_{u,t} - \frac{1}{2}h\bigg(\frac{1}{4}hW_{u,t} + \frac{1}{2}hH_{u,t}\bigg)\\
&\quad= \frac{1}{2}h^2H_{s,t} - \bigg(\frac{1}{4}h^2H_{s,u} + \frac{1}{4}h^2H_{u,t} + \frac{1}{8}h^2W_{u,t} - \frac{1}{8}h^2W_{s,u}\bigg) + \frac{1}{8}h^2H_{s,u} - \frac{1}{8}h^2H_{u,t}.
\end{aligned}$$

By equation (5.30) in the previous proof, we see that the first two terms cancel. Thus

$$\mathbb{E}\big[K_{s,t}\,\big|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\big] = \frac{1}{8}N_{s,t},$$

and so the desired result (5.40) now follows as

$$\begin{aligned}
\mathbb{E}\big[K_{s,t}\,|\,W_{s,t}, H_{s,t}, n_{s,t}\big] &= \mathbb{E}\big[\mathbb{E}\big[K_{s,t}\,\big|\,W_{s,u}, W_{u,t}, H_{s,u}, H_{u,t}\big]\,|\,W_{s,t}, H_{s,t}, n_{s,t}\big]\\
&= \frac{1}{8}\mathbb{E}\big[N_{s,t}\,|\,W_{s,t}, H_{s,t}, n_{s,t}\big]\\
&= \frac{1}{8\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}},
\end{aligned}$$

by the independence of $\big(W_{s,t}, H_{s,t}, N_{s,t}\big)$ and equation (5.37), which were established in the proof of Theorem 5.3.1. Finally, we note that $K_{s,t}^2$ does not change if W is replaced by $-W$, whereas $n_{s,t}$ changes sign when the Brownian motion is 'flipped'. So by the symmetry of W, the random variables $K_{s,t}^2$ and $n_{s,t}$ are uncorrelated. Thus

$$\underbrace{\mathbb{E}\big[K_{s,t}^2 n_{s,t}\big]}_{=\,0} = \frac{1}{2}\mathbb{E}\big[K_{s,t}^2 \,|\, n_{s,t} = 1\big] + \frac{1}{2}\mathbb{E}\big[-K_{s,t}^2 \,|\, n_{s,t} = -1\big],$$

$$\underbrace{\mathbb{E}\big[K_{s,t}^2\big]}_{=\,\frac{1}{720}h} = \frac{1}{2}\mathbb{E}\big[K_{s,t}^2 \,|\, n_{s,t} = 1\big] + \frac{1}{2}\mathbb{E}\big[K_{s,t}^2 \,|\, n_{s,t} = -1\big],$$

gives the desired conditional moment (5.41). □

Finally, using these optimal estimators for $L_{s,t}$ and $K_{s,t}$, we give the theoretical justification for the choices of piecewise linear paths previously used in (5.5) and (5.9). These paths match $\mathbb{E}\big[L_{s,t}|W_{s,t}, H_{s,t}, n_{s,t}\big]$ and correlate with $\mathbb{E}\big[K_{s,t}|W_{s,t}, H_{s,t}, n_{s,t}\big]$.

Theorem 5.3.3. *Consider the $(W_{s,t}, H_{s,t}, n_{s,t})$-measurable piecewise linear paths $\gamma = (\gamma^\tau, \gamma^\omega) : [0,1] \to \mathbb{R}^2$, $\widetilde{\gamma} = (\widetilde{\gamma}^\tau, \widetilde{\gamma}^\omega) : [0,1] \to \mathbb{R}^2$ given by $\gamma_0 = \widetilde{\gamma}_0 = (s, W_s)$ and*

$$\gamma_{r_i, r_{i+1}} := \begin{cases} \big(0, A_{s,t}\big), & \textit{if } i = 0 \\ \big(h, B_{s,t}\big), & \textit{if } i = 1 \\ \big(0, W_{s,t} - A_{s,t} - B_{s,t}\big), & \textit{if } i = 2, \end{cases} \tag{5.42}$$

$$\widetilde{\gamma}_{r_i, r_{i+1}} := \begin{cases} \big(0, C_{s,t}\big), & \textit{if } i = 0 \\ \big(\frac{1}{2}h, 0\big), & \textit{if } i = 1 \\ \big(0, D_{s,t}\big), & \textit{if } i = 2 \\ \big(\frac{1}{2}h, 0\big), & \textit{if } i = 3 \\ \big(0, W_{s,t} - C_{s,t} - D_{s,t}\big), & \textit{if } i = 4, \end{cases} \tag{5.43}$$

where $h = t - s$ and

$$\begin{aligned}
&\big(A_{s,t}, B_{s,t}\big)\\
&= \underset{\substack{(A,B)\in\mathbb{R}^2 \text{ s.t. constraints}\\ (5.44),(5.45),(5.46) \text{ hold}}}{\arg\min}\left|\int_0^1 \gamma_{0,r}^{\tau}\gamma_{0,r}^{\omega}\, d\gamma_r^{\tau} - \mathbb{E}\bigg[\int_s^t (u-s)W_{s,u}du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\right|,\\
&\big(C_{s,t}, D_{s,t}\big)\\
&= \underset{\substack{(C,D)\in\mathbb{R}^2 \text{ s.t. constraints}\\ (5.44),(5.45),(5.46) \text{ hold}}}{\arg\min}\left|\int_0^1 \widetilde{\gamma}_{0,r}^{\tau}\widetilde{\gamma}_{0,r}^{\omega}\, d\widetilde{\gamma}_r^{\tau} - \mathbb{E}\bigg[\int_s^t (u-s)W_{s,u}du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\right|.
\end{aligned}$$

with the constraints (5.44), (5.45) and (5.46) for the paths γ and $\widetilde{\gamma}$ given by

$$\gamma_1^{\omega} - \gamma_0^{\omega} = \widetilde{\gamma}_1^{\omega} - \widetilde{\gamma}_0^{\omega} = W_{s,t}, \tag{5.44}$$

$$\int_0^1 \big(\gamma_r^{\omega} - \gamma_0^{\omega}\big)\, d\gamma_r^{\tau} = \int_0^1 \big(\widetilde{\gamma}_r^{\omega} - \widetilde{\gamma}_0^{\omega}\big)\, d\widetilde{\gamma}_r^{\tau} = \int_s^t W_{s,u}du, \tag{5.45}$$

$$\int_0^1 \big(\gamma_r^{\omega} - \gamma_0^{\omega}\big)^2 d\gamma_r^{\tau} = \int_0^1 \big(\widetilde{\gamma}_r^{\omega} - \widetilde{\gamma}_0^{\omega}\big)^2 d\widetilde{\gamma}_r^{\tau} = \mathbb{E}\bigg[\int_s^t W_{s,u}^2 du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]. \tag{5.46}$$

Then the first increments, $A_{s,t}$ and $C_{s,t}$, of the piecewise linear paths γ and $\widetilde{\gamma}$ are

$$\begin{aligned}
A_{s,t} &:= \frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}B_{s,t},\\
C_{s,t} &:= \frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}D_{s,t},
\end{aligned}$$

where the second increments, $B_{s,t}$ and $D_{s,t}$, of the paths are given by the formulae

$$\begin{aligned}
B_{s,t} &:= \epsilon_{s,t}\bigg(W_{s,t}^2 + \frac{12}{5}H_{s,t}^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}h^{\frac{1}{2}}n_{s,t}W_{s,t}\bigg)^{\frac{1}{2}},\\
D_{s,t} &:= \epsilon_{s,t}\bigg(\frac{1}{3}W_{s,t}^2 + \frac{4}{5}H_{s,t}^2 + \frac{4}{15}h - \frac{1}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t}\bigg)^{\frac{1}{2}},\\
\epsilon_{s,t} &:= \operatorname{sgn}\bigg(W_{s,t} - \frac{3}{\sqrt{24\pi}}h^{\frac{1}{2}}n_{s,t}\bigg).
\end{aligned}$$

Proof. Since γ and $\widetilde{\gamma}$ are piecewise linear, it is simple to compute the integrals

$$\begin{aligned}
\int_0^1 (\gamma_r^\omega - \gamma_0^\omega)\,d\gamma_r^\tau &= h\Big(A_{s,t} + \frac{1}{2}B_{s,t}\Big),\\
\int_0^1 (\gamma_r^\omega - \gamma_0^\omega)^2\,d\gamma_r^\tau &= h\Big(A_{s,t}^2 + A_{s,t}B_{s,t} + \frac{1}{3}B_{s,t}^2\Big),\\
\int_0^1 (\widetilde{\gamma}_r^\omega - \widetilde{\gamma}_0^\omega)\,d\widetilde{\gamma}_r^\tau &= h\Big(C_{s,t} + \frac{1}{2}D_{s,t}\Big),\\
\int_0^1 (\widetilde{\gamma}_r^\omega - \widetilde{\gamma}_0^\omega)^2\,d\widetilde{\gamma}_r^\tau &= h\Big(C_{s,t}^2 + C_{s,t}D_{s,t} + \frac{1}{2}D_{s,t}^2\Big).
\end{aligned}$$

It follows from the constraints (5.45) and (5.46) with equations (5.28) and (5.29) that

$$\begin{aligned}
h\Big(A_{s,t} + \frac{1}{2}B_{s,t}\Big) &= \frac{1}{2}hW_{s,t} + hH_{s,t},\\
h\Big(A_{s,t}^2 + A_{s,t}B_{s,t} + \frac{1}{3}B_{s,t}^2\Big) &= \frac{1}{3}hW_{s,t}^2 + hW_{s,t}H_{s,t} + 2\mathbb{E}\big[L_{s,t}\,\big|\,W_{s,t}, H_{s,t}, n_{s,t}\big],\\
h\Big(C_{s,t} + \frac{1}{2}D_{s,t}\Big) &= \frac{1}{2}hW_{s,t} + hH_{s,t},\\
h\Big(C_{s,t}^2 + C_{s,t}D_{s,t} + \frac{1}{2}D_{s,t}^2\Big) &= \frac{1}{3}hW_{s,t}^2 + hW_{s,t}H_{s,t} + 2\mathbb{E}\big[L_{s,t}\,\big|\,W_{s,t}, H_{s,t}, n_{s,t}\big],
\end{aligned}$$

So by Theorem 5.3.1, substituting in the formula for the conditional expectation yields

$$\begin{aligned}
A_{s,t} + \frac{1}{2}B_{s,t} &= \frac{1}{2}W_{s,t} + H_{s,t},\\
A_{s,t}^2 + A_{s,t}B_{s,t} + \frac{1}{3}B_{s,t}^2 &= \frac{1}{3}W_{s,t}^2 + W_{s,t}H_{s,t} + \frac{1}{15}h + \frac{6}{5}H_{s,t}^2 - \frac{1}{4\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t},\\
C_{s,t} + \frac{1}{2}D_{s,t} &= \frac{1}{2}W_{s,t} + H_{s,t},\\
C_{s,t}^2 + C_{s,t}D_{s,t} + \frac{1}{2}D_{s,t}^2 &= \frac{1}{3}W_{s,t}^2 + W_{s,t}H_{s,t} + \frac{1}{15}h + \frac{6}{5}H_{s,t}^2 - \frac{1}{4\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t},
\end{aligned}$$

Since $a^2 + ab + \frac{1}{3}b^2 = \big(a + \frac{1}{2}b\big)^2 + \frac{1}{12}b^2$ and $c^2 + cd + \frac{1}{3}d^2 = \big(c + \frac{1}{2}d\big)^2 + \frac{1}{4}d^2$, this gives

$$\begin{aligned}
\frac{1}{12}B_{s,t}^2 &= \frac{1}{3}W_{s,t}^2 + W_{s,t}H_{s,t} + \frac{1}{15}h + \frac{6}{5}H_{s,t}^2 - \frac{1}{4\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t} - \Big(\frac{1}{2}W_{s,t} + H_{s,t}\Big)^2,\\
\frac{1}{4}D_{s,t}^2 &= \frac{1}{3}W_{s,t}^2 + W_{s,t}H_{s,t} + \frac{1}{15}h + \frac{6}{5}H_{s,t}^2 - \frac{1}{4\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t} - \Big(\frac{1}{2}W_{s,t} + H_{s,t}\Big)^2,
\end{aligned}$$

and so there are two possible values of $B_{s,t}$ and $D_{s,t}$ where (5.45) and (5.46) hold,

$$B_{s,t} = \pm\sqrt{W_{s,t}^2 + \frac{12}{5}H_{s,t}^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t}}\,,$$

$$D_{s,t} = \pm\sqrt{\frac{1}{3}W_{s,t}^2 + \frac{4}{5}H_{s,t}^2 + \frac{4}{15}h - \frac{1}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t}}\,.$$

Thus, if equations (5.45) and (5.46) are satisfied, we have

$$\begin{aligned}\int_0^1 (\gamma_r^\tau - \gamma_0^\tau)(\gamma_r^\omega - \gamma_0^\omega)\,d\gamma_r^\tau &= \frac{1}{2}h^2 A_{s,t} + \frac{1}{3}h^2 B_{s,t}\\ &= \frac{1}{4}h^2 W_{s,t} + \frac{1}{2}h^2 H_{s,t} \pm \frac{1}{12}h^2\sqrt{W_{s,t}^2 + \frac{12}{5}H_{s,t}^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t}}\,,\\ \int_0^1 (\widetilde{\gamma}_r^\tau - \widetilde{\gamma}_0^\tau)(\widetilde{\gamma}_r^\omega - \widetilde{\gamma}_0^\omega)\,d\widetilde{\gamma}_r^\tau &= \frac{1}{2}h^2 C_{s,t} + \frac{3}{8}h^2 D_{s,t}\\ &= \frac{1}{4}h^2 W_{s,t} + \frac{1}{2}h^2 H_{s,t} \pm \frac{1}{8}h^2\sqrt{\frac{1}{3}W_{s,t}^2 + \frac{4}{5}H_{s,t}^2 + \frac{4}{15}h - \frac{1}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t}}\,.\end{aligned}$$

Using Theorem 5.3.2, we can estimate the corresponding integral of Brownian motion.

$$\begin{aligned}&\mathbb{E}\bigg[\int_s^t (u-s)W_{s,u}\,du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\\ &\quad= \mathbb{E}\bigg[\frac{1}{2}h\int_s^t W_{s,u}\,du - \int_s^t \frac{u-s}{h}\,W_{s,t}\Big(\frac{1}{2}h - (u-s)\Big)du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\\ &\qquad\quad - \mathbb{E}\bigg[\int_s^t \Big(W_{s,u} - \frac{u-s}{h}\,W_{s,t}\Big)\Big(\frac{1}{2}h - (u-s)\Big)du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\\ &\quad= \frac{1}{3}h^2 W_{s,t} + \frac{1}{2}h^2 H_{s,t} - h^2\,\mathbb{E}\big[K_{s,t}\,|\,n_{s,t}\big]\\ &\quad= \frac{1}{3}h^2 W_{s,t} + \frac{1}{2}h^2 H_{s,t} - \frac{1}{8\sqrt{6\pi}}n_{s,t}h^{\frac{5}{2}}.\end{aligned}$$

Taking the difference between these integrals gives

$$\begin{aligned}&\bigg|\int_0^1 (\gamma_r^\tau - \gamma_0^\tau)(\gamma_r^\omega - \gamma_0^\omega)\,d\gamma_r^\tau - \mathbb{E}\bigg[\int_s^t (u-s)W_{s,u}\,du\,\bigg|\,W_{s,t}, H_{s,t}, n_{s,t}\bigg]\bigg|\\ &= \bigg| -\frac{1}{12}h^2 W_{s,t} + \frac{1}{8\sqrt{6\pi}}n_{s,t}h^{\frac{5}{2}} \pm \frac{1}{12}h^2\sqrt{W_{s,t}^2 + \frac{12}{5}H_{s,t}^2 + \frac{4}{5}h - \frac{3}{\sqrt{6\pi}}n_{s,t}h^{\frac{1}{2}}W_{s,t}}\,\bigg|,\end{aligned}$$

and

$$\begin{aligned}
&\left|\int_0^1 (\widetilde{\gamma}_r^\tau - \widetilde{\gamma}_0^\tau)(\widetilde{\gamma}_r^\omega - \widetilde{\gamma}_0^\omega)\, d\widetilde{\gamma}_r^\tau - \mathbb{E}\bigg[\int_s^t (u-s) W_{s,u}\, du \,\bigg|\, W_{s,t}, H_{s,t}, n_{s,t}\bigg]\right| \\
&= \left| -\frac{1}{12}h^2 W_{s,t} + \frac{1}{8\sqrt{6\pi}} n_{s,t} h^{\frac{5}{2}} \pm \frac{1}{8}h^2 \sqrt{\frac{1}{3}W_{s,t}^2 + \frac{4}{5}H_{s,t}^2 + \frac{4}{15}h - \frac{1}{\sqrt{6\pi}} n_{s,t} h^{\frac{1}{2}} W_{s,t}}\,\right|.
\end{aligned}$$

Since we would like the path γ to minimise this quantity, the optimal choice of sign for the square root term is $\epsilon_{s,t} := \operatorname{sgn}\big(W_{s,t} - \frac{3}{\sqrt{24\pi}} h^{\frac{1}{2}} n_{s,t}\big)$, and the result follows. □

Chapter 6

Numerical examples

We present here numerical examples deploying the splitting paths presented in Section 5.1, including derivations of the resulting numerical schemes. We consider the stochastic FitzHugh-Nagumo (additive noise) and the stochastic Lotka-Volterra (multiplicative noise) models, and as an example of higher order schemes applied to Multi-level Monte Carlo we consider the problem of pricing a basket option for a model of interacting assets. As a small experiment, we also consider the long time behaviour of splittings for the stochastic anharmonic oscillator, for comparison with results in [39]. More examples can be found in our paper [19]. For the majority of these examples, the 'non-diffusion' ODEs coming from the SDE splitting will not admit a closed-formed solution and thus must be further discretized. We will show how such ODEs can be resolved.

Throughout, we shall compare methods using the following strong error estimator:

Definition 6.0.1 (Strong error estimator for SDEs)**.** *For $N \geqslant 1$, let Y_N denote a numerical solution to the SDE (2.1) computed at time T with a fixed step size $h = \frac{T}{N}$. Then we define the following estimator for quantifying the strong convergence of Y_N:*

$$S_N := \sqrt{\mathbb{E}\Big[\|Y_N - Y_T^{\textit{fine}}\|^2\Big]}, \tag{6.1}$$

where $Y_T^{\textit{fine}}$ denotes a numerical solution to (2.1) computed with a finer step size, $h^{\textit{fine}} \leqslant \frac{1}{10}h$, but using the same Brownian motion (so that Y_N and $Y_T^{\textit{fine}}$ are close). In our examples, the expectation in (6.1) will be estimated by standard Monte Carlo.

When defining numerical methods, we sometimes use W_k as shorthand for $W_{t_k,t_{k+1}}$, (and similarly H_k and n_k instead of $H_{t_k,t_{k+1}}$ and $n_{t_k,t_{k+1}}$).

6.1 FitzHugh-Nagumo model

We consider a stochastic FitzHugh-Nagumo (FHN) model which has been used for describing the spike activity of neurons [4, 40]. The stochastic FHN model follows the two-dimensional additive noise SDE (where Itô and Stratonovich coincide) is given by

$$d\begin{pmatrix} v_t \\ u_t \end{pmatrix} = \begin{pmatrix} \frac{1}{\epsilon}\big(v_t - v_t^3 - u_t\big) \\ \theta v_t - u_t + \beta \end{pmatrix} dt + \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} dW_t. \tag{6.2}$$

To discretize the stochastic FHN model, we apply the splitting path (5.5) and, similar to [4], apply a Strang splitting to approximate the resulting drift ODE. This leads to the splitting method:

$$\begin{aligned}
\begin{pmatrix} V_k^{(1)} \\ U_k^{(1)} \end{pmatrix} &:= \begin{pmatrix} V_k \\ U_k \end{pmatrix} + \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} \frac{1}{2}W_k^1 + H_k^1 - \frac{1}{2}C_k^1 \\ \frac{1}{2}W_k^2 + H_k^2 - \frac{1}{2}C_k^2 \end{pmatrix}, \\
\begin{pmatrix} V_k^{(2)} \\ U_k^{(2)} \end{pmatrix} &:= \varphi_{\frac{1}{2}h}^{\textsf{Strang}} \begin{pmatrix} V_k^{(1)} \\ U_k^{(1)} \end{pmatrix} + \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} C_k^1 \\ C_k^2 \end{pmatrix}, \\
\begin{pmatrix} V_{k+1} \\ U_{k+1} \end{pmatrix} &:= \varphi_{\frac{1}{2}h}^{\textsf{Strang}} \begin{pmatrix} V_k^{(2)} \\ U_k^{(2)} \end{pmatrix} + \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} \frac{1}{2}W_k^1 - H_k^1 - \frac{1}{2}C_k^1 \\ \frac{1}{2}W_k^2 - H_k^2 - \frac{1}{2}C_k^2 \end{pmatrix},
\end{aligned} \tag{6.3}$$

where the Strang splitting steps are given by

$$\varphi_{\frac{1}{2}h}^{\textsf{Strang}} \begin{pmatrix} v \\ u \end{pmatrix} := \begin{pmatrix} \widetilde{v}\Big(e^{-\frac{h}{2\epsilon}} + \widetilde{v}^2\big(1 - e^{-\frac{h}{2\epsilon}}\big)\Big)^{-\frac{1}{2}} \\ \widetilde{u} + \frac{1}{4}\beta h \end{pmatrix}, \tag{6.4}$$

with $\widetilde{v}$ and $\widetilde{u}$ defined by

$$\begin{pmatrix} \widetilde{v} \\ \widetilde{u} \end{pmatrix} := \exp\left(\frac{1}{2}h\begin{pmatrix} 0 & -\frac{1}{\epsilon} \\ \theta & -1 \end{pmatrix}\right) \begin{pmatrix} v\Big(e^{-\frac{h}{2\epsilon}} + v^2\big(1 - e^{-\frac{h}{2\epsilon}}\big)\Big)^{-\frac{1}{2}} \\ u + \frac{1}{4}\beta h \end{pmatrix}.$$

and the explicit formula for the above matrix exponential is given in [4, Section 6.2]. We note that, similar to the CIR model, the stochastic FHN model is challenging to accurately simulate due to the vector field not being globally Lipschitz continuous. That said, as the drift does have polynomial growth and satisfies a one-sided Lipschitz condition, there are numerical methods for (6.2) with strong convergence guarantees. We will compare our scheme (6.3) against two such methods; the Strang splitting scheme proposed in [4] and the Tamed Euler-Maruyama method introduced in [28].

6.1.1 Derivation of FitzHugh-Nagumo splitting

The derivation of the splitting scheme (6.3) is as follows. Proceeding as described in Example 2.2.1, the splitting path (5.5) applied to (6.2) gives the following series of ODEs:

$$d\begin{pmatrix} v_r^{(0)} \\ u_r^{(0)} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}\left(\frac{1}{2}W_{s,t} + H_{s,t} - \frac{1}{2}C_{s,t}\right) dr\ , \qquad \begin{pmatrix} v_0^{(0)} \\ u_0^{(0)} \end{pmatrix} = \begin{pmatrix} v_0 \\ u_0 \end{pmatrix}\ , \tag{6.5}$$

$$d\begin{pmatrix} v_r^{(1)} \\ u_r^{(1)} \end{pmatrix} = \frac{1}{2}h\begin{pmatrix} \frac{1}{\epsilon}(v_r^{(1)} - (v_r^{(1)})^3 - u_r^{(1)}) \\ \theta v_r^{(1)} - u_r^{(1)} + \beta \end{pmatrix} dr\ , \qquad \begin{pmatrix} v_0^{(1)} \\ u_0^1 \end{pmatrix} = \begin{pmatrix} v_1^{(0)} \\ u_1^{(0)} \end{pmatrix}\ , \tag{6.6}$$

$$d\begin{pmatrix} v_r^{(2)} \\ u_r^{(2)} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} C_{s,t}\ dr\ , \qquad \begin{pmatrix} v_0^{(2)} \\ u_0^{(2)} \end{pmatrix} = \begin{pmatrix} v_1^{(1)} \\ u_1^{(1)} \end{pmatrix}\ , \tag{6.7}$$

$$d\begin{pmatrix} v_r^{(3)} \\ u_r^{(3)} \end{pmatrix} = \frac{1}{2}h\begin{pmatrix} \frac{1}{\epsilon}(v_r^{(3)} - (v_r^{(3)})^3 - u_r^{(3)}) \\ \theta v_r^{(3)} - u_r^{(3)} + \beta \end{pmatrix} dr\ , \qquad \begin{pmatrix} v_0^{(3)} \\ u_0^{(3)} \end{pmatrix} = \begin{pmatrix} v_1^{(2)} \\ u_1^{(2)} \end{pmatrix}\ , \tag{6.8}$$

$$d\begin{pmatrix} v_r^{(4)} \\ u_r^{(4)} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}\left(\frac{1}{2}W_{s,t} - H_{s,t} - \frac{1}{2}C_{s,t}\right) dr\ , \qquad \begin{pmatrix} v_0^{(4)} \\ u_0^{(4)} \end{pmatrix} = \begin{pmatrix} v_1^{(3)} \\ u_1^{(3)} \end{pmatrix}\ , \tag{6.9}$$

where $C_{s,t}$ is defined as in Example 5.1.4. Solving to find $(v_1^{(4)}, u_1^{(4)})^\top$ produces our estimate of a solution to (6.2) at time h. The ODEs (6.5), (6.7) and (6.9) are exactly solvable. The terms (6.6) and (6.8) require further analysis, proving the source of discretization error in our spitting scheme. Following [4], we note that we may rewrite these terms in the form

$$d\begin{pmatrix} v_r \\ u_r \end{pmatrix} = \frac{1}{2}h\left(\begin{pmatrix} 0 & -\frac{1}{\epsilon} \\ \theta & -1 \end{pmatrix}\begin{pmatrix} v_r \\ u_r \end{pmatrix} + \begin{pmatrix} \frac{1}{\epsilon}(v_r - v_r^3) \\ \beta \end{pmatrix}\right) dr\ . \tag{6.10}$$

To maintain the potential order of the splitting scheme (global $O(h^{1.5})$ in the commutative case), we require that our discretisation of the resulting sequence of ODEs has at least local strong error smaller than $O(h^2)$. A Strang splitting of (6.10) provides this as the two separate components in (6.10) admit closed form solutions. An ODE Strang splitting of (6.10) gives:

$$\begin{aligned}
d\begin{pmatrix} \tilde{v}_r^{(0)} \\ \tilde{u}_r^{(0)} \end{pmatrix} &= \frac{1}{2}h\begin{pmatrix} \frac{1}{\epsilon}(\tilde{v}_r^{(0)} - (\tilde{v}_r^{(0)})^3) \\ \beta \end{pmatrix} dr\ , & \begin{pmatrix} \tilde{v}_0^{(0)} \\ \tilde{u}_0^{(0)} \end{pmatrix} &= \begin{pmatrix} v \\ u \end{pmatrix}\ , \\
d\begin{pmatrix} \tilde{v}_r^{(1)} \\ \tilde{u}_r^{(1)} \end{pmatrix} &= \frac{1}{2}h\begin{pmatrix} 0 & -\frac{1}{\epsilon} \\ \theta & -1 \end{pmatrix}\begin{pmatrix} \tilde{v}_r^{(1)} \\ \tilde{u}_r^{(1)} \end{pmatrix} dr\ , & \begin{pmatrix} \tilde{v}_0^{(1)} \\ \tilde{u}_0^{(1)} \end{pmatrix} &= \begin{pmatrix} \tilde{v}_{0.5}^{(0)} \\ \tilde{u}_{0.5}^{(0)} \end{pmatrix}\ , \\
d\begin{pmatrix} \tilde{v}_r^{(2)} \\ \tilde{u}_r^{(2)} \end{pmatrix} &= \frac{1}{2}h\begin{pmatrix} \frac{1}{\epsilon}(\tilde{v}_r^{(2)} - (\tilde{v}_r^{(2)})^3) \\ \beta \end{pmatrix} dr\ , & \begin{pmatrix} \tilde{v}_0^{(2)} \\ \tilde{u}_0^{(2)} \end{pmatrix} &= \begin{pmatrix} \tilde{v}_1^{(1)} \\ \tilde{u}_1^{(1)} \end{pmatrix}\ .
\end{aligned}$$

The value of $(\tilde{v}_{0.5}^{(2)}, \tilde{u}_{0.5}^{(2)})^\top$ is the Strang splitting estimate. As $v' = (v - v^3)$ admits a closed form solution, solving the ODEs for initial values $u, v \in \mathbb{R}$ leads to (6.4). Combining everything, we obtain (6.3).

6.1.2 Numerical results

For our numerical experiments, we select the following parameters for the FHN model

$$\epsilon = 1\ , \quad \theta = 1\ , \quad \beta = 1\ , \quad \sigma_1 = 1\ , \quad \sigma_2 = 1\ , \quad (v_0, u_0) = (0, 0)\ , \quad T = 2\ .$$

We compare our proposed splitting with the Strang splitting presented in [4] and the Tamed Euler method of [28]. We see in Figure 6.1 that our proposed splitting exhibits a $3/2$ strong convergence rate and is significantly more accurate than the other schemes (for a fixed h).

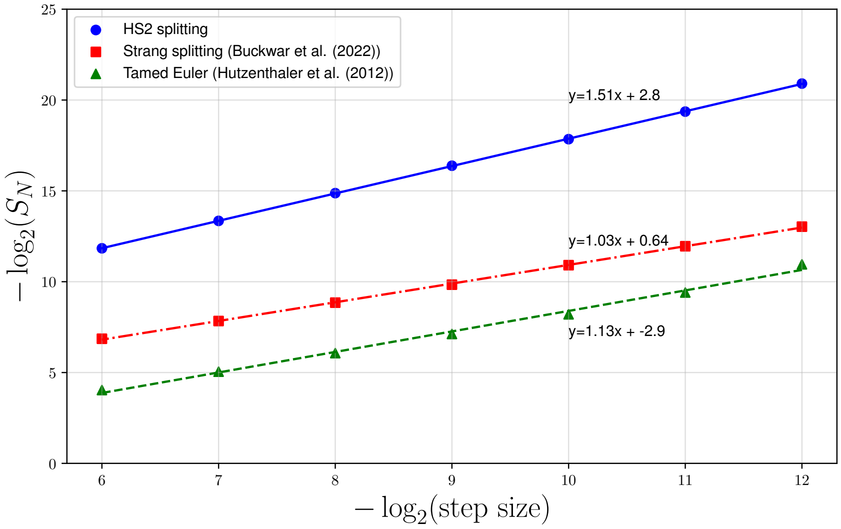


Fig. 6.1: S_N estimated for (6.2) using 1,000 sample paths as a function of step size $h = \frac{T}{N}$.

The higher order scheme will be more computationally expensive than the lower order schemes, this is quantified in Table 6.1: Where we see that, for the same value of h, to simulate 1000 sample paths the HS2 splitting takes approximately 3 times as long as the Strang splitting, and just under 5 times as long as the Tamed Euler scheme.

HS2 splitting (6.3)	Strang splitting [4]	Tamed Euler [28]
8.22	2.74	1.75

Table 6.1: Compute time to simulate 1000 sample paths of (6.2) with 100 steps (seconds)

While the high order splitting takes longer to run, as observed in Figure 6.1, it is significantly more accurate. We quantify this in Table 6.2, where we see that the HS2 splitting is an order of magnitude faster than the other two methods – to achieve the same level of precision. We thus conclude that the proposed high order splitting method (6.3) gives the best performance for the FHN model.

HS2 splitting (6.3)	Strang splitting [4]	Tamed Euler [28]
4.42	29.67	85.39

Table 6.2: Estimated compute time to produce 1000 sample paths of (6.2) with an error of $S_N = 10^{-3}$ (seconds)

6.2 Lotka-Volterra model

As an example of an SDE with multiplicative noise, we consider the stochastic Lotka-Volterra (LV) model, which is a classical model of predator-prey population dynamics [1, 66]. The stochastic LV model follows the two-dimensional additive noise SDE given by

$$d\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} x_t(b_1 - a_{11}x_t - a_{12}y_t) \\ y_t(b_2 - a_{21}x_t - a_{22}y_t) \end{pmatrix} dt + \begin{pmatrix} G_1 x_t & 0 \\ 0 & G_2 y_t \end{pmatrix} \circ dW_t \ , \tag{6.11}$$

with $x_0, y_0 = x, y > 0$. Where y denotes the predator population and x the prey. Depending on the choice of coefficients, this system can cover three standard classifications [66]:

Predator-prey:	$b_1 > 0 \ , \ b_2 < 0 \ , \ a_{12} > 0 \ , \ a_{21} < 0$
Cooperation:	$b_1 > 0 \ , \ b_2 > 0 \ , \ a_{12} < 0 \ , \ a_{21} < 0$
Competition:	$b_1 > 0 \ , \ b_2 > 0 \ , \ a_{12} > 0 \ , \ a_{21} > 0$

Assuming that $G_1, G_2 > 0$ then the solution is guaranteed positive [46, Theorem 2.1]. Under the ‘predator-prey’ classification we observe dynamics like those displayed in Figure 6.2: a growth in the population of the prey is followed by an increase in the predator population, followed by a subsequent drop in both before the cycle repeats. For positive values of a_{11} and a_{22} the dynamics will spiral towards a stationary point. The observed cycles are also quite sensitive to the initial condition.

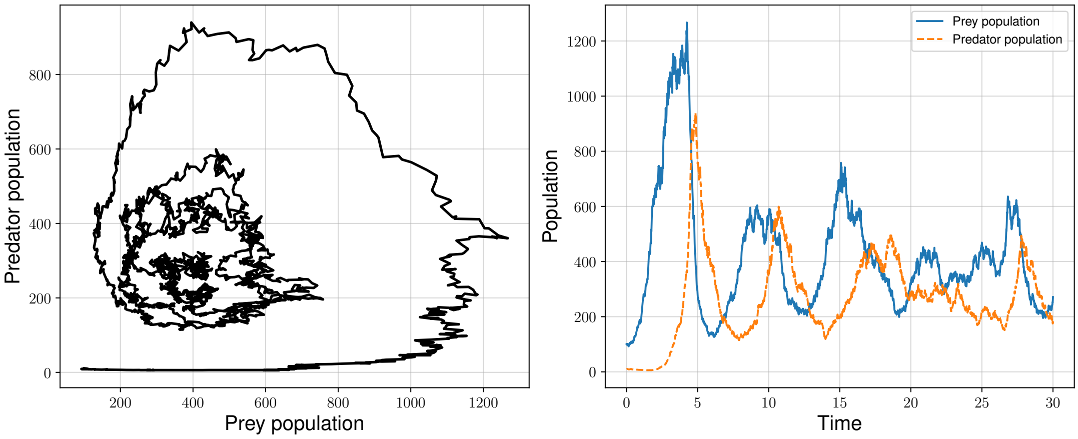


Fig. 6.2: Example dynamics for the LV model, with parameter values $a_{11} = a_{22} = 0.0005$, $a_{12} = -a_{21} = 0.003$, $b_1 = -b_2 = 1$ and $G_1 = G_2 = 0.2$. The left plot displays the dynamics in phase space, and the right plot shows how the populations evolve through time.

To discretize the stochastic LV model we choose the path γ^{HS1} (5.4) and an ODE Strang splitting to deal with several of the resulting ODEs. This leads to the following splitting method:

$$\begin{aligned}
\begin{pmatrix} X_k^{(1)} \\ Y_k^{(1)} \end{pmatrix} &= \phi^{\text{Strang}}_{\frac{3-\sqrt{3}}{6}h} \begin{pmatrix} X_k \\ Y_k \end{pmatrix} , \\
\begin{pmatrix} X_k^{(2)} \\ Y_k^{(2)} \end{pmatrix} &= \begin{pmatrix} X_k^{(1)} \exp\left(G_1(W_k^1/2 + \sqrt{3}H_k^1)\right) \\ Y_k^{(1)} \exp\left(G_2(W_k^2/2 + \sqrt{3}H_k^2)\right) \end{pmatrix} , \\
\begin{pmatrix} X_k^{(3)} \\ Y_k^{(3)} \end{pmatrix} &= \phi^{\text{Strang}}_{\frac{\sqrt{3}}{3}h} \begin{pmatrix} X_k^{(2)} \\ Y_k^{(2)} \end{pmatrix} , \\
\begin{pmatrix} X_k^{(4)} \\ Y_k^{(4)} \end{pmatrix} &= \begin{pmatrix} X_k^{(3)} \exp\left(G_1(W_k^1/2 - \sqrt{3}H_k^1)\right) \\ Y_k^{(3)} \exp\left(G_2(W_k^2/2 - \sqrt{3}H_k^2)\right) \end{pmatrix} , \\
\begin{pmatrix} X_{k+1} \\ Y_{k+1} \end{pmatrix} &= \phi^{\text{Strang}}_{\frac{3-\sqrt{3}}{6}h} \begin{pmatrix} X_k^{(4)} \\ Y_k^{(4)} \end{pmatrix} , \qquad (6.12)
\end{aligned}$$

where $\phi^{\text{Strang}}_{ch}$ is defined as in (6.13) below. The Strang splitting steps enforce an additional restriction that $a_{12} > 0$, restricting us to 'predator-prey' and 'competition' classifications. The Strang splitting also assumes that x_t and y_t are strictly positive, thus additional care should be taken when this may break down.

Remark 6.2.1. *For the sake of numerical comparison with schemes designed for Itô SDEs, we note that the SDE* (6.11) *may equivalently be written in Itô form as*

$$d\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} x_t\left(b_1 + G_1^2/2 - a_{11}x_t - a_{12}y_t\right) \\ y_t\left(b_2 + G_2^2/2 - a_{21}x_t - a_{22}y_t\right) \end{pmatrix} dt + \begin{pmatrix} G_1 x_t & 0 \\ 0 & G_2 y_t \end{pmatrix} dW_t\ .$$

6.2.1 Derivation of Lotka-Volterra splitting

Applying the splitting path (5.4) we obtain the following series of ODEs

$$\begin{aligned}
d\begin{pmatrix} x_r^{(0)} \\ y_r^{(0)} \end{pmatrix} &= \frac{3-\sqrt{3}}{6}h\begin{pmatrix} x_r^{(0)}(b_1 - a_{11}x_r^{(0)} - a_{12}y_r^{(0)}) \\ y_r^{(0)}(b_2 - a_{21}x_r^{(0)} - a_{22}y_r^{(0)}) \end{pmatrix} dr\ , & \begin{pmatrix} x_0^{(0)} \\ y_0^{(0)} \end{pmatrix} &= \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \\
d\begin{pmatrix} x_r^{(1)} \\ y_r^{(1)} \end{pmatrix} &= \begin{pmatrix} G_1(\frac{1}{2}W_{s,t}^1 + \sqrt{3}H_{s,t}^1)x_r^{(1)} \\ G_2(\frac{1}{2}W_{s,t}^2 + \sqrt{3}H_{s,t}^2)y_r^{(1)} \end{pmatrix} dr\ , & \begin{pmatrix} x_0^{(1)} \\ y_0^{(1)} \end{pmatrix} &= \begin{pmatrix} x_1^{(0)} \\ y_1^{(0)} \end{pmatrix} \\
d\begin{pmatrix} x_r^{(2)} \\ y_r^{(2)} \end{pmatrix} &= \frac{\sqrt{3}}{3}h\begin{pmatrix} x_r^{(2)}(b_1 - a_{11}x_r^{(2)} - a_{12}y_r^{(2)}) \\ y_r^{(2)}(b_2 - a_{21}x_r^{(2)} - a_{22}y_r^{(2)}) \end{pmatrix} dr\ , & \begin{pmatrix} x_0^{(2)} \\ y_0^{(2)} \end{pmatrix} &= \begin{pmatrix} x_1^{(1)} \\ y_1^{(1)} \end{pmatrix} \\
d\begin{pmatrix} x_r^{(3)} \\ y_r^{(3)} \end{pmatrix} &= \begin{pmatrix} G_1(\frac{1}{2}W_{s,t}^1 - \sqrt{3}H_{s,t}^1)x_r^{(3)} \\ G_2(\frac{1}{2}W_{s,t}^2 - \sqrt{3}H_{s,t}^2)y_r^{(3)} \end{pmatrix} dr\ , & \begin{pmatrix} x_0^{(3)} \\ y_0^{(3)} \end{pmatrix} &= \begin{pmatrix} x_1^{(2)} \\ y_1^{(2)} \end{pmatrix} \\
d\begin{pmatrix} x_r^{(4)} \\ y_r^{(4)} \end{pmatrix} &= \frac{3-\sqrt{3}}{6}h\begin{pmatrix} x_r^{(4)}(b_1 - a_{11}x_r^{(4)} - a_{12}y_r^{(4)}) \\ y_r^{(4)}(b_2 - a_{21}x_r^{(4)} - a_{22}y_r^{(4)}) \end{pmatrix} dr\ . & \begin{pmatrix} x_0^{(4)} \\ y_0^{(4)} \end{pmatrix} &= \begin{pmatrix} x_1^{(3)} \\ y_1^{(3)} \end{pmatrix}
\end{aligned}$$

The second and fourth ODEs are relatively simple to deal with resulting in exponentials. To deal with the first, third and fifth ODEs we again propose an ODE Strang splitting. We note that these ODEs may equivalently be written in the following form:

$$d\begin{pmatrix} x_r \\ y_r \end{pmatrix} = ch\left(\begin{pmatrix} x_r(b_1 - a_{11}x_r) \\ y_r(b_2 - a_{22}y_r) \end{pmatrix} + \begin{pmatrix} -a_{12}x_r y_r \\ -a_{21}x_r y_r \end{pmatrix}\right) dr\ .$$

A Strang splitting applied to this then produces the following sequence of ODEs

$$\begin{aligned}
d\begin{pmatrix} \tilde{x}_r^{(0)} \\ \tilde{y}_r^{(0)} \end{pmatrix} &= ch\begin{pmatrix} \tilde{x}_r^{(0)}(b_1 - a_{11}\tilde{x}_r^{(0)}) \\ \tilde{y}_r^{(0)}(b_2 - a_{22}\tilde{y}_r^{(0)}) \end{pmatrix} dr\ , & \begin{pmatrix} \tilde{x}_0^{(0)} \\ \tilde{y}_0^{(0)} \end{pmatrix} &= \begin{pmatrix} x \\ y \end{pmatrix}\ , \\
d\begin{pmatrix} \tilde{x}_r^{(1)} \\ \tilde{y}_r^{(1)} \end{pmatrix} &= ch\begin{pmatrix} -a_{12}\tilde{x}_r^{(1)}\tilde{y}_r^{(1)} \\ -a_{21}\tilde{x}_r^{(1)}\tilde{y}_r^{(1)} \end{pmatrix} dr\ , & \begin{pmatrix} \tilde{x}_0^{(1)} \\ \tilde{y}_0^{(1)} \end{pmatrix} &= \begin{pmatrix} \tilde{x}_{0.5}^{(0)} \\ \tilde{y}_{0.5}^{(0)} \end{pmatrix}\ , \\
d\begin{pmatrix} \tilde{x}_r^{(2)} \\ \tilde{y}_r^{(2)} \end{pmatrix} &= ch\begin{pmatrix} \tilde{x}_r^{(2)}(b_1 - a_{11}\tilde{x}_r^{(2)}) \\ \tilde{y}_r^{(2)}(b_2 - a_{22}\tilde{y}_r^{(2)}) \end{pmatrix} dr\ , & \begin{pmatrix} \tilde{x}_0^{(2)} \\ \tilde{y}_0^{(2)} \end{pmatrix} &= \begin{pmatrix} \tilde{x}_1^{(1)} \\ \tilde{y}_1^{(1)} \end{pmatrix}\ ,
\end{aligned}$$

with the last ODE being solved for time $r = 0.5$. These three ODEs are each solvable explicitly, yielding the following solutions (we will not comment on the existence/uniqueness of the ODE solutions here)

$$\begin{pmatrix} \tilde{x}_{0.5}^{(0)} \\ \tilde{y}_{0.5}^{(0)} \end{pmatrix} = \left(\frac{b_1 x \exp(chb_1/2)}{b_1 - a_{11}x + a_{11}x \exp\left(chb_1/2\right)} \, , \quad \frac{b_2 y \exp(chb_2/2)}{b_2 - a_{22}y + a_{22}y \exp(chb_2/2)} \right)^\top \, ,$$

$$\begin{pmatrix} \tilde{x}_{1}^{(1)} \\ \tilde{y}_{1}^{(1)} \end{pmatrix} = \left(\frac{-\tilde{a}_{12}c_1 \exp(\tilde{a}_{12}c_1c_2)}{\tilde{a}_{21}\exp(\tilde{a}_{12}c_1c_2) - \exp(\tilde{a}_{12}c_1)} \, , \quad c_1 - \frac{\tilde{a}_{21}c_1 \exp(\tilde{a}_{12}c_1c_2)}{\tilde{a}_{21}\exp(\tilde{a}_{12}c_1c_2) - \exp(\tilde{a}_{12}c_1)} \right)^\top \, ,$$

$$\begin{pmatrix} \tilde{x}_{0.5}^{(2)} \\ \tilde{y}_{0.5}^{(2)} \end{pmatrix} = \left(\frac{b_1 \tilde{x}_1^{(1)} \exp(chb_1/2)}{b_1 - a_{11}\tilde{x}_1^{(1)} + a_{11}\tilde{x}_1^{(1)} \exp(chb_1/2)} \, , \quad \frac{b_2 \tilde{y}_1^{(1)} \exp(chb_2/2)}{b_2 - a_{22}\tilde{y}_1^{(1)} + a_{22}\tilde{y}_1^{(1)} \exp(chb_2/2)} \right)^\top \, ,$$

where $\tilde{a}_{12} = ch \times a_{12}$, $\tilde{a}_{21} = ch \times a_{21}$,

$$c_1 := \tilde{y}_{0.5}^{(0)} - \frac{a_{21}}{a_{12}}\tilde{x}_{0.5}^{(0)} \, , \quad \text{and} \quad \tilde{a}_{12}c_1c_2 := \log\left(\tilde{x}_{0.5}^{(0)}\right) - \log\left(\tilde{a}_{12}\,\tilde{y}_{0.5}^{(0)}\right) \, .$$

We then define

$$\phi_{ch}^{\text{Strang}} \begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} \tilde{x}_{0.5}^{(2)} \\ \tilde{y}_{0.5}^{(2)} \end{pmatrix} \, . \tag{6.13}$$

Combining everything we obtain the splitting method given above.

6.2.2 Numerical results

For our numerical experiments, we select the following parameters for the LV model

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} \, , \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 0.0001 & 0.01 \\ -0.01 & 0.0001 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \, ,$$

with initial value $x_0 = y_0 = 10$. The model was simulated out to time $T = 1$ (in part as blow up was observed for the Euler scheme for times larger than $T = 5$). These parameters fall under the 'predator-prey' classification. We compare the splitting derived above with a Strang splitting (5.3), a Lie Trotter splitting (5.1) and an Euler scheme. For the Strang and Lie Trotter splitting, we again used the ODE Strang splitting (6.13). The observed strong error rates are presented in Figure 6.3, where we observe that the proposed splitting achieves a $3/2$ strong order convergence rate and is more accurate than the other methods.

The timings for the compared numerical schemes, for a fixed h, are given in Table 6.3.
In Table 6.4 we compare the times required by each of the schemes to reach an accuracy of $S_N = 10^{-3}$, again we see that the proposed high order splitting performs the best out of the compared schemes. The HS1 splitting is an order of magnitude faster than the other

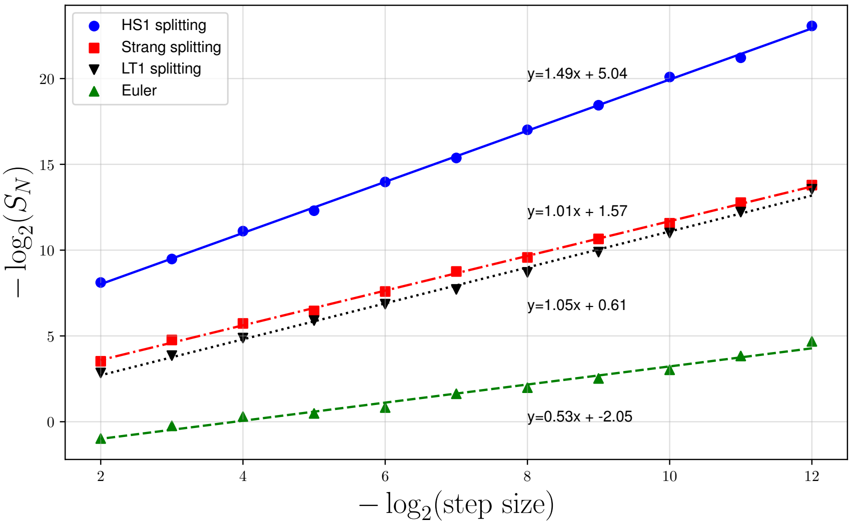


Fig. 6.3: Strong errors for (6.11), estimated from 1000 sample paths.

HS1 splitting (6.12)	Strang splitting	LT1 splitting	Euler
12.26	6.83	3.66	0.87

Table 6.3: Compute time to simulate 1000 sample paths of (6.11) with 100 steps (seconds)

splitting schemes and we note the relative inadequacy of Euler scheme for the LV model. Interestingly, here the LT1 splitting requires roughly 2 times the number of time steps to achieve the same accuracy as the Strang splitting, but takes half as long to run. Making the LT1 and Strang splittings comparable in terms of strong error performance for this example.

HS1 splitting (6.12)	Strang splitting	LT1 splitting	Euler
1.26	21.11	21.29	$\sim 146,000$*

Table 6.4: Estimated compute time to produce 1000 sample paths of (6.11) with an error of $S_N = 10^{-3}$ (seconds). $(*)$ based on Figure 6.3 the Euler method would require $\sim 2^{24}$ steps to achieve an accuracy of $S_N = 10^{-3}$, in combination with Table 6.3 this gives an approximate time to run of 40.5 hours.

Remark 6.2.2. *It is worth noting the instability of the Euler scheme with the Lotka-Volterra model, which was observed to quickly lead to numerical blowups. The choice of model parameters was in part made to avoid this instability.*

6.3 Multilevel Monte Carlo

As described in Section 1.3, MLMC is a variance reduction technique often used to replace the standard Monte Carlo estimator. The 'multilevel' part of the MLMC estimator involves coupled 'coarse' and 'fine' estimates. For an increment only numerical scheme it is clear how to couple the levels as $W_{s,t} = W_{s,u} + W_{u,t}$. In order to use the high order paths we have introduced we note the following relation for the space-time Lévy area which allows us to use the same Brownian paths across step sizes.

$$H_{s,t} = \frac{1}{4}\big(W_{s,u} - W_{u,t}\big) + \frac{1}{2}\big(H_{s,u} + H_{u,t}\big),$$

which follows by (5.26). Thus, we are able to generate a Brownian path for $h_l = 2^{-l}$ and subsample to $h_{l-1} = 2^{-(l-1)}$. The space-time Lévy swing can also easily be coupled by simply using the definition $n_{s,t} := \operatorname{sgn}(H_{s,u} - H_{u,t})$. We also note that the space-time Lévy swings over the sub intervals $n_{s,u}$ and $n_{u,t}$ are independent of $n_{s,t}$.

6.3.1 Example: Interacting stock model

As noted in the introduction, the variance of the correction terms in the MLMC are controlled by the strong error. In the following example, we compare the HS1 splitting with the Strang splitting. As shown in Table 5.1 both these splittings achieve a weak order of $O(h^2)$, but the HS1 splitting achieves a strong order of $O(h^{1.5})$ compared with the Strang splitting's $O(h)$. We are thus interested to explore what benefit can be gained using a higher order (strong) scheme with MLMC.

For demonstration, we consider the following toy model. Let us consider a financial market of M assets, with (risk-neutral) prices $\{S^{(i)}\}_{i=1}^{M}$ each obeying the following dynamics

$$dS_t^{(i)} = \left(rS_t^{(i)} - \beta_i\left(S_t^{(i)} - \sum_{j=1}^{M} c_j^{(i)} S_t^{(j)}\right)\right)dt + \sigma_i S_t^{(i)} dW_t^{(i)}\ , \tag{6.14}$$

with initial value $S_0^{(i)}$, where r is the risk free rate, $\beta_i \geqslant 0$ controls the reversion speed, $c_j^{(i)} \geqslant 0$ and $\sigma_i \geqslant 0$. The idea behind this model is that different assets may derive their prices from each other. For example, you could have a two asset model of a commodity and a manufactured good using that commodity. In which case the commodity may be assumed to follow a geometric Brownian motion (GBM) and the manufactured good is expected to cost some multiple of the commodity. We are then interested in pricing some derivative whose price depends on the price of the M assets. For example, we may wish to price the equal-weighted basket call option with strike price K and payoff

$$\left(\sum_{i=1}^{M} S_T^{(i)} - K\right)^{+} . \tag{6.15}$$

Remark 6.3.1. *As the model we introduce in* (6.14) *is linear, it would also be possible to derive an exact solution. However, the resulting integrals would likely be difficult to couple within the MLMC framework.*

Numerical experiments

For our numerical experiments, we consider a three asset model. We label the assets X, Y and Z. Assuming common reversion speed β, the system has the following dynamics

$$d\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \left\{ r\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} - \beta\left(\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} - \Theta\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix}\right)\right\} dt + \begin{pmatrix} \sigma_X X_t & 0 & 0 \\ 0 & \sigma_Y Y_t & 0 \\ 0 & 0 & \sigma_Y Y_t \end{pmatrix} dW_t \ ,$$

where Θ denotes the interaction matrix defined by

$$\Theta := \begin{pmatrix} 1 & 0 & 0 \\ c_{yx} & 0 & 0 \\ c_{zx} & c_{zy} & 0 \end{pmatrix} ,$$

The solvability of the resulting ODEs is quite dependent on the choice of coefficients in our interaction matrix. So, for simplicity, we set many values to zero and assume that c_{yx}, c_{zx} and c_{zy} are positive constants. This system then models a primary good X (following GBM), a secondary good Y and some product derived from both: Z. Applying the splitting path γ^{HS2} (5.5) we arrive at the following splitting scheme:

$$\begin{pmatrix} X_k^{(1)} \\ Y_k^{(1)} \\ Z_k^{(1)} \end{pmatrix} = \begin{pmatrix} X_k \exp\left(\sigma_X\left(W_k^1/2 + H_k^1 - C_k^1/2\right)\right) \\ Y_k \exp\left(\sigma_Y\left(W_k^2/2 + H_k^2 - C_k^2/2\right)\right) \\ Z_k \exp\left(\sigma_Z\left(W_k^3/2 + H_k^3 - C_k^3/2\right)\right) \end{pmatrix} , \qquad \begin{pmatrix} X_k^{(2)} \\ Y_k^{(2)} \\ Z_k^{(2)} \end{pmatrix} = \phi\begin{pmatrix} X_k^{(1)} \\ Y_k^{(1)} \\ Z_k^{(1)} \end{pmatrix} ,$$

$$\begin{pmatrix} X_k^{(3)} \\ Y_k^{(3)} \\ Z_k^{(3)} \end{pmatrix} = \begin{pmatrix} X_k^{(2)} \exp\left(\sigma_X C_k^1\right) \\ Y_k^{(2)} \exp\left(\sigma_Y C_k^2\right) \\ Z_k^{(2)} \exp\left(\sigma_Z C_k^3\right) \end{pmatrix} , \qquad \begin{pmatrix} X_k^{(4)} \\ Y_k^{(4)} \\ Z_k^{(4)} \end{pmatrix} = \phi\begin{pmatrix} X_k^{(3)} \\ Y_k^{(3)} \\ Z_k^{(3)} \end{pmatrix} ,$$

$$\begin{pmatrix} X_{k+1} \\ Y_{k+1} \\ Z_{k+1} \end{pmatrix} = \begin{pmatrix} X_k^{(4)} \exp\left(\sigma_X\left(W_k^1/2 - H_k^1 - C_k^1/2\right)\right) \\ Y_k^{(4)} \exp\left(\sigma_Y\left(W_k^2/2 - H_k^2 - C_k^2/2\right)\right) \\ Z_k^{(4)} \exp\left(\sigma_Z\left(W_k^3/2 - H_k^3 - C_k^3/2\right)\right) \end{pmatrix} ,$$

where, for $\tilde{\beta} := \beta h/2$,

$$\phi\begin{pmatrix} x \\ y \\ z \end{pmatrix} := e^{rh/2}\begin{pmatrix} x \\ e^{-\tilde{\beta}}y + c_{yx}(1-e^{-\tilde{\beta}})x \\ e^{-\tilde{\beta}}z + \tilde{\beta}c_{zy}e^{-\tilde{\beta}}y + \big\{c_{zx}(1-e^{-\tilde{\beta}}) + c_{yx}c_{zx}\big(1-(1+\tilde{\beta})e^{-\tilde{\beta}}\big)\big\}x \end{pmatrix},$$

which may not be the most aesthetically pleasing splitting scheme we have presented, but has the benefit that each ODE term from the splitting path is exactly solvable. As the splitting scheme above is not difficult to derive, we will not present its derivation and instead proceed straight to our numerical experiments.

We choose the following set of parameter values

$$r = 0.05\ , \quad \beta = 1\ , \quad c_{yx} = 1.5 \quad c_{zx} = c_{zy} = 1\ , \quad \sigma_X = \sigma_Y = \sigma_Z = 0.2\ ,$$

with initial values $X_0 = \$100, Y_0 = \$150$ and $Z_0 = \$250$. With these parameters, Y is expected to cost $1.5X$ and Z is expected to cost $Y + X$. An example realisation of the model is shown in Figure 6.4.

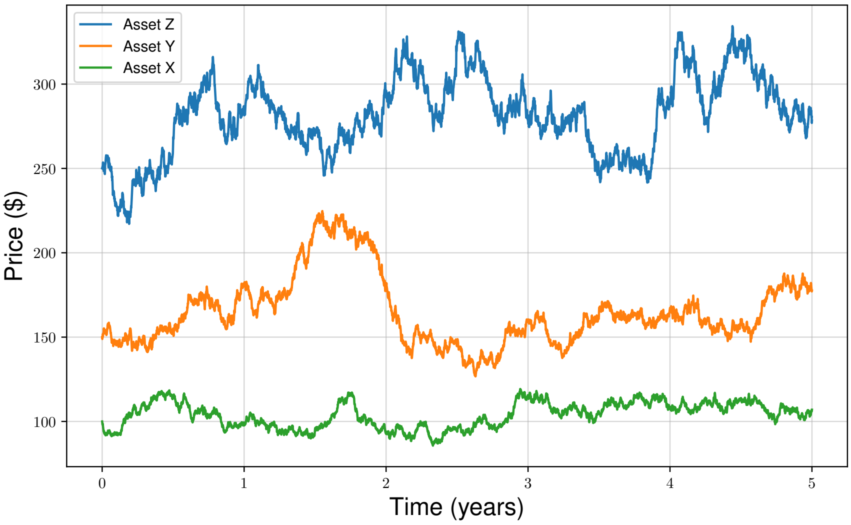
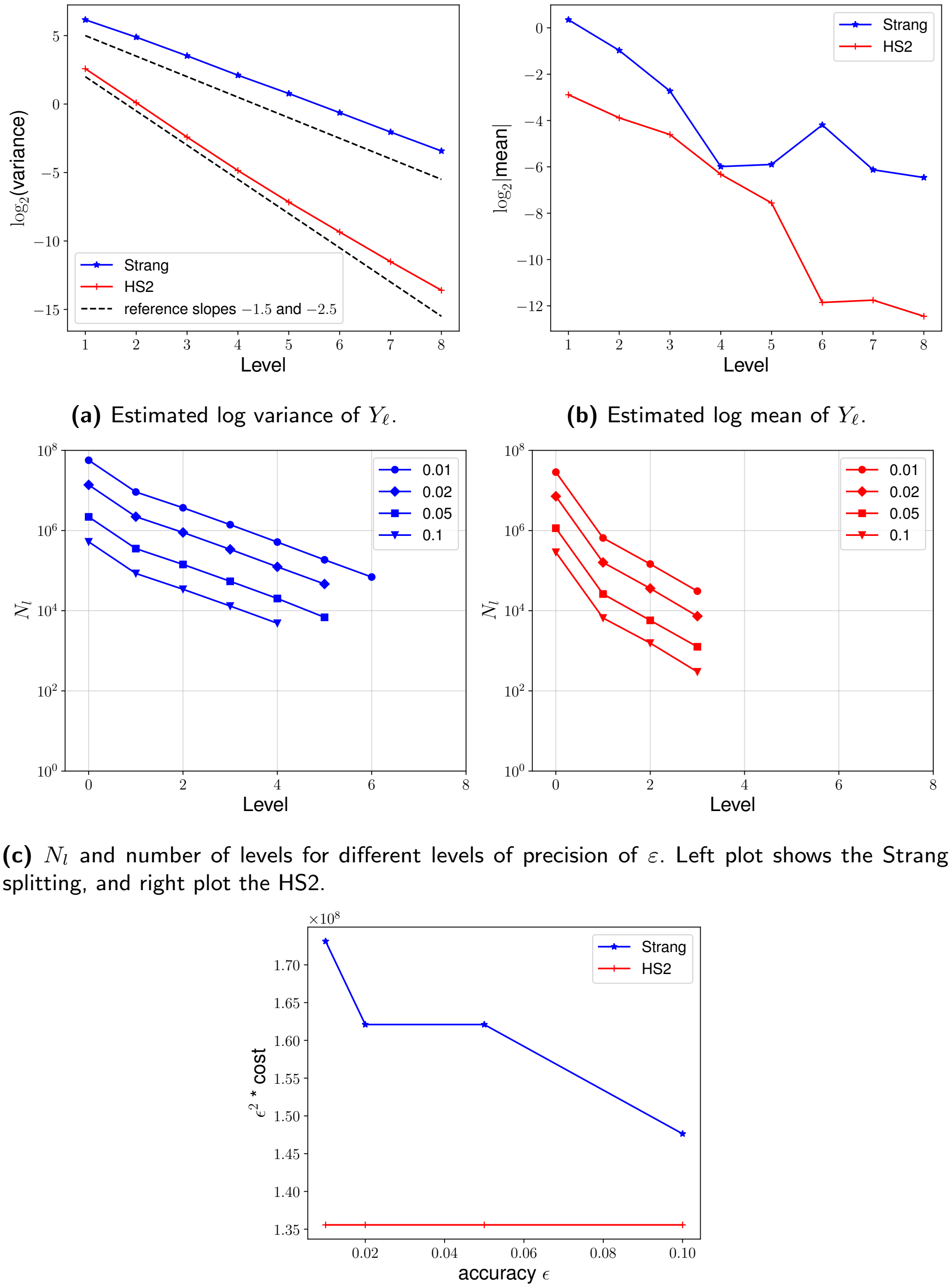


Fig. 6.4: Example price path for three interacting assets.

We price the 'out of the money' basket option (6.15) with strike $K = 550$ and expiry $T = 1$. We compare the results using a standard Monte Carlo estimator and the MLMC estimator. To assess the impact of using a higher order strong scheme, we compare with the Strang splitting. Both the Strang and the HS2 splitting are weak order $O(h^2)$. For our numerical experiments, we estimate the computational cost of each level C_ℓ using the computational time and apply the MLMC algorithm as described in [24, Algorithm 1].



(a) Estimated log variance of Y_ℓ.

(b) Estimated log mean of Y_ℓ.

(c) N_l and number of levels for different levels of precision of ε. Left plot shows the Strang splitting, and right plot the HS2.

(d) Cost comparison of HS2 and Strang splitting

Fig. 6.5: Plots comparing MLMC performance for Strang splitting and HS2 splitting.

In Figure 6.5 we plot the results of applying the MLMC algorithm with the HS2 splitting and a Strang splitting. From plot 6.5a we see that the HS2 splitting achieves a lower variance and a faster rate of decrease across the different levels, when compared with the Strang slitting. From plot 6.5b we see that the correction terms are smaller at each level for the HS2 splitting. In Figure 6.5c, we see that the HS2 splitting requires fewer levels and fewer samples per level than the Strang splitting to achieve a set accuracy of ε. As displayed in plot 6.5d, to achieve a given level of accuracy, the total computational cost (as defined in (1.10)) of the HS2 splitting is thus lower than the Strang splitting. We thus see a clear benefit from using the higher order splitting $HS2$.

6.4 Testing Long-time integration

As a small experiment, we consider here the example of the scalar anharmonic oscillator

$$dy_t = \sin(y_t)dt + dW_t \ . \tag{6.16}$$

The motivation for this experiment is the paper [39] where it was shown that a non-Markovian Euler scheme achieves a high order of convergence in the long time horizon, transitioning from a 1st order weak scheme to a second order weak scheme (see their Figure 2). We are thus interested if such behaviour can be observed for our splitting schemes.

Using either of the paths γ^{SO1} (5.8) or γ^{SO2} (5.9), we propose the following splitting method for the anharmonic oscillator:

$$\begin{aligned}
\widetilde{Y}_k^{\mathsf{SR}} &:= Y_k^{\mathsf{SR}} + C_1, \\
\widetilde{Y}_{k+\frac{2}{3}}^{\mathsf{SR}} &:= \widetilde{Y}_k^{\mathsf{SR}} + \frac{2}{3}\Big(f\big(\widetilde{Y}_k^{\mathsf{SR}}\big)h + C_2\Big), \\
Y_{k+1}^{\mathsf{SR}} &:= Y_k^{\mathsf{SR}} + \frac{1}{4}f\big(\widetilde{Y}_k^{\mathsf{SR}}\big)h + \frac{3}{4}f\big(\widetilde{Y}_{k+\frac{2}{3}}^{\mathsf{SR}}\big)h + W_k,
\end{aligned} \tag{6.17}$$

where $C_1, C_2 \in \mathbb{R}^d$ are the first two increments of the driving piecewise linear path γ. Both paths γ^{SO1} and γ^{SO2} will achieve a strong order convergence of $3/2$. For our numerical experiments we use the path γ^{SO2}. This splitting method uses Ralston's method [60] to discretize the ODEs resulting from the splitting path. The derivation of (6.17) is presented in our paper [19, Section 5.2]. The non-Markovian Euler scheme used in [39], follows a simple modification of the Euler method (Definition 1.4):

$$Y_{k+1} := Y_k + f(Y_k)h + \frac{1}{2}\big(W_{t_{k-1},t_k} + W_{t_k,t_{k+1}}\big) \ .$$

In Figure 6.6 we present the result of our experiment. The 'L_2 distribution error' used to compare the scheme is as described in [39], where a histogram describing the distribution is constructed at each point in time. This estimated distribution is then compared using an L_2 distance with a histogram estimated with a finer time step. We see a similar trend between the splitting and the non-Markovian Euler, where the error for both decreases with time. However, in contrast to our previous numerical experiments, we do not see a benefit from using the splitting, and it in fact performs worse than the non-Markovian Euler. An interesting question for future research is to see if the non-Markovian Euler can be improved upon by including additional random variables such as the space-time Lévy area $H_{s,t}$ we have used here.

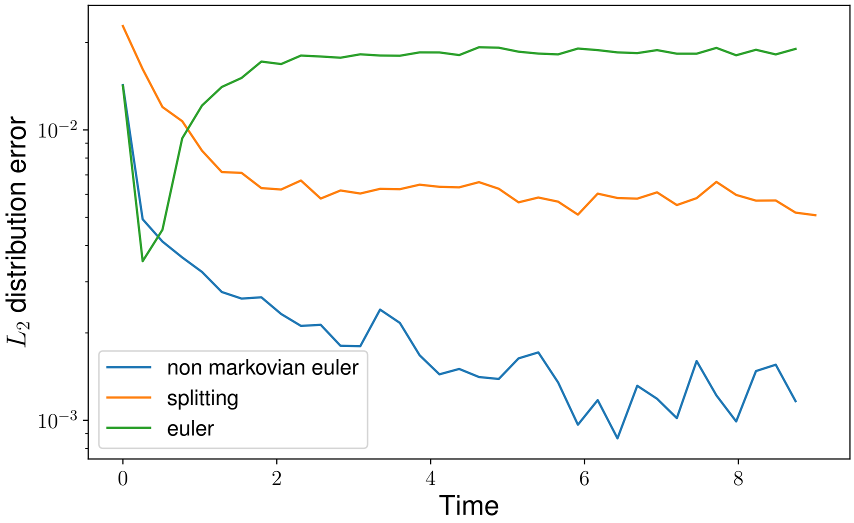


Fig. 6.6: Comparison of splitting (6.17) with the non-Markovian Euler scheme, based on 10^6 samples.

Chapter 7

Conclusions and future work

In this thesis, we have presented a new simple methodology for designing and analysing splitting methods for SDE simulation. The key idea is to replace the system's Brownian motion with a piecewise linear path. Moreover, for SDEs satisfying a commutativity condition, we developed several high order splitting methods which displayed state-of-the-art convergence in experiments. As part of this investigation, we also detailed how recently developed estimators for iterated integrals of Brownian motion can be directly incorporated into such methods. Since these estimators were simply obtained as the expectation of iterated integrals, conditional on the generatable random variables, they are optimal in an $L^2(\mathbb{P})$ sense.

Furthermore, the results in this thesis may lead to several areas of future research:

- **Development and analysis of methods inspired by splitting paths**

  For example, in the additive noise setting, the following Strong 1.5 Stochastic Runge-Kutta method is based on the splitting path (5.11), but with $K_{s,t} = 0$.

$$\begin{aligned}\widetilde{Y}_k^{\mathsf{SRK}} &:= Y_k^{\mathsf{SRK}} + \sigma H_k,\\ \widetilde{Y}_{k+\frac{5}{6}}^{\mathsf{SRK}} &:= \widetilde{Y}_k^{\mathsf{SRK}} + \frac{5}{6}\Big(f\big(\widetilde{Y}_k^{\mathsf{SRK}}\big)h + \sigma W_k\Big),\\ Y_{k+1}^{\mathsf{SRK}} &:= Y_k^{\mathsf{SRK}} + \frac{2}{5}f\big(\widetilde{Y}_k^{\mathsf{SRK}}\big)h + \frac{3}{5}f\big(\widetilde{Y}_{k+\frac{5}{6}}^{\mathsf{SRK}}\big)h + \sigma W_k. \qquad (7.1)\end{aligned}$$

  As (7.1) does not follow a high order splitting path or use Ralston's method, it was not included in Section 6.4 and thus we leave its analysis as future work. Similarly, conducting error analyses and further numerical investigations for the Shifted Euler and Runge-Kutta methods described in our paper [19, Section 5] is a future topic.

- **Development of high order splitting methods for general SDEs**

  For example, the below method is a combination of the Strang splitting (5.3) and the log-ODE method from rough path theory [50, Appendices A and B].

$$Y_{k+1} := \exp\left(\frac{1}{2}f(\cdot)h\right)\exp\left(g(\cdot)W_k + \sum_{i<j}\big[g_i, g_j\big](\cdot)A_k^{ij}\right)\exp\left(\frac{1}{2}f(\cdot)h\right)Y_k,$$

where $[g_i, g_j](\cdot) = g_j'(\cdot)g_i(\cdot) - g_i'(\cdot)g_j(\cdot)$ is the standard vector field Lie bracket and $A_k = \{A_k^{ij}\}_{1\leqslant i,j\leqslant d}$ is the Lévy area of the Brownian motion over $[t_k, t_{k+1}]$. If A_k is replaced by a random matrix $\widetilde{A}_k$, with the same mean and covariance, we expect the resulting splitting method to achieve $O(h^2)$ weak convergence. Similarly, we expect that the Ninomiya-Ninomiya [53] and Ninomiya-Victoir [55] weak second order schemes can be reinterpreted as path-based splittings.

- **Incorporating adaptive step sizes into $(W_k\ , H_k\ , n_k)$-based methods**

  Since it is possible to generate both $(W_{s,u}, H_{s,u}, n_{s,u})$ and $(W_{u,t}, H_{u,t}, n_{u,t})$ conditional on $(W_{s,t}, H_{s,t}, n_{s,t})$, where $u = s + \frac{1}{2}h$ is the midpoint of $[s,t]$, the proposed splitting methods can be applied using an adaptive step size. Such a methodology was detailed and initially investigated in [14, Chapter 6].

- **Application to high-dimensional SDEs in physics and data science**

  High-dimensional SDEs have seen a variety of real-world applications, ranging from molecular dynamics [38, 48] to machine learning [31, 33, 41, 63, 67, 69]. Therefore, it would be interesting to investigate whether the splitting methods developed in this paper could improve algorithms used in these applications.

- **Systematic comparison with cubature methods on Wiener space for weak approximations of SDEs**

  As discussed briefly in Section 4.4 our analysis draws inspiration from previous work on Cubature on Wiener space [26, 44]. Indeed, the path based perspective and emphasis on the algebraic structure of the iterated integrals is shared with our work; However, a systematic methodological and numerical comparison with Cubature is outside the scope of this thesis and so is left as a topic for future research. One would expect Cubature to outperform Monte Carlo methods given a smooth payoff and a limited number of steps, but (multilevel) Monte Carlo is much easier to use in practice.

Bibliography

[1] M. Arató. 'A famous nonlinear stochastic equation (Lotka-Volterra model with diffusion)'. In: *Mathematical and Computer Modelling* 38.7 (2003). Hungarian Applied Mathematics, pp. 709–726. ISSN: 0895-7177. DOI: https://doi.org/10.1016/S0895-7177(03)90056-2. URL: https://www.sciencedirect.com/science/article/pii/S0895717703900562.

[2] Christian Bayer. 'The geometry of iterated Stratonovich integrals'. In: *preprint* (2006). URL: https://www.wias-berlin.de/people/bayerc/files/strat_geom.pdf.

[3] Sani Biswas et al. 'An explicit Milstein-type scheme for interacting particle systems and McKean–Vlasov SDEs with common noise and non-differentiable drift coefficients'. In: *https://arxiv.org/2208.10052* (2022).

[4] Evelyn Buckwar et al. 'A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model'. In: *Applied Numerical Mathematics* 179 (2022), pp. 191–220.

[5] Thomas Cass and Peter Friz. 'Densities for rough differential equations under Hörmander's condition'. In: *Annals of mathematics* (2010), pp. 2115–2141.

[6] Ilya Chevyrev and Andrey Kormilitzin. 'A primer on the signature method in machine learning'. In: *arXiv preprint arXiv:1603.03788* (2016).

[7] J. M. C. Clark and R. J. Cameron. *The maximum rate of convergence of discrete approximations for stochastic differential equations*. in Stochastic Differential Systems Filtering and Control, ed. by Grigelionis (Springer, Berlin), 1980.

[8] Alexander Davie. 'KMT theory applied to approximations of SDE'. In: *Stochastic Analysis and Applications*. Vol. 100. Springer Proceedings in Mathematics and Statistics. Springer, 2014, pp. 185–201.

[9] Andrew S. Dickinson. 'Optimal Approximation of the Second Iterated Integral of Brownian Motion'. In: *Stochastic Analysis and Applications* 25.5 (2007), pp. 1109–1128.

[10] Tanja Eisner et al. 'Markov Operators'. In: *Operator Theoretic Aspects of Ergodic Theory*. Cham: Springer International Publishing, 2015, pp. 249–271. ISBN: 978-3-319-16898-2. DOI: 10.1007/978-3-319-16898-2_13. URL: https://doi.org/10.1007/978-3-319-16898-2_13.

[11] Regina C. Elandt. 'The Folded Normal Distribution: Two Methods of Estimating Parameters from Moments'. In: *Technometrics* 3.4 (1961), pp. 551–562.

[12] Wei Fang and Michael B. Giles. 'Adaptive Euler-Maruyama method for SDEs with nonglobally Lipschitz drift'. In: *Annals of Applied Probability* 30.2 (2020), pp. 526–560.

[13] Istvan Farago and Agnes Havasi. 'On the convergence and local splitting error of different splitting schemes'. In: *Progress in Computational Fluid Dynamics, an International Journal* 5.8 (2005), pp. 495–504.

[14] James Foster. 'Numerical approximations for stochastic differential equations'. PhD thesis. University of Oxford, 2020. URL: `https://ora.ox.ac.uk/objects/uuid:775fc3f5-501c-425f-8b43-fc5a7b2e4310`.

[15] James Foster and Karen Habermann. 'Brownian bridge expansions for Lévy area approximations and particular values of the Riemann zeta function'. In: *To appear in Combinatorics, Probability and Computing* (2021). URL: `https://arxiv.org/abs/2102.10095`.

[16] James Foster, Terry Lyons and Vlad Margarint. 'An asymptotic radius of convergence for the Loewner equation and simulation of SLE traces via splitting'. In: *Journal of Statistical Physics* 189.18 (2022).

[17] James Foster, Terry Lyons and Harald Oberhauser. 'An Optimal Polynomial Approximation of Brownian Motion'. In: *SIAM Journal on Numerical Analysis* 58.3 (2020), pp. 1393–1421.

[18] James Foster, Terry Lyons and Harald Oberhauser. 'The shifted ODE method for underdamped Langevin MCMC'. In: *https://arxiv.org/abs/2101.03446* (2021).

[19] James Foster, Goncalo dos Reis and Calum Strange. *High order splitting methods for SDEs satisfying a commutativity condition*. 2022. DOI: `10.48550/ARXIV.2210.17543`. URL: `https://arxiv.org/abs/2210.17543`.

[20] Peter K Friz and Nicolas B Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Vol. 120. Cambridge University Press, 2010.

[21] Peter K. Friz and Martin Hairer. *A Course on Rough Paths: With an Introduction to Regularity Structures*. Springer, 2020.

[22] Jessica G Gaines and Terry J Lyons. 'Random generation of stochastic area integrals'. In: *SIAM Journal on Applied Mathematics* 54.4 (1994), pp. 1132–1146.

[23] Saadia Ghazali. 'The global error in weak approximations of stochastic differential equations'. In: (2006).

[24] Michael B Giles. 'Multilevel monte carlo methods'. In: *Acta numerica* 24 (2015), pp. 259–328.

[25] L Gyurko. 'Numerical methods for approximating solutions to rough differential equations'. PhD thesis. University of Oxford, 2008.

[26] Lajos Gergely Gyurkó and Terry J Lyons. 'Efficient and practical implementations of cubature on Wiener space'. In: *Stochastic analysis 2010* (2011), pp. 73–111.

[27] Zhengmian Hu, Feihu Huang and Heng Huang. 'Optimal Underdamped Langevin MCMC Method'. In: *Advances in Neural Information Processing Systems* (2021).

[28] Martin Hutzenthaler, Arnulf Jentzen and Peter E. Kloeden. 'Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients'. In: *Annals of Applied Probability* 22.4 (2012), pp. 1611–1641.

[29] Yuga Iguchi and Toshihiro Yamada. 'Operator splitting around Euler-Maruyama scheme and high order discretization of heat kernels'. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 55 (2021), pp. 323–367.

[30] Patrick Kidger. 'On Neural Differential Equations'. PhD thesis. University of Oxford, 2021. URL: `https://ora.ox.ac.uk/objects/uuid:af32d844-df84-4fdc-824d-44bebc3d7aa9`.

[31] Patrick Kidger et al. 'Efficient and Accurate Gradients for Neural SDEs'. In: *Advances in Neural Information Processing Systems* (2021).

[32] Patrick Kidger et al. 'Neural Controlled Differential Equations for Irregular Time Series'. In: *Advances in Neural Information Processing Systems* (2020).

[33] Patrick Kidger et al. 'Neural SDEs as Infinite-Dimensional GANs'. In: *Proceedings of the 38th International Conference on Machine Learning* (2021).

[34] Jack PC Kleijnen, Ad Ridder and Reuven Rubinstein. 'Variance reduction techniques in Monte Carlo methods'. In: (2010).

[35] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.

[36] Raphael Kruse and Yue Wu. 'A randomized and fully discrete Galerkin finite element method for semilinear stochastic evolution equations'. In: *Mathematics of Computation* 88 (2019), pp. 2793–2825.

[37] Raphael Kruse and Yue Wu. 'A randomized Milstein method for stochastic differential equations with non-differentiable drift coefficients'. In: *Discrete and Continuous Dynamical Systems. Series B. A Journal Bridging Mathematics and Sciences* 24.8 (2019), pp. 3475–3502.

[38] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*. Interdisciplinary Applied Mathematics, Springer, 2015.

[39] Benedict Leimkuhler, Charles Matthews and MV Tretyakov. 'On the long-time integration of stochastic gradient systems'. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 470.2170 (2014), p. 20140120.

[40] José R. León and Adeline Samson. 'Hypoelliptic stochastic FitzHugh-Nagumo neuronal model: mixing, up-crossing and estimation of the spike rate'. In: *Annals of Applied Probability* 28.4 (2018), pp. 2243–2274.

[41] Ruilin Li, Hongyuan Zha and Molei Tao. 'Sqrt(d) dimension dependence of Langevin Monte Carlo'. In: *Proceedings of the 10th International Conference on Learning Representations* (2022).

[42] Christian Litterer and Terry Lyons. 'High order recombination and an application to cubature on Wiener space'. In: (2012).

[43] Terry Lyons, Michael Caruana and Thierry Lévy. *Differential Equations Driven by Rough Paths*. Vol. 1908 of Lecture Notes in Mathematics. Springer, 2007.

[44] Terry Lyons and Nicolas Victoir. 'Cubature on Wiener space'. In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 460.2041 (2004), pp. 169–198.

[45] Shev MacNamara and Gilbert Strang. 'Operator splitting'. In: *Splitting methods in communication, imaging, science, and engineering* (2016), pp. 95–114.

[46] Xuerong Mao, Glenn Marion and Eric Renshaw. 'Environmental Brownian noise suppresses explosions in population dynamics'. In: *Stochastic Processes and their Applications* 97.1 (2002), pp. 95–110.

[47] Aaron Meurer et al. 'SymPy: symbolic computing in Python'. In: *PeerJ Computer Science* 3 (Jan. 2017), e103. ISSN: 2376-5992. DOI: `10.7717/peerj-cs.103`. URL: `https://doi.org/10.7717/peerj-cs.103`.

[48] Grigori N. Milstein and Michael V. Tretyakov. *Stochastic Numerics for Mathematical Physics*. Springer, 2004.

[49] Tetsuya Misawa. 'Numerical integration of stochastic differential equations by composition methods'. In: *Dynamical systems and differential geometry (Japanese)*. 1180. 2000, pp. 166–190.

[50] James Morrill et al. 'Neural Rough Differential Equations for Long Time Series'. In: *Proceedings of the 38th International Conference on Machine Learning* (2021).

[51] James Morrill et al. 'On the Choice of Interpolation Scheme for Neural CDEs'. In: *Transactions on Machine Learning Research* (2022).

[52] Jan Mrongowius and Andreas Rößler. 'On the approximation and simulation of iterated stochastic integrals and the corresponding Lévy areas in terms of a multidimensional Brownian motion'. In: *Stochastic Analysis and Applications* 40.3 (2022), pp. 397–425.

[53] M. Ninomiya and S. Ninomiya. 'A new higher-order weak approximation scheme for stochastic differential equations and the Runge–Kutta method'. In: *Finance and Stochastics* 13.3 (2009), pp. 415–443.

[54] SYOITI Ninomiya and YUJI Shinozaki. 'On implementation of high-order recombination and its application to weak approximations of stochastic differential equations'. In: *Proceedings of the NFA 29th Annual Conference*. 2021.

[55] Syoiti Ninomiya and Nicolas Victoir. 'Weak approximation of stochastic differential equations and application to derivative pricing'. In: *Applied Mathematical Finance* 15.2 (2008), pp. 107–121.

[56] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[57] Gilles Pagès. 'Numerical probability'. In: *Universitext*. Springer, 2018.

[58] Sahani Pathiraja. 'L2 convergence of smooth approximations of Stochastic Differential Equations with unbounded coefficients'. In: *https://arxiv.org/abs/2011.13009* (2020).

[59] Eckhard Platen Peter E. Kloeden. *Numerical Solution of Stochastic Differential Equations*. 1st ed. Springer Berlin, Heidelberg, 2013. ISBN: 9783662126165; 3662126168.

[60] Anthony Ralston. 'Runge-Kutta methods with minimum error bounds'. In: *Mathematics of Computation* 269 (1962), pp. 431–437.

[61] A. Rößler. 'Runge–Kutta methods for the strong approximation of solutions of stochastic differential equations'. In: *SIAM Journal on Numerical Analysis* 8.3 (2010), pp. 922–952.

[62] Ruoqi Shen and Yin Tat Lee. 'The Randomized Midpoint Method for Log-Concave Sampling'. In: *Advances in Neural Information Processing Systems* (2019).

[63] Yang Song et al. 'Score-Based Generative Modeling through Stochastic Differential Equations'. In: *Proceedings of the International Conference on Learning Representations* (2021).

[64] Gilbert Strang. 'On the Construction and Comparison of Difference Schemes'. In: *SIAM Journal on Numerical Analysis* 5.3 (1968), pp. 506–517.

[65] Irene Tubikanec et al. 'Qualitative properties of different numerical methods for the inhomogeneous geometric Brownian motion'. In: *Journal of Computational and Applied Mathematics* 406 (2022).

[66] Fernando Vadillo. 'Comparing stochastic Lotka–Volterra predator-prey models'. In: *Applied Mathematics and Computation* 360 (2019), pp. 181–189. ISSN: 0096-3003. DOI: `https://doi.org/10.1016/j.amc.2019.05.002`. URL: `https://www.sciencedirect.com/science/article/pii/S0096300319303893`.

[67] Max Welling and Yee Whye Teh. 'Bayesian Learning via Stochastic Gradient Langevin Dynamics'. In: *Proceedings of the 28th International Conference on Machine Learning* (2011).

[68] Magnus Wiktorsson. 'Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions'. In: *Annals of Applied Probability* 11.2 (2001), pp. 470–487.

[69] Qinsheng Zhang and Yongxin Chen. 'Path Integral Sampler: A Stochastic Control Approach For Sampling'. In: *Proceedings of the International Conference on Learning Representations* (2022).

PART II

Machine learning for battery lifetime prognostics

Nomenclature

BMS	Battery management system
CALCE	Centre for Advanced Life Cycle Engineering
CC	Constant-current
CC-CV	Constant-current constant-voltage
CV	Constant-voltage
DOD	Depth of discharge
DOE OE	U.S. Department of Energy's Office of Electricity
DST	Dynamic stress test
ECM	Equivalent circuit model
EIS	Electrochemical impedance spectroscopy
EOL	End of life
EV	Electric vehicle
eVTOL	Electric vertical takeoff and landing
FUDS	Federal Urban Driving Schedule
HEV	Hybrid electric vehicles
HPPC	Hybrid power pulse characterisation
HWFET	Highway Fuel Economy Driving Schedule
IR	Internal resistance
LCO	Lithium cobalt oxide ($LiCoO_2$)
LFP	Lithium iron phosphate ($LiFePO_4$)
LMO	Lithium ion manganese oxide ($LiMn_2O_4$)
NASA	National Aeronautics and Space Administration
NCA	Lithium nickel cobalt aluminium oxide ($LiNiCoAlO_2$)
NMC	Lithium nickel manganese cobalt oxide ($LiNiMnCoO_2$)
OCV	Open-circuit voltage
RPT	Reference performance tests
RUL	Remaining useful life
SOC	State of charge
SOH	State of health
UDDS	Urban Dynamometer Driving Schedule

Chapter 8

A brief outline of part II

As a complete departure from the first part of this thesis, we now turn our attention to the data-driven analysis of Lithium-ion batteries. The work contained here comes from a series of papers published in the area. As each paper contains its own introduction, we point the reader to these for a description of relevant advances and background knowledge for this study.

Chapter 9 presents an extensive review of battery datasets in the public domain.

In Chapter 10, we begin our efforts to understand the degradation of batteries: introducing the novel concept of 'elbows' to describe a key inflection point for the internal resistance (IR) of cells and presenting a machine learning model for their prediction. We also present here a model to estimate the IR of cells from cycling data, this has the additional output of completing an existing dataset which did not contain IR data.

In Chapter 11, we continue our investigation into the degradation of batteries. We show that the full lifetime capacity and IR curves can be described by a few key points. We then present a model which can predict these key values from any single individual cycle of data. From this prediction, we can then predict the full degradation of a battery over its lifetime. Prediction from one cycle as presented here was a novel idea and has a distinct advantage over previous approaches: which require multiple cycles of data and are largely restricted to early life prediction. The results we present here have particular application to the assessment of 'second life' batteries, which is a growing area of interest because of the necessity to reduce the waste associated with replacing used cells. At the time of publication our results represented an improvement over past results both in terms of required input data, versatility and accuracy.

In Chapter 12, we return to the question of data with a particular emphasis on experimental design. We answer the question of how many cells need to be tested to accurately capture cell-to-cell variability, up to specified levels of accuracy and confidence. This analysis is intended to assist the rigorous design of experiments while reducing operational costs and maximising informational gain. We supplement our analysis by presenting its application to several disparate datasets, and proposing two different approaches to experimental design which leverage our results.

Chapter 9

Review of publicly available data for Li-ion batteries.

The work in this chapter is taken from our paper [58], which was a joint work with Dr. Shawn Li, my supervisor Prof. Gonçalo dos Reis and Mohit Yadav MSc.

Abstract

Lithium-ion batteries are fuelling the advancing renewable-energy based world. At the core of transformational developments in battery design, modelling and management is data. In this work, the datasets associated with lithium batteries in the public domain are summarised. We review the data by mode of experimental testing, giving particular attention to test variables and data provided. Alongside highlighted tools and platforms, over 30 datasets are reviewed.

9.1 Introduction

Lithium batteries currently dominate the battery market and the associated research environment. They display favourable properties when compared to other existing battery types: high energy efficiency, low memory effects and proper energy density for large scale energy storage systems and for battery/hybrid electric vehicles (HEV) [136]. Given these facts, lithium production has been expanding rapidly and the use of lithium batteries is wide spread and increasing [56].

From design and sale to deployment and management, and across the value chain [105], data plays a key role informing decisions at all stages of a battery's life. During design, data-informed approaches have been used to accelerate slower discovery processes such as component development and production optimisation (for electrodes, electrolytes, additives and formation) [128, 129]. At sale, they can classify batteries based on expected lifetime [78, 197]. At deployment, data on the expected lifetime and performance of batteries – for a range of chemistries, geometries, capacities and manufacturers – can help to determine the best battery for a given application: under different ageing stresses such as various charge/discharge currents [23, 197, 205], operating temperatures [54, 161, 185], depth of discharges (DODs) [29, 193] and periods of disuse [111, 171]. In use, the battery management system (BMS), controlling the battery's operation, relies heavily on data both for its own design and for the training and calibration of the models it uses.

Data driven approaches are showing great promise and proof of this is the growing body of literature exploring the interplay between data-driven techniques and battery applications [7, 106, 144, 233]. The approach has been deployed in the design of new models for the estimation of state of health (SOH) [112, 166, 179, 183], state of charge (SOC) [36, 39] and internal resistance (IR) [123, 172]; the prediction of remaining useful life (RUL) under cycling degradation [78, 197, 214], calendar ageing [130] and from electrochemical impedance spectroscopy (EIS) data [246]; the identification and prediction of phase change-points in capacity fade curves (knees) [78] and IR rise curves (elbows) [210]; new general online estimation methods for advanced BMSs [150]. Moreover, the data-driven paradigm has been used to improve fault detection [40, 151, 238], charge management [127, 135], thermal management [198] and so much more: from materials development based on atomistic principles [50] to techno-economic analysis [175–177, 215] and approaches to recycling [250].

Batteries are subjected to a wide range of operating conditions in turn influencing their performance, and thus, data covering these conditions is fundamental to the design and validation of accurate models. Physics-based and empirical models, often used in the BMS or 'in the cloud' with new 'digital twin' approaches [121, 233], require careful calibration of model parameters; and, machine learning and statistical based approaches require large amounts of data for training and perform poorly when predicting 'out of distribution' (in circumstances which differ greatly from those present in the data used to train the models). Within their vast scope of deployment, batteries undergo application specific degradation: the demands placed on an electric vehicle (EV) battery – periods of high, varying, load followed by extended rest – are quite different from those placed on powertools, laptops, cellphones, stationary energy storage, aeroplanes or satellites. For this reason, application specific data is needed and we bring attention to this in our discussion.

Well formatted and easily accessible public datasets will bring 'fresh eyes' to problems. Not everyone has access to a Lab to run experiments or the funds required to purchase data. Data that remains local to its generating lab can be leveraged only by a tiny fraction of a wide community of experts. The benefits of public data are numerous: researchers performing experiments gain a reference for their design and new insights into their data as other researchers with cross-domain expertise employ it; modellers and industry profit greatly from the ability to validate results and speed up discovery on public data; and, the barrier to entry is lowered for those new to an area. More data means more research and research is essential for economic growth, job creation and societal progress [76].

The **main contribution** of this work is to provide an actionable summary of publicly available lithium-ion battery data, giving particular attention to explored test variables and provided data. With this information, we hope to inform future research and experimental design, and encourage the sharing of new, accessible and well formatted datasets. To assist the reader, at the end of the main sections we provide tables summarising the presented datasets by cell, test variables, given data and number of cells with hyperlinks.

This work is organised as follows. The accessible testing data is categorised in Section 9.2 according to type and includes datasets available on request. Tools, libraries, platforms and a perspective on current limitations are covered in Section 9.3. Section 9.4 contains the conclusion of this review work and is followed by a nomenclature listing.

Links to data: All web links have been verified (at final submission). The links are given with bibliographical number and direct hyperlinks attached to the word 'URL'.

License: Datasets are provided under certain license attributions mainly according to Creative Commons [46, URL], the Open Database License [149, URL] and the Database Contents License [148, URL]. We refer to the supplementary material Section 1 for a summary description of the shorthand nomenclature.

Reference for 18650 type cells: Where full cell descriptions for a dataset were not given by the generating authors we refer to the resource [147, URL] which provides an extensive reference for the identification of 18650 type batteries.

9.2 Where is the Data?

Historically, interest in different cell chemistries, testing conditions and procedures evolved reflecting the technological improvements batteries underwent. The first significant public battery dataset can be traced back to 2008 published by NASA [47]. As new battery chemistries appeared, the interest shifted from lithium iron phosphate (LFP) to lithium nickel manganese cobalt oxide (NMC) and lithium nickel cobalt aluminium oxide (NCA) batteries. Both NMC and NCA chemistries are better suited for power tools, e-bikes and other electric powertrains as they offer higher specific energy, reasonably good specific power and long lifespan. In Fig.9.1, a hierarchical architecture of existing battery datasets across time is given. The number of cells tested and the variety of testing variables explored has increased with growing interest in data-driven techniques and a desire to understand more complex interactions.

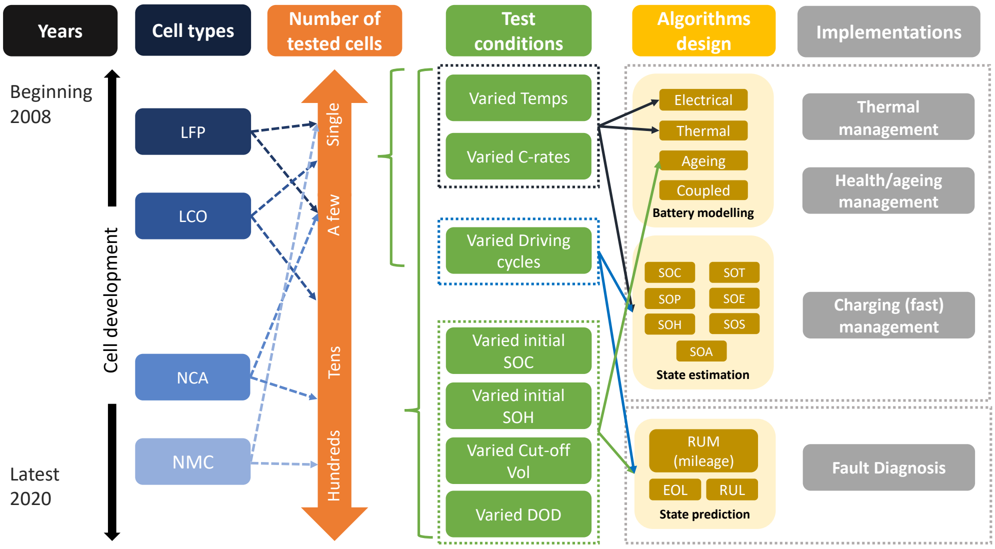


Fig. 9.1: Hierarchical architecture of the existing battery datasets from an historical point of view.

Cell chemistry, number of tested cells and testing conditions are key to determine the usefulness of a specific battery dataset. We provide a comprehensive examination of the available datasets, in particular, highlighting these three elements.

9.2.1 Cycle ageing data

The generation of cycling data from the beginning to the end of a battery's life requires a significant investment of time and resources spanning many months or years. Experiments are run to investigate the influence of in-cycle factors (charging current, discharging current, temperature and DOD) on the capacity retention and (sometimes) rise in the internal resistance of batteries. Typically, cycle ageing datasets include in-cycle measurements of current, voltage and temperature, and per-cycle measurements of capacity and IR or impedance.

Models are then developed according to the recorded cycling dataset to, among other things, predict future capacity retention, internal resistance growth and other health metrics. An overview of the typical recorded data and modelling pipeline for cycling (in particular, high-throughput) degradation datasets is illustrated in Fig.9.2.

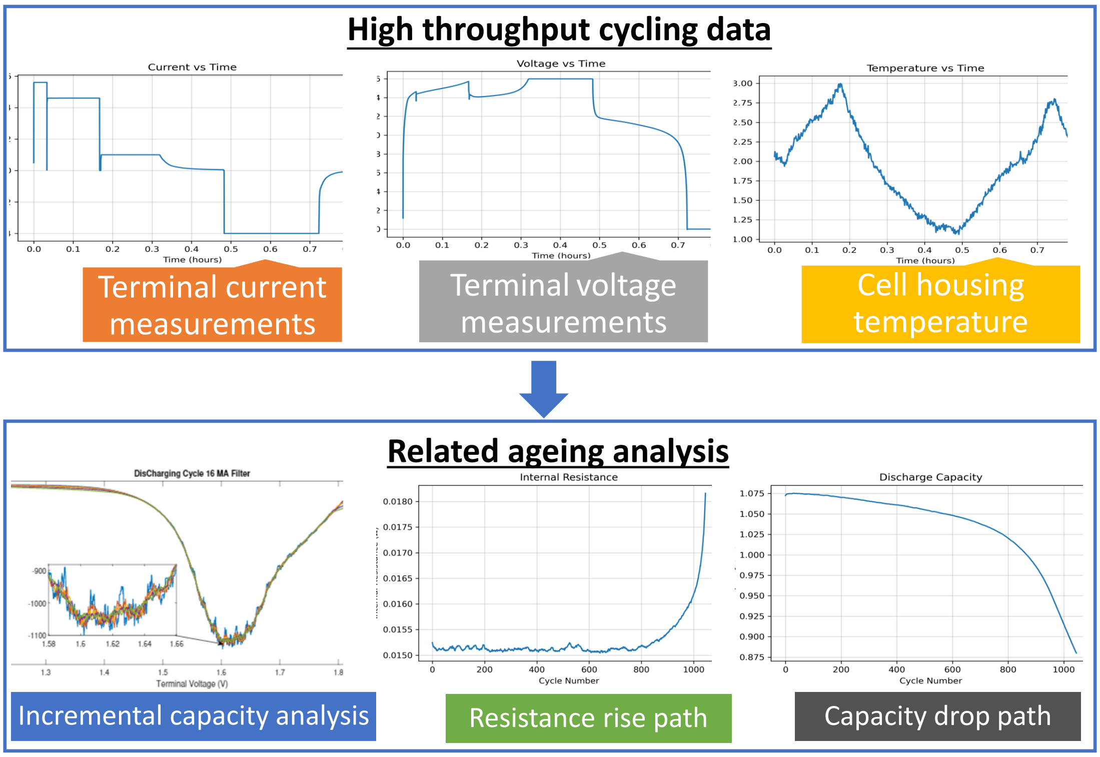


Fig. 9.2: The typical plots of a high-throughput cycling dataset encompassing measured terminal current, voltage and temperature variations. Capacity, IR, voltage and temperature can then be used for the ageing analysis.

We prioritise in this section datasets with multiple cells, frequent in-cycle measurements and labs with multiple datasets. Smaller datasets (with only a few cells) and datasets without any in-cycle measurements are left to the end of this section (section 9.2.1). The reader is invited to consult Table 9.2, at the end of the section, for an overview of the datasets discussed here.

National Aeronautics and Space Administration

NASA hosts two high-throughput battery datasets on their website [163, URL] totalling 62 cells. We provide here a brief description of the datasets, for a full cell-by-cell experimental description see the 'ReadMe' file accompanying the datasets.

The first of these datasets 'Battery Data Set' [185] contains data for 34 Li-ion 18650 cells with a nominal capacity of 2Ah (we were unable to confirm the chemistry of these cells). This dataset was also the first publicly available battery dataset and has had a profound impact on the field; Table 9.1 summarises representative research work drawing on this dataset, giving a glimpse at its influence. Cells were cycled in a range of ambient temperatures (4 °C, 24°C, 43°C), charged with a common CC-CV protocol and with different discharging regimes. The dataset includes in-cycle measurements of terminal current, voltage and cell temperature, and cycle-to-cycle measurements of discharge capacity and EIS impedance readings. The dataset is provided in '.mat' format under a double-attribution license[1]. The experiments were ended when cell capacity fell below 30% or 20% of nominal capacity.

The second dataset hosted by NASA, the 'Randomised Battery Usage Data Set' [23], contains data for 28 lithium cobalt oxide (LCO) 18650 cells with a nominal capacity of ∼2.2Ah. The cells in this dataset were continuously operated. The dataset consists of 7 groups of 4 cells each group cycled at a set ambient temperature (room temp, 40°C); for 5 of these groups the cells were CC-charged to a fixed voltage and then discharged with currents selected at random from the group's discharge distribution table (7 different regimes). The other two groups were randomly charged and discharged. The dataset includes in-cycle measurements of terminal current, voltage and cell temperature, and measurements of discharging capacity and EIS impedance readings at 50 cycle intervals. The dataset is provided in a '.mat' format and measurements appear to have been taken until the cells reached between 80% to 50% SOH.

Category	SOH estimation and RUL prediction	Health prognostics and fault diagnostics	Battery modelling	Algorithms introduction and comparison
Ref	[24, 101, 146, 159, 186, 187, 249]	[73, 124, 213]	[132, 234]	[192]

Table 9.1: NASA dataset repository: Related papers and the corresponding research conducted. (See additionally Supplementary material Table 3 for full details.)

Centre for Advanced Life Cycle Engineering

The Centre for Advanced Life Cycle Engineering (CALCE) battery group has carried out substantive cycling tests for a diverse range of LCO/graphite cells. These datasets are hosted on their website [31, URL] – publications using the data should cite the corresponding CALCE article(s). Data is grouped by cell specification and not all data for a given specification comes from the same publication. We provide here a brief description of the datasets, for a full experimental description see the description on the website and the associated papers.

[1] As per the NASA description: 'Publications making use of databases obtained from this [the NASA] repository are requested to acknowledge both the assistance received by using this repository and the donors of the data.'

CALCE hosts data for 15 LCO prismatic CS2 cells grouped by experimental conditions (and publication) into ‘Type-1’ to ‘Type-6’. ‘Type-1’ and ‘Type-2’ accompany one paper [94] and ‘Type-3’ to ‘Type-6’ another [232]. Type-1’ consists of four 0.9Ah cells, ‘Type-2’ of four 1.1Ah cells and ‘Type-3’ to ‘Type-6’ each contain between one and two 1.1Ah cells. The cells appear to have been cycled at room temperature (23°C) and the experiments investigate different depths and ranges of partial charge and discharging, with a variety of C-rates. The dataset provides the cell cycler logs in Excel or ‘.txt’ format containing measurements of current, voltage, discharge/charge capacity and energy, internal resistance and impedance. For each cell there are multiple files each containing the data for multiple cycles; the files are named according to the date at which they were recorded and, in our opinion, a significant amount of pre-processing is required to use this dataset. The data was recorded until batteries had (at least) passed their end of life (EOL), 80% SOH, with less than 200 cycles of data for the ‘Type-1’ batteries and approximately 800 cycles for the other cells.

The second set of cells tested by CALCE are 12 LCO prismatic CX2 cells with a rated capacity of 1.35Ah. Which, similarly to the CS2 cells, are grouped into ‘Type-1’ to ‘Type-6’. ‘Type-1’ and ‘Type-2’ (four cells each) were cycled in the same way as ‘Type-1’ of the CS2 cells [235]. The other four groups each have a single cell cycled with a variety of charge/discharge protocols; one of the cells was cycled at a range of temperatures (25°C, 35°C, 45°C, 55°C). The datasets are provided in the same format as the CS2 data with the same measurements.

In subsequent battery experiments [193], the group examined the influence of different depths of discharge (DOD) and discharging current stresses on the ageing of pouch cells: testing 16 LCO 1.5Ah pouch cells in a ‘semi-temperature controlled’ room (25$\pm$2°C) [193]. The dataset is grouped by DOD and discharging protocol, provided in ‘.mat’ format, containing cycler voltage, current and charge/discharge capacity data for between 400 and 800 ‘equivalent cycles’.

Toyota Research Institute in partnership with MIT and Stanford

In partnership with MIT and Stanford, the Toyota Research Institute (TRI) has published two substantial and easy to use high-throughput cycling datasets. Combined, these datasets contain data for 357 $(= 124 + 233)$ commercial LFP/graphite cells manufactured by A123 Systems (APR18650M1A) with a rated capacity of 1.1Ah. These two datasets are hosted online [216, URL], with accompanying experimental descriptions, under ‘CC BY 4.0’[2]. The

[2] To avoid confusion with Constant Current (CC), we add quotation marks when referring to a Creative Commons License.

datasets are provided in '.csv', MATLAB struct and (second dataset only) JSON struct formats and a link to a GitHub repository with initial scripts is provided with the data. We point to the file structure of these datasets as a reference for future work: organised by cell → cycle → recorded data. Papers utilising the data should cite the appropriate publication.

The first of these datasets [197] (124 cells) was designed to explore the influence of fast charging protocols on cell ageing. Each cell was cycled with one from a range (72 different profiles) of one or two step fast charging protocols and a common CC-discharge protocol. The cells were cycled in a temperature controlled environment (30°C). Data was logged from cycle 2 until a cell reached its EOL (80% SOH) – between 150 to 2300 cycles. The dataset contains in-cycle measurements of temperature, current, voltage, charge and discharge capacity, as well as per-cycle measurements of capacity, internal resistance and charge time. The data is split into three batches corresponding to three blocks of experiments carried out separately. In the accompanying paper [197] a feature based model is built on data from the first 100 cycles to predict the EOL. Since the dataset's release, numerous other papers have been published working with this data.

The second of these datasets [6] (233 cells) builds on the first: designing an approach to quickly optimise fast charging protocols. Again, cells were cycled in a temperature controlled environment (30°C) with a common discharging protocol. The dataset is split into five batches of between 45 and 48 cells each; these batches were tested sequentially: for the first batch one of 224 different six-step charging protocols was chosen at random for each cell, the cells were tested for 100 cycles and then a model (trained on previously collected data) was used to predict the EOL based on this data. This prediction was used to inform the selection of charging protocols for the next batch of cells. This was repeated with the first four batches; the final batch was then tested until past the EOL comparing the selected optimal charging protocols with several other protocols. The dataset contains the same readings as the first dataset of 124 cells [197] except for the exclusion of IR readings. An attempt has been made to recover this missing data [210] where the IR has been predicted with a CNN model trained on the first dataset; this predicted IR data can be found online [211, URL].

Sandia National Lab

The Sandia National Lab has performed testing for three chemistries of 18650 form cells: 'LFP from A123 Systems (APR18650M1A, 1.1Ah), NCA from Panasonic (NCR18650B, 3.2Ah), and NMC from LG Chem (18650HG2, 3Ah)' [161]. In total there are 86 cells (30 LFP, 24 NCA and 32 NMC). The data from this study has been made available on the Battery Archive website [189, URL] – see Section 9.3.1 below. The data is shared under a double attribution license and on the website is denoted by the 'SNL' keyword. The experimental description is available on the Battery Archive page and in the relevant publication [161].

The cells were cycled at a range of temperatures (15°C, 25°C and 35°C) with different DODs (0-100%, 20-80% and 40-60%) and discharge currents (0.5C, 1C, 2C and 3C); at least 2 cells from each chemistry were cycled in each combination of temperature, DOD and discharge current (12 groups) apart from the 3C discharge for the NCA cells. All cells were charged with a fixed rate of 0.5C. The cells were cycled until reaching their EOL (80% SOH) – at the time of publication cycling was still ongoing. The dataset contains in-cycle measurements of current, voltage, temperature and energy (Wh), and per-cycle measurements of charged/discharged capacity (bottom to top of DOD range) and other summary statistics. Periodically (roughly every 3% capacity loss), EIS measurements were taken measuring the full capacity of the cell. The data is provided in '.csv' format

Battery Intelligence Lab at the University of Oxford

The Battery Intelligence Lab at the University of Oxford hosts several battery degradation datasets on their homepage [104, URL]. We review here the 'Path dependence battery degradation dataset' [171] which is made of three parts. The files are provided under '.mat' format and all are licensed under Open Data Commons' ODbL v1.0 & DbCL v1.0 license. The dataset parts can be found as follows: Part 1 [104, URL] or [168, URL]; Part 2 [169, URL]; and Part 3 [170, URL].

The 3-year long project [171], spanning 2017-2020, studied ageing 'path dependence' of Li-ion cells by subjecting them to combined load profiles comprising fixed periods of calendar and cyclic ageing. The path dependence phenomena reflects the ageing sensitivity of cells to the order and periodicity of calendar ageing and cyclic ageing. The study analysed 28 commercial 3Ah 18650 NCA/graphite cells (NCR18650BD). The dataset is provided in 3 parts (Part 1, 2 & 3) with the 28 cells split among ten groups (9 groups of 3 cells; 1 group of 1 cell), all tested at 24°C. We provide a small breakdown for reference and point to the informative 'ReadMe' files. The data provided includes time, current, voltage, capacity and temperature, and the RPT and EIS testing data.

Group 1-4, 3 cells per group, were aged through cycling at a low C rate (C/2 and C/4) followed by 5 or 10 days of calendar ageing with RPTs run every 48 cycles. The first 18-months of experimental data is presented in 'Part 1' with months 19-36 presented in 'Part 2'. Additional to cell Groups 1-4, in Part 2 one finds Group 5 & 6 as control experiments. The cells of Group 5 are exposed to continuous C/2 cycling while Group 6 is exposed only to calendar ageing (at 90% SOC). Group 7-10 are presented in the dataset's 'Part 3' and parallels Group 1-4. Here, each group is cycled with CC-CV profiles then 5 or 10 days of calendar ageing. Reference performance tests (RPT) and EIS tests are used periodically to characterise the cells to differentiate the influence of different storage times and C-rates on battery degradation.

Hawaii Natural Energy Institute

Researchers from the Hawaii Natural Energy Institute (HNEI) investigated the variability of cell degradation across 51 cells through cycling [51]. Data for 15 of these cells is shared on the Battery Archive website [10, URL] (denoted by 'HNEI' dataset). These 15 cells are commercial 2.8Ah NMC-LCO/graphite 18650 cells (LG Chem, model 'ICR18650 C2'). The cells were cycled with fixed 1.5C discharge and C/2 charge protocols at 25°C for $\sim$1000 cycles. The dataset contains in-cycle measurements of current, voltage and charged/discharged capacity and energy, and per cycle measurements of charge/discharge capacity. Roughly every 100 cycles RPTs were run which are also present in the data. Files are in '.csv' format and shared under 'CC BY 4.0' plus 'source attribution' to Battery Archive. Additional experimental details and cell summary statistics (e.g. initial cell weight and received SOC) can be found in the accompanying paper [51].

EVERLASTING project

The recent European Commission funded project 'Electric Vehicle Enhanced Range, Lifetime And Safety Through INGenious battery management' (EVERLASTING) [77, URL] has published several battery related datasets on the '4TU.ResearchData' website [102, URL]. Of particular interest are the three datasets connected with the technical report produced by the project [218]. The report explores ageing from three different angles: drive cycle, calendar and CC-CV ageing at a range of temperatures; the datasets are described in the relevant sections of our paper.

Of these datasets, one experiment 'Lifecycle ageing' was carried out to investigate the interactions between temperature, charge/discharge C-rates and capacity loss. These experiments were performed on 28 Li-ion 18650 3.5Ah commercial cells for a range of temperatures (0°C, 10°C, 25°C and 45°C), discharge C-rates (0.5C, 3C) and charge C-rates (0.5C, 1C). Two cells were tested at each possible pairing of temperature/charge-rate and temperature/discharge rate (except for 0°C discharge). All 'charge' ('discharge') experiments had a common discharge (charge) profile. The data is hosted separately grouped by temperature (0°C and 10°C) [90, URL] and (25°C and 45°C) [219, URL]. The provided data is in '.csv' format with cycler logs (including voltage, current, charge/discharge capacity and energy) from characteristic cycles run roughly every two months – it is unclear if the data is complete.

Others

The Karlsruhe Institute of Technology (KIT) provides cycling data for 4 battery packs each consisting of 11 NMC/graphite 40Ah cells on their website [206, URL] (under 'CC BY 4.0'). The batteries were cycled, at room temperature, in series with a range of charge/discharge profiles (detailed in the relevant paper [205]). The dataset provides high frequency (cell-by-cell and battery wise) measurements of voltage, temperature and inverter current/voltage for each of the tested charge/discharge profiles. The dataset is provided in well structured folders with '.csv' files and a starter MATLAB script.

Provided on the University College London (UCL) data website [96, URL] is cycling data for a single 3.5Ah LG Chem NCA INR18650 MJ1 cell, given under 'CC0 1.0'. The cell was cycled according to the manufacturers recommendations in a fixed ambient temperature (24°C) for 400 cycles [97]. The dataset provides in-cycle measurements of temperature, voltage and capacity, and per-cycle measurements of charge/discharge capacity, given in '.csv' format.

Berkeley provides data from a single Sanyo 18650 3.7V 2.6Ah LCO/graphite cell on the Dryad Data website [92, URL] (under 'CC BY 4.0'). The cell was cycled with a variety of non-standard fast charging protocols. The dataset contains in-cycle measurements of voltage, current, temperature and charge/discharge capacity for 46 consecutive cycles and is provided in '.csv' format.

Researchers from Xi'an Jiaotong University [244, 245] deploy the Coulomb counting method in combination with data-driven techniques to propose methods for SOC calibration and estimation. Both works use the same cycling data for battery cells under a regime of *fast capacity degradation*. The cells full physical description is found in the relevant paper [245, Table 1]. To summarise, two lithium-ion pouch cells with chemistry NMC/graphite and nominal capacity of 27Ah were cycled from new until reaching 80% capacity. The cells were cycled with a CC-CV charge and CC discharge followed by a 30min relaxation period between cycles; the chamber's temperature was fixed at 40°C and a total of about 400 cycles is recorded. The full cycling data and description file can be found in the paper's supplementary material [245] and it is unclear under which sharing license it is offered. However, provided separately [243, URL] under 'CC0 1.0' are the first 100 cycles of data (in '.xlsx' format). The experimental data recorded is: battery voltage, current, charging/discharging capacity and energy.

Diao et al. [54] provide a dataset considering the influence of ambient temperature, discharging current stress and cut-off points of charging current for CC-CV cell ageing. This dataset is hosted on the Mendeley data platform [52, URL] and shared under 'CC BY-NC 3.0'. In the experiment, 192 LCO/graphite pouch-type 3.36Ah cells were tested using the above three stress factors. The dataset contains capacity measurements taken at 50 cycle intervals, is given in '.mat' format and only 182 of the 192 cells appear to be listed.

Researchers at Poznan University of Technology provide data for 28 Samsung NMC/carbon 2.6Ah 18650 cells on the Mendeley data platform [30, URL] (under 'CC BY 4.0'). The cells were cycled at a variety of temperatures, DODs and charging/discharging currents until reaching 80% SOH. The dataset consists of 'learning data' from 28 cells containing summary measurements of ambient temperature, discharging current, DOD, average charging current and number of equivalent cycles for cells at a range of SOH values (9 measurements from 100% to 80% SOH), given in '.xlsx' format. This data was used in the paper [29] to train several models to predict the cell's current SOH.

Lastly, we mention data, shared by researchers at the University of Oviedo under 'CC BY 4.0', for two LFP pouch cells [68, URL]. The cells were tested at room temperature (23°C) for a single full charge/discharge cycle at a constant current rate of C/25 [67]. The dataset contains voltage, current and temperature readings, from the charge/discharge cycle, sampled every 2s for a total experimental time of 60 hours (details can be found in the associated paper [67, Section 5]).

Location with weblink	Paper Ref	Cell (form size chemistry)	Test variables	Data given	No. of cells
NASA [163, URL]	[185]	18650 2Ah (?)	Dhrg, T	Q, IR, V, I, T	34
	[23]	18650 2.2Ah LCO	Chrg, Dhrg, T	Q, IR, V, I, T	28
CALCE [31, URL]	[94, 235]	prismatic 1.1Ah LCO	Chrg, Dhrg	Q, IR, E, V, I, T	15
	[94, 235]	prismatic 1.35Ah LCO	Chrg, Dhrg, T	Q, IR, E, V, I, T	12
	[193]	pouch 1.5Ah LCO	Chrg, DOD	Q, V, I	16
TRI [216, URL]	[197]	18650 1.1Ah LFP/gr	Chrg	Q, IR, V, I, T	124
	[6]		Chrg	Q, V, I, T	233
Sandia [189, URL]	[161]	18650 multiple	Dhrg, DOD, T	Q, E, V, I, T	86
Oxford [104, URL]	[171]	18650 3Ah NCA/gr	Chrg, Cal	Q, E, V, I, T	28
HNEI [10, URL]	[51]	18650 2.8Ah NMC-LCO/gr	–	Q, E, V, I	15
EVERLASTING [90, URL] [219, URL]	[218]	18650 3.5Ah NCA/gr	Chrg, Dhrg, T	Q, E, V, I	28
KIT [206, URL]	[205]	— 40Ah NMC/gr	Chrg, Dhrg	V, I, T	44
UCL [96, URL]	[97]	18650 3.5Ah NCA/gr	–	Q, V, T	1
Berkeley [92, URL]	–	18650 2.6Ah LCO/gr	Chrg	Q, V, I, T	1
Xi'an Jiaotong [243, URL]	[244, 245]	pouch 27Ah NMC/gr	–	Q, E, V, I	2
Diao et al.[52, URL]	[54]	pouch 3.36Ah LCO/gr	Chrg, Dhrg, T	Q	192
Poznan [30, URL]	[29]	18650 2.6Ah NMC/carbon	Chrg, Dhrg, DOD, T	Q, I, T	28

Table 9.2: Overview of cycle ageing datasets. 'gr' stands for 'graphite', 'Cal' denotes calendar ageing, 'Chrg' charge protocol and 'Dhrg' discharge, 'E' denotes 'energy'. Here, we use 'IR' to denote both internal resistance and impedance. No 'test variables' indicates that all cells in the experiment were cycled in the same way.

9.2.2 Drive cycle data

Energy is required to propel an automobile. With a conventional internal combustion engine the combustion of fossil fuels, converted to mechanical energy, drives the vehicle forward. However, with global concern surrounding greenhouse gas levels there is an urgent push for the automobile industry to reduce carbon emissions. For this reason, standardised testing procedures capturing the dynamic power demands of driving are indispensable: allowing the relative efficiency and performance of engines to be compared. These standard test procedures are referred to as driving cycles.

A driving cycle is a standardised dynamic vehicle drive schedule encoded by a velocity-time table/profile. The velocity and acceleration are pre-scheduled per time step, and thus the required mechanical power is a function of time. The integral of mechanical power over the duration of the driving schedule represents the total energy required for a specific driving cycle. For electric vehicles the battery system generates this required mechanical energy. Datasets collected by cycling batteries according to the drive schedules can be used to compare the efficiency of EVs with traditional vehicles and to test the performance of derived battery models and SOC estimation algorithms under realistic conditions.

The globally recognised driving cycle tables can be divided into three groups: European driving cycles, US driving cycles and Asian (Japanese, Chinese -Beijing) driving cycles [27, 74]. For example, the Urban Dynamometer Driving Schedule (UDDS) [44] is commonly used for 'city-based EV driving cycle tests' representing light-duty city driving conditions. US06 represents an aggressive driving cycle with high engine loads. The european drive cycle ARTEMIS [3] contains 12 driving cycles that range across several driving conditions: congested urban, free-flow urban, secondary roads, main roads and motorways. The Highway Fuel Economy Driving Schedule (HWFET) is used to describe cars cruising under 60mph on a highway. And, the Air Resources Board LA92 dynamometer driving profile was developed to depict a driving cycle with higher top and average speed, lower idle time, fewer stops per mile and a higher maximum rate of acceleration when compared with UDDS.

An overview of driving cycle data reviewed in this section can be found in Table 9.5.

University of Wisconsin-Madison & McMaster University

The battery research group at the University of Wisconsin-Madison offers a battery testing dataset covering four typical driving cycles: US06, HWFET, UDDS and LA92. The dataset, published on the Mendeley data website [116, URL] (under 'CC BY 4.0'), contains data from a single 2.9Ah NCA Panasonic 18650PF cell. The cell was cycled according to the above driving cycles and an additional 'neural network driving cycle' systematically through a range of temperatures (25°C, 10°C, 0°C, -10°C, and -20°C, in that order). A full experimental description can be found in the accompanying 'ReadMe' file. The dataset includes characterisation data from Hybrid Power Pulse Characterisation (HPPC) and EIS tests, and in-cycle measurements from the driving cycles including voltage, current, capacity, energy and temperature. The data is presented in '.mat' and '.csv' files with a well structured format sorted by temperature, test type and drive cycle.

The same group, but operating at McMaster University, provides another driving schedule test dataset for a series of battery tests carried out for a single 3Ah LG Chem INR18650HG2 NMC cell [114, URL]. The cell was cycled at six different ambient temperatures (40°C, 25°C, 10°C, 0°C, -10°C, and -20°C) according to the same mix of drive cycles as the Panasonic cell. The dataset contains the same data as for the Panasonic cell (in a similar format) with the addition of 'prepared data' which has been processed in order to train and test a provided SOC estimator.

The above two driving cycle datasets, hosted by University of Wisconsin-Madison and McMaster University, provide a benchmark for driving cycle tests and are at the heart of crucial contributions in the development of SOC estimation algorithms and battery models; some of these works are reviewed in Table 9.3 and Table 9.4.

Category	Ref	Detail
SOC estimation	[38]	This paper introduces a data-driven approach for State of Charge (SOC) estimation of Li-ion batteries using a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM).
	[15]	This paper proposed a stacked bidirectional LSTM neural network for SOC estimation of lithium-ion batteries.
Battery modelling	[247]	In this paper, the potential of applying advanced machine learning techniques to model lithium-ion batteries is explored. Rather than using the more common ECM and physics-based models, a data-driven approach is used to build battery models.
	[115]	In this paper, ECMs of lithium-ion batteries are built to capture various the electrochemical properties of the battery. The ECMs are validated by a series of five automotive drive cycles performed at temperatures ranging from -20°C to 25°C.

Table 9.3: Panasonic 18650PF Li-ion Battery Data: Related paper and the corresponding research conducted.

Category	Ref	Detail
SOC estimation	[225]	This paper introduces a data-driven approach for State of Charge (SOC) estimation of Li-ion batteries using a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM).
	[224]	This paper proposed a stacked bidirectional LSTM neural network for SOC estimation of lithium-ion batteries.

Table 9.4: LG 18650HG2 Li-ion Battery Data: Related paper and the corresponding research conducted.

Others

Researchers from the University of Science and Technology of China (USTC) have explored the co-estimation of model parameters and SOC for batteries and ultra-capacitors [230]. The data accompanying this research has been shared in the journal 'Data in Brief' [229, URL] along with an experimental description, under 'CC BY 4.0'. A lithium battery pack (LFP-1665130-10Ah, produced by Fujian Brother Electric CO., LTD of China – 4 prismatic cells in series) and an ultra-capacitor (BCAP3000 P270 2.7V/3.0Wh, produced by Maxwell

Technologies, Inc.) were each cycled once according to two different driving cycles (DST and UDDS) at room temperature. The dataset, provided in '.xlsx' format, contains per second measurements of current and voltage for the battery and ultracapacitor during the two drive cycle profiles.

The EVERLASTING project provides two drive cycling datasets both shared under 'CC BY-NC 4.0'. The first of these datasets [48, URL] contains data for two battery modules each built from 16 NCA/graphite 3.5Ah LG Chem INR18650 MJ1 cells. The modules were cycled at a variety of temperatures according to an 'adapted real driving profile'. The dataset contains in-cycle measurements of pack voltage, current, charge/discharge capacity, ambient temperature and per-cell temperature. The second of these datasets, described in the EVERLASTING report [218], contains data for 16 NCA 3.5Ah LG Chem INR18650 MJ1 cells cycled according to a recorded city drive profile for two DOD ranges (70-90% and 10-90%) and at a variety of temperatures (0°C, 10°C, 25°C and 45°C) – 2 cells per combination. In addition, 2 cells were cycled according to a recorded highway drive profile at 25°C (10-90% DOD). This dataset is stored in two locations according to temperature: (10°C and 0°C) [90, URL] and (45°C and 25°C) [219, URL]. The datasets are both in '.csv' format but with different information depending on the temperature. The cells cycled at 25°C and 45°C include measurements of voltage, current and charge/discharge capacity and energy; whereas, the data for the cells at 0°C and 10°C has a different file structure and additionally includes temperature readings.

The Oxford Battery Intelligence Laboratory provides the 'Battery Degradation Dataset 1' [18] on their website [104, URL], licensed under Open Data Commons' ODbL & DbCL. This dataset contains data for eight 740mAh lithium-ion pouch cells manufactured by Kokam (part number SLPB533459H4). The cells were cycled at a constant ambient temperature (40°C) using a CC-CV charging regime and the ARTEMIS [3] driving cycle discharging profiles until their EOL (80% SOH). The dataset is provided in '.mat' format containing voltage, temperature and discharge capacity (mAh) measurements. These measurements (taken at 10 minute intervals) were recorded during characterisation tests performed every 100 cycles. A full experimental description can be found in the PhD thesis of C. Birkl [18, Chapter 5.2].

The Institute for Power Electronics and Electrical Drives at Aachen University hosts drive cycling data for 28 Samsung 18650 NCA/carbon+silicon cells with a nominal capacity of 3.4Ah on their website [110, URL] (under 'CC-BY-4.0'). The cells were cycled at a fixed ambient temperature (25°C) with a CC-CV charging regime and a recorded drive cycling discharge profile. The dataset contains in-cycle measurements of voltage, current and temperature, and checkup tests every 30 cycles with capacity, quasi open-circuit voltage (OCV) and pulse tests (at 80%, 50% and 20% SOC). The dataset is provided in '.csv' format and a detailed experimental description can be found in the accompanying 'MetaData' file.

For completeness, we mention a dataset available to 'IEEE DataPort'[3] subscribers [133, URL], under 'CC BY 4.0', containing data from simulated driving cycles composed according to the Federal Test Procedure repository. We point the reader to the data description given on their website.

Location with URL	Cell (form size chemistry)	Test variables	Data recorded	No. of cells
Madison [116, URL]	18650 2.9Ah NCA	cycle, T	Q,V, I, E, T, EIS	1
McMaster [114, URL]	18650 3Ah NMC	cycle, T	Q,V, I, E, T, EIS	1
USTC [229, URL]	prismatic 10Ah LFP	cycle	V, I	1
EVERLASTING	18650 3.5Ah NCA/gr			
[48, URL]	16 cell modules	T	Q,V, I, T	2
[90, URL] [219, URL]	18650 3.5Ah NCA/gr	cycle, DOD, T	Q,V, I, E, T	18
Oxford [104, URL]	pouch 0.74Ah —	–	Q,V, T	8
Aachen [110, URL]	18650 3.4Ah NCA/C+Si	–	Q,OCV, V, I, T	28

Table 9.5: Overview of Driving cycle data. 'cycle' here denotes the use of different drive cycle profiles, 'E' denotes 'energy'. No 'test variables' indicates that all cells in the experiment were cycled in the same way.

9.2.3 Characterisation data for cell modelling

The cycling performance of different lithium battery chemistries is varying and highly dependent on operating conditions (temperature, current load, age). To evaluate the viability of lithium batteries to a given application, features of battery performance such as the OCV-SOC table, impedance and IR are necessary. These cycling features can then be used to model the electrical dynamics and cycling performance of a battery. The experimental data collected for this purpose mainly targets the short-term responses of current and voltage, and focuses on the impedance variance at different battery SOC levels and temperatures.

In-cycle battery data

The CALCE battery group has piloted the research in terms of battery modelling and internal state estimation providing a series of baseline in-cycle datasets for cells from different types of lithium batteries [31, URL] (under 'attribution' license). Two experiments, namely, low-current OCV and incremental-current OCV, have been deployed to collect OCV for commercial INR 18650-20R 2Ah NMC/graphite cells. The OCV dataset includes different OCV-SOC tables achieved at three ambient temperatures (0°C, 25°C and 45°C). Voltage responses under different dynamic current profiles, such as, the Dynamic Stress Test (DST), Federal Urban Driving Schedule (FUDS), US06 Highway Driving Schedule and Beijing

[3] This dataset is not 'public' but we are aware that many readers may have IEEE memberships. We have not verified the contents of this dataset.

Dynamic Stress Test, are provided to test the accuracy of the proposed SOC estimation algorithms [95, 236] and analyse the dependence of SOC estimation on OCV variations due to temperature [248]. Since the OCV-SOC table is temperature-sensitive, further investigation has been conducted by CALCE battery group providing a temperature-dependent OCV-SOC dataset for 2.23Ah A123 LFP/graphite cells. The temperature-dependent OCV-SOC dataset is collected from low current OCV tests for a wide range of temperatures spanning from -10°C to 50°C with an interval of 10°C. Additionally, the experimental data of DST and FUDS tests performed in the corresponding ambient temperatures is available. All data is given in '.xlsx' format and provided is the data from the OCV tests and in-cycle data from the drive cycles (including voltage, current, charge/discharge capacity and energy, IR and impedance).

In order to develop an advanced model which reproduces the thermal and electrical dynamics of the battery, Planella et al. [160] at Warwick University tested the cycling behaviours of commercial 5Ah LG INR21700 M50 NMC cells with a range of ambient temperatures (0°C, 10°C and 25°C) and C-rates (0.1C, 0.5C, 1C and 2C). In their experiments, four cells are tested per specific C-rate and temperature. The dataset, along with experimental description and additional scripts, is hosted on Github [79, URL] (under BSD 3-Clause License). The data is the cycler logs given in '.csv' format containing voltage, current, capacity, energy and temperature readings from cycles run to compare the derived model with real data.

Few works have been conducted to test the discharging power behaviour of cells. One publicly available dataset [141, URL], under 'CC BY-NC 3.0', has investigated the behaviour of 4 types of 18650 Li-ion cells, produced by 4 different cell manufacturers (LG 18650-HB6 1.5Ah NMC, Panasonic NCR18650B 3.35Ah NCA , Shenzhen IFR18650 1.5Ah LFP and Efest IMR18650 3.1Ah LMO – between 8 and 13 cells per manufacturer), at different constant power discharge rates [142] and a constant ambient temperature (25°C). In particular, the experiments were designed to capture the available power response of cells at high (out of specification) current loads. The provided data, given in '.csv' and '.mat' format, appears to contain cycler logs for each cell (spanning $\sim$6 hours) with voltage, current, temperature, charge/discharge capacity and power measurements, however, no column headings or 'ReadMe' file are given.

Impedance Spectroscopy data

Applying electrochemical impedance spectroscopy (EIS) to measure the impedance of lithium batteries is widely accepted in battery research [223]. EIS can separate the dependence of different components by varying the frequency of applied AC currents. Allowing the contribution of the solution resistance, charge transfer/polarisation resistance, double layer capacity, wire inductance etc., to be interpreted from the measured responses. The EIS provides a tool to understand and model the complicated non-linear electrochemical process occurring inside a battery. A typical characterisation process for a lithium battery, using EIS measurements according to the frequency domain analysis and modelling, can be found [80]; the frequency setting of EIS inputs are standard for most systems: ranging from 20mHz to 10kHz. In general, high-frequency EIS responses are considered indicative of inductive behaviour and low-frequency responses indicative of capacitive behaviour.

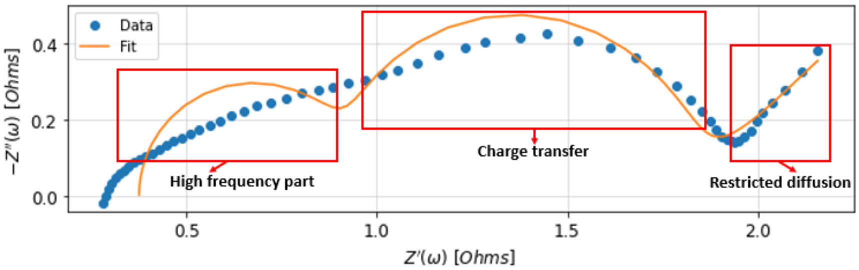


Fig. 9.3: A typical Nyquist plot: Battery characterisation using EIS measurements

As given in Fig. 9.3, the Nyquist representation of an impedance spectrum (acquired from the lithium battery test) is used to fit an equivalent circuit model (ECM) – an ECM provides a simplified battery model as a circuit of standard components whose parameters are fitted to approximately replicate the measured response. The lithium Nyquist plot can, in general, be divided into three parts based on the frequency responses. In the high-frequency segment the inductive behaviour of wires is dominant, contributed to by the cables impedance, fittings, connectors, cell tabs and current collectors. Over the middle-frequency segment the shape of the Nyquist plot behaves like a depressed semi-circle, representing the charge transfer resistance and double-layer capacitance, its shape affected by the temperature. For the low-frequency segment the slower ion-diffusion process dominates the cell dynamics and measuring the resistance response here takes longer to perform. In addition, this process can be influenced by many factors, such as electrode material, porosity, operating temperatures, SOC and voltages. Typically, cell capacitance has a very steep slope (around 90 degrees). As illustrated in Fig.9.3, at points the frequency responses behave closely to those of a capacitor.

One of the largest broad-scale datasets of EIS measurements has been shared on the ‘zenodo’ platform [246, URL] (under ‘CC BY 4.0’) containing over 20,000 EIS readings collected from 12 Eunicell LR2032 45mAh LCO/graphite cells. The cells were cycled at a range of different temperatures (25°C, 35°C and 45°C) with multiple frequencies of EIS measurements taken at different SOC levels. The experiment was stopped when cells reached their EOL (80% SOH). The dataset was recorded to accompany research exploring the prediction of RUL and SOH from EIS data [246]. The provided data contains the EIS measurements (resistance, impedance and phase at a range of frequencies) and, separately, measurements of capacity in ‘.txt’ format.

In Section 9.2.7 below, we refer to another EIS dataset [156] available upon request.

9.2.4 Calendar ageing

Calendar ageing 'comprises all ageing processes that lead to the degradation of a battery cell independent of charge-discharge cycling' [111]. Such ageing is most pronounced in applications where periods of idleness are longer than operation, such as with electric vehicles. It is argued that calendar ageing may also play a role in cycle ageing studies where cycle depths and current rates are low [200]. In this section we overview datasets dedicated to calendar ageing see Table 9.6 for an overview of the datasets.

Outside the battery cycling data, the CALCE group has also studied calendar ageing and a dataset appears on their website [31, URL] ('Pouch Cells: Storage Data and Test Description'): 144 LCO/graphite 1.5Ah pouch-type cells with three different initial SOC values (0%, 50% and 100%) were calendar aged at four different storage temperatures (-40°C, -5°C, 25°C and 50°C). There are three testing groups, 48 cells per group, with capacity and impedance measurements taken every 3 weeks, 3 months and 6 months, respectively. The dataset is provided in '.xls' format and contains the cycler data (current, voltage, charge/discharge capacity and energy, internal resistance and impedance) from the periodic characterisation cycles.

Group 6 in 'Part 2' [212, URL] of the Oxford 'Path dependence battery degradation dataset' [171] contains the data of one single cell exposed to continuous calendar ageing at 90% SOC. We do not provide further details here and refer the reader to the description given already in Section 9.2.1.

As part of the EVERLASTING project [77] (see Section 9.2.1) calendar ageing was performed on several NCA/graphite 18650 3.5Ah LG Chem cells (model INR18650 MJ1). The testing was carried out for a range of temperatures (0°C, 10°C, 25°C and 45°C) and the cells were stored at OCV with different initial SOC levels (10%, 70% and 90%). The data shared online does not appear to be complete; however, data for 2 cells stored at 25°C [217, URL], 3 cells stored at 0°C and 3 cells stored at 10°C [89, URL] is available. The provided data is in '.csv' format, shared under license 'CC BY-NC 4.0', and contains cycler data (voltage, current, capacity and energy) from characterisation tests performed periodically.

Lastly, we refer the reader to Section 9.2.7 regarding 'Data on demand'. We mention the dataset made available by Dr Dhammika Widanalage (Warwick Manufacturing Group, Warwick University (UK)) which contains many cells tested under calendar ageing.

Location with URL	Cell (form size chemistry)	Test variables	Data recorded	No. of cells
CALCE [31, URL]	pouch 1.5Ah LCO/gr	SOC, T, time	Q, IR, E, V, I	144
Oxford [212, URL]	18650 3Ah NCA/gr	–	Q, E, V, I, T	1
EVERLASTING [217, URL] [89, URL]	18650 3.5Ah NCA/gr	SOC, T	Q, E, V, I	8

Table 9.6: Overview of Calendar Ageing degradation data. Here, 'time' denotes the frequency at which ageing was interrupted to take measurements.

9.2.5 Aeroplanes, satellites and energy storage

Beyond traditional cycling, calendar and drive cycle ageing, there are a few public datasets containing battery cycle data from more specialist applications. We review here four datasets relating to usage in aeroplanes, satellites and energy storage.

Aeroplane usage battery data

Another NASA Ames Prognostics Data Repository [163, URL] dataset is the 'HIRF Battery DataSet' [118]. It contains usage data from one single battery powering a small electric unmanned aerial vehicle [101]. The data in provided in '.mat' format under a double attribution license (see Section 9.2.1) a 'reference document' is provided on the NASA website explaining the file structure and experimental details.

Researchers from Carnegie Mellon University provide the 'eVTOL Battery Dataset' [16, URL] (shared under 'CC BY-NC-SA 4.0'). The dataset consists of discharge data from 22 Sony-Murata 18650 3Ah VTC-6 cells cycled with simulated Electric Vertical Takeoff and Landing (eVTOL) duties [17]. The data is provided in '.csv' format and includes voltage, temperature, current and charge/discharge capacity and energy measurements. The provided 'ReadMe' file and corresponding paper [17] give a full experimental description. This dataset is the first of its kind: providing public eVTOL data.

Simulated satellite operation profile battery data

The last dataset hosted by NASA [163, URL] that we report on is the 'Small Satellite Power Simulation Data Set' [117]. The dataset is provided in '.mat' format under a double attribution license (see Section 9.2.1). It contains data for two BP930 batteries (off-the-shelf 18650 Li-ion cells rated at 2.1Ah) run 'continuously with a simulated satellite operation profile completion for a single cycle' – experimental description in the corresponding paper [32, Section IV]. Additional details and data descriptions can be found by consulting the 'reference document' provided on the NASA website.

Stationary energy storage

Researchers from the University of Oxford and 'EnergyVill', with data provided on the Oxford Research Archive [174, URL], built a battery ageing model to serve a techno-economic analysis for grid-connected batteries [175–177]. Six Kokam 16Ah lithium polymer cells (model SLPB-78205130H) were aged following profiles corresponding to optimal trading strategies for stationary batteries in the Belgian day-ahead market of 2014. Experimental details can be found in the mentioned references and the PhD thesis of J. Reniers [177]. The cycling ageing tests were performed for up to one year to record the entire battery degradation process from the beginning of life to EOL. This dataset contains measured current, voltage and operating temperatures at $\sim$200 second intervals, and monthly capacity measurements (details provided in 'ReadMe' file). Files are given in '.csv' format and the database is shared under both the ODbL v1.0 and DbCL v1.0 license.

9.2.6 Synthetic data

Data driven approaches require data; thus, a lack of data is a significant barrier to their use. The obvious solution of collecting more data, covering a wide range of operating conditions, is expensive and time-consuming. Another approach is to use the available data to generate more data. This can be achieved by perturbing the data (data augmentation) or by generating artificial data. In this subsection we will review examples of the latter: producing so-called *synthetic data* for battery cells. Synthetic data can enhance existing datasets improving the performance of trained models and allowing for interpolation between cycling conditions not included in the experimental data. This interpolation step may be particularly important to data driven approaches enabling prediction 'outside the distribution' of the experimental data.

Here, we briefly describe one approach [165] to generate synthetic current and voltage data. For the generation of current curves a Markov chain approach can be used: transition probability matrices are constructed from real EV cycling data and then by iterating through the matrices (Markov chain propagation) synthetic current data can be obtained. From the generated current profile 'voltage cluster centroids' (the average value of temporally local voltage clusters obtained via k-means clustering) can then be predicted by a neural network trained on real data. These clusters have been shown to provide an effective feature for the prediction of SOC [164].

A comprehensive synthetic diagnostic dataset containing more than 500,000 individual voltage vs. capacity curves has been generated alongside a prognostic dataset with more than 130,000 individual degradation paths for a commercial graphite based LFP battery [64]. The diagnostic datasets [60, URL] and the prognostic dataset [61, URL] are both available on the Mendeley data website under a 'CC BY 4.0' license. The data is given in '.mat' format.

9.2.7 Data by request

Research projects are often subject to restrictions on the public release of generated datasets, however, upon publication some authors make their data 'available upon request'. This section briefs on such works and the corresponding datasets.

We mention here research carried out at the University of Warwick (UK). Dr Dhammika Widanalage, the principal investigator for the project, has provided us with the following description: 'Warwick University (UK) has been conducting thorough ageing tests on a batch of commercial LG M50 21700 cells (graphite/Si-NMC811). These tests consider two types of cell ageing: calendar and cycling. The calendar ageing tests cover four different ambient temperatures (0°C, 25°C, 45°C, and 60°C) and thirteen different initial SOC settings. Three cells were tested for each combination of ambient temperature and initial SOC. The cycling ageing tests consist of cells cycled at a variety of current C-rates for two low ambient temperatures (0°C and 10°C); the cells were immersed in an oil bath for thermal management. For all experiments (calendar and cycling) RPT were performed periodically to measure capacity losses, IR growth, and to log the pseudo-OCV values. In detail, first the discharge capacity was measured by the CC discharging protocol then the resistance at five different SOC levels (100%, 80%, 50%, 20%, 5%) was measured using pulse charge/discharge HPPC tests. The RPT procedures were run every four weeks for the calendar ageing tests and approximately every two weeks for the cycling ageing tests.' The above described experiments provide a comprehensive ageing dataset and set a benchmark for future data collection. The ageing data can be used for the analysis, modelling, prediction and tracing of ageing trajectories. Unfortunately, an external link to freely access the data cannot be offered. However, the datasets and on-going research progress (corresponding experimental cell data) are available for academic use, on request. If interested, please contact Dr Dhammika Widanalage via email `Dhammika.Widanalage@warwick.ac.uk`.

Other researchers at the University of Warwick have performed an ageing investigation based on EIS measurements for four NCA 18650 cells [155, 156]. Their EIS dataset has been deposited onto the university data repository [154, URL] and is accessible by request. Additional research manuscripts making use of datasets that remain private but whose authors point that the research datasets are available by request are listed in Table 9.7. A fuller description of their data and experimental work is detailed in the works themselves and summarised in Supplementary material Section 3 (expanding Table 9.7).

Applications	Ref	Data features
SOC/SOH estimation	[202]	real-word EV data; bulk datasets (300 EV & 400 HEV); battery pack health; NMC batteries; long-term test (over 12 months);used for big data analysis and machine learning method
	[203]	single cell tests (3 cells); SOC/SOH; statistical data-driven model fusion; 18650 LCO; DST and capacity tests
Battery modelling	[181]	a small batch of cells tested (51 cells/20 cells); cylindrical and pouch cells; NMC and NMC-LMO batteries; varied temperatures and accelerated ageing tests; electrical/thermal/ageing modelling
	[126]	a few cells tested (27 cells); calendar ageing test; long-term tests (over one year); NCA batteries; varied temperatures and initial SOC; calendar ageing modelling

Fault identification	[40]	31 NMC cells; charge profiles (rate ranged from 4C-9C); data driven method; Li-Plating; 450 cycles; 30°C test temperature; capacity, end-of-charge rest voltage (EOCV), open circuit voltage (OCV), and Coulombic efficiency (CE) were recorded
Capacity related early heath prognostics/ RUL prediction	[53]	a small batch of cells tested (35 cells); NMC batteries; early fault detection; real data collected on production lot samples; data-driven methods
	[158]	a few of cells tested; varied temperatures; incremental capacity analysis; LFP, NCM capacity data

Table 9.7: Non-publicly available Battery Data: Related paper and the corresponding research conducted. (See additionally Supplementary material Section 3.)

9.3 Data governance, repositories, tools and future outlook

Lithium batteries have been widely deployed and a vast quantity of battery data is generated daily from end-users, battery manufacturers, BMS providers and other original equipment manufacturers. Two elements are key in enabling the value of data: accessibility and ease of use. If no one can find or understand a public dataset it has no value. And, much of the time spent pre-processing data could be saved given a widely used standard publication format.

In this section, we review data platforms and online repositories that can be used to host data; tools for data validation and processing; and community maintained living reviews

9.3.1 Data repositories and platforms

Data storage platforms provide a common and easily navigable location to find and (possibly) share data. They also promote standardisation in data format and descriptions. We point here to several repositories hosting public battery data.

- **Scholarly usable, citable and freely accessible**
  1. **Battery Archive [10, URL]**: Battery archive, developed at the City University of New York Energy Institute, provides a free repository of battery testing data which is easily searchable by cell chemistry, form, capacity and test variables. Different datasets, shared by various institutions, have common file formats and the website provides easy access to the data. We highlight their 'rules for metadata' section proposing a common nomenclature to use for descriptions of cells and cycling conditions.
  2. **DOE OE [57, URL]**: The U.S. Department of Energy's Office of Electricity (DOE OE) has collaborated with two national labs, *Sandia National Laboratories* and *Pacific Northwest National Lab*, to carry out battery research addressing energy storage risk assessment and mitigation. Their website provides free access to the resulting research data including abuse tests, cycling tests and EIS measurements.
  3. **NREL [145, URL]**: The National Renewable Energy Laboratory is a national laboratory of the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy, operated by the Alliance for Sustainable Energy, providing free battery datasets to aid in the development of cell models and tools to

facilitate the deployment of renewable energy. Regarding their battery research, well-rounded testing data encompassing the failure data collected from hundreds of abuse tests (nail penetration, thermal abusing, and internal short-circuiting), ageing cycling data, driving cycle data and other commercial oriented battery operating data (collected from EV operation) has been provided.

- **Public Digital data repositories**

  There are several curated data platforms that make research data discoverable, freely reusable and citable. A non-exhaustive list of the publicly accessible data repositories where battery data has been deposited is outlined as follows.

  1. Dryad [59, URL]
  2. Zenodo [240, URL]
  3. European federation of data driven innovation hubs [75, URL]
  4. Mendeley data centre [140, URL]
  5. 4TU.ResearchData [102, URL]
  6. Google Database [88, URL]

9.3.2 Community maintained reviews and standards

There are a few community maintained online resources listing publicly available battery datasets. The approaches taken to curate such lists differ but represent a critical initial step from the community to make public datasets more accessible and understandable. This review includes, at the time of publication, the datasets in these referred community maintained resources and several other datasets with corresponding descriptions. Researchers with knowledge on where to find battery datasets are heartily invited to contribute to the living reviews listed below.

- Community databases of publicly available battery datasets maintained by
  1. Dr. Valentin Sulzer (University of Michigan): [222, URL]
     (by way of private communication, this resource is no longer maintained)
  2. Dr. Bolfazl Shahrooei (Iranian Space Research Center): [8, URL]
     (community maintained and active)
- Standards and identification references
  1. BatteryStandards.info [12, URL]: Website containing information on around 400 standards for rechargeable batteries including: battery test standards across categories such as characterisation tests, safety tests, performance tests and requirements.
  2. An extensive identification reference for lithium-ion Battery of size-type 18650 covering brand, model, capacity, chemistry, max charge/discharge and link to product specification datasheet is presented in: [147, URL].

9.3.3 Data processing and validation tools

Battery cycling data is highly complex. Different cycling protocols, cycler manufacturers and experimental configurations make it difficult to compare datasets and validate models. As a result, several high quality open source packages have been created to perform data processing, parsing and validation. We provide a non-exhaustive summary of available tools.

- **Tools for data management and validation**[4]
    1. **BEEP (Battery Evaluation and Early Prediction) [14, URL]**: a package for parsing and featurizing battery cycling data specifically geared towards cycle life prediction [100].
    2. **cellpy [34, URL]**: a package which parses Arbin cycler data and enables manipulation of cycling data using pandas dataframes. In addition, it enables incremental capacity (dQ/dV) analysis and the extraction of open circuit relaxation points.
    3. **impedance.py [108, URL]**: a package for the analysis of electrochemical impedance spectroscopy (EIS) data. Core functionality includes plotting experimental impedance spectra, fitting impedance spectra to equivalent circuit models, computing and plotting the impedance spectra of equivalent circuit models and validation of impedance spectra using the Kramers-Kronig relations.
    4. **Bayesian Hilbert transform [87, URL]**: Python implementation of [125] providing validation of EIS data via the Kramers-Kronig relation recast under a Bayesian framework.

Lastly, we point to means of extracting numerical data from data visualisations, for instance, the open-source software *WebPlotDigitizer* [182, URL]. Given an image of a plot the raw data points are identified in a semi-automated manner. Numerical data is extracted based on the identified data points and user-defined calibration points marked on the plot. Such an approach has been used in [179] (with MATLAB's GRABIT tool) to extract capacity fade curve data from published work.

9.3.4 Current limitations

Effective energy storage is critical. Improvements in safety, density and longevity mean more reliable devices, vehicles requiring less frequent charging and replacement, and efficient and long lasting stationary energy solutions. Currently, the communication of data between end-users, manufacturers, distributors and providers is weak. Greater transparency in this aspect would accelerate scientific progress in all areas. Fig.9.4 illustrates the wide ranging deployment of batteries across industries.

[4] The authors kindly thank J. Koeller [11] for his assistance in developing this list.

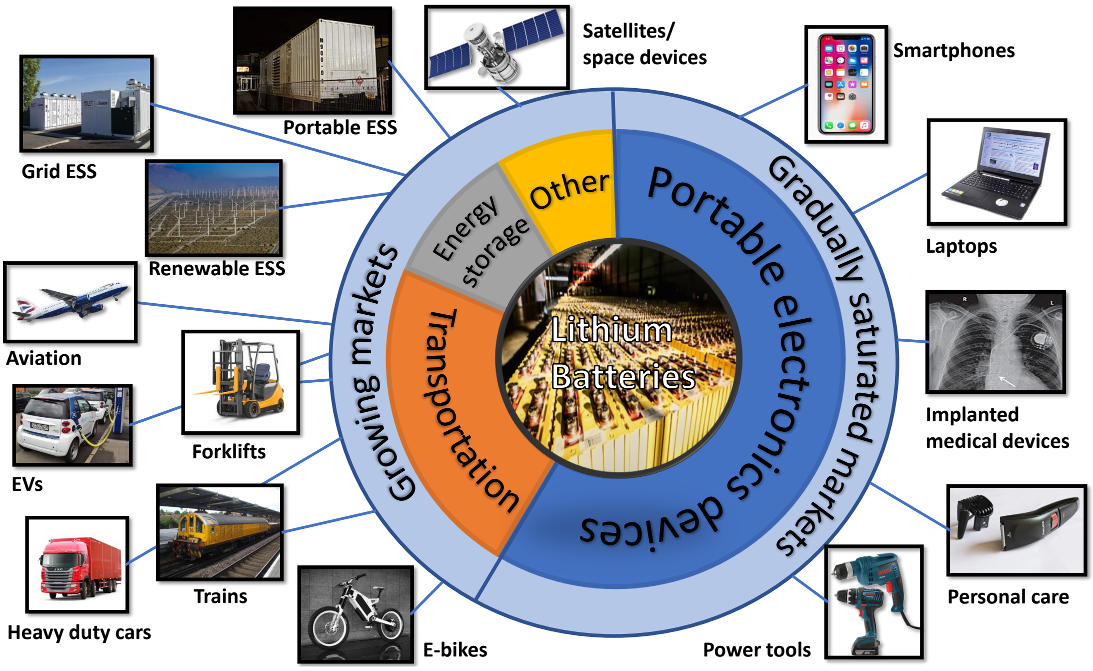


Fig. 9.4: Lithium battery sample applications.

Regarding the data reviewed in the manuscript, we failed to find many examples of ‘on field data’ where the varying conditions of battery usage can be seen. Examples of such data would be: data regarding aerospace applications either from the perspective of aeroplane electrification or simply from satellite usage where batteries are a mission critical element; battery usage data for energy storage systems (either at home-owner level or at the electric grid level); data regarding electric heavy-duty vehicles (e.g., firetrucks or buses); data linking material science data to cycling data or data connecting manufacturing to degradation; data that can be used to optimise the cell selection process for the purpose of battery pack formation. Moreover, all the data reviewed in this manuscript is from first life applications where the battery was tested from new. We have not found any data on the so-called *battery 2nd life* where the battery, say, was redeployed from a EV into a stationary application like grid energy storage. Lastly, left out of this study was a review of data relating to *abuse testing* and data *containing mechanical measurements*. A representative of the latter would be datasets that include mechanical measurements, e.g., cell dilation or weight.

Battery testing is costly and lengthy, and this is unlikely to change: how can one understand the life-cycle of a battery that lives for 10 years without carrying out 10 years of testing? A sub-problem in this context is the sparsity of recorded data, for example, cells are usually tested within a (dotted) range of fixed temperatures and with fixed cycling conditions. These conditions do not reflect the variability of real use. And, many approaches fall short when interpolating between recorded data: how does one predict cell degradation at 25°C from data recorded at 40°C and 10°C. Methodologies addressing this problem are beginning to emerge [167, 214].

From a holistic point of view, the publicly available datasets come in all shapes and sizes. Files appear in '.mat', '.txt', '.csv' or 'Excel' format ('.mat' and '.csv' being the more common) with wildly varying file structures: from raw cycler data – split by cycle, week, month or not at all – to structured data with explanatory scripts and text to assist the user. From our understanding, there is a general lack of consensus on the way to present data. For instance, different brands of cell cycling machines output data in different ways including varying nomenclature for the same quantities. In this regard, we highlight again the open-source Python-based framework BEEP (Battery Evaluation and Early Prediction) [14] for the management and processing of high-throughput battery cycling data and the Battery Archive's 'Rules for Metadata' section [10] proposing a common nomenclature for the descriptions of cells and cycling conditions. In the Author's opinion, exemplary datasets for file format and description are those provided by P. Kollmeyer in Section 9.2.2 and the Toyota Research Institute data in Section 9.2.1. We leave a suggestion for any group sharing data: to provide a basic accompanying script (MATLAB or Python) that plots the uploaded data (time series/EIS or capacity/resistance change) and text explaining file structure. This, on its own, would expedite the understanding of datasets; however, there is a clear and greater benefit which could be gained from researchers adopting a uniform file format.

9.4 Conclusions

Comprehensive battery datasets play a critical role for battery research both in academia and industry. However, publicly available datasets are distributed sporadically as battery testing is costly and lengthy. In this work, a review of the existing battery datasets in the public domain is provided with a category-type break-down covering the testing regimes, cell specifications and provided data. This informs a long view on the available datasets hinting at gaps in the experimental space which in itself presents an opportunity for further work. Lastly, high-quality open source packages for a variety of battery-related tools are also reviewed.

With this work we wish to convey two further messages,

1. the academic community is starved for research data, and
2. we strongly encourage any person or group (academic or industrial) to share their data.

Chapter 10

Elbows of Internal resistance rise curves

The work presented in this chapter is taken from our paper [210], which was a joint work with Prof. Gonçalo dos Reis, Dr. Shawn Li, and Richard Gilchrist MSc.

Abstract

The degradation of lithium-ion cells with respect to increases of internal resistance (IR) has negative implications for rapid charging protocols, thermal management and power output of cells. Despite this, IR receives much less attention than capacity degradation in Li-ion cell research. Building on recent developments on 'knee' identification for capacity degradation curves, we propose the new concepts of 'elbow-point' and 'elbow-onset' for IR rise curves, and a robust identification algorithm for those variables. We report on the relations between capacity's knees, IR's elbows and end of life for the large dataset of the study. We enhance our discussion with two applications. We use neural network techniques to build independent state of health capacity and IR predictor models achieving a mean absolute percentage error (MAPE) of 0.4% and 1.6%, respectively, and an overall root mean squared error below 0.0061. A relevance vector machine, using the first 50 cycles of life data, is employed for the early prediction of elbow-points and elbow-onsets achieving a MAPE of 11.5% and 14.0%, respectively.

10.1 Introduction

Sales of electric vehicles (EVs) and energy storage systems are undergoing a marked growth as battery costs continue to fall and with the introduction of increasingly strict regulations on CO_2 and NO_x emissions, deadlines on the decommissioning of fossil fuel power stations and bans on the sale of internal combustion engines. Lithium-ion (Li-ion) batteries are widely deployed in EVs and energy storage systems due to their outstanding characteristics, such as low maintenance requirements, high Coulombic efficiency and market-leading energy density; however, in operation, Li-ion batteries are sensitive to over-charging/discharging, high current stresses, over-temperature and under-temperature. Even when cycled under moderate operating conditions, solid-electrolyte interphase (SEI) layer growth on anodes gradually consumes active material, leading to poor cyclability. Extreme operating conditions will further accelerate ageing processes, potentially resulting in high-risk failure scenarios such as gassing, mechanical cracking of electrodes, internal short circuits and thermal runaway [19, 33, 85, 120, 131, 137, 152, 221, 228]. Furthermore, the degradation rates of identical chemistry cells differ due to disparities in manufacturing quality and operating conditions [63, 84, 201, 228]. The accurate prognosis of cell degradation is therefore imperative. This is referred to as the State of Health (SOH) of the cell and can be defined with respect to its capacity or its internal resistance (IR). A cell's capacity fades as its calendar and cycle age increase, and degradation mechanisms take place within the cell that reduce the available lithium inventory and accessible active material in the electrodes [82, 98]. Conversely, as the cell is cycled, IR increases due to the thickening formation of the SEI, and the consumption of electrolyte and lithium in this process [85, 228].

Given the importance of driving range, capacity is the primary SOH measurement for pure EVs; naturally, capacity-based SOH measurement is less important for hybrid electric vehicles (HEVs), instead, importance is placed on a cell's ability to supply high operating currents. With the increase of IR, the current deliverability of a cell is diminished, making IR a key SOH measurement for hybrid vehicles. For a given current, increased IR can raise the terminal voltage during the charging phase. As a result, the imposed charging current must be taped down to avoid the battery voltage from exceeding its maximum limit; thus, leading to longer charging times and poor rapid charging ability [1, 131, 180, 239]. In addition, the growth of IR values will incur more heat generation for a given load creating more work for the thermal management system. To the best of our knowledge, the majority of EV manufacturers only provide a battery warranty securing that the capacity shall remain above 70% of its initial value, but ignore a battery warranty based on IR. With a greater understanding of expected IR growth such warranties could be provided. There is thus significant value to be gained from the prognosis of IR growth trends; however, the prediction of IR degradation using data from early cycles remains largely unexplored. There is substantive research e.g. [78, 197] conducted for early prediction of capacity but not for IR.

As discussed in-depth in [78], a cell's capacity does not degrade linearly throughout its lifetime: degradation is path-dependent [171], and a strong association exists between capacity and internal resistance [9]. While the cell's capacity typically starts to degrade in a linear manner, there eventually comes a point, called the 'knee-point', after which the rate of capacity degradation increases considerably [55, 69, 93, 143, 191, 196, 241]. In [78] one can find a review of knee-point identification methods [55, 191, 241], and, crucially, the additional variable 'knee-onset' is introduced (along with an alternative identification mechanism) to provide a useful indication of the beginning of a sharp increase in the capacity degradation trend. However, the corresponding notion of ‘knee-point’ and ‘knee-onset’ in IR degradation curves is absent from the current literature. In this paper, we bridge this gap by addressing the IR rise curve and the corresponding change points: the ‘elbow-onset’ for when the IR curve becomes nonlinear and the midpoint of the accelerated IR increase which we call the ‘elbow-point’.

There are three main contributions of this work. Firstly, at a data pre-processing level, we create an accurate IR predictor utilising machine learning convolutional neural network (CNN) techniques. This predictor is then used to complete the dataset of [6] (for which no IR readings were logged). Secondly, underpinned by the completed dataset, the concepts of elbow-point and elbow-onset points for IR rise curves are proposed along with corresponding identification methods. Thirdly, we showcase a working example of using the predicted and real IR data for the early prediction of elbow-point and elbow-onset using only the first 50-cycles of the cell’s lifespan data.

The rest of this paper is organised as follows. Section 10.2 introduces the data pool and the data pre-processing approach addressing a missing IR data problem. In Section 10.3, we propose the elbow-onset and -point concepts and identification algorithms, concluded by a study of the numerous relationships between these quantities. Section 10.4 presents the relevance vector machine (RVM)-based machine learning approach for the early prediction of elbows. Results, contributions and future work are summarised in Section 10.5.

10.2 Battery Data Framework and Data Pre-Processing Procedures

In this Chapter we work with the TRI dataset [6, 197] (see Section 9.2.1). We recall that the TRI dataset contains 8 batches of cells, which we refer to sequential as batches 1–8. To distinguish individual cells within the dataset, we refer to cell Y of batch X as bXcY.

10.2.1 Data Pre-Processing via a Machine Learning Approach: Completing the Missing IR Data

Our first goal, to increase the scope of our analysis, is to address the missing IR data of batch 8. We draw on machine learning techniques and build an IR prediction model (on the data from batches 1–3) to predict the missing IR data of batches 4–8. Increasing the number of matched capacity-IR curves from 124 pairs to 357 (= 124 + 233). Of these 357 pairs 169 (= 124 + 45) contain measurements up to or past the EOL. This will enhance our later analysis comparing elbows, knees and the EOL, as well as the early prediction of elbows. For statistical reasons, we build a simple yet accurate capacity predictor to test for distributional dissimilarity between batches 1–3 and batches 4–8.

Pre-Processing and Modelling Pipeline

We split the cells of the dataset into training and test sets: grouping by batch so that our test set contains an equal percentage of cells from each batch. As input our capacity and IR prediction models take one charge/discharge cycle of voltage, current and SOC data (the integral of the current from one full cycle). This data was cleaned, standardised to have values between 0 and 1, interpolated using the 'SciPy' [109] function 'interp1d' to one measurement every four seconds and zero-extended so that the data for each cycle of each cell was of equal length and consistent time step. The median filter, averaging five nearest time instants, was applied to smooth the measurements of capacity and IR prior to prediction.

To design our models for IR and capacity prediction we utilised K-fold cross validation. A validation set of cells was chosen at random from the training set, our models fitted to the remaining training set and evaluated on the validation set throughout training. This step was then repeated K-times with a new validation set and corresponding model. The average performance of the validation sets was used to optimise model design and choice of hyper-parameters. K-fold cross validation is particularly useful when working with small datasets: mitigating the risk of over-fitting a particular validation set [113]. After settling on the model's architecture and hyper-parameters (described next), a copy of the model was fitted to the whole training set and then evaluated on the test set to calculate performance metrics.

Model for IR Prediction

We propose a model consisting of a convolutional 'feature extraction' block followed by two densely connected layers displayed in Figure 10.1 and described in Table 10.1. Our model was implemented in Python using TensorFlow via the Keras API [43]. All layer names given in Table 10.1 refer to the corresponding Keras layers. The model was trained on the data from batches b1–b3 using the adam optimiser for 50 epochs with a batch size of 526 and the mean absolute error – Equation (10.1) – as its loss function.

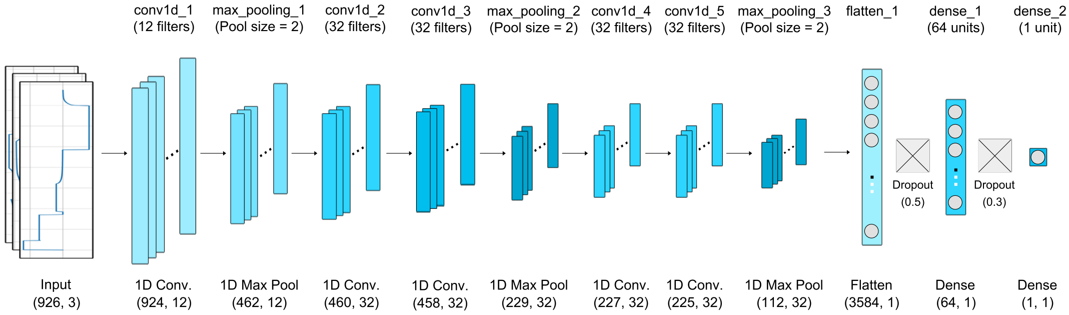


Fig. 10.1: Schematic of machine learning model for internal resistance (IR) prediction.

Table 10.1: Proposed architecture of convolutional neural network (CNN) model for prediction of IR. Hyper-parameters are given in the format: filters, kernel size, activation for conv1d layers; pool size for max_pooling; dropout rate for dropout; nodes, activation for dense layers.

Layer Name	Input Size	Hyper-Parameters	Output Size
conv1d_1	926×3	12, 3, *ReLU*	924×12
max_pooling_1	924×12	2	462×12
conv1d_2	462×12	32, 3, *ReLU*	460×32
conv1d_3	460×32	32, 3, *ReLU*	458×32
max_pooling_2	458×32	2	229×32
conv1d_4	229×32	32, 3, *ReLU*	227×32
conv1d_5	227×32	32, 3, *ReLU*	225×32
max_pooling_3	225×32	2	112×32
flatten_1	112×32	-	3584
dropout_1	3584	0.5	3584
dense_1	3584	64, *ReLU*	64
dropout_2	64	0.3	64
dense_2	64	1, *linear*	1

The machine learning performance scores selected for this work are the mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE) defined as follows. For $\boldsymbol{y}$ the vector of true values and $\hat{\boldsymbol{y}}$ the vector of predicted values

$$\text{MAE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} |\hat{y}_i - y_i| \,, \qquad \text{MAPE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{100\%}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \frac{|\hat{y}_i - y_i|}{y_i} \,, \tag{10.1}$$

$$\text{and} \quad \text{RMSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} (\hat{y}_i - y_i)^2} \,. \tag{10.2}$$

Our model's performance metrics for IR prediction can be found in Table 10.2. We are unaware of works using the A123 dataset for IR estimation. Nonetheless, the estimation of IR has been addressed for other datasets [86, 91, 123, 172, 188, 220, 242]. We obtain an RMSE of 0.00035 and an MAPE of 1.6% which is low (if nominally compared with capacity estimation accuracy in the literature).

Table 10.2: Average performance of model to predict IR with 95% prediction intervals.

	RMSE		MAPE (%)	
	Train	**Test**	**Train**	**Test**
IR	$0.00029 \pm 6.2 \times 10^{-5}$	$0.00035 \pm 5.0 \times 10^{-5}$	1.19 ± 0.22	1.60 ± 0.24

Validation Step via a Model for Capacity Prediction

We have shown that our model for IR prediction is effective on batches 1–3. To see if we can trust the predictions that this model makes on batches 4–8 we check for non-similarity between the datasets. We do this by extrapolating on capacity – a variable present for all batches. This is a standard process in imputation (simple or multiple). To this end, we utilise a simple feed-forward neural network consisting of three densely-connected layers: the first two layers containing 32 neurons with the rectified linear unit (ReLU) activation function and the final layer consisting of a single neuron with a linear activation. The model was trained for 100 epochs with a batch size of 512 using the adam optimiser and the mean squared error as its loss function. During training, a dropout of 0.2 was used between the middle and last layer. Trained on all of the data from batches 1–3 and tested on batches 4–8, the model obtained the performance metrics displayed in Table 10.3 with an MAPE of 0.51%. This test gives us confidence that both datasets [6, 197] are indeed not dissimilar.

Table 10.3: Average performance of capacity model trained on batches 1–3 tested on batches 4–8, with 95% prediction intervals.

	RMSE		MAPE (%)	
	Train	**Test**	**Train**	**Test**
Capacity	$0.0053 \pm 4.2 \times 10^{-3}$	$0.0095 \pm 4.6 \times 10^{-3}$	0.37 ± 0.30	0.51 ± 0.26

The prediction of capacity (and SOH) is of wider interest than our discussion of elbows, so we briefly compare these results with those found in the literature. We point to Table 1 in [166] (MAPE and RMSE error given) and Table 2 in [144] (error type not given) for a review/comparative work on capacity estimation. We cannot directly compare our results, as the data is different. However, from a strictly numerical point of view, our RMSE of 0.0095 and MAPE 0.51% errors for capacity (Table 10.3) are lower than the values of [166] (Table 1)—for a fair comparison, one would need to test the varying approaches on a common dataset.

Predicting the missing IR data

In order to address the missing data issue, we trained the IR model on batches 1–3 multiple times and an ensemble of these models was used to predict on batches 4–8. This predicted IR data is available at `https://doi.org/10.7488/ds/2957` (accessed on 21/02/2021) . Figure 10.2 shows the IR for sample cell b8c4 and we strongly emphasise to the reader that the extrapolation of the IR data past EOL (80% capacity) is, as fully expected, not reliable; this stems from the limitation of the training dataset (batches 1–3) with data only up to the EOL. Prediction outside that range of input data is not reliable as can be seen in Figure 10.2 where we observe a strong widening in the prediction intervals past the EOL. The prediction intervals provided throughout the text are calculated in a frequentist manner. A given model is fitted to data multiple times and performance metrics/predictions recorded. The empirical average and variance-value of predictions are calculated and under the assumption of normality one uses those values to produce prediction intervals (at any given probability quantile level q, e.g. in Figure 10.2 we have $q = 95\%$ and $q = 80\%$).

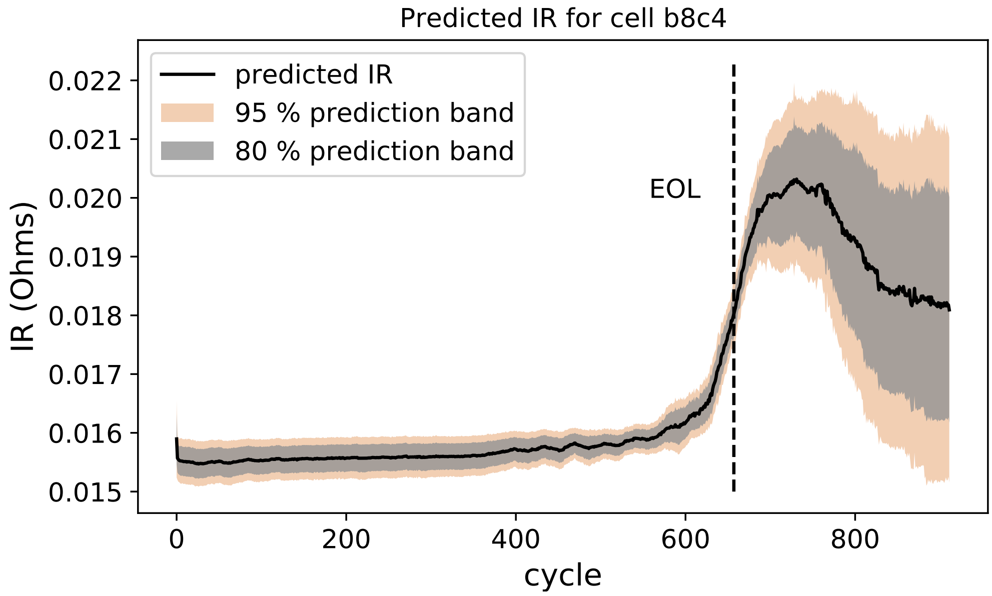


Fig. 10.2: The predicted IR data for cell b8c4 are given by the black continuous line and is formed from the average of 20 predictions. We display 80% and 95% prediction intervals. Beyond the intuition of extrapolation, these intervals show that predictions past the EOL (capacity) should not be trusted.

Algorithmic Framework

The proposed algorithmic framework takes full advantage of machine learning-based approaches to solve the missing IR data problem in the raw data pool allowing the generation of artificial IR data to complete the life cycle data. The predicted IR data can be used for elbow-point and -onset identification and is able to assist the early prediction of the elbow-point and elbow-onset in IR curves.

The schematic framework of the algorithms is illustrated in Figure 10.3. Section 10.2 introduces the data pre-processing procedure, where a CNN-based predictor has been trained on the data from batches 1–3 to predict the missing IR of batches 4–8. For validation, a capacity estimator was trained to test for dissimilarity between the two datasets. In Section 10.3, using the completed data we confirm significant linear relationships between knee/elbow-points, -onsets and EOL. Further tests are carried out in Section 10.4 relating to the early prediction of elbow-points and -onsets. In particular, the straightforward RVM-based quantitative method is applied for the early prediction of elbows.

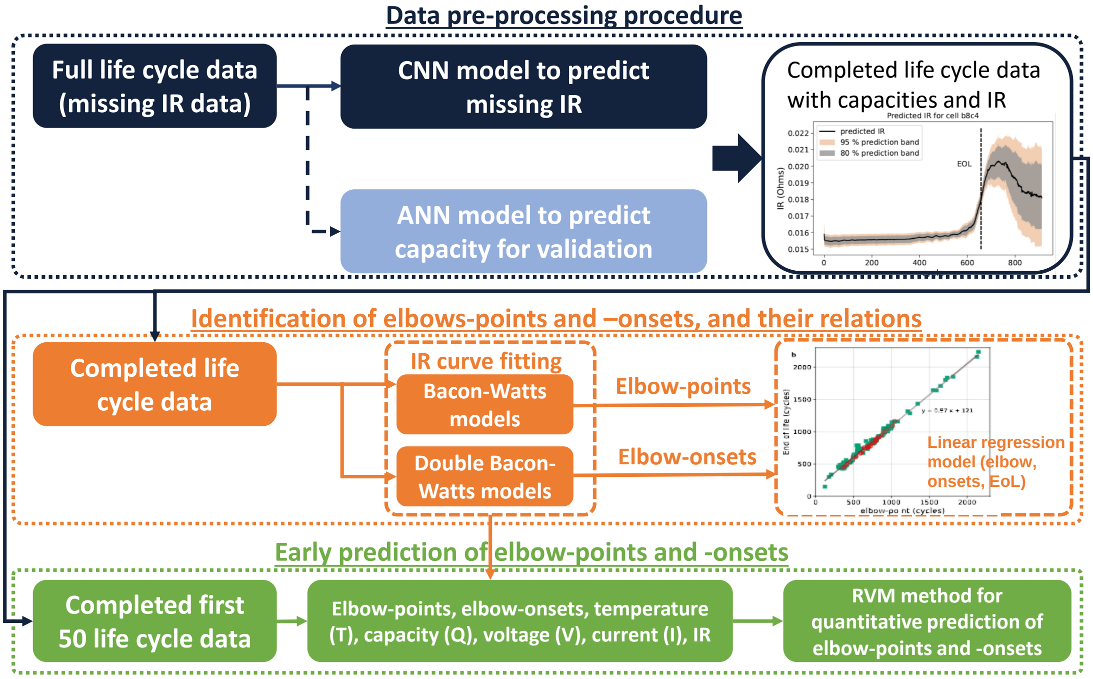

Data pre-processing procedure
Full life cycle data (missing IR data)
CNN model to predict missing IR
ANN model to predict capacity for validation
Completed life cycle data with capacities and IR
Predicted IR for cell b8c4
predicted IR
95 % prediction band
80 % prediction band
EOL
IR (Ohms)
0.022
0.021
0.020
0.019
0.018
0.017
0.016
0.015
0
200
400
600
800
Identification of elbows-points and –onsets, and their relations
Completed life cycle data
IR curve fitting
Bacon-Watts models
Double Bacon-Watts models
Elbow-points
Elbow-onsets
b
End of life (cycles)
elbow-po nt (cycles)
y = 0.97 x + 121
2000
1500
1000
500
0
Linear regression model (elbow, onsets, EoL)
Early prediction of elbow-points and -onsets
Completed first 50 life cycle data
Elbow-points, elbow-onsets, temperature (T), capacity (Q), voltage (V), current (I), IR
RVM method for quantitative prediction of elbow-points and -onsets

Fig. 10.3: Graphical abstract for the proposed algorithmic framework.

10.3 Identification of Elbows, Knees and Their Relations

10.3.1 Methodology

Fermín et. al. [78] proposed the use of the Bacon–Watts (Equation (10.5)) and the double Bacon–Watts model (Equation (10.6)) for the identification of knee-point and knee-onset, respectively. We will use the same basic methodology, with the addition of several steps to account for noise in the data and potential sigmoid behaviour. The high level of noise present in the IR data, see Figures 10.4a and 10.5a, prevents the Bacon–Watts model from neatly fitting the data, and this is overcome via a smoothing step as described in Algorithm 1 (block 1) below. We report that this noise also causes issues for the alternative knee identification methods proposed in [55, 78, 191], see Figure 10.5. In addition, we observed sigmoid-type capacity fade curves for some cells in batch 8, and hence, we employ a subroutine to isolate the knee/elbow identification from the right-most plateau. We present first the algorithm and afterwards reason its several steps.

Algorithm 1 'Smoothed Bacon–Watts': Identification of knee/elbow-point and -onset

```
Block 1: Data smoothing.
1.  Fit isotonic regression to (capacity/IR degradation) lifespan data (across the full curve).
2.  Determine data-truncation cycle-point $n^*$:
    (a) Fit (10.3) (Asymmetrical sigmoidal) to isotonic regression curve,
    (b) Find cycle-number $n^*$: cycle at which 2nd derivative of fitted (10.3) changes
        sign, else last cycle in series.
3.  Fit (10.4) (line-plus-exponential) to isotonic regression curve up to cycle $n^*$.
Block 2: Identification.
4.  Fit Bacon–Watts model (10.5) to (10.4). Identify knee/elbow-point.
5.  Fit double Bacon–Watts model (10.6) to (10.4). Identify knee/elbow-onset.
```

The isotonic regression step, Step 1, solves several issues: it annuls the behaviour of capacity increase or IR decrease across the first few cycles and removes the influence of sharp movements where the IR decreases or increases due to measurement errors. From first principles, our choice reflects the fact that the electrochemical degradation mechanisms within the cell are irreversible. For a given load and set of ambient conditions, IR increase may be caused by the thickening of the SEI on the anode which irreversibly consumes lithium and electrolyte. Additionally, IR increase can be caused by a loss of anode and cathode material which can result from many factors, such as electrode particle cracking and loss of electrical contact as a result of mechanical expansion/contraction during cycling, corrosion of current collectors at low cell voltage and binder decomposition at high cell voltage. These same mechanisms also lead to an irreversible reduction in capacity and, as such, the monotonicity of the model is reflective of the real-world evolution of a cell's capacity over its lifespan. The isotonic regression is performed using the Scikit-learn Python package [157] and the procedure is described in [35].

Throughout the manuscript, and the following equations, the generic ε variable denotes the errors/residuals of its associated model, indicated as a superscript, and is a normal random variable with zero mean and finite (unknown) variance.

The asymmetrical sigmoidal fitting step, Step 2. The asymmetrical sigmoidal ('as') model is described by Equation (10.3)

$$Y^{as} = d + \frac{a-d}{\left[1+\left(\frac{x}{c}\right)^{b}\right]^{m}} + \varepsilon^{as}, \tag{10.3}$$

where a and d associate to the top and bottom plateau of the curve respectively, b controls the slope between plateaus, m the level of asymmetry and c determines the inflexion point. For given data, the constants are estimated by straightforward least-squares estimation (similarly for subsequent parametric models).

In several cells from batch 8, we observe a sigmoid-type capacity fade curve where, after passing the knee and then degrading linearly for some time, the degradation approaches a plateau (e.g. cell b8c4). To isolate the detection of knees/elbows from this behaviour, we propose the fitting of the asymmetrical sigmoidal model to then truncate the data before said plateau (point n^{*}) via the 2nd derivative truncation rule.

The final smoothing step, Step 3, involves fitting the parametric line-plus-exponential ('le') model of Equation (10.4) to the isotonic data (from Step 2) up to cycle n^{*}. This idea can be traced back to [106] (Section 2.2.1) under the name of 'Exponential/linear hybrid model' – [214, 220] discuss other parametric models. The line-plus-exponential is described by the following model:

$$Y^{le} = \beta_0 + \beta_1 x + \beta_2 \exp(\lambda x - \theta) + \varepsilon^{le}, \tag{10.4}$$

where β_0, β_1 and β_2 control the intersection point and slope of the line, and the size of the exponential, respectively. The quantity λ controls the 'speed' of the exponential and θ controls where the impact of the exponential starts. The main motivation for model (10.4) is that for many cells, the degradation of IR is very close to linear until close to the elbow-onset followed by a sharp elbow-point.

For the identification of the knee/elbow-point, Step 4, we use the Bacon–Watts model. Fermín et al. [78] (Equation (1)) describe the Bacon–Watts ('bw') model (10.5) as a two straight-line relationships around the transition point x_1:

$$Y^{bw} = \alpha_0 + \alpha_1(x - x_1) + \alpha_2(x - x_1)\tanh\{(x - x_1)/\gamma\} + \varepsilon^{bw}, \tag{10.5}$$

where α_0, α_1 and α_2 control the slopes of the intersecting lines and the intercept-weigh of the leftmost segment respectively and γ controls the abruptness of the transition. We fix γ as a small value to obtain an abrupt transition. After optimisation, the fitted value of x_1 is defined as the knee/elbow-point.

The identification of the knee/elbow-onset, Step 5, is performed by the double Bacon–Watts model ('dbw') (10.6) (also [78] Equation (2)) modifying Bacon–Watts to identify two transition points, concretely:

$$Y^{dbw} = \hat{\alpha}_0 + \hat{\alpha}_1(x - x_0) + \hat{\alpha}_2(x - x_0)\tanh\{(x - x_0)/\hat{\gamma}\} + \hat{\alpha}_3(x - x_2)\tanh\{(x - x_2)/\hat{\gamma}\} + \varepsilon^{dbw}, \tag{10.6}$$

as in Equation (10.5), the parameters $\hat{\alpha}_i$ and x_j are estimated and $\hat{\gamma}$ is chosen as a small value to produce abrupt transitions at x_0 and x_2. The knee/elbow-onset is defined as the change point x_0.

Figure 10.4 displays the output of Algorithm 1 applied to the IR curve of cell b1c29 (non-predicted data). Elbow-point and its onset are identified, and the smoothing steps are illustrated showing the fitted isotonic regression and line-plus-exponential model against the input data (for this cell, Step 2 yields n^* as the final cycle number). Figure 10.5 displays the performance of other known algorithms for knee identification applied to the elbow identification problem. We find that [55, 78, 191]'s algorithms are too sensitive to noise to provide consistent identification results. Our approach addresses the noise issue allowing for coherent elbow identification. From a statistical point of view, any identification approach will be affected by the noise in the data; thus, the identified elbows will be less exact than the identified knees, for which the data is much smoother. For comparison, the non-parametric bootstrap procedure was used to calculate 95% confidence intervals (CI) for the knee/elbow-points and -onsets identified by Algorithm 1. The average CI's width was 24 cycles for the elbow-point, 4 cycles for the knee-point, 35 cycles for the elbow-onset and 5 cycles for the knee-onset; this difference is a direct consequence of the noise present in the IR data. Finally, Algorithm 1 applied to knee identification recovers fully the results of [78] (we omit these results).

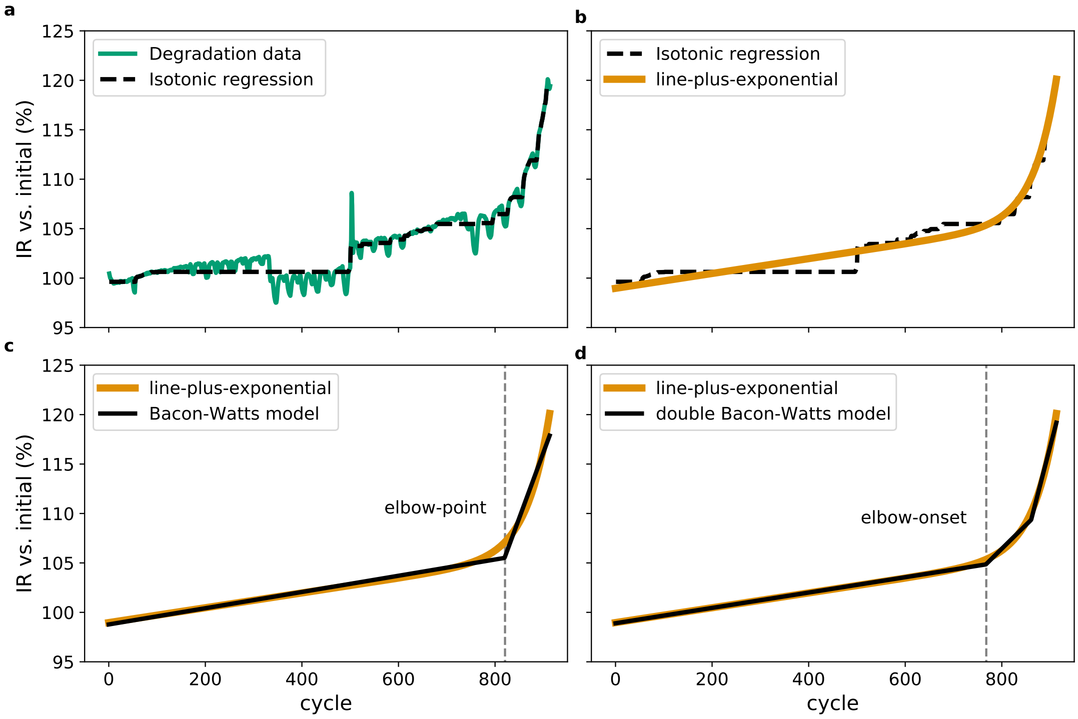


Fig. 10.4: Steps of Algorithm 1 applied to the internal resistance degradation curve of cell b1c29 (non-predicted data). (**a**), step 1. (**b**), step 3. (**c**), step 4. (**d**), step 5. Step 2 is omitted as it has no impact here: n^* is chosen as the final cycle number. The width of the 95% confidence interval (computed by the non-parametric bootstrapping procedure) for the elbow-point of this curve is 23 cycles, and for the elbow-onset it is 38 cycles.

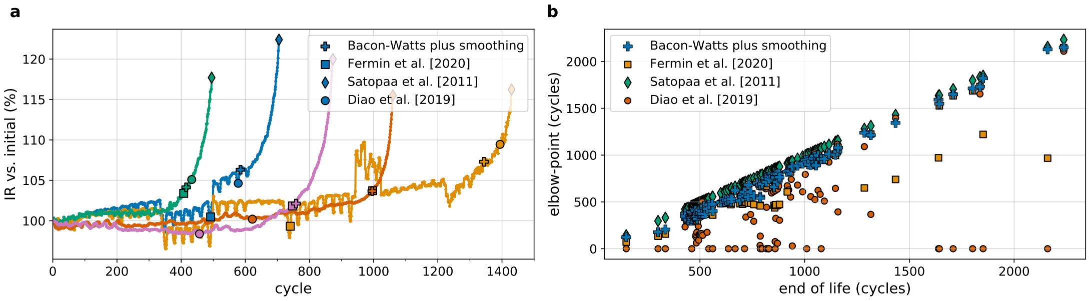


Fig. 10.5: (**a**), Comparison of elbow-points obtained with Algorithm 1, [78]'s Bacon–Watts, maximum curvature and slope changing ratio methods on a sample of cells from the A123 dataset (from left to right b2c34, b1c30, b3c15, b3c1, b1c3). (**b**), Comparison of elbow-points for all cells in the A123 dataset. One expects to see a linear relationship between EOL and elbow-point; of the methods compared only Algorithm 1 and the algorithm of Satopaa et al. [2011] recover a linear relationship reliably, however, by examining plot (**a**), we see that Satopaa's algorithm selects the end point as the elbow.

Zhang et al. [241] report for a dataset of nickel-manganese-cobalt cells that the knee-point appeared at between 90–95% nominal capacity; in [78], it was reported that the knee-point, for batches 1–3 of the A123 dataset, appeared on average at 95% nominal capacity and the knee-onset at 97.1% nominal capacity, with an average gap of 108 cycles between the knee and its onset. We report that, for the A123 dataset batches 1, 2, 3 and 8, on average, the elbow-onset appears at 103.0% initial IR (93.6% nominal capacity) and the elbow-point at 104.7% initial IR (91.3% nominal capacity), with the elbow-onset and its point on average 52 cycles apart; on average, both elbows appear after the knee-point. These reported figures are calculated from the smoothed exponential curve as described in Algorithm 1.

10.3.2 Linear Relations

Figure 10.6 illustrates the strong linear relationships observed between the calculated knee/elbow-points and the EOL, making it possible to estimate each point given a measurement or prediction of another point(s). These linear relations are obtained using a standard linear regression model $y = c_0 + c_1 x + \varepsilon$, where y denotes the dependent variable, x the independent variable, ε represents the residuals, and c_0 and c_1 control the intercept and slope of the linear model, respectively. The obtained coefficient values along with their confidence intervals are presented in Table 10.4, where the knee relations agree with those found in [78] (Table 1).

We present the linear relationships obtained when including the predicted IR data. From viewing Figure 10.6, comparing the green squares and black circles, the reader will appreciate that their inclusion did not significantly influence the linear relationship obtained. This observation lends a second layer of credibility to the predicted IR data: the elbows displayed in the predicted IR match closely with what one would expect given the linear relationships observed on batches 1–3.

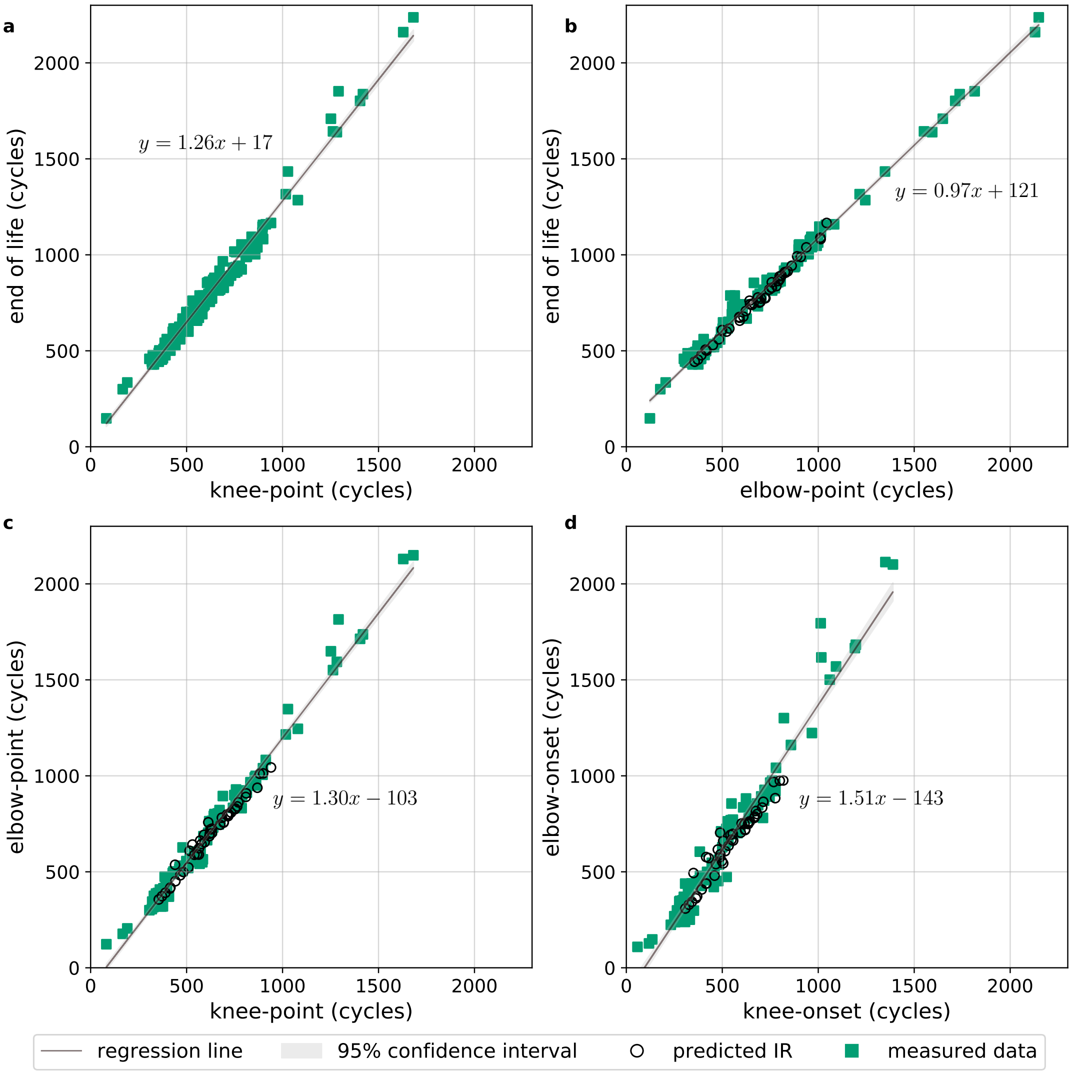


Fig. 10.6: (**a**), Linear regression model linking the knee-point to end of life. (**b**), elbow-point to end of life. (**c**), knee-point to elbow-point. (**d**), knee-onset to elbow-onset. Every linear model is presented with a 95% confidence band on the plotted regression line; all linear relations here are calculated from the A123 dataset enriched with the predicted IR data for batch 8. Elbow points derived from the predicted IR data are highlighted as open black circles; the reader will appreciate that their inclusion did not significantly influence the linear regression results obtained.

Table 10.4: Coefficients of four linear regression models relating the knee-point (**a**) and the elbow-point (**b**) to the end of life, the knee-point to the elbow-point (**c**) and the knee-onset to elbow-onset (**d**), respectively. The p-values for β_1 were computed using the Wald test, and the small values allow the rejection of the null hypothesis that a linear relationship does not exist. The 95% confidence intervals for the estimated coefficients are calculated via bootstrapping. The coefficient of determination, R^2, of these linear regression models is (**a**) 0.9822, (**b**) 0.9896, (**c**) 0.9818 and (**d**) 0.9520; all close to 1, showing that the fitted models explain the observed data well.

(a) **Knee-Point to EOL**			(b) **Elbow-Point to EOL**		
Coefficient	Estimate	*p*-value	Coefficient	Estimate	*p*-value
Intercept (β_0)	17 ± 21		Intercept (β_0)	121 ± 11	
Slope (β_1)	1.26 ± 0.04	4.0×10^{-148}	Slope (β_1)	0.97 ± 0.02	4.5×10^{-162}
EOL $= 1.26 \times \text{knee-point} + 17$			EOL $= 0.97 \times \text{elbow-point} + 121$		
(c) **Knee-Point to Elbow-Point**			(d) **Knee-Onset to Elbow-Onset**		
Coefficient	Estimate	*p*-value	Coefficient	Estimate	*p*-value
Intercept (β_0)	-103 ± 28		Intercept (β_0)	-143 ± 42	
Slope (β_1)	1.30 ± 0.05	3.6×10^{-147}	Slope (β_1)	1.51 ± 0.08	4.5×10^{-112}
elbow-point $= 1.30 \times \text{knee-point} - 103$			elbow-onset $= 1.51 \times \text{knee-onset} - 143$		

10.4 Early Prediction of Elbows

A real-word challenge is how to predict the trajectory of IR growth, e.g. the elbow points in IR curves as to detect early signs of unacceptable degradation. For example, to filter out cell production lots that will exhibit faster increases in IR or to schedule HEV battery replacement/maintenance. We complement the previous section in scope of the findings of [78] (Section 3). We apply the quantitative knee prediction algorithm developed there to the early prediction of elbows without any additional optimisation, i.e. 'as is'. A full description of the model and feature extraction process can be found in [78] and supplementary material; however, we provide a brief overview. It is outside the scope of this paper to revisit the early prediction of knees.

The quantitative prediction of the elbows is performed by a RVM [20], a type of linear regression mechanism, taking features extracted from the early life of the cells. The feature extraction process takes as input the first 50-cycles of the available per-cycle and in-cycle measurements (capacity, IR, charging-times, voltage, current, temperature) and draws on time-series analysis to calculate a vast collection of summary statistics without input from domain expertise (see [78] Supplementary Figure 5). Then, a sequential feature selection funnel is deployed to select around 100 features to train the RVM [78] (Supplementary Figures 6 and 7). When using batch 8 the input IR is the predicted IR from Section 10.2.1 – the cases with/without batch 8 are distinguished. The model is trained on data from all but one cell and tested on the remaining cells (leave-one-out framework); this process is independently repeated such that each cell is used for testing once. The performance metrics displayed in Table 10.5 are the average of the test performances.

The resultant early predictions are reported in Table 10.5, where two points should be made salient. Firstly, on elbows vs. knees prediction, when compared to [78], the model performs worse predicting elbows than when predicting knees: MAPE 13.8% vs. 12.0%, elbow-onset vs. knee-onset, and MAPE 10.7% vs. 9.4%, elbow-point vs. knee-point – overall, the elbow prediction is up to 2% worse when compared with knee prediction. This lower accuracy in elbow (vs. knee) prediction was expected as the input IR measurements are much noisier than the capacity measurements, and hence, the identification of elbows is inherently less exact, which in turn affects the predictive performance – as argued in Section 10.3.1, the confidence intervals for the elbow identification are significantly wider than those for the knees. Due to this higher noise in the elbows, when predicting elbows from input data, the relationship between input data and elbows will be weaker/noisier than when predicting the knees.

Secondly, the inclusion of the predicted IR data leads to a marginally worse average performance of our model: the MAPE worsens by 0.2% for the elbow-onset prediction and by less than 0.8% for the elbow-point prediction, see Table 10.5. This critically showcases that the generated IR data may be used for the prediction of elbows – which we emphasise was an input feature to the RVM.

Table 10.5: Result of RVM regressor for elbow-onset (**a**) and elbow-point (**b**) when predictions are made from the first 50 cycles. The 90% confidence intervals (CI) were calculated via bootstrapping. The entry 'With b8?' refers to results computed with ('Yes') and without ('No') the inclusion of the artificially predicted IR data of batch 8.

(a) Elbow-Onset Prediction				**(b) Elbow-Point Prediction**			
With b8?	Metric	Score	CI ($\alpha = 0.1$)	With b8?	Metric	Score	CI ($\alpha = 0.1$)
No	MAE (cycles)	89.1	[77.0, 101.8]	No	MAE (cycles)	76.3	[64.5, 88.6]
	MAPE (%)	13.8	[12.4, 15.3]		MAPE (%)	10.7	[9.5, 12.0]
Yes	MAE (cycles)	91.3	[79.4, 104.0]	Yes	MAE (cycles)	83.4	[72.8, 94.6]
	MAPE (%)	14.0	[12.6, 15.5]		MAPE (%)	11.5	[10.4, 12.8]

From a methodological point of view, we employed the simple RVM algorithm of [78] in a direct manner: without any additional optimisation to take into account the noisier IR data or the predicted IR data. This was a choice to prove that the generated IR data can be used for early prediction. There is indeed room for future improvements in the early prediction of IR elbows and such is left for future research. Lastly, increasing the number of cells displaying elbows by prediction to 169 (= 124 + 45) will benefit approaches that are highly dependent on the size of a dataset.

10.5 Conclusions and Future Work

In this original work, the IR rise curve of Li-ion cells is characterised by the novel concept of ‘elbow-point’ and ‘elbow-onset’; a generalist identification algorithm is then proposed. The proposed approach is able to handle not only measurement noises but also sigmoid-type patterns in capacity fade and IR rise curves. The findings highlight significant linear relationships between EOL, capacity knee-point/IR elbow-point and capacity knee-onset/IR elbow-onset for the data under study.

Two machine learning-related goals were achieved. The first, part of the data pre-processing step, draws on neural network techniques to build independent IR and capacity SOH predictors achieving a small MAPE of 1.6% and 0.4% respectively – these results are of wider general interest. The proposed IR estimator has been deployed to complete an existing cell cycling dataset with missing IR measurements resulting in a well-rounded life cycle dataset encompassing both capacity and IR data. The generated data is publicly available. Such datasets can be used for both identification and the early prediction of elbows in IR curves. We then provided an illustrative example for such an early predictor of IR elbows. Furthermore, the cells with predicted IR are shown to be usable for the early prediction of elbows: resulting in only slightly worse average performance than when they are excluded (the MAPE worsens by less than 0.8%).

The methods of elbow identification and prediction, in this work, have commercial value to battery manufacturers as well as end users such as fleet managers and energy storage utility operators. Accurate early forecasting of the IR elbows will allow manufacturers to set appropriate performance and lifetime warranties for their products. Additionally, elbow forecasting allows battery users to accurately and conveniently schedule battery maintenance and replacement, or adjust the duty cycle to accommodate the reduced performance of the battery pack as it degrades.

In the future, the accuracy of the early prediction will be enhanced. Multiple dimensions of inputs encompassing the predicted IR data and other measurements will be used to train the model with an improved tolerance for noisy data. Overall, elbow identification and elbow early prediction can be used to influence the design of the thermal management system: accounting for the additional heat dissipated by cells as they approach their EOL. A study comparing the relations between knee/elbow-onset and -point across more datasets is left to future work.

Chapter 11

One cycle prediction

The work presented in this chapter is taken from our paper [208]. We work with the datasets of Seversion [197] and Attia [6] (from the TRI see Section 9.2.1 of Chapter 9). In addition, we make use of the synthetic IR data developed in Chapter 10.

Abstract

There is a large demand for models able to predict the future capacity retention and internal resistance (IR) of Lithium-ion battery cells with as little testing as possible. We provide a data-centric model accurately predicting a cell's entire capacity and IR trajectory from one single cycle of input data. This represents a significant reduction in the amount of input data needed over previous works. Our approach characterises the capacity and IR curve through a small number of key points, which, once predicted and interpolated, describe the full curve. With this approach the remaining useful life is predicted with an 8.6% mean absolute percentage error when the input-cycle is within the first 100 cycles.

11.1 Introduction

Sales of electric vehicles and energy storage systems are undergoing a marked growth as battery costs continue to fall and governments around the world introduce increasingly strict emissions regulations.

Of importance to all applications is a cell's state-of-health (SOH). In many applications the key metric for cell health is capacity retention. In this regard, SOH is often interpreted as the current capacity of a cell as a percentage of its rated capacity. As the capacity degrades over time so does the cell's usefulness eventually reaching a point at which the cell is no longer deemed useful for its current application. This point, called the end-of-life (EOL), is often a predefined capacity level. Another key health indicator is the internal resistance (IR) of the cell: as the cell degrades its IR increases, impairing the cell's ability to provide and receive charge. Capacity degradation and IR rise of a Li- ion cell are often not linear throughout its lifetime [78, 210]. Cell capacity typically starts to degrade in a linear manner until reaching a critical point, called the ‘knee’ (referred to henceforth as the knee-point), at which the rate of capacity degradation increases considerably [55, 69, 143]. In [78] the additional variable ‘knee-onset’ is introduced (along with an alternative identification mechanism) to provide a useful indication of the start of this rate of increased degradation. Building on this idea, [210] introduced the variables ‘elbow-onset’ and ‘elbow-point’ describing the same phenomena as the knee-onset and -point but for the IR rise curve. Accurate identification and prediction of the occurrence of knee-onset and -point can provide essential guidance for scheduling of replacements and cell maintenance to prolong service life. However, knee-points (and knee-onset) may appear before or after the EOL is reached and their occurrences are also cell chemistry dependent [45]. The same holds for elbow-onsets and points. Other degradation metrics, such as the remaining-useful-life (RUL) or the whole capacity trajectory, thus have to be considered collaboratively for a comprehensive view.

Much research has been dedicated to the modelling of Li-ion cells and, in particular, lifetime prediction such as EOL and RUL. Broadly, there are two approaches to this problem either *model-based* or *data-centric*. The model-based approach encapsulates empirical models, Equivalent circuit models and physics-based models. It includes electrochemical type models where the cell’s internal physical degradation mechanisms are simulated (see [226] for a review), and parametric/semi-parametric type models where empirical models are fitted to realised capacity fade curves and combined with advanced filtering techniques to predict future degradation [37, 179]. The data-centric approach consists of machine learning and statistical models trained on in-cycle and cycle-to-cycle measurement data such as voltage, current, capacity, temperature and internal resistance. Feature based approaches allow for expert input on essential features [66, 173, 178, 197] but may also take a purely data-driven feature selection approach. Feature free approaches use deep learning techniques such as Convolutional neural networks (CNN) to process ‘raw’ cycle data. The data-centric approach typically requires larger data sets for training than model-based approaches, nonetheless, this approach is showing great potential [42, 144, 233]. Physics-informed models need to be calibrated to cell data (a non-trivial problem) and if too simplified lose the inherent conservation laws. The computational requirements of semi-parametric type modelling are much lower than those of electrochemical-models or the data-centric approach. For both data-centric and semi-parametric modelling, the available data for training/calibration is a methodological limitation in itself, be it on the quality of fitting, extrapolation or simply the method one can use.

Research on data-driven modelling to address lifetime modelling of Li-ion cells has mainly focused on the prediction of EOL or RUL [78, 197] in contrast the literature on the prediction of the complete capacity trajectory is sparse. We point notably to [214] who make use of a simple feed forward neural network to enhance the *slope and bias correction* model migration technique. At its base, this approach uses a parametric model for the capacity fade curve then a neural network is trained to migrate this fitted model from the first 30% (50-150 cycles depending on cell) of data into a prediction at a given future cycle. With this approach, they are able to accurately describe the full capacity fade curve of their 4 test cells.

Our study aims to determine the smallest number of cycles needed to accurately predict the whole capacity/IR trajectory of a cell. We found that the information contained in any one cycle of charge/discharge data suffices. This speaks directly to cell manufacturers who need to grade batteries and buyers of cells who need to test samples of purchased batches of cells for quality control.

This aim stems from several gaps in the current literature. Firstly, we address a gap that electrochemical/physics-based modelling is yet to close. Concretely, the prediction of the lifespan of a cell from a single cycle of input data. Take for instance the well-known Doyle-Fuller-Newman (DFN) model for lithium-ion batteries and its variants (see [28] for a review). Parameterising a DFN model from cycling data is impossible (much less from one single cycle) without many material assumptions: one would need stoichiometries of the two electrodes which cannot be obtained from cycle data [41]; one would need to dissemble the cell to carry specific measurements which may take around 3 months [71, 72]. Without disassembling the cell, one needs to rely on current-voltage response, for which case many of the parameters are not well identifiable – we point that work on sensitivity analysis and optimal excitation for parameter identification can be found in [153]. Secondly, to the best of our knowledge, the machine learning prediction models developed so far require gradient information for prediction. This implies longitudinal data spanning a large number of cycles, e.g., 50 to 100 cycles to predict EOL are needed [78, 197] and usually via a feature generation step. In terms of the amount of input data, the current best art for quantitative early prediction of RUL is using only 4 cycles of data achieving a 10.6 % *mean absolute percentage error* (MAPE) [103], and this result marks a non-trivial improvement over early work. However, reducing this number further would represent a further reduction in testing times and costs. Thirdly, the majority of the literature deals solely with the prediction of EOL or RUL. RUL is a key indicator of cell health and the EOL of cell quality but neither is a complete picture, both fail to capture non-linear dynamics in the capacity fade trajectory. And lastly, the vast majority of work focuses on capacity retention and ignores questions on IR degradation, both are important SOH indicators and neither is a complete picture on its own.

In fact, the 80%-capacity level for EOL is an industry postulation while the knee/elbow-point (and knee/elbow-onset) [78, 210] reflect better traceable physics/eletrochemical causal changes.

There is thus space for new approaches to predict the full capacity and IR trajectory, and to do so from a reduced number of measurements.

The main contribution of this work addresses the above four limitations from the data-driven modelling point of view. It avoids lengthy testing, the disassembling of cells and populates a space for which physics-based modelling is yet to provide an answer. The model proposed uses a CNN which jointly predicts, from a single cycle of data, the full capacity fade trajectory and the full IR rise trajectory.

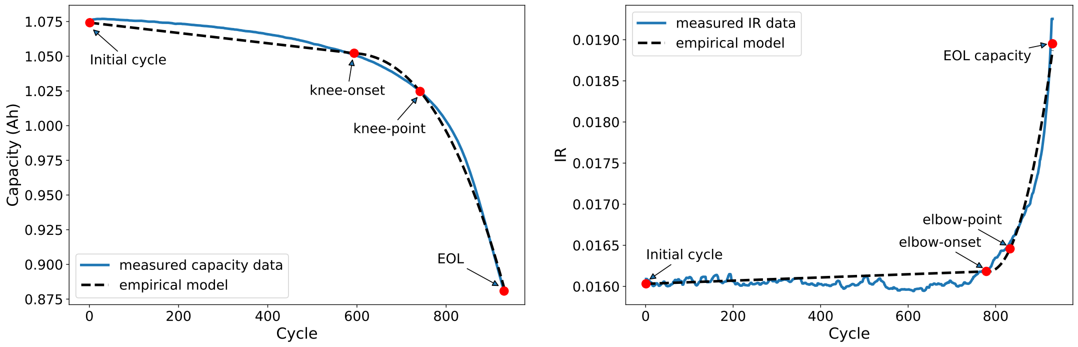
The rest of this study is organised as follows. Section 2 describes this study's datasets and Section 11.2 contains a full description of the proposed modelling approaches and insights leading to it. An account of the model's performance is given in Section 11.3 including a comparison with existing art: methods and approaches used, presented results, used features, mode of feature selection and the number of cycles used for prediction. Section 11.4 concludes this work.

11.2 Predicting future capacity and internal resistance

A question present in all battery applications is: what does the future degradation of a particular cell look like, when will a cell no longer be suitable for its current application and at what speed will this degradation occur? When solely considering cycle-ageing, ideally one would know the future capacity and internal resistance of a cell any number of cycles into the future, up to (and perhaps beyond) the EOL.

Following on from previous work [78, 210], we describe the capacity degradation by use of the knee-onset, knee-point and EOL, and the IR rise curve by the elbow-onset and elbow-point. Fermin et al. [78] proposed the use of the Bacon-Watts and double Bacon-Watts model to identify the cycle at which the knee-point and knee-onset occur, respectively. Strange et al. [210] proposed an additional smoothing process prior to deploying the Bacon-Watts models – this process involves fitting an empirical line-plus-exponential model to an isotonic regression of the data. The linear relationships between these points are also explored in the cited papers. Here, we use the second approach (with smoothing). Additionally, as we are interested in describing the full curves, we must select the capacity and IR values at the knees and elbows. Since the recorded data (in particular the IR) is noisy, we take these capacity/IR values from the smoothed (line-plus-exponential) curves and not the raw data.

We now propose a simple empirical model with which we can describe the full capacity and IR curves. In addition to the knees (for the capacity) and elbows (for the IR) we need the first and last points of both curves. As described previously, data was recorded until the cells reached 80% of their nominal capacity (~ 0.88Ah). So, we describe the capacity curve by four points: the current cycle (measured capacity), knee-onset (empirical capacity), knee-point (empirical capacity) and EOL (0.88Ah). And, we describe the IR curve by the current cycle (measured IR), elbow-onset (empirical IR), elbow-point (empirical IR) and capacity-EOL (empirical IR). Our proposed approach, assuming the current cycle is sampled ahead of the knee/elbow-onset and -point is as follows, shown in Fig. 11.1: 1) fit a cubic spline between the four points; 2) take the cubic spline as the approximation between the last three points; 3) approximate between the first two points by a straight line. Applied to the measured capacity fade curves of batches 1, 2, 3 and 8 (up to EOL), and taking cycles 1 to 100 as the 'current cycle', this model obtains a root mean square error (RMSE) of 0.0039 and a coefficient of determination (R^2) value of 0.9931, and applied to the measured IR curves it obtains a RMSE of 0.00015 and an R^2 value of 0.9838 showing a strong agreement with the measured values. Up to the onset points the degradation is linear, and thus the straight line approximation (step 3) performs well. Comparing after the knee-onset the model obtains a RMSE of 0.0046 and an R^2 value of 0.9902, and restricted to after the elbow-onset the model obtains a RMSE of 0.00017 and an R^2 value of 0.9832. So, our simple interpolation accurately describes the true curves. An example of the approximated capacity and IR curves is given in Fig. 11.1.



(a) capacity

(b) internal resistance

Fig. 11.1: Empirical model fitted to the capacity and IR curves of cell b3c12.

Then, in order to predict the entire capacity fade and IR rise curves, it is enough to have a measurement (or prediction) of the current capacity/IR value and predictions of the number of cycles until (and the remaining capacity/IR values at) the knee-onset, knee-point, EOL, elbow-onset and elbow-point. To illustrate these quantities, we predict the '*time to knee-onset*' (ttk-o), '*time to knee-point*' (ttk-p), RUL, '*time to elbow-onset*' (tte-o) and '*time to elbow-point*' (tte-p). In addition, we predict the remaining capacities at knee-onset (Q@k-o) and knee-point (Q@k-p), and the IR values at elbow-onset (IR@e-o), elbow-point (IR@e-p) and the IR at (capacity) EOL (IR@EOL). The retained capacity at EOL is known and thus does not need prediction. The empirical model described above can then be fitted through the predicted points giving a prediction of the entire capacity degradation and IR rise curves. It should be pointed out that the characteristic points we select are stylistic in nature and that not all cells display knees or elbows. However, the authors believe that the basic idea of identifying and predicting key points before fitting an empirical model should be applicable to a wide range of ageing modes. Different ageing modes may require a different stylised model (this is left to future research).

We restrict most of our discussion to 'early' prediction, here defined as the first 100 cycles of data (*initial* setting). This is a more difficult setting than prediction at later points (*full* setting) and allows for a direct comparison with previous works in the literature. Indeed, as expected, our model performs better as the cycle from which predictions are made approaches the actual cycle of the predicted quantity. For illustrative purposes, our model's performance predicting the RUL versus the distance from the EOL is presented in Fig. 11.5.

11.2.1 Modelling approach

Ideally, the testing required to make a prediction should be limited in time, making prediction fast and convenient. Thus, we restrict to the prediction from a single cycle of data. This removes the need for past knowledge of a cell, a problem faced in so-called 'second-life' applications, where the historical cycling profiles of most cells are unknown. In addition, this restriction massively multiplies the available data for training. For each cell (when restricting to the first n-cycles of data) we have n-cycles of data and n corresponding distinct values for each item we are predicting. Predicting from one cycle of data, we thus have n examples from each train/test cell. This multiplication of the training set allows us to use deep learning techniques.

We will now describe our proposed model. We take a data-driven and feature-free approach. We propose a model consisting of a convolutional 'feature extraction' block followed by two densely connected layers, displayed in Fig. 11.2 and described in Table 11.1. As output, this model can be trained to predict values jointly (*joint prediction*) or separately. As input our model takes a single cycle of voltage, current and SOC data (obtained by the coulomb counting method, from one full cycle). Our model was implemented in Python using TensorFlow via the Keras API [43]. All layer names given in Fig. 11.2 refer to the corresponding Keras layers. To assist training, batch normalisation was used before each MaxPooling1D layer. The model was trained using the *adam* optimiser for 100 epochs with a batch size of 512. And the *mean absolute error* was set as the loss function. A *learning rate scheduler* (described in Eq. (11.1)) was used during training with a 'decay rate' of 0.9 and a 'decay step' of 5 epochs.

Since the in-cycle data was not recorded at consistent time intervals or for the same number of time-steps, after cleaning, the data was interpolated and *nan*-extended to a consistent length and time-step. The interpolation was performed with the *SciPy* [109] function *interp1d* and interpolated to one measurement every four seconds. The data was then allocated at random (by cell) into an 80-20 train-test split. The training and testing sets were then

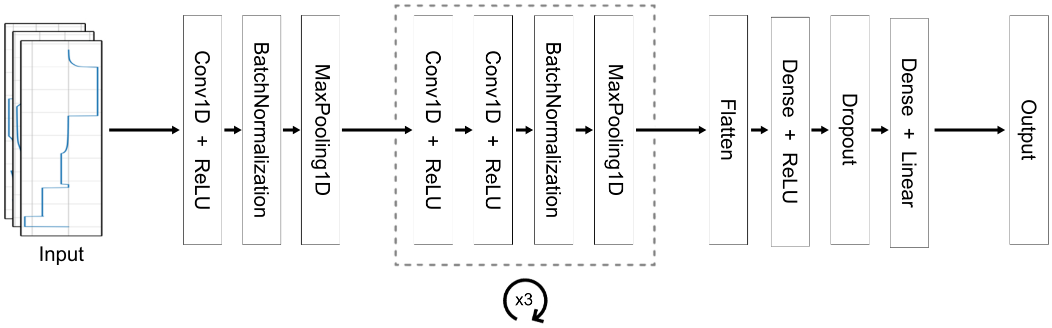


Fig. 11.2: Representation of the CNN architecture. The '×3' notation denotes three repeated blocks with the displayed configuration.

layer name	input size	hyper-parameters	output size
conv1d_1	926×3	24, 6, *ReLU*	921×24
batch_normalization_1	921×24	-	921×24
max_pooling_1	921×24	2	460×24
conv1d_2	460×24	32, 3, *ReLU*	458×32
conv1d_3	458×32	32, 3, *ReLU*	456×32
batch_normalization_2	456×32	-	456×32
max_pooling_2	456×32	2	228×32
conv1d_4	228×32	32, 3, *ReLU*	226×64
conv1d_5	226×64	32, 3, *ReLU*	224×64
batch_normalization_3	224×64	-	224×64
max_pooling_3	224×64	2	112×64
conv1d_6	112×64	32, 3, *ReLU*	110×64
conv1d_7	110×64	32, 3, *ReLU*	108×64
batch_normalization_4	108×64	-	108×64
max_pooling_4	108×64	2	54×64
flatten_1	54×64	-	3456
dense_1	3456	64, *ReLU*	64
dropout_1	64	0.4	64
dense_2	64	$n\_outputs$, *linear*	$n\_outputs$

Table 11.1: Proposed architecture of the CNN model for the prediction of ttk-o, Q@k-o, ttk-p, Q@k-p, and RUL. The model can be trained to predict multiple points at once (joint prediction) or separately. Hyper-parameters are given in the format: filters, kernel size, activation for conv1d layers; pool size for max_pooling; dropout for dropout; nodes, activation for dense layers.

restricted to the first 100 cycles of data before being standard scaled (when the model was trained in the full setting, the restriction step was dropped). Hyper parameters were optimised prior to testing using randomly selected validation subsets of cells from the training set.

When using the model to predict future capacity (Q@k-o and Q@k-p) and IR (IR@e-o, IR@e-p and IR@EOL) the model was trained to predict the *loss in capacity* and *rise in IR*, respectively. Given a measurement (or prediction) of current capacity and current IR, these can then easily be converted into predictions of future capacity and IR. The loss in capacity from the start to EOL is of the order of 0.1. Similarly, the loss in capacity to knee-onset and knee-point is quantitatively small. The IR rises are even smaller, of order 0.001. Small target values can mean small values in a model's loss function which can negatively impact training. Thus to improve performance, our model was trained to predict $10000\times$(loss in capacity) and $2000000\times$(rise in IR). These multiplications were accounted for when converting to a prediction of future capacity and IR. The multiplicative constants selected here are largely arbitrary (chosen to roughly match the range of RUL values) and their exact specification did not significantly impact model performance.

11.2.2 Prediction intervals via the forward-dropout method

Here we briefly describe the approach taken to provide prediction intervals. A simple approach is to train and predict with multiple independent copies of a model calculating prediction intervals from the independent predictions. Here by 'independent' we mean models trained separately: due to the stochastic nature of the training each trained model will provide different predictions. This approach is often referred to as an 'ensemble' approach, and this is the approach taken to produce the performance metrics displayed in this paper. However, there are notable issues which may make such an approach unattractive or unfeasible. Firstly, there is the computational cost and time associated with training a model repeatedly and independently. And secondly, there is the cost of storing multiple models in memory, which poses a particular barrier in storage limited applications such as integrated chipsets.

Another approach, superior to the ensemble approach in both aspects described (although not necessarily in terms of accuracy), is to deploy *dropout* during the forward pass of a network (forward dropout). That is, predicting with a trained model multiple times each time with a random dropout (of a pre-specified rate) applied and calculating prediction intervals from these predictions. This approach can be viewed as a Bayesian approximation of a Gaussian process [83]. The rate at which dropout is applied during prediction is optimised such that the distribution of 'residuals', from dropout prediction to the median dropout prediction, matches the distribution of residuals from the model without dropout's prediction to the true value. This optimisation can be performed prior to deployment (on a validation set), or 'on the fly' after deployment as predictions are made and then compared with realised results. This is our preferred approach when using the model to predict the full capacity fade and IR rise curves. When this approach was applied to our model the dropout layer present in Table 11.1 and Fig 11.2 was used to apply dropout in training with our selected training dropout and then in prediction with a separately optimised prediction dropout rate.

11.3 Model performance

11.3.1 Performance metrics

We now present our model's performance metrics when predicting each quantity in isolation. The figures for the prediction of ttk-o, ttk-p, RUL, IR@e-o and IR@e-p are presented in Table 11.2. For the capacity related predictions, the cycle error (MAE and RMSE) is lower for the points which are temporarily closer to the cycle from which the prediction is made: the knee-onset and knee-point. However, in percentage terms, the model performs better predicting the RUL than the ttk-o or ttk-p. This is explained by the larger target value and thus smaller percentage error for a given cycle error. The larger percentage error for the knee-onset prediction is explained by the smaller target value. The prediction of the IR related quantities is quantitatively worse than the capacity related ones; this is explained by the (much) noisier IR measurements which in turn effect the elbow identification [210, page 8]. A more granular view of the errors can be found in Fig. 11.4b (for the RUL) and Fig. 11.3.

	RMSE (cycles)		**MAE (cycles)**		**MAPE (%)**	
	Train	Test	Train	Test	Train	Test
ttk-o	38 ± 5.0	84 ± 12.0	21 ± 2.7	55 ± 6.5	4.9 ± 0.72	12.6 ± 1.44
ttk-p	41 ± 4.8	83 ± 14.4	26 ± 3.4	55 ± 6.1	4.2 ± 0.48	9.7 ± 0.94
RUL	50 ± 4.7	100 ± 19.3	32 ± 3.1	66 ± 7.8	3.8 ± 0.33	8.6 ± 0.95
tte-o	51 ± 3.7	112 ± 23.8	30 ± 2.6	71 ± 10.2	4.6 ± 0.37	11.9 ± 1.87
tte-p	50 ± 4.5	105 ± 28.3	31 ± 3.1	68 ± 11.7	4.2 ± 0.40	10.1 ± 1.54

Table 11.2: Performance of proposed model to predict Capacity's: ttk-o, ttk-p and RUL; and IR's: tte-o and tte-p.

The results of comparing the predicted Q@k-o, Q@k-p, IR@e-o, IR@e-p and IR@EOL with the values from the empirical fitted curve are presented in Table 11.3 where it is shown that our model can accurately predict these values from a single cycle of data.

	RMSE		**MAE**		**MAPE (%)**	
	Train	Test	Train	Test	Train	Test
Q@k-o	$0.0018 \pm 2.4\text{e}{-4}$	$0.0082 \pm 6.7\text{e}{-4}$	$0.0013 \pm 1.9\text{e}{-4}$	$0.0041 \pm 2.6\text{e}{-4}$	0.13 ± 0.02	0.40 ± 0.03
Q@k-p	$0.0022 \pm 2.8\text{e}{-4}$	$0.0075 \pm 4.6\text{e}{-4}$	$0.0017 \pm 2.5\text{e}{-4}$	$0.0040 \pm 2.1\text{e}{-4}$	0.17 ± 0.02	0.41 ± 0.02
IR@e-o	$5.4\text{e}{-5} \pm 6.2\text{e}{-6}$	$0.00019 \pm 2.1\text{e}{-5}$	$3.3\text{e}{-5} \pm 3.8\text{e}{-6}$	$0.00014 \pm 1.6\text{e}{-5}$	0.20 ± 0.02	0.84 ± 0.10
IR@e-p	$6.7\text{e}{-5} \pm 6.7\text{e}{-6}$	$0.00021 \pm 2.5\text{e}{-5}$	$4.4\text{e}{-5} \pm 5.0\text{e}{-6}$	$0.00015 \pm 2.1\text{e}{-5}$	0.26 ± 0.03	0.85 ± 0.12
IR@EOL	$0.00020 \pm 4.9\text{e}{-5}$	$0.00041 \pm 5.5\text{e}{-5}$	$0.00014 \pm 2.0\text{e}{-5}$	$0.00032 \pm 4.0\text{e}{-5}$	0.72 ± 0.10	1.71 ± 0.21

Table 11.3: Performance of model to predict future capacity and IR, when current capacity/IR is known.

11.3.2 Full curve prediction

We now inspect the performance of the model to predict the full capacity and IR curves. For this we trained three models to predict jointly related quantities. These three models were a 'time to' model (predicting ttk-o, ttk-p, RUL, tte-o and tte-p), a 'capacity' model (Q@k-o and Q@k-p) and an 'IR' model (IR@e-o, IR@e-p and IR@EOL). In this way we can recover the performance metrics of individual prediction and avoid issues such as the knee-point being predicted before the knee-onset. For each of these models forward dropout rates for each of their outputs were optimised by training and testing on subsets of the training data. For the 'time to' model the selected dropout rates were 0.45, 0.325, 0.35, 0.35 and 0.30 for the ttk-o, ttk-p, RUL, tte-o and tte-p, respectively; for the 'capacity' model 0.3 and 0.15, Q@k-e and Q@k-p; and, for the 'IR' model 0.75, 0.7 and 0.5, IR@e-o, IR@e-p and IR@EOL. Final models were then trained on the full training set and multiple predictions made with the selected dropout rates. Example plots of the predictions produced by this model are presented in Fig 11.3, where we present plots for 9 randomly chosen cells from the 35 test cells at random cycles from between cycle 1 and 10. The full curve prediction intervals presented in this plot were calculated by fitting our empirical model (Fig. 11.1b) to the prediction intervals calculated from the dropout predictions of knee/elbow -onsets and points, and the EOL. We do not address the impact of measurement noise on our input data. A simple approach to address this issue would be to allow a normal distribution around the measured capacity/IR with variance calibrated to the training data, or to take capacity values from several cycles as input.

It is clear from Fig. 11.3 and Fig. 11.4b that, on average, our model performs worse on cells with longer cycle lives and Fig. 11.4c shows the reason why: a significantly lower amount of training data is available for cells with an EOL above 1200 cycles.

A related problem to the prediction of RUL is the classification of cells by expected lifetime. For example, manufacturers may wish to select only the best performing cells to place in a battery pack. In the context of the Severson dataset we point to [197] who report a 4.9% test error classifying cells by 'long-lived' (EOL > 550 cycles) and 'short lived' (EOL < 550 cycles) from 5 cycles of data, and [78] who achieve an accuracy of 88% classifying the batteries into 'short' (knee-point < 500 cycles), 'medium' (knee-point between 500–1100 cycles) and 'long-range' (knee-point > 1100 cycles) from 5 cycles of data. For comparison with these results Fig. 11.4a displays a confusion matrix obtained from converting our model's cycle predictions into three classes 'short' (EOL < 550 cycles), 'medium' (550-1200 cycles) and 'long' lived (> 1200 cycles). We see that our model achieves a comparable level of accuracy, while performing the classification from a single cycle of data. This shows that, while not performing well on long-lived cells (in the regression problem), the prediction is competitive for the classification problem. The barrier of 550 matches that used in [197], and the barrier of 1200 is in line with that used in [78] (as the EOL occurs somewhat after the knee-point).

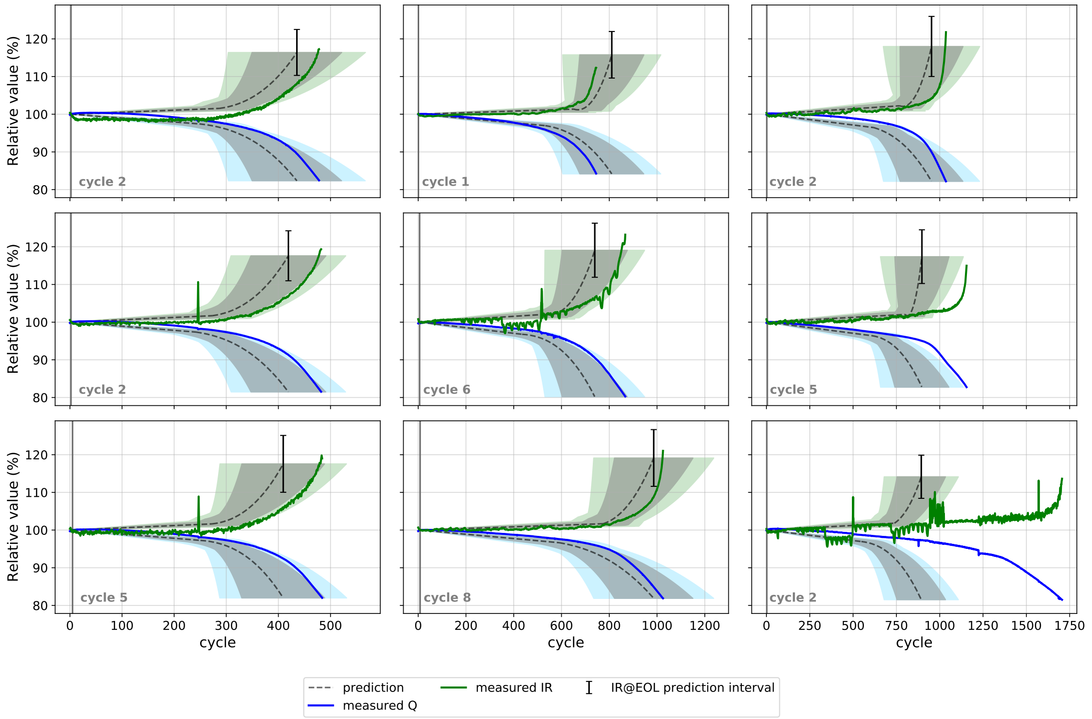


Fig. 11.3: Plots of model predictions for nine randomly chosen test cells at randomly selected input cycles. Model predictions are produced from data of a single cycle, given the measured capacity/IR at that cycle. The prediction intervals (95% and 80%) are calculated under a normality assumption from 100 predictions with forward dropout applied. Moving from left to right, and then down a row repeating, the 'test cell' displayed is b2c30, b8c20, b3c9, b2c14, b1c7, b3c34, b2c45, b3c26 and b1c4. Here, we present the IR and capacity values relative to the values measured at the input cycle.

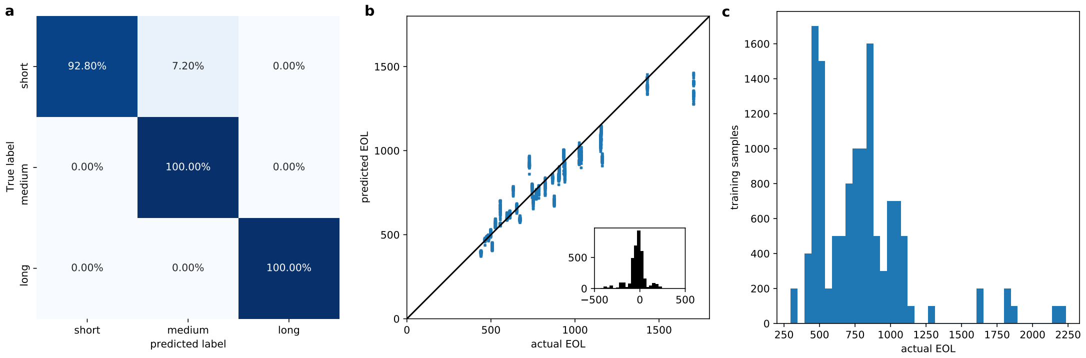


Fig. 11.4: a. Confusion matrix for prediction of EOL classes 'short life' (EOL < 550), 'medium life' ($550 <$ EOL < 1200) and 'long life' (EOL > 1200), based on quantitative prediction for test set. **b.** Actual EOL vs predicted EOL with distribution of errors, showing a slight bias to under-predict EOL especially for longer lived cells. **c.** Distribution of training data for EOL showing a significantly lower amount of training data for long lived cells.

In Fig. 11.5 we note that the model performance improves in a largely linear manner as the cycle from which the prediction was made approaches the EOL.

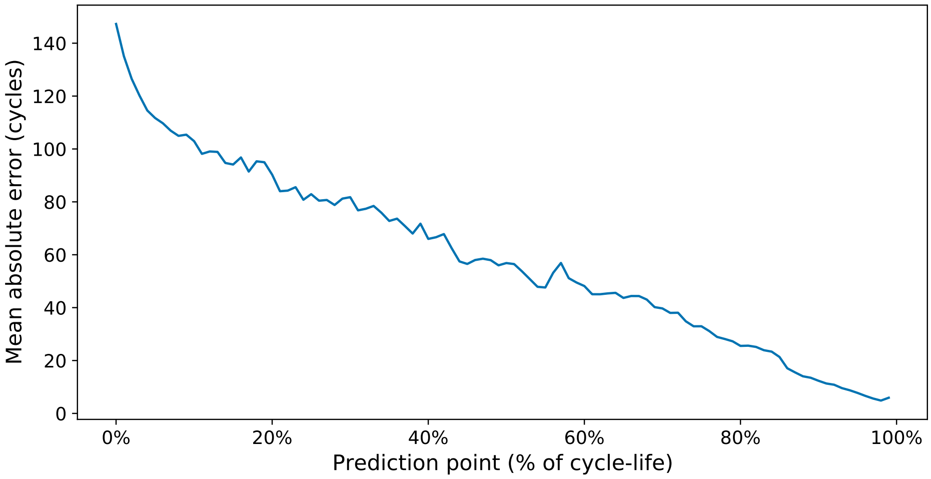


Fig. 11.5: Plot of model's performance to predict RUL in the full setting, showing a mostly linear improvement in performance as the prediction point approaches the EOL.

11.3.3 Comparison with prior art

We report a comparative study of our work to existing art utilising the Severson dataset. A review of works presenting models for prediction of EOL and RUL utilising different datasets can be found in the introduction (and references therein). Model performance may vary due to different employed datasets. In this regards, we can only meaningfully compare our results to these in terms of methodology. However, we choose to provide a quantitative comparison against literature drawing on the the same dataset analysed in this work (Severson dataset). In particular, we discuss [103, 134, 197, 199, 237] to review methods and approaches used, results presented, features selected, mode of feature selection and the number of cycles required for prediction. A summary of this comparison can be found in Table 11.4. We separate EOL and RUL prediction as these problems are quantitatively distinct.

Prediction of RUL (summarized in Table 11.4 (a))

Hong et al. [103] propose a Dilated CNN to predict the RUL. Their approach is feature free. As input their model takes 4 cycles of voltage, current and temperature data. They introduce two settings: the 'initial' setting where they restrict to data from the first 100 cycles, and the 'full' setting, where they restrict to data before EOL. They obtain a reported MAPE of 10.6% and 19.7% in the initial and full settings respectively.

In order to compare our model's performance with that of [103] in Table 11.5 we present our model's performance in the initial and full setting. Where we obtain an MAPE of 9.6% and 12.8% respectively. Outperforming the model of [103] in both settings using fewer cycles of data. We emphasise the methodological difference here to [103]: their use of 4 cycles is explicitly to capture inter-cycle cross-data correlations and temporal patterns; our results need only 1 cycle worth of information which does not contain any type of gradient information.

Prediction of EOL (summarised in Table 11.4 (b))

Severson et al. [197] predict the EOL using a Regularised Linear Model (RLM) trained on features extracted from the first 100 cycles of data. They emphasise the importance of including voltage data in their regression models, in particular capacity as a function of voltage (Q(V)); proposing three candidate models to predict EOL, all utilising features extracted from the gradients of voltage discharge curves. The best performing of these models (utilising the most features) obtains a reported MAPE of 9.1%

Ma et al. [134] propose a new 'Broad Learning-Extreme Learning Machine', this model is tested to predict the capacity and EOL on three data sets. For the Severson dataset they present results only for the prediction of EOL. Like Severson their model takes as input features extracted from the first 100 cycles of data. With this model they obtain a reported MAPE of 9%.

Shen et al. [199] predict EOL using a Relevance Vector Machine (RVM) to enhance the dataset by generating 'artificial cells' with long cycle-lives then the enhanced dataset is used to train a CNN. As input the CNN takes $\Delta Q_{100-10}(V)$, thus we consider it a 'feature based' CNN. Evaluating their model on a primary and secondary test set, they report an average MAPE of 11.7%.

In contrast to the approaches described above we take a feature extraction free approach, utilising a convolutional neural network to learn the 'optimal' features. As input our model takes a single cycle of voltage and current data, thus our model sees no gradient information. Predicting from a single cycle of data we then greatly increase the available training and testing data, enabling us to utilise deep learning techniques. The performance of our model to predict the EOL and RUL when restricted to batches 1,2 and 3 is presented in Table 11.5.

Prediction of the IR rise curve

To the authors' knowledge the only other work predicting the IR rise curve for the Severson dataset is [210]. Where a RVM was used to predict the elbow-onset and -point from the first 50 cycles of data. For the prediction of elbow-onset they achieved a MAPE of 14.0% and a MAE of 91.3, and for the elbow-point a MAPE of 11.5% and a MAE of 83.4. So we have improved on this previous work in terms of accuracy and number of input cycles.

Prediction of the entire capacity fade curve

Paper	Inputs	Cycles used	MAPE	What is predicted
This work	V, I	**1**	**8.8** % **9.6** %	EOL RUL
[103]	V, I, T	4	10.6 %	RUL
[134]	SOH, Q(V), IR, ttc	100	9.0 %	EOL
[197]	SOH, Q(V), IR, ttc, T	100	9.1 %	
[199]	Q(V)	100	11.7 %	
[237]	SOH, Q(V), V, T	250	7.0 %	

Table 11.4: Comparison of results from works using the Severson dataset. For comparison purposes, reported results exclude data from batch 8. Ordered by number of cycles used and reported MAPE. Inputs listed are in-cycle measurements of voltage (V), current (I), temperature (T), capacity as a function of voltage (Q(V)); and cycle-to-cycle measurements of capacity (SOH), internal resistance (IR) and time to charge (ttc).

	RMSE (cycles)		**MAE (cycles)**		**MAPE (%)**	
	Train	Test	Train	Test	Train	Test
EOL	55.0 ± 5.8	110 ± 24.4	33 ± 3.4	73 ± 12.4	3.5 ± 0.35	8.8 ± 1.43
RUL (initial)	55.0 ± 5.8	110 ± 24.4	33 ± 3.4	73 ± 12.4	3.7 ± 0.43	9.6 ± 1.47
RUL (full)	38 ± 2.7	99 ± 34.8	23 ± 2.5	59 ± 12.6	5.3 ± 0.47	12.8 ± 1.26

Table 11.5: Performance of proposed model to predict EOL and RUL when restricted to batches 1, 2 and 3. We report performance both in the initial setting (input cycle < 100 cycles) and the full setting.

To the authors' knowledge the only other work predicting the entire capacity fade curve for the Severson dataset can be found in Herring et al. [100]. Presenting a python library for the prognosis and cycle life prediction of Li-ion cells. As an example of their libraries performance, they predict the evolution of cell capacity for the Severson dataset. They train a multi-task linear model to predict the number of cycles until a cell reaches a range of SOH levels. This model takes as input the features in [197] covering 100 cycles of data. No performance metrics were provided.

11.4 Conclusions

The prediction of future capacity loss and IR rise is a problem of great importance. Current capacity prediction algorithms demand input data across many tens of full charge-discharge cycles to work and IR rise prediction has received little attention in the literature. In our framework, the remaining-useful-life and the entire capacity/IR trajectory are accurately predicted from a single input data cycle. This reduction entails a significant increase in prognostics procedures' affordability through reduced testing times, and stands to benefit academics and industry.

Differentiating from existing methods, we use key quantities as a dimension reduced description of the capacity fade and IR rise curve, which combined with an empirical model describe the full curves. Regarding model selection and simplification, we effectively demonstrate that gradient information is not required for the prediction of future capacity degradation. To the best of our knowledge this is in stark contrast to all previous work in this domain, which explicitly or implicitly require gradient information for prediction. Lastly, our model shows competitive performance compared with prior art, demonstrating the power of deep learning unlocked by considering each data cycle individually.

In terms of future work, the methodology we present can be deployed to electrochemical impedance spectroscopy (EIS) data which, a priori, is easier to gather.

Methods

Learning rate scheduler starting from the default Keras learning rate, the learning rate scheduler updates the learning every 'decay step' number of epochs as described in Eq. (11.1)

$$\text{new learning rate} = \text{previous learning rate} \times \text{decay rate}\ . \tag{11.1}$$

Chapter 12

Sample sizes required to understand cell-to-cell variability

The work contained in this section is from our paper [209], which was a joint work with Dr. Michael Allerhand, Dr. Philipp Dechent and Prof. Gonçalo dos Reis.

Abstract

The testing of battery cells is an expensive and long process, and hence understanding how large a test set needs to be is very useful. This work proposes an automated methodology to estimate the smallest sample size of cells required to capture the cell-to-cell variability seen in a larger population. We define cell-to-cell variation based on the slopes of a linear regression model applied to capacity fade curves. Our methodology determines a sample size which estimates this variability within user specified requirements on precision and confidence. The sample size is found using the distributional properties of the slopes under a normality assumption. The implementation is available on GitHub.

For the five datasets in the study, we find that a sample size of 8-10 cells (at a prespecified precision and confidence) captures the cell-to-cell variability of the larger datasets. We show that prior testing knowledge can be leveraged with machine learning models to operationally optimise the design of new cell-testing leading up to a 75% reduction in experimental costs.

12.1 Introduction

Current lithium-ion cells do not yet meet some applications' required power and energy densities. Therefore, research on new materials is dynamic, and ageing behaviour is an essential part of the evaluation. Furthermore, new applications for batteries with particular requirements - such as the electrification of ships, trucks [81], aircraft [204], tractors and construction machinery - often bring new load profiles, which will most likely induce a different ageing behaviour. Therefore, adapted test and evaluation methods are necessary for a reliable lifetime prediction. Batteries are highly complex systems with physical, chemical and electrical effects taking place simultaneously requiring a lot more effort to accurately model these effects themselves. And thus, in the short and medium term data-driven and empirical models will be used. These methods include diagnostic methods such as deep learning or neural networks based on systematically generated data from accelerated ageing tests in the laboratory [122].

The ageing of lithium-ion batteries depends on the complex interaction of numerous stress factors such as current rate and temperature, which necessitates an extensive test matrix. In addition, the transfer of test results to new batteries with varying materials and dimensions to create models for new cells is very limited. Currently, complex testing is carried out on a small scale on random samples due to the lack of testing resources. This limits the scope of a test regarding the number of different stress factors, the resolution of the influence, and the statistical aspects of cell-to-cell variation.

The ageing is primarily noticeable to the user as lower capacity and thus shorter operating time [19]. Many different stress factors need to be considered for ageing prediction and testing. These factors are, e.g. temperature, storage voltages for calendar ageing as well as cycle depth, state of charge (SoC) range, mechanical pressure, current rate and charge throughput for (charge/discharge) cycle ageing [70, 195].

Ageing takes place in all components of a battery, not only in the electrodes and electrolyte, but also in the casing and separator [4]. The mentioned stress factors influence ageing in electrodes and electrolytes. For example, the dissolution of electrolyte and binder as well as the reduction of the active surface in anodes are accelerated by high temperature and high state-of-charge. These ageing effects lead to capacity and power losses. In contrast, low temperature and high current accelerate the deposition of metallic lithium on the anode surface [227]. Fast development cycles only allow short testing periods, but the longevity under multiple scenarios must also be guaranteed. Therefore, ageing prediction with accelerated ageing is possible and necessary [65]. For a meaningful acceleration of the lifetime tests, the intensified ageing conditions should not trigger additional ageing mechanisms (e.g. lithium deposition) and the share of irreversible ageing and reversible capacity effects from the inhomogeneity of the lithium distribution and the anode overhang must be separated [119]. Studies combining ageing tests, cell-to-cell variation and post-mortem analyses to investigate the effects of ageing in the material over time require a high number of cells to be aged under comparable conditions and investigated in post-mortem analyses.

The fundamental patterns of how these factors influence ageing are known, and there are already investigations on this subject [19]. The challenge, however, is that each cell type is different, and thus the impact of those factors varies. This means that every new cell has to be investigated in extensive tests to be able to estimate the long-term behaviour. All these measures require very large test capacities. Furthermore, massive parallel tests are necessary to obtain results in a shorter time compared to sequential tests. Naturally, the goal is to test as efficiently and as little as possible and still achieve solid predictability. To uphold high utilisation, channels should be used in succession, testing new stress factors on a new cell when channels become available. In addition, batteries already start to age at the time of production [21]. Therefore, cells that enter a test at different times may already behave differently. Cells, therefore, need to be stored with minimal degradation at low temperatures and medium-low state-of-charge levels.

Furthermore, there are variations between individual cells of the same cell type [13]. They can be attributed to the tolerances in the production and cannot be avoided. Thus, it is not sufficient to test one cell, but all tests must be repeated with multiple cells. In a recent study by the 3rd author, it was shown that more cells should be tested to accurately capture variability than what is typically done today [49]. So far, publications of Design-of-Experiment include only either stress factors [162, 194] or cell-to-cell variation [51, 190]. And, feedback-based experiments include only a minimal design space of stress factors and extremely accelerated ageing (around 30 days of testing per cell) [6]. Therefore, the published posterior methods cannot be used on ageing tests aimed at predicting lifetime at up to 10-15 years of operation.

12.2 Design of a sequential analysis

All of the above discussed aspects render the testing of batteries very costly. Therefore, it is crucial to consider which stress factors of the measuring matrix and what number of cells are necessary for the intended purpose and how to adjust the design of the experiment during the test phase to incorporate knowledge gained on the fly and in a feedback loop for additional tests.

Battery degradation prediction is also limited by the amount of data available for either creating empirical models or parameterising physics based or data-based models. Furthermore, due to the vast parameter space of stress factors influencing battery degradation, tests can only provide meaningful data when those stress factors are consistently considered.

When a test is finished, and the end-of-life of the cell is reached, the testing equipment is freed for further use, and a question arises: Should you do more of the same testing or consider different stress factors, in order to get the maximum information in a given time? While conducting an ageing study, the result of the study can change with additional tests.

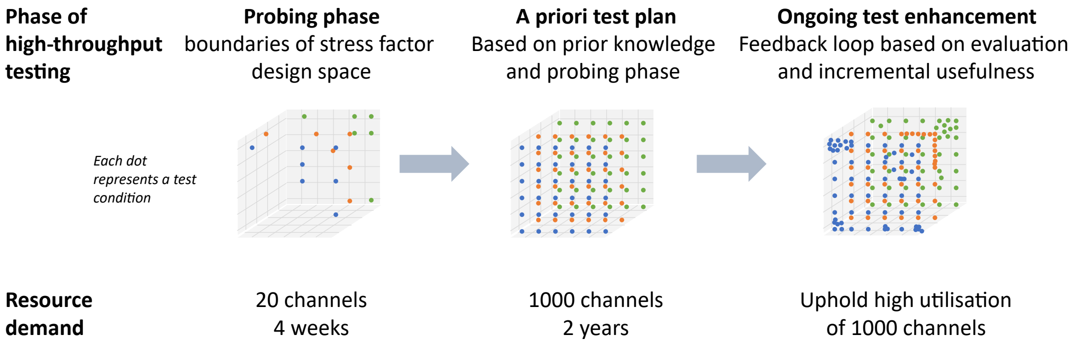

Phase of high-throughput testing
Probing phase
boundaries of stress factor design space
A priori test plan
Based on prior knowledge and probing phase
Ongoing test enhancement
Feedback loop based on evaluation and incremental usefulness
Each dot represents a test condition
Resource demand
20 channels
4 weeks
1000 channels
2 years
Uphold high utilisation of 1000 channels

Fig. 12.1: Sketch of high-throughput test example using 1000 channels and different phases of the design of experiment and continuous test enhancement.

Figure 12.1 shows the idea of the underlying testing concept as an example in this work. Each dot represents a test condition with the stress factors as the axis of the design space. First, within a probing phase, only a few cells (20 in this example) are tested to identify the boundaries of the stress factors under investigation for the given cell. The aim is to collect this data within a short time – for example 4 weeks. Then, based on this data and additional prior knowledge transfer from previous ageing tests, an a priori test plan is created and rolled out on a massive test infrastructure. In this second phase, individual stress factors are tested on 1000 channels parallel. Finally, in the last phase, channels become available due to cells degrading faster with some stress factors, additional cells are tested to increase the data available at areas of interest with the most amount of information gained at those conditions. The second and third stages overlap and will continue for as long as the equipment is available or a sufficiently high accuracy and diversity of the measuring data is reached. This can be up to 2-3 years of testing.

Figure 12.2 shows a number of channels and their usage over time. Images a) and b) show example decisions made after one year based on an automatic usefulness calculation: in a) different testing is given priority, while in b) the calculation showed more tests were necessary for the same testing condition since the desired level of confidence was not yet reached.

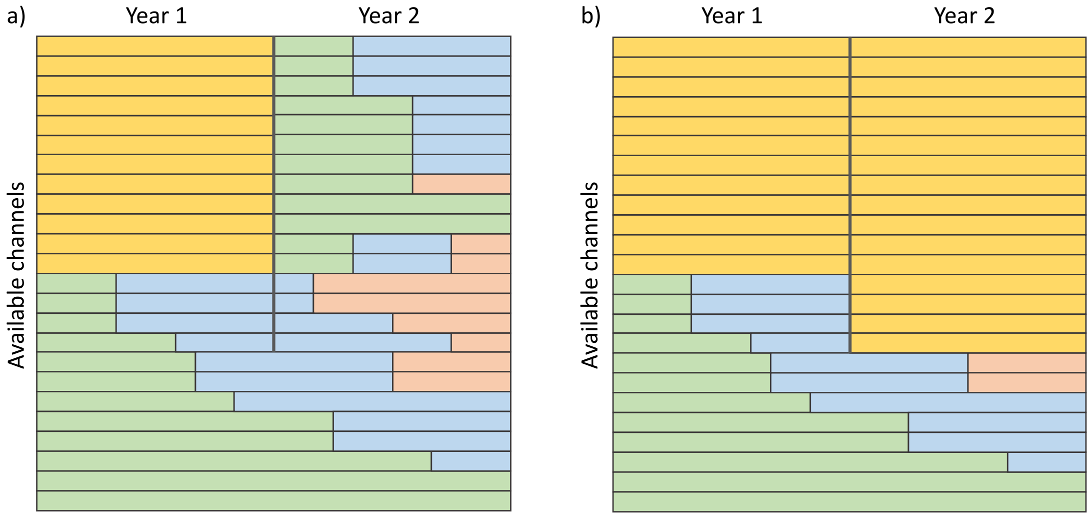


Fig. 12.2: Example for an ageing test plan for 2 years with 24 channels available. Each row represents a channel and a) and b) describe two scenarios with a change of the test plan after one year depending of having tested "enough" to capture variability. Yellow denotes channels used to capture the cell-to-cell variability, green shows the initial run of other stress factors, and blue and red show sequential tests on the same channels. In a) after 1 year enough cells have been tested under the same conditions, so additional test conditions can be tested with the available channels. In b) the amount of data collected is not yet enough and additional tests are started with the same ageing conditions.

12.3 Datasets overview

In terms of data for this work, we use the same data as [49, Section 2.1] but with novel techniques. For ease of comparison we adopt their nomenclatures. The description next follows closely that in [49, Section 2.1]. For a general overview of publicly available battery data, see [58]. The datasets were chosen for study based on the necessity of testing as many cells as possible within each dataset. All datasets are open source. Each dataset features a single type of commercially available Li-ion cells, however the manufacturers, chemistries, and cell sizes vary from one dataset to the next. Although the methods outlined below can be applied to different form factors, all datasets used 18650 cylindrical cells. Some datasets had identical experimental settings, meaning that each cell was tested in the same manner, whereas others changed the stress factors somewhat beyond the expected uncontrollable experimental variability. The datasets are as follows, and notation-wise we reserve the letter N to denote the total amount of cells in a dataset.

Baumhöfer 2014	48 cells, Sanyo/Panasonic UR18650E, NMC/graphite, 1.85 Ah
Dechent-2020	22 cells, Samsung INR18650-35E, NCA/graphite, 3.5 Ah
Dechent-2017	21 cells, Samsung NR18650-15 L1, NMC/graphite, 1.5 Ah
Severson-2019	67 out of 124 cells, A123 APR18650 M1 A, LFP/graphite, 1.1 Ah
Attia-2020	45 cells, A123 APR18650 M1 A, LFP/graphite, 1.1 Ah
Attia-predicted	45 cells, Predicted data for Attia-2020 using model proposed in [208].

The capacity fade curves in Baumhöfer-2014, Severson-2019 and Attia-2020 (also Attia-predicted) exhibit the so-called *Knee* phenomena of rapid non-linear degradation [5, 78]. The Dechent-2017 and Dechent-2020 contain linear capacity fade trajectories over time. The capacity fade trajectories (y-axis) plotted against time (x-axis) can be found in Figure 12.3 below. For all the datasets, the capacity is normalised to the nominal capacity, and hence, expressed as a percentage – we work with state of health (SOH).

The Attia-predicted dataset is data generated considering the first 20 cycles of the Attia-2020 dataset and using the one-cycle predictor model proposed in [208].

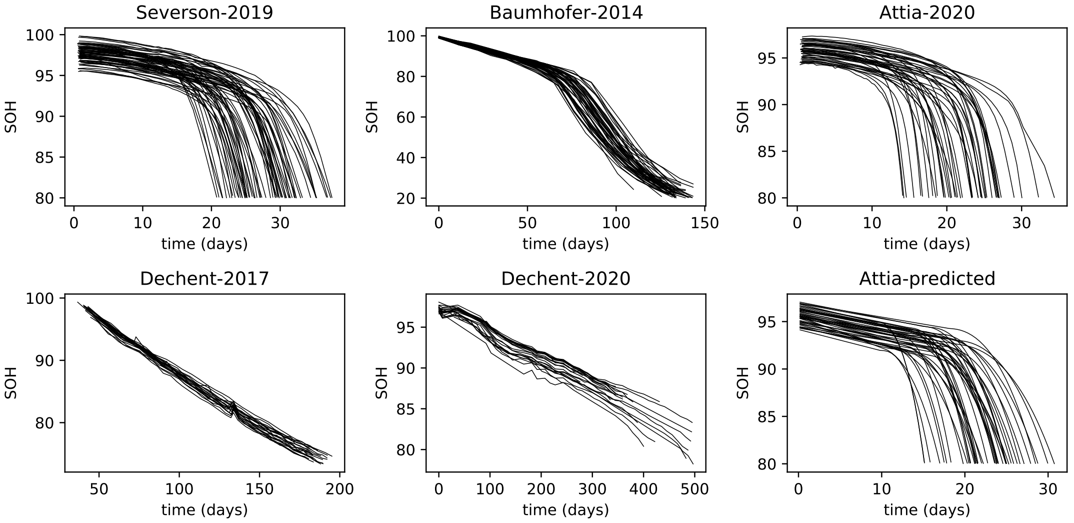


Fig. 12.3: Capacity fade trajectories (y-axis) over time (x-axis) for the six datasets mentioned in Section 12.3

A critical remark on cell-to-cell variability across datasets.

It should be noted that the datasets considered do not all share the same driving factor of variability. For Baumhöfer-2014 and Dechent-2020 (within each dataset), the cells were cycled identically in the same environment. Therefore, the variability observed in the data is essentially the intrinsic *manufacturing variability* of the cells. In contrast, Severson-2019 and Attia-2020 consider a wide range of charge protocols, and this is an additional driving factor for cell-to-cell variability in the datasets. Thus, the variability observed in these datasets is driven by intrinsic and extrinsic factors. However, as in [49], this paper works with a restricted subset of these datasets where the variability of extrinsic factors is lower – in practise, this translated into selecting only cells with a life cycle of between 23 and 40 days, and excluding *Batch 2* of Severson's [197] original dataset. Dechent-2017 also shows extrinsic and intrinsic factors, but with small differences ($< 15\%$ difference of charge current) between the tested cells. The ability to observe solely intrinsic cell-to-cell variability is, of course, experimentally dependent.

It should be emphasised that the methodology proposed in this paper is built on a certain assumption of normality (see next section). Thus it is best suited for experiments whose main source of variability is intrinsic or where the extrinsic variability is lesser. Experiments designed with large levels of extrinsic variability (as with Batch 2 of Severson's [197] dataset compared to Batches 1 and 3) may be multi-modal in nature (e.g., 50 cells tested at $-10°C$ and 50 cells at $40°C$). In such cases, for the purpose of estimating variability, one would either require a multidimensional methodology accounting for extrinsic factors or to cluster the experiment into distinct datasets where this in less of a factor. The methodology proposed in this manuscript follows this latter framework. At the end of the next section we discuss the current difficulty with the multidimensional methodology.

12.4 Methodology and estimation

The measure of cell variation for a dataset

Following Dechent et al. [49], we measure variation between cells as variation in the slopes of straight lines (12.1) fitted through the cell's repeated capacity measures,

$$\text{Model } \textit{Linear-2}: \qquad y(t) = \alpha + \beta t + \varepsilon, \tag{12.1}$$

where t is time, y is capacity, ε is a normal random variable with zero mean and finite (unknown) variance denoting the errors/residuals. The slope β and intercept α are fitted to the data (via standard least squares). Each slope β represents a cell's rate of capacity fade over successive cycles (the parameter α is discarded).

This manuscript focuses on a one-parameter model for variability and it will be shown below that the number of cells necessary to capture variability suggested by this method is already high (e.g., half the total number of cells of the Dechent datasets). More complex models could be explored with a greater availability of data. In general, our methodology can be applied to other normally distributed summary statistics. For clarification, this work improves the statistical methodology deployed by [49] for this problem and does not propose a new measure of variability. This is left for future research.

For a sample of n slopes $\{\beta_i\}_{i=1}^n$ we define the *sample mean* (denoted $\bar{\beta}_n$) and the *sample standard deviation* (denoted $\hat{\sigma}_n$) as

$$\bar{\beta}_n = \frac{1}{n}\sum_{i=1}^{n}\beta_i \quad \text{and} \quad \hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\beta_i - \bar{\beta}_n\right)^2}. \tag{12.2}$$

The sample standard deviation of the slopes β is the measure of cell-to-cell variation chosen for this work.

We make the following working assumption.

Assumption: For each given dataset, the population of slopes β is normally distributed, i.e., slopes $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$, where both the *population mean* μ_β and the *population standard deviation* σ_β are unknown parameters (that differ dataset to dataset).

Under this assumption, $\hat{\sigma}_n$ follows a Chi-distribution with $n-1$ degrees of freedom ([2]), namely

$$\frac{\sqrt{n-1}}{\sigma_\beta}\hat{\sigma}_n \sim \chi_{n-1} \ .$$

In order to specify a confidence level that $\hat{\sigma}_n$ is close to σ_β as a function of n, working with the χ_{n-1} distribution is inconvenient. Nonetheless, it turns out that as n increases the chi-distribution is well approximated by a normal distribution ([2], [139], [26]). We thus further assume that the distribution of $\hat{\sigma}_n$ in (12.2) can be approximated by a normal distribution with mean σ_β and *standard error* s_n. That is, we assume $\hat{\sigma}_n \sim \mathcal{N}(\sigma_\beta, s_n)$.

Capturing representative cell-to-cell variation for a dataset

The closeness of the estimate $\hat{\sigma}_n$ to the true value σ_β is quantified by the standard error s_n.

The number of cells required to capture cell-to-cell variation is thus given by the value of n for which s_n is small enough to ensure a given precision with a given level of confidence. However, the standard error is scale dependent, so it is instead more convenient to work with the *relative standard error* (RSE) defined as the percentage ratio of the standard error to the standard deviation

$$\mathrm{RSE} := 100\frac{s_n}{\sigma_\beta}. \tag{12.3}$$

For a concrete prospective: with a normal distribution, roughly 68% of samples are expected to fall within one standard deviation of the mean. Assuming that the sampling distribution of standard deviations is approximately normal, the RSE can thus be viewed as an upper bound on how far the sample standard deviation $\hat{\sigma}_n$ is expected to differ from the population standard deviation σ_β, with a confidence of 68% that the bound will not be exceeded.

As the RSE is defined in relation to s_n, it is quite easy to obtain confidence levels that our estimate $\hat{\sigma}_n$ will differ from σ_β by no more than any percentage level $k\%$. Simply dividing k by the measured RSE will give the number of standard errors that a deviation of $k\%$ would correspond to. And then, the number $q := -k/\mathrm{RSE}$ can be compared with the CDF of a standard normal to yield the confidence level that it will not be exceeded.

For a given sample size n the RSE can be obtained in two ways: theoretically and empirically. From the theoretical perspective, under the asymptotic regime of $n > 10$ ([2], [26]) the RSE is given by a *deterministic expression*:

$$\text{RSE} = 100\frac{1}{\sqrt{2(n-1)}} \quad \Rightarrow \quad n = 1 + \frac{1}{2}\frac{1}{\text{RSE}^2}. \tag{12.4}$$

The inversion shown above gives a sample size n which (for the reasons given above) corresponds to a confidence level of approximately 68%. The reader can compare the results this equation gives with Table 12.1.

To measure the RSE empirically, the quantities in formula (12.3) must be replaced with estimates: σ_β can be approximated by taking the empirical standard deviation of the largest available sample (the whole dataset), and, s_n by using a bootstrapping procedure to construct a distribution of sample standard deviations and then taking its standard deviation. Concretely, for a given sample size n and a number of bootstrap samples b (say $b = 1000$), sample (with replacement) b sets of n slopes. Taking the standard deviation for each set of slopes produces a distribution of b standard deviations; taking the standard deviation of this distribution gives an estimate for s_n.

Making use of these results in practice

We can now describe concretely our approach to calculate the required number of cells to maintain an accurate picture of cell-to-cell variability. Two elements must be prespecified: a maximum acceptable deviation $k\%$ for the estimate $\hat{\sigma}_n$ of σ_β and the level of confidence required that this k will not be exceeded. Firstly, the linear regression (12.1) is fitted to the capacity data giving a list of slopes. Then, using this list of slopes, for sample sizes from $n = 2$ up to the full size N of the dataset the RSE is calculated as described above - examples of the resulting values can be seen in Figure 12.4. For the acceptable deviation level $(k\%)$ the probability that it will not be exceeded is then calculated for each sample size - this is presented in Figure 12.5 for $k = 25\%$. The required sample size is then the smallest sample size providing the required confidence level. The theoretical and empirical results will not always agree and (after checking for outliers and normality as described below in relation to Table 12.4) we recommend selecting the larger of the two sample sizes.

In Section 12.5 we compare the empirical and theoretical sample sizes our methodology recommends for the datasets selected for this work. In Table 12.1 we present the theoretical required number of samples for a range of maximum acceptable deviations (relative to σ_β) and confidence levels that this maximum will not be exceeded.

For example, to obtain an estimate of standard deviation s_n that deviates from σ_β by not more than 25% at a confidence level of 68%, Table 12.1 indicates a sample of at least $n = 9$ cells (see also [26, p120] or [25, p103]).

		Maximum acceptable deviation (%)								
		5	10	15	20	25	30	35	40	50
Confidence (%)	50	92	24	12	7	5	4	3	3	2
	60	143	37	17	10	7	5	4	4	3
	68	199	51	23	14	$\boxed{9}$	7	6	5	3
	75	266	68	31	18	12	9	7	6	4
	80	330	84	38	22	15	11	8	7	5
	85	416	105	48	27	18	13	10	8	6
	90	543	137	62	35	23	17	13	10	7
	95	770	194	87	50	32	23	17	14	9
	99.7	1763	442	197	112	72	50	37	29	19

Table 12.1: Theoretical number of samples required to estimate the population standard deviation for a range of maximum acceptable deviations (relative to σ_β) and confidence levels that this maximum will not be exceeded. The 68% confidence row corresponds to the asymptotic $\mathtt{RSE}$ result of Equation (12.4) with the boxed $n = 9$ stands for $\mathtt{RSE} = 0.25$.

Comparison to prior art

The main contribution of this work, in comparison to [49], is a statistically quantified choice of the required sample size n. While the definitions chosen for variability are of the same form as there, the methodology developed for choosing the sample size is different. The methodology[1] of [49] requires the selection of a manual threshold limit for each dataset and the exact statistical meaning of this is unclear. In contrast, the parameters of maximum acceptable deviation and level of confidence used by our approach have clear statistical interpretations.

The methodology here focuses solely on a standard linear regression model and models capturing non-linear degradation are not included (e.g., the *line-exponential* model highlighted in [49] or the Bacon-Watts model [78]). The three parameters of the line-exponential model in [49] are not easily interpretable and the model suffers from a lack of robustness. Additionally, the line-exponential model requires the full longitudinal data to work well (see [5, 210]) which limits its usability in online applications as is discussed below (see Section 12.6.2).

[1] From a bird's eye perspective, both here and in [49], the starting point are models like (12.1) and a variation metric is build from their parameters. To work with the subsampled distributions, we use bootstrapping while [49] uses a hierarchical Bayesian approach. The final aspect, and the main difference of approach, requires a technical explanation. The method explained in the final paragraph of [49, Section 3] is to linearise the relationship between variation and sample size by taking logs – this tacitly assumes an unstated power curve relationship between variation and sample size – then identify a "stable region" of the linearised relationship by extrapolation from *manually* chosen points. It is not clear how this could be automated or which statistical interpretation it has. Finally they threshold deviations from the line to find the smallest sample size n. We compare sampling distributions using the $\mathtt{RSE}$ (12.3) as a general scale-invariant measure. The user then specifies a statistically interpretable and justified threshold on the percentage of $\mathtt{RSE}$ at a confidence level and the sample size is found without any further (manual) choice.

This project's implementation under CC BY 4.0 copyright license is publicly available on GitHub (see Additional Information section) as an open invitation for further testing and experimenting.

12.5 Results

For sake of exposition, this section is presented under the choice of a maximum acceptable deviation of at most 25% and a confidence level of 68%. Figure 12.4 shows the relative standard error of sample standard deviation as a function of the cell sample size, (starting with the smallest sample with any variation $n = 2$ until the total number N of samples available). The empirical estimates (shown as open circles) and the theoretical estimates (shown as lines) are obtained as described in Section 12.4.

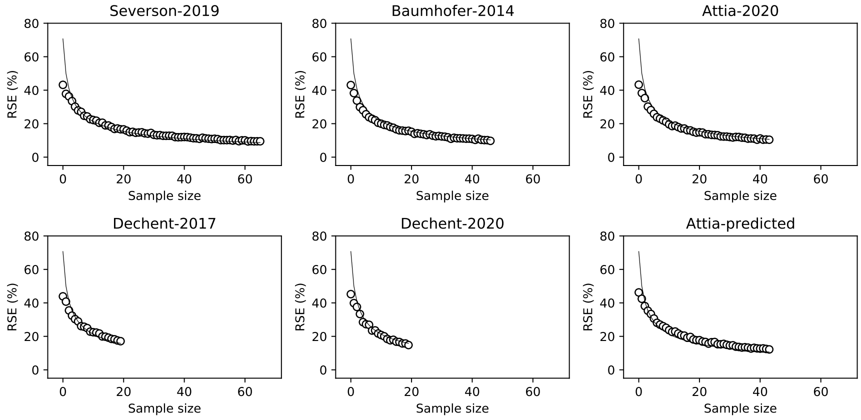


Fig. 12.4: Relative standard error (RSE) of sample standard deviation as a function of sample size. Black continuous line given by deterministic RSE asymptotic approximation of (12.4).

Figure 12.5 shows the corresponding confidence levels for a threshold on maximum acceptable deviation specified at $k = 25\%$. These confidence levels are obtained by comparison with a normal distribution as described in Section 12.4. The figure shows that the theoretical estimates are generally a close fit to the empirical estimates. Comparison with empirical data, such as the Severson-2019 data shown in Figure 12.5, shows a good agreement in that sample size $n = 9$ is the smallest sample where the percentage RSE is not more than 25% with a confidence level of 68%.

Table 12.2 shows empirical estimates of the sample size needed to estimate standard deviation that deviates from σ_β by not more than $k = 25\%$ with a confidence of 68%. The theoretical estimate is $n = 9$. Where there are differences between the theoretical and empirical results, (for example the Dechent-2017 dataset), it is most probably because the assumption of normal sampling was not met.

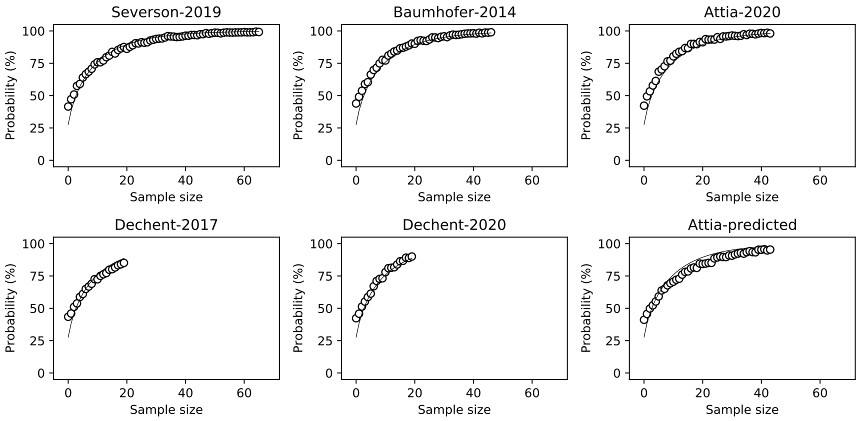


Fig. 12.5: Probability of a random sample estimate with relative error not more than 25%.

		Required Sample Size	
	N	Empirical	Theoretical
Baumhöfer-2014	48	8	9
Severson-2019	67	9	
Attia-2020	45	9	
Dechent-2017	21	10	
Dechent-2020	21	10	

Table 12.2: Sample sizes for the datasets of this study (at 25% maximum acceptable deviation and 68% confidence) per dataset, and theoretical sample size estimate (see Table 12.1). N is the total number of cells tested.

Figure 12.6 shows a Q-Q-plot graphical assessment of distribution normality of cells with Linear-2 (12.1) slopes in each dataset. There could be several reasons for departures from normality in the Dechent-2017 dataset. One possibility is simply that the dataset has too few cells, for example, both Dechent-2017 and Dechent-2020 have just 21 cells. As a general evaluation, for Severson-2019, Baumhöfer-2014 and Attia-2020, there is a very good agreement of the quantiles (large majority of samples) but there is evident left- and right-skew hinting at a non-symmetric distribution. Figure 12.3 shows that the Dechent datasets do not display capacity fade curves with knee-points.

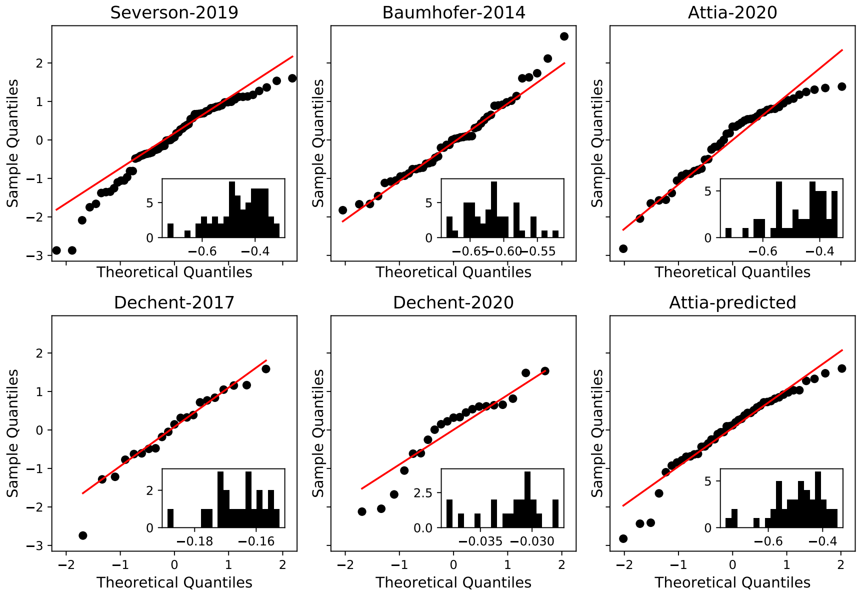


Fig. 12.6: Q-Q plots for the standardised distribution of cell slopes β. The inset histogram plots show the true (non-standardised) distribution of slopes for each dataset.

12.6 Two applied examples

12.6.1 Using prior knowledge to inform new testing

From an historical perspective, Attia-2020 [6] appears in the literature one-year after Severson-2019 [197]. Both datasets report cycling data from similar battery cells, and we argue that knowledge gleaned from Severson-2019 could have been used to inform testing for Attia-2020. This example explores this idea.

Imagine an experiment as follows: take from the public sphere the existing Severson-2019 [197] dataset and train on it the machine learning *one-cycle predictor model* of [208] (the one-cycle model is a model designed to predict the remaining capacity degradation trajectory of a Li-ion cell from any single input cycle). Then, start the cycling experiment on the cells of Attia-2020 over a short amount of time (the first 20-cycles) and, on that information, apply the [197] trained one-cycle model to build predicted trajectories for all the cells of Attia-2020. Finally, let the rest of the Attia-2020 experiment take its course. The paths of the capacity fade curves for the three datasets can be found in Figure 12.3. The methodology of the previous section is then used.

To the obtained linear model slopes of the Attia-predicted dataset apply the sample size methodology developed in Section 12.4. Table 12.3 reports the estimated sample sizes for a 25% maximum acceptable deviation and a confidence level of 68% at which a theoretical sample size is computed to be $n = 9$ (see Table 12.1). The sample size estimate for the Attia-2020 experiment is $n = 9$ while for the Attia-predicted, computed using only very early life data, is $n = 11$.

		Required Sample Size	
	N	Empirical	Theoretical
Severson-2019	67	9	9
Attia-2020	45	9	
Attia-predicted	45	11	

Table 12.3: Sample size for Attia-predicted (at 25% maximum acceptable deviation and 68% confidence) per dataset, and sample size theoretical estimate (see Table 12.1). Information on Severson-2019 and Attia-2020 kept for comparison.

We argue that this example validates the idea of using prior information to inform the design of a future experiment. From the calculations, having used $n \approx 10$ cells in the Attia-2020 experiment instead of 45 cells would have sufficed to create a representative sample of cells to capture cell-to-cell variability for that dataset. This reduction in sample size for the testing equates to a $\approx 75\%$ reduction in experimental costs and, in view of Figure 12.2, would free about 35 cell-cycler channels after just 20-cycles (time-equivalent) of testing (see the concept of Figure 12.2).

12.6.2 Generating test cases in an online format to inform larger and longer experiments

In this example, we employ the estimation procedure of Section 12.4 under a segmentation of input longitudinally across time (recall Figure 12.1 and 12.3).

Imagine an experiment as follows: a cell cycling experiment having N cells is allocated to a cell-cycler and it is to last a T amount of time (say 10 weeks). All cycling data is collected[2]. Once the experiment runs through 20% of its allocated time (two weeks), the procedure described in Section 12.4 is applied to the data available and the representative sample size, say $n_{20\%}$, is determined. Once the experiment runs through 40% of its allocated time the procedure is applied again (to all data since the beginning of the experiment) and $n_{40\%}$ is estimated. This is then repeated at increments of 20% time until the end of the experiment is reached yielding the estimates $n_{60\%}$, $n_{80\%}$ and $n_{100\%}$.

[2] We ignore the possibility of having prior knowledge of the cells, otherwise one can easily leverage the ideas of Section 12.6.1 by applying, e.g., the one-cycle model at further judiciously chosen time points.

The estimated samples sizes $n_{x\%}$ per dataset can be seen in Table 12.4 and these values need to be understood in partnership with a verification of the normality assumption underlying our methodology. This latter element is given in Figure 12.7 in the form of Q-Q plots at each stage of our theoretical experiment.

		Sample size for percentage of input				
Dataset	N	20%	40%	60%	80%	100%
Baumhöfer-2014	48	7	12	7	7	8
Severson-2019	67	45	17	16	6	9
Attia-2020	45	11	27	6	10	9
Dechent-2017	21	-	7	11	8	10
Dechent-2020	21	9	7	7	9	10
Severson-2019*	66	7	12	18	6	10

Table 12.4: Required sample size given the first $20, 40, \ldots, 100\%$ of input data (at 25% maximum acceptable deviation and 68% confidence). It should be noted that the theoretical number of required cells is 9 regardless of input size (see Table 12.1). Severson-2019* denotes the results after an outlier cell is removed from the Severson-2019 dataset (as justified below).

For both Dechent-datasets, which exhibit linear degradation fade curves (Figure 12.3), the estimated sample sizes $n_{x\%}$ are stable across the longitudinal increments in time of the curves and deviate slightly from the theoretical estimate ($n = 9$). The empty $n_{20\%}$-entry for Dechent-2017 is due to insufficient datapoints on the capacity fade curve over that time interval (see also Figure12.7). We thus suggest that data is recorded at a higher frequency.

For Attia-2020 and Severson-2019 (see Figure 12.7), there is a high variability of the data across the 20% to 60% input marks and, prominent, are the few but heavy outliers (on the left tail) that strongly influence the estimate for the sample size. Thus, the results in Table 12.4 at 20%-60% percentages of input are not inline with the theoretical result. It is also important to note the strong non-normal nature of the slope distributions around 60% for Severson-2019 and 40% for Attia-2020. For this range, cells are transitioning through the inflection point of their capacity fade curve, i.e., some cells have passed their knees (experiencing rapid capacity loss) and others are still maintaining a linear decay. The data thus display a left skew at these percent levels, violating the normality assumption. At the 80% and 100% marks, the data conforms to normality and this is reflected in the estimated n in Table 12.4 being closer to the theoretical one.

For Severson-2019 there is a cell which decays notably faster during the early life (easily identifiable in Figure 12.7 at the 20% mark on the left tail). This results in a large variability in estimates of the standard deviation. For this reason, our methodology recommends keeping a large percentage of the cells at this stage. In Table 12.4, the row Severson-2019* displays the results of our methodology after removing altogether the outlier cell (hence $N = 66$

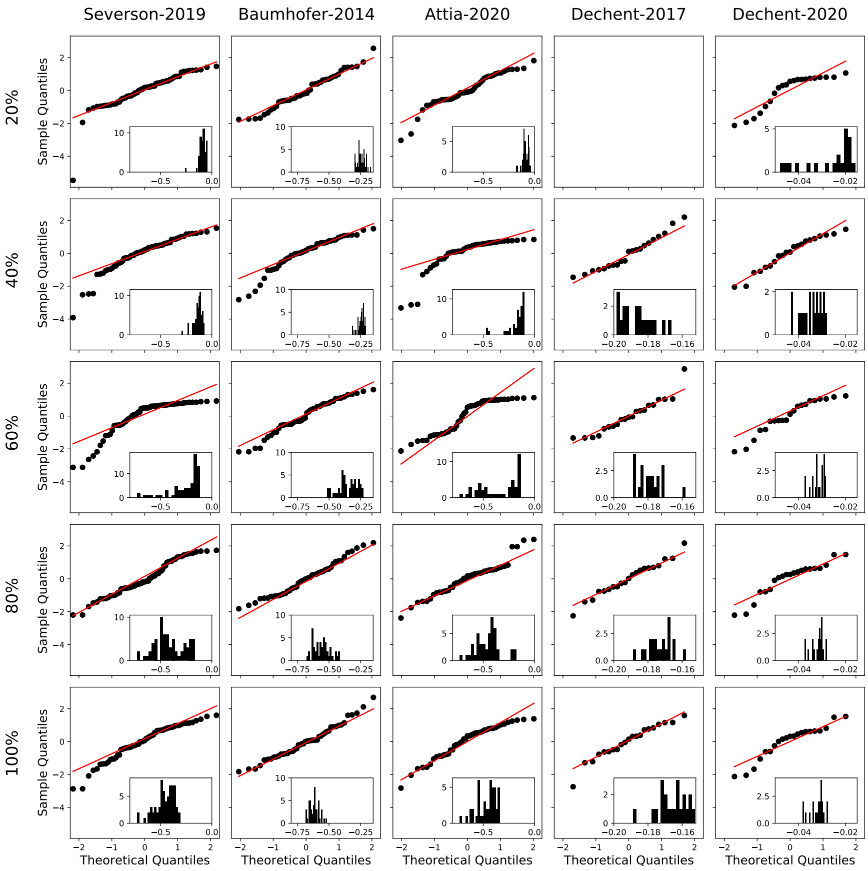


Fig. 12.7: Q-Q plots for the standardised distribution of cell slopes β across the percent input of time data as according to Table 12.4. Datasets are left-to-right and percent input from top-to-bottom starting at 20% until 100%. The inset histogram plots show the true (non-standardised) distribution of slopes for each dataset and percentage of input data.

instead of $N = 67$). The estimated sample sizes then conform to those observed for the other datasets (Attia-2020 in particular). This removal can be justified in practice as one cell degrading much faster than all others is likely to be faulty (accounting for significant differences in testing protocol).

For Baumhöfer-2014, the results follow the trend of the two Dechent datasets even though the capacity fade curves exhibit knees. This is explained by the less extreme (more gradual) nature of the knees displayed in the Baumhöfer-2014 dataset: there is no abrupt cliff (as in Severson-2019 and Attia-2020; Figure 12.3) and thus no large break from normality. We do notice some effect at the 40% level (see Figure 12.7), where there is a noticeable left skew in the data which accounts for the larger estimated value of n in Table 12.4. This effect is small in comparison to that observed for Severson-2019 and Attia-2019.

Recognisably, this example is not as conclusive as the previous one, nonetheless, it entails the critical conclusion that the underlying modelling assumption need to be verified for conclusions to be drawn. We strongly believe that this idea warrants further exploration given its potential and hope to revisit it in future research. Lastly, it is very unclear if the line-exponential model would yield better results (see discussion in Section 12.4) – improving upon this is left for future research.

12.7 Conclusions and outlook

The goal of this work was to propose a methodology to determine the smallest sample size that captures in a justified and automated way the cell-to-cell variation seen in a larger population. This manuscript improves upon the contribution of [49] by studying anew and re-thinking the underpinning statistical methodology. Under a normality assumption, that needs to be validated as part of the usage, our automatic methodology recommends a choice of $n = 9$ for a maximum acceptable deviation of 25% with a confidence level of 68%.

In future work it would be helpful to model and better explain a representative sub-population able to capture the cell-to-cell variation via the shape of the cell capacity fade trajectory. For clarity of ideas, the manuscript's focus was placed on a linear-regression method and not on models able to capture the non-linear degradation (the reason for this is argued above). As an outlook, with new larger datasets becoming available, this analysis could be performed with more complex health indicators in mind for example derived from OCV, DVA or ICA [62].

One idea for future exploration is that Internal Resistance profiles data can be included as follows: find the β^{Q} for the capacity (Q) curves according to the linear model (12.1); find the β^{IR} from the Internal Resistance (IR) curves; assume both sets β^{Q} and β^{IR} are normally distributed. Then *sum both*. I.e., define $\beta := \beta^{\mathrm{Q}} + \beta^{\mathrm{IR}}$; since the sum of Normal random variables is a normal random variable then the analysis carries through. This is contingent on Internal Resistance being included in the datasets which is often not the case [58].

Many laboratories have at their disposal large datasets across a rich test matrix where each entry has at most 3 battery cells [58, Section 2.7]. How to incorporate the findings of this work on such small datasets is still an open question – one possibility is to clump together entries of the test matrix to increase the number of available cells. If one has several of these datasets available (created at different timelines, institutions, testing machines), then how to combine them is also unknown. Critically, the message of [231] needs to be emphasised here: the metadata of test sets needs to be sufficiently complete (for instance, adding cell weights be useful for variation analysis). Otherwise, it will be difficult to credibly state that such datasets are sufficiently alike that they can be seen as an independent sample from the same statistical distribution.

Lastly and for perspective, the quantification of cell-to-cell variability is an open research topic and this work joins hands with [49] as educated first steps towards a general solution.

Funding

This project was funded by an industry-academia collaborative grant *EPSRC EP/R511687/1* awarded by *EPSRC & University of Edinburgh* program *Impact Acceleration Account (IAA)*.

G. dos Reis acknowledges support from the *Fundaçao para a Ciencia e a Tecnologia* (Portuguese Foundation for Science and Technology) through the project UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications, CMA/FCT/UNL).

P. Dechent was supported by Bundesministerium für Bildung und Forschung (BMBF 03XP0302C).

Acknowledgements

All authors thank N. Augustin, V. Inacio (both at University of Edinburgh) and S. Greenbank (University of Oxford) for the helpful comments.

Author contributions

All authors provided domain expertise, edited and reviewed the manuscript. M.A.: methodology, software, visualisation; writing - technical report. C.S..: methodology, software, data curation, visualisation; writing, review and editing. P.D.: conceptualisation, visualisation; writing, review and editing. G.d.R.: conceptualisation, supervision, funding acquisition; writing, review and editing.

Competing interest declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

This work has no supplementary information file.

The code used to produce this research is available under CC BY 4.0 at

https://github.com/calum-strange/sample_sizes_for_batteries .

Chapter 13

Conclusions and future work

The key limitation of data driven approaches is the availability of data. Throughout this thesis we have attempted to tackle that problem. In Chapter 9 we reviewed the battery data available in the public domain, seeing that while there exists a substantial volume of data it still remains somewhat limited in scope. Most available datasets focus on one cell chemistry for a set and constant cycling profile, or for a limited range of cycling profiles. Some datasets also contain data for varying temperatures, pressures, depth of discharge and calendar ageing, but limited data exists covering an extensive range of variables (see the "Test variables" column in Table 9.2). The main reason for this is cost and the fact that datasets are created with a specific goal in mind. It should also be noted that the original experimental design may not align with subsequent research. For example, the Severson dataset [197] was designed to compare different fast charging protocols, but has been used widely (including in the papers contained within this thesis) for the design of RUL models.

In the subsequent chapters we have also dealt with the issue of limited data. In Chapter 10 we introduced the novel concept of elbows in IR rise curves and dealt with the issue of missing IR data for the Attia dataset [6]. In Chapter 11 we sought to improve on the practicality of existing RUL models which required the collection of many sequential cycles of data. This would rule out their use in many important applications such as the testing of cells for second life usage. We proved that useful predictions can in fact be made from a single cycle of data. And, in Chapter 12 we looked at experimental design and at specifying how many cells should be tested in order to capture cell to cell variability.

Following the path of this thesis, we propose several areas for future research:

- **Synthetic data and extrapolation across datasets**

  The available data is specific to certain applications with experiments designed towards a specific end. As we displayed in Figure 10.2, data driven approaches may perform poorly when predicting out of sample. For this reason, it is important that the data used for training data driven models for deployment in commercial applications is representative of the real world usage. This means that the use of many data driven approaches is limited to those with access to a large quantity of high quality application specific data. The collection of such data is naturally expensive and thus may prove prohibitive for many applications. Many users will thus gravitate towards more physics based and simpler data driven approaches. One approach to solving this problem would be the generation of realistic synthetic datasets which could potentially prove fruitful for interpolating between, and supplementing, real experimental data [99, 138, 165].

- **Developing interpretable one cycle models**

  Another issue with many data driven approaches is their lack of explainability and interpretability. For example, the approach we took in Chapter 11 of using CNNs provides competitive performance from a vastly reduced number of input cycles; However, the CNN architecture means the model operates largely as a black box. There is a large body of literature in the wider Machine learning community looking at the problem of explaining black box models [22], but some would argue against using deep models in the first place [184]. An interesting area of future research could look at replicating the results we found using an interpretable model.

- **Hybrid models**

  Hybrid approaches also seek to provide interpretable models while maintaining predictive power. They do this by combining explainable models with less explainable (but more predictive) models. The explainable model could be a physics based model or an interpretable data driven model. For example, [199] combines an explainable RVM with a CNN model. This is an interesting area for future research.

- **Incorporating one cycle prediction in an online framework**

  As the one cycle model can predict from any single cycle during the lifetime of a cell, it fits naturally within an online prediction framework. The paper [207] made a start in this direction: combining a few different models in the online framework and proposing an exponential smoothing of the one cycle prediction, but there is still room for further research here.

- **Prediction with inconsistent cell usage**

  In many applications, cells undergo unpredictable and inconsistent usage; However, it is likely that some portion of the cell usage is somewhat consistent and predictable over time. This portion could be during charging (for example with consumer electronics and EVs) or discharging (for example in solid state energy storage). In such applications it makes sense to perform prognostics during the most controlled stage of cell usage. We refer to [107] as an example of a recent work predicting RUL from the consistent discharge portion of cell usage. While the dataset explored in that work (the Severson dataset [197]) contained data for cells with a range of different charging protocols these protocols were held constant over time. Future work could look at predicting cell lifetime where cells cycling is not held constant.

- **Design of experiments for second life applications**

The one cycle model we developed is "historyless" in that it doesn't see the history of the cell usage when making its prediction; However, the dataset it was tested on is not historyless, as all cells were cycled in a consistent manner from their BOL to EOL. Future experiments could look at predicting the RUL for cells with unknown history. One potential experiment could split cell usage into two phases: say with a range of cycling profiles for the first phase and one common profile for all cells in the second phase. This is designed to emulate something like second life applications for EV batteries where past usage of cells differ but all cells are likely destined for a similar second life.

- **Verifying the one cycle approach on additional datasets**

  The one cycle model should be tested further on different datasets with different experimental conditions. In [207] the one cycle approach was verified on a different dataset than the Severson-Attia dataset and this should be expanded upon.

- **Extrapolation of model prediction to different use cases**

  For a given cell with a known (or unknown) history it should be possible to predict its expected lifetime for a range of potential use cases. One potential way to achieve this would be to make a prediction using a data driven model (based on the assumption of a certain future usage) and to adjust this prediction up or down using some informed interpolation between different future use cases. This would help to address the limitations with data driven methodologies being restricted to prediction within sample.

- **Prediction from reference performance tests**

  Predictions made from data driven models should be incorporated with periodic reference performance testing of cells. To this end there is scope for research looking at making predictions for existing datasets containing data from RPT cycles.

- **Prediction from direct current pulse testing**

  It would be interesting to recast the one-cycle methodology taking as input the data associated to direct current pulse testing. As this data can be gathered in less than a minute, even predicting the EOL from pulse test data would represent a massive step forward from the one-cycle model presented in this thesis.

Bibliography

[1] Shabbir Ahmed et al. ‘Enabling fast charging - A battery technology gap assessment’. In: *Journal of Power Sources* 367 (2017), pp. 250–262.

[2] Sangtae Ahn and Jeffrey A Fessler. ‘Standard errors of mean, variance, and standard deviation estimators’. In: *EECS Department, The University of Michigan* (2003). https://web.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf, pp. 1–2.

[3] Michel André. ‘The ARTEMIS European driving cycles for measuring car pollutant emissions’. In: *Science of the total Environment* 334 (2004), pp. 73–84.

[4] Selcuk Atalay et al. ‘Theory of battery ageing in a lithium-ion battery: Capacity fade, nonlinear ageing and lifetime prediction’. In: *Journal of Power Sources* 478 (Dec. 2020), p. 229026. DOI: 10.1016/j.jpowsour.2020.229026. URL: https://doi.org/10.1016/j.jpowsour.2020.229026.

[5] Peter M Attia et al. ‘Review—”Knees” in Lithium-Ion Battery Aging Trajectories’. In: *Journal of The Electrochemical Society* (2022). URL: http://iopscience.iop.org/article/10.1149/1945-7111/ac6d13.

[6] Peter M. Attia et al. ‘Closed-loop optimization of fast-charging protocols for batteries with machine learning’. In: *Nature* 578.7795 (Feb. 2020), pp. 397–402. DOI: 10.1038/s41586-020-1994-5. URL: https://doi.org/10.1038/s41586-020-1994-5.

[7] Muratahan Aykol et al. ‘Perspective–Combining Physics and Machine Learning to Predict Battery Lifetime’. In: *Journal of The Electrochemical Society* (2021).

[8] *B. Shahrooei’s online spreadsheet of battery datasets*. URL: https://docs.google.com/spreadsheets/d/10w5yXdQtlQjTTS3BxPP233CiiBScIXecUp2OQuvJ_JI.

[9] Yun Bao, Wenbin Dong and Dian Wang. ‘Online internal resistance measurement application in lithium ion battery capacity and state of charge estimation’. In: *Energies* 11.5 (2018), p. 1073.

[10] Battery Archive. *Homepage of Battery Archive*. URL: https://www.batteryarchive.org/study_summaries.html.

[11] Battery Bits. *The Battery Software Open Source Landscape*. Feb. 2021. URL: https://medium.com/batterybits/the-battery-software-open-source-landscape-933b88957ef5.

[12] batterystandards.info. *Homepage of batterystandards.info*. Feb. 2021. URL: https://www.batterystandards.info/intro.

[13] David Beck et al. 'Inhomogeneities and Cell-to-Cell Variations in Lithium-Ion Batteries, a Review'. In: *Energies* 14.11 (June 2021), p. 3276. DOI: `10.3390/en14113276`. URL: `https://doi.org/10.3390/en14113276`.

[14] BEEP. *BEEP GitHub repository*. Aug. 2020. URL: `https://github.com/tri-amdd/beep`.

[15] Chong Bian, Huoliang He and Shunkun Yang. 'Stacked bidirectional long short-term memory networks for state-of-charge estimation of lithium-ion batteries'. In: *Energy* 191 (2020), p. 116538. ISSN: 0360-5442. DOI: `https://doi.org/10.1016/j.energy.2019.116538`. URL: `http://www.sciencedirect.com/science/article/pii/S0360544219322339`.

[16] Alexander Bills et al. *eVTOL Battery Dataset*. Mar. 2021. DOI: `10.1184/R1/14226830.v1`. URL: `https://dx.doi.org/10.1184/R1/14226830.v1`.

[17] Alexander Bills et al. 'Universal Battery Performance and Degradation Model for Electric Aircraft'. In: *arXiv preprint arXiv:2008.01527* (2020).

[18] Christoph Birkl. 'Diagnosis and prognosis of degradation in lithium-ion batteries'. PhD thesis. University of Oxford, 2017.

[19] Christoph R. Birkl et al. 'Degradation diagnostics for lithium ion cells'. In: *Journal of Power Sources* 341 (Feb. 2017), pp. 373–386. DOI: `10.1016/j.jpowsour.2016.12.011`. URL: `https://doi.org/10.1016/j.jpowsour.2016.12.011`.

[20] Christopher M. Bishop. 'Sparse Kernel Machines'. In: *Pattern recognition and machine learning*. Springer, 2006. Chap. 7, pp. 325–353.

[21] I Bloom et al. 'An accelerated calendar and cycle life study of Li-ion cells'. In: *Journal of Power Sources* 101.2 (Oct. 2001), pp. 238–247. DOI: `10.1016/s0378-7753(01)00783-2`. URL: `https://doi.org/10.1016/s0378-7753(01)00783-2`.

[22] Francesco Bodria et al. 'Benchmarking and survey of explanation methods for black box models'. In: *Data Mining and Knowledge Discovery* (2023), pp. 1–60.

[23] B. Bole, C. Kulkarni and M. Daigle. 'Randomized Battery Usage Data Set'. In: *NASA Ames Prognostics Research Center,* (2009).

[24] Brian Bole, Chetan S Kulkarni and Matthew Daigle. *Adaptation of an electrochemistry-based Li-ion battery model to account for deterioration observed under randomized use*. Tech. rep. SGT, Inc. Moffett Field United States, 2014.

[25] George E. P. Box, J. Stuart Hunter and William G. Hunter. *Statistics for experimenters*. Second. Wiley Series in Probability and Statistics. Design, innovation, and discovery. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2005, pp. xviii+633.

[26] George E. P. Box, William G. Hunter and J. Stuart Hunter. *Statistics for experimenters*. Wiley Series in Probability and Mathematical Statistics. An introduction to design, data analysis, and model building. John Wiley & Sons, New York-Chichester-Brisbane, 1978, pp. xviii+653. ISBN: 0-470-09315-7.

[27] DD Brandt. 'Driving cycle testing of electric vehicle batteries and systems'. In: *Journal of power sources* 40.1-2 (1992), pp. 73–79.

[28] Ferran Brosa Planella, Muhammad Sheikh and W. Dhammika Widanage. 'Systematic derivation and validation of a reduced thermal-electrochemical model for lithium-ion batteries using asymptotic methods'. In: *Electrochimica Acta* 388 (2021), p. 138524. ISSN: 0013-4686. DOI: `https://doi.org/10.1016/j.electacta.2021.138524`. URL: `https://www.sciencedirect.com/science/article/pii/S0013468621008148`.

[29] Damian Burzyński and Leszek Kasprzyk. 'A novel method for the modeling of the state of health of lithium-ion cells using machine learning for practical applications'. In: *Knowledge-Based Systems* 219 (2021), p. 106900.

[30] Damian Burzyński and Leszek Kasprzyk. *NMC cell 2600 mAh cyclic aging data V1*. DOI: `10.17632/k6v83s2xdm.1`. URL: `https://dx.doi.org/10.17632/k6v83s2xdm.1`.

[31] *CALCE Battery Research Group Homepage*. URL: `https://web.calce.umd.edu/batteries/data.htm`.

[32] Zachary Cameron et al. 'A battery certification testbed for small satellite missions'. In: *2015 IEEE AUTOTESTCON*. IEEE. 2015, pp. 162–168.

[33] Ian D Campbell et al. 'How Observable Is Lithium Plating? Differential Voltage Analysis to Identify and Quantify Lithium Plating Following Fast Charging of Cold Lithium-Ion Batteries'. In: *Journal of The Electrochemical Society* 166.4 (2019), A725–A739.

[34] cellpy. *cellpy GitHub repository*. Aug. 2020. URL: `https://github.com/jepegit/cellpy`.

[35] Nilotpal Chakravarti. 'Isotonic median regression: a linear programming approach'. In: *Mathematics of operations research* 14.2 (1989), pp. 303–308.

[36] Wen-Yeau Chang. 'The state of charge estimating methods for battery: A review'. In: *International Scholarly Research Notices* 2013 (2013).

[37] Yang Chang, Huajing Fang and Yong Zhang. 'A new hybrid method for the prediction of the remaining useful life of a lithium-ion battery'. In: *Applied energy* 206 (2017), pp. 1564–1578.

[38] E. Chemali et al. 'Long Short-Term Memory Networks for Accurate State-of-Charge Estimation of Li-ion Batteries'. In: *IEEE Transactions on Industrial Electronics* 65.8 (2018), pp. 6730–6739. DOI: `10.1109/TIE.2017.2787586`.

[39] Ephrem Chemali et al. 'State-of-charge estimation of Li-ion batteries using deep neural networks: A machine learning approach'. In: *Journal of Power Sources* 400 (2018), pp. 242–255.

[40] Bor-Rong Chen et al. 'A machine learning framework for early detection of lithium plating combining multiple physics-based electrochemical signatures'. In: *Cell Reports Physical Science* (2021), p. 100352.

[41] Chang-Hui Chen et al. 'Development of experimental techniques for parameterization of multi-scale lithium-ion battery models'. In: *Journal of The Electrochemical Society* 167.8 (2020), p. 080534.

[42] Chi Chen et al. 'A critical review of machine learning of energy materials'. In: *Advanced Energy Materials* 10.8 (2020), p. 1903242.

[43] François Chollet et al. 'Keras: The python deep learning library'. In: *ascl* (2018), ascl–1806.

[44] GH Cole. *A Simplified Battery Discharge Profile Based Upon the Federal Urban Driving Schedule*. Tech. rep. EG and G Idaho, Inc., Idaho Falls, ID (USA), 1988.

[45] RW Cook, LG Swan and KP Plucknett. 'Failure mode analysis of lithium ion batteries operated for low Earth orbit CubeSat applications'. In: *Journal of Energy Storage* 31 (2020), p. 101561.

[46] Creative Common. *Homepage of Creative Commons licenses*. URL: `https://creativecommons.org/licenses`.

[47] *DASHlink - Li-ion Battery Aging Datasets*. URL: `https://c3.nasa.gov/dashlink/resources/133/`.

[48] Klaas De Craemer and Khiem Trad Trad. 'Cyclic ageing with driving profile of a lithium ion battery module'. In: (Feb. 2021). DOI: `10.4121/14096567`. URL: `https://doi.org/10.4121/14096567`.

[49] Philipp Dechent et al. 'Estimation of Li-Ion Degradation Test Sample Sizes Required to Understand Cell-to-Cell Variability'. In: *Batteries & Supercaps* 4.12 (Sept. 2021), pp. 1821–1829. DOI: `10.1002/batt.202100148`. URL: `https://doi.org/10.1002/batt.202100148`.

[50] Volker L Deringer. 'Modelling and understanding battery materials with machine-learning-driven atomistic simulations'. In: *Journal of Physics: Energy* 2.4 (2020), p. 041003.

[51] Arnaud Devie, George Baure and Matthieu Dubarry. 'Intrinsic Variability in the Degradation of a Batch of Commercial 18650 Lithium-Ion Cells'. In: *Energies* 11.5 (Apr. 2018), p. 1031. DOI: `10.3390/en11051031`. URL: `https://doi.org/10.3390/en11051031`.

[52] Weiping Diao. *Data for: Accelerated Cycle Life Testing and Capacity Degradation Modeling of LiCoO2-graphite Cells*. 2021. URL: `http://dx.doi.org/10.17632/c35zbmn7j8.1`.

[53] Weiping Diao, Ijaz Haider Naqvi and Michael Pecht. 'Early detection of anomalous degradation behavior in lithium-ion batteries'. In: *Journal of Energy Storage* 32 (2020), p. 101710.

[54] Weiping Diao, Saurabh Saxena and Michael Pecht. 'Accelerated cycle life testing and capacity degradation modeling of LiCoO2-graphite cells'. In: *Journal of Power Sources* 435 (2019), p. 226830.

[55] Weiping Diao et al. 'Algorithm to Determine the Knee Point on Capacity Fade Curves of Lithium-Ion Cells'. In: *Energies* 12 (July 2019), p. 2910.

[56] Yuan-Li Ding et al. 'Automotive Li-ion Batteries: Current Status and Future Perspectives'. In: *Electrochemical Energy Reviews* 2 (Mar. 2019), pp. 1–28. DOI: 10.1007/s41918-018-0022-z.

[57] DOE OE. *Homepage of DOE OE*. Aug. 2020. URL: https://www.sandia.gov/energystoragesafety-ssl/research-development/research-data-repository/.

[58] Goncalo Dos Reis et al. 'Lithium-ion battery data and where to find it'. In: *Energy and AI* 5 (2021), p. 100081.

[59] dryad. *Homepage of Dryad*. URL: https://datadryad.org/stash.

[60] Matthieu Dubarry. *Graphite//LFP synthetic training diagnosis dataset*. 2020. URL: https://data.mendeley.com/datasets/bs2j56pn7y/1.

[61] Matthieu Dubarry. *Graphite//LFP synthetic training prognosis dataset*. 2020. URL: https://data.mendeley.com/datasets/6s6ph9n8zg/1.

[62] Matthieu Dubarry and George Baure. 'Perspective on Commercial Li-ion Battery Testing, Best Practices for Simple and Effective Protocols'. In: *Electronics* 9.1 (Jan. 2020), p. 152. DOI: 10.3390/electronics9010152. URL: https://doi.org/10.3390/electronics9010152.

[63] Matthieu Dubarry, George Baure and Arnaud Devie. 'Durability and reliability of EV batteries under electric utility grid operations: path dependence of battery degradation'. In: *Journal of The Electrochemical Society* 165.5 (2018), A773–A783.

[64] Matthieu Dubarry and David Beck. 'Big data training data for artificial intelligence-based Li-ion diagnosis and prognosis'. In: *Journal of Power Sources* 479 (2020), p. 228806.

[65] Matthieu Dubarry et al. 'Identifying battery aging mechanisms in large format Li ion cells'. In: *Journal of Power Sources* 196.7 (Apr. 2011), pp. 3420–3425. DOI: 10.1016/j.jpowsour.2010.07.029. URL: https://doi.org/10.1016/j.jpowsour.2010.07.029.

[66] Ali Al-Dulaimi et al. 'Hybrid deep neural network model for remaining useful life estimation'. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3872–3876.

[67] Yuviny Echevarría, Cecilio Blanco and Luciano Sánchez. 'Learning human-understandable models for the health assessment of Li-ion batteries via multi-objective genetic programming'. In: *Engineering Applications of Artificial Intelligence* 86 (2019), pp. 1–10.

[68] Yuviny Echevarría Cartaya, Luciano Sanchez and Cecilio Blanco Viejo. *Li-ion Battery charge/discharge benchmark*. Jan. 2017. URL: `http://dx.doi.org/10.17632/r4n22f4jfk.1`.

[69] Madeleine Ecker et al. 'Calendar and cycle life study of Li(NiMnCo)O2-based 18650 lithium-ion batteries'. In: *Journal of Power Sources* 248 (2014), pp. 839–851.

[70] Madeleine Ecker et al. 'Development of a lifetime prediction model for lithium-ion batteries based on extended accelerated aging test data'. In: *Journal of Power Sources* 215 (Oct. 2012), pp. 248–257. DOI: `10.1016/j.jpowsour.2012.05.012`. URL: `https://doi.org/10.1016/j.jpowsour.2012.05.012`.

[71] Madeleine Ecker et al. 'Parameterisation of a Physico-Chemical Model of a Lithium-Ion Battery Part I: Determination of Parameters'. In: *J. Electrochem. Soc* 162.9 (2015), A1836–A1848.

[72] Madeleine Ecker et al. 'Parameterization of a physico-chemical model of a lithium-ion battery: II. Model validation'. In: *Journal of The Electrochemical Society* 162.9 (2015), A1849.

[73] Ömer Faruk Eker, Faith Camci and Ian K Jennions. 'Major challenges in prognostics: study on benchmarking prognostic datasets'. In: *Proceedings of the 1st European Conference of the Prognostics and Health Management Society. Dresden , Germany , 3-5 July 2012*. PHM Society, 2012, pp. 148–155.

[74] United States Environmental Protection Agency EPA. *Criteria Air Pollutants*. Nov. 2020. URL: `https://www.epa.gov/criteria-air-pollutants`.

[75] euhubs4data. *Homepage of euhubs4data*. URL: `https://euhubs4data.eu/datasets/`.

[76] European data portal. *A European Strategy for data*. URL: `https://digital-strategy.ec.europa.eu/en/policies/strategy-data`.

[77] *EVERLASTING project homepage*. 2021. URL: `https://everlasting-project.eu//`.

[78] Paula Fermín et al. 'Identification and machine learning prediction of knee-point and knee-onset in capacity degradation curves of lithium-ion cells'. In: *Energy and AI* (2020), p. 100006.

[79] Muhammad Sheikh Ferran Brosa Planella and W. Dhammika. *Systematic derivation and validation of a reduced thermal-electrochemical model for lithium-ion batteries using asymptotic methods*. 2020. URL: `https://github.com/brosaplanella/TEC-reduced-model`.

[80] Y Firouz et al. 'Lithium-ion capacitor–Characterization and development of new electrical model'. In: *Energy* 83 (2015), pp. 597–613.

[81] William L. Fredericks et al. ‘Performance Metrics Required of Next-Generation Batteries to Electrify Vertical Takeoff and Landing (VTOL) Aircraft’. In: *ACS Energy Letters* 3.12 (Nov. 2018), pp. 2989–2994. DOI: 10.1021/acsenergylett.8b02195. URL: https://doi.org/10.1021/acsenergylett.8b02195.

[82] Franziska Friedrich et al. ‘Capacity Fading Mechanisms of NCM-811 Cathodes in Lithium-Ion Batteries Studied by X-ray Diffraction and Other Diagnostics’. In: *Journal of the Electrochemical Society* 166 (2019), A3760–A3774.

[83] Yarin Gal and Zoubin Ghahramani. ‘Dropout as a Bayesian approximation: Representing model uncertainty in deep learning’. In: *international conference on machine learning*. 2016, pp. 1050–1059.

[84] Yang Gao et al. ‘Lithium-ion battery aging mechanisms and life model under different charging stresses’. In: *Journal of Power Sources* 356 (2017), pp. 103–114.

[85] James A Gilbert, Ilya A Shkrob and Daniel P Abraham. ‘Transition metal dissolution, ion migration, electrocatalytic reduction and capacity loss in lithium-ion full cells’. In: *Journal of The Electrochemical Society* 164.2 (2017), A389–A399.

[86] Giuseppe Giordano et al. ‘Model-based lithium-ion battery resistance estimation from electric vehicle operating data’. In: *IEEE Transactions on Vehicular Technology* 67.5 (2018), pp. 3720–3728.

[87] Bayesian Hilbert Transform GitHub. *Bayesian Hilbert Transform GitHub*. Feb. 2021. URL: https://github.com/ciuccislab/BHT.

[88] google. *Homepage of the google database*. URL: https://blog.google/products/search/discovering-millions-datasets-web/.

[89] Jaykanth Govindarajan. ‘Calendar ageing test results on commercial 18650 Li ion cell @ 10°C and 0°C’. In: (Apr. 2021). DOI: 10.4121/14377184.v1. URL: https://dx.doi.org/10.4121/14377184.v1.

[90] Jaykanth Govindarajan. *Lifecycle ageing tests on commercial 18650 Li ion cell @ 10°C and 0°C*. Apr. 2021. DOI: 10.4121/14377295. URL: https://dx.doi.org/10.4121/14377295.

[91] Arijit Guha and Amit Patra. ‘State of health estimation of lithium-ion batteries using capacity fade and internal resistance growth models’. In: *IEEE Transactions on Transportation Electrification* 4.1 (2017), pp. 135–146.

[92] Defne Gun, Hector Perez and Scott Moura. *Berkeley: eCAL fast charging test data*. Aug. 2015. URL: https://datadryad.org/stash/dataset/doi:10.6078/D1MS3X.

[93] Xuebing Han et al. ‘Cycle Life of Commercial Lithium-Ion Batteries with Lithium Titanium Oxide Anodes in Electric Vehicles’. In: *Energies* 7 (Aug. 2014), pp. 4895–4909.

[94] Wei He et al. 'Prognostics of lithium-ion batteries based on Dempster–Shafer theory and the Bayesian Monte Carlo method'. In: *Journal of Power Sources* 196.23 (2011), pp. 10314–10321.

[95] Wei He et al. 'State of charge estimation for Li-ion batteries using neural network modeling and unscented Kalman filter-based error cancellation'. In: *International Journal of Electrical Power & Energy Systems* 62 (2014), pp. 783–791.

[96] Thomas Heenan et al. 'Lithium-ion Battery INR18650 MJ1 Data: 400 Electrochemical Cycles (EIL-015)'. In: (May 2020). DOI: `10.5522/04/12159462.v1`. URL: `https://dx.doi.org/10.5522/04/12159462.v1`.

[97] TMM Heenan et al. 'An Advanced Microstructural and Electrochemical Datasheet on 18650 Li-ion Batteries with Nickel-Rich NMC811 Cathodes and Graphite-Silicon Anodes'. In: *Journal of the Electrochemical Society* 167.14 (2020), p. 140530.

[98] Christopher Hendricks et al. 'A failure modes, mechanisms, and effects analysis (FMMEA) of lithium-ion batteries'. In: *Journal of Power Sources* 297 (2015), pp. 113–120.

[99] Aniruddh Herle, Janamejaya Channegowda and Dinakar Prabhu. 'Overcoming limited battery data challenges: A coupled neural network approach'. In: *International Journal of Energy Research* 45.14 (2021), pp. 20474–20482.

[100] Patrick Herring et al. 'BEEP: A Python library for Battery Evaluation and Early Prediction'. In: *SoftwareX* 11 (2020), p. 100506.

[101] Edward F Hogge et al. 'Verification of a remaining flying time prediction system for small electric aircraft'. In: *Annual conference of the phm society (Vol.7, No.1*. 2015.

[102] *Homepage of 4TU.ResearchData Repository*. URL: `https://data.4tu.nl/`.

[103] Joonki Hong et al. 'Towards the swift prediction of the remaining useful life of lithium-ion batteries with end-to-end deep learning'. In: *Applied Energy* 278 (2020), p. 115646.

[104] David Howey. *Oxford battery team Data and code*. 2011. URL: `http://howey.eng.ox.ac.uk/data-and-code/`.

[105] David A Howey et al. 'Free Radicals: Making a Case for Battery Modeling'. In: *The Electrochemical Society Interface* 29.4 (2020), p. 30.

[106] Chao Hu et al. 'Remaining useful life assessment of lithium-ion batteries in implantable medical devices'. In: *Journal of Power Sources* 375 (2018), pp. 118–130.

[107] Rasheed Ibraheem, Calum Strange and Gonçalo dos Reis. 'Capacity and Internal Resistance of lithium-ion batteries: Full degradation curve prediction from Voltage response at constant Current at discharge'. In: *Journal of Power Sources* 556 (2023), p. 232477. ISSN: 0378-7753. DOI: `https://doi.org/10.1016/j.jpowsour.2022.232477`. URL: `https://www.sciencedirect.com/science/article/pii/S0378775322014549`.

[108] impedance.py. *impedance.py GithHub repository*. Aug. 2020. URL: `https://github.com/ECSHackWeek/impedance.py`.

[109] Eric Jones, Travis Oliphant, Pearu Peterson et al. *SciPy: Open source scientific tools for Python*. 2001. URL: `http://www.scipy.org/`.

[110] Dominik Jöst et al. *Timeseries data of a drive cycle aging test of 28 high energy NCA/C+Si round cells of type 18650*. 2021. DOI: `10.18154/RWTH-2021-02814`. URL: `https://publications.rwth-aachen.de/record/815749`.

[111] Peter Keil et al. 'Calendar aging of lithium-ion batteries'. In: *Journal of The Electrochemical Society* 163.9 (2016), A1872.

[112] Jungsoo Kim et al. 'Estimation of Li-ion Battery State of Health based on Multilayer Perceptron: as an EV Application'. In: *IFAC-PapersOnLine* 51.28 (2018), pp. 392–397.

[113] Ron Kohavi et al. 'A study of cross-validation and bootstrap for accuracy estimation and model selection'. In: *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*. Vol. 2. Montreal, Canada. Aug. 1995, pp. 1137–1143.

[114] P Kollmeyer et al. 'LG 18650HG2 Li-ion Battery Data and Example Deep Neural Network xEV SOC Estimator Script'. In: *Mendeley Data* 3 (2020). URL: `http://dx.doi.org/10.17632/cp3473x7xv.3`.

[115] P. Kollmeyer, A. Hackl and A. Emadi. 'Li-ion battery model performance for automotive drive cycles with current pulse and EIS parameterization'. In: *2017 IEEE Transportation Electrification Conference and Expo (ITEC)*. 2017, pp. 486–492. DOI: `10.1109/ITEC.2017.7993319`.

[116] Phillip Kollmeyer. *Panasonic 18650PF Li-ion Battery Data*. June 2018. URL: `http://dx.doi.org/10.17632/wykht8y7tg.1`.

[117] C. Kulkarni and A. Guarneros. 'Small Satellite Power Simulation Data Set'. In: *NASA Ames Prognostics Research Center,* (2015).

[118] C. Kulkarni et al. 'HIRF Battery Data Set'. In: *NASA Ames Prognostics Research Center,* (2015).

[119] Meinert Lewerenz et al. 'Differential voltage analysis as a tool for analyzing inhomogeneous aging: A case study for LiFePO4|Graphite cylindrical cells'. In: *Journal of Power Sources* 368 (Nov. 2017), pp. 57–67. DOI: `10.1016/j.jpowsour.2017.09.059`. URL: `https://doi.org/10.1016/j.jpowsour.2017.09.059`.

[120] Jing Li et al. 'Study of the failure mechanisms of LiNi0. 8Mn0. 1Co0. 1O2 cathode material for lithium ion batteries'. In: *Journal of The Electrochemical Society* 162.7 (2015), A1401–A1408.

[121] Weihan Li et al. 'Digital twin for battery systems: Cloud battery management system with online state-of-charge and state-of-health estimation'. In: *Journal of Energy Storage* 30 (2020), p. 101557.

[122] Weihan Li et al. ‘One-shot battery degradation trajectory prediction with deep learning’. In: *Journal of Power Sources* 506 (Sept. 2021), p. 230024. DOI: `10.1016/j.jpowsour.2021.230024`. URL: `https://doi.org/10.1016/j.jpowsour.2021.230024`.

[123] Kaizhi Liang et al. ‘Data-driven Ohmic resistance estimation of battery packs for electric vehicles’. In: *Energies* 12.24 (2019), p. 4772.

[124] Datong Liu et al. ‘Data-driven prognostics for lithium-ion battery based on Gaussian Process Regression’. In: *Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing)*. IEEE. 2012, pp. 1–5.

[125] Jiapeng Liu, Ting Hei Wan and Francesco Ciucci. ‘A Bayesian view on the Hilbert transform and the Kramers-Kronig transform of electrochemical impedance data: Probabilistic estimates and quality scores’. In: *Electrochimica Acta* 357 (2020), p. 136864.

[126] Kailong Liu et al. ‘An evaluation study of different modelling techniques for calendar ageing prediction of lithium-ion batteries’. In: *Renewable and Sustainable Energy Reviews* 131 (2020), p. 110017.

[127] Kailong Liu et al. ‘Charging pattern optimization for lithium-ion batteries with an electrothermal-aging model’. In: *IEEE Transactions on Industrial Informatics* 14.12 (2018), pp. 5463–5474.

[128] Kailong Liu et al. ‘Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification’. In: *IEEE/ASME Transactions on Mechatronics* (2021).

[129] Kailong Liu et al. ‘Mass load prediction for lithium-ion battery electrode clean production: a machine learning approach’. In: *Journal of Cleaner Production* 289 (2021), p. 125159.

[130] Kailong Liu et al. ‘Modified Gaussian process regression models for cyclic capacity prediction of lithium-ion batteries’. In: *IEEE Transactions on Transportation Electrification* 5.4 (2019), pp. 1225–1236.

[131] Qianqian Liu et al. ‘Understanding undesirable anode lithium plating issues in lithium-ion batteries’. In: *RSC Advances* 6 (2016), pp. 88683–88700.

[132] Weilin Luo et al. ‘Study on impedance model of Li-ion battery’. In: *2011 6th IEEE Conference on Industrial Electronics and Applications*. IEEE. 2011, pp. 1943–1947.

[133] Massimiliano Luzi. *Automotive Li-ion Cell Usage Data Set*. 2018. DOI: `10.21227/ce9q-jr19`. URL: `http://dx.doi.org/10.21227/ce9q-jr19`.

[134] Yanying Ma et al. ‘The capacity estimation and cycle life prediction of lithium-ion batteries using a new broad extreme learning machine approach’. In: *Journal of Power Sources* 476 (2020), p. 228581.

[135] Leonardo KK Maia et al. ‘Expanding the lifetime of Li-ion batteries through optimization of charging profiles’. In: *Journal of Cleaner Production* 225 (2019), pp. 928–938.

[136] Mario Marinaro et al. 'Bringing forward the development of battery cells for automotive applications: Perspective of R&D activities in China, Japan, the EU and the USA'. In: *Journal of Power Sources* 459 (2020), p. 228073. ISSN: 0378-7753. DOI: https://doi.org/10.1016/j.jpowsour.2020.228073. URL: https://www.sciencedirect.com/science/article/pii/S0378775320303761.

[137] Tomoyuki Matsuda et al. 'Investigation of the influence of temperature on the degradation mechanism of commercial nickel manganese cobalt oxide-type lithium-ion cells during long-term cycle tests'. In: *Journal of Energy Storage* 21 (2019), pp. 665–671.

[138] Karthik S Mayilvahanan et al. 'Supervised learning of synthetic big data for Li-ion battery degradation diagnosis'. In: *Batteries & Supercaps* 5.1 (2022), e202100166.

[139] I. McLeod. *Sampling Distribution of the Mean and Standard Deviation in Various Populations*. Demonstration-WebPage-Link, Wolfram Demonstrations Project, published: March 7 2011, Accessed: 2010-09-30.

[140] mendeley. *Homepage of Mendeley*. URL: https://data.mendeley.com/.

[141] Christoph Nebl et al. *Data for: Prediction of Constant Power Delivery of Lithium-Ion Cells at High Loads*. Version V1. 2020. DOI: 10.17632/ptxpzt876r.1. URL: https://dx.doi.org/10.17632/ptxpzt876r.1.

[142] Christoph Nebl et al. 'Prediction of constant power delivery of lithium-ion cells at high loads'. In: *Journal of Energy Storage* 30 (2020), p. 101552.

[143] Jeremy Neubauer and Ahmad Pesaran. 'The ability of battery second use strategies to impact plug-in electric vehicle prices and serve utility energy storage applications'. In: *Lancet* 196 (Dec. 2011), pp. 10351–10358.

[144] Man-Fai Ng et al. 'Predicting the state of charge and health of batteries using data-driven machine learning'. In: *Nature Machine Intelligence* (2020), pp. 1–10.

[145] NREL. *Homepage of the National Renewable Energy Laboratory of the U.S. Department of Energy*. Aug. 2020. URL: https://www.nrel.gov/research/data-tools.html.

[146] Benjamín E Olivares et al. 'Particle-filtering-based prognosis framework for energy storage devices with a statistical characterization of state-of-health regeneration phenomena'. In: *IEEE Transactions on Instrumentation and Measurement* 62.2 (2012), pp. 364–376.

[147] *Online identification reference spreadsheet for 18650 Li-ion cells*. spreadsheet Link.

[148] Open Knowledge Foundation. *Homepage of Database Contents License (DbCL) v1.0*. URL: http://opendatacommons.org/licenses/dbcl/1.0/.

[149] Open Knowledge Foundation. *Homepage of Open Data Commons Open Database License (ODbL)*. URL: https://opendatacommons.org/licenses/odbl/.

[150] Haihong Pan et al. ‘Novel battery state-of-health online estimation method using multiple health indicators and an extreme learning machine’. In: *Energy* 160 (2018), pp. 466–477.

[151] Yue Pan et al. ‘Internal short circuit detection for lithium-ion battery pack with parallel-series hybrid connections’. In: *Journal of Cleaner Production* 255 (2020), p. 120277.

[152] Kang-Joon Park et al. ‘Degradation Mechanism of Ni-Enriched NCA Cathode for Lithium Batteries: Are Microcracks Really Critical?’ In: *ACS Energy Letters* 4 (2019), pp. 1394–1400.

[153] Saehong Park et al. ‘Optimal experimental design for parameterization of an electrochemical lithium-ion battery model’. In: *Journal of The Electrochemical Society* 165.7 (2018), A1309.

[154] Carlos Pastor-Fernandez. *Data for A comparison between electrochemical impedance spectroscopy and incremental capacity-differential voltage as li-ion diagnostic techniques to identify and quantify the effects of degradation modes within battery management systems*. Aug. 2016. URL: `http://wrap.warwick.ac.uk/87247/`.

[155] C Pastor-Fernández et al. ‘A study of cell-to-cell interactions and degradation in parallel strings: implications for the battery management system’. In: *Journal of Power Sources* 329 (2016), pp. 574–585.

[156] Carlos Pastor-Fernández et al. ‘A comparison between electrochemical impedance spectroscopy and incremental capacity-differential voltage as Li-ion diagnostic techniques to identify and quantify the effects of degradation modes within battery management systems’. In: *Journal of Power Sources* 360 (2017), pp. 301–318.

[157] F. Pedregosa et al. ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[158] Pucheng Pei et al. ‘Capacity estimation for lithium-ion battery using experimental feature interval approach’. In: *Energy* (2020), p. 117778.

[159] José Affonso Moreira Penna, Cairo Lúcio Nascimento and Leonardo Ramos Rodrigues. ‘Health monitoring and remaining useful life estimation of lithium-ion aeronautical batteries’. In: *2012 IEEE Aerospace Conference*. IEEE. 2012, pp. 1–12.

[160] Ferran Brosa Planella, Muhammad Sheikh and W Dhammika Widanage. ‘Systematic derivation and validation of a reduced thermal-electrochemical model for lithium-ion batteries using asymptotic methods’. In: *arXiv preprint arXiv:2011.01611* (2020).

[161] Yuliya Preger et al. ‘Degradation of Commercial Lithium-Ion Cells as a Function of Chemistry and Cycling Conditions’. In: *Journal of The Electrochemical Society* 167.12 (2020), p. 120532.

[162] Wenzel Prochazka, Gudrun Pregartner and Martin Cifrain. 'Design-of-Experiment and Statistical Modeling of a Large Scale Aging Experiment for Two Popular Lithium Ion Cell Chemistries'. In: *Journal of The Electrochemical Society* 160.8 (2013), A1039–A1051. DOI: 10.1149/2.003308jes. URL: https://doi.org/10.1149/2.003308jes.

[163] 'Prognostics center of excellence - data repository'. In: *NASA Ames Prognostics Research Center,* (). URL: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository.

[164] Moinak Pyne, Benjamin J Yurkovich and Stephen Yurkovich. 'Capacity fade estimation using supervised learning'. In: *2017 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE. 2017, pp. 872–878.

[165] Moinak Pyne, Benjamin J Yurkovich and Stephen Yurkovich. 'Generation of Synthetic Battery Data with Capacity Variation'. In: *2019 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE. 2019, pp. 476–480.

[166] Taichun Qin, Shengkui Zeng and Jianbin Guo. 'Robust prognostics for state of health estimation of lithium-ion batteries based on an improved PSO–SVR model'. In: *Microelectronics Reliability* 55.9-10 (2015), pp. 1280–1284.

[167] Yan Qin, Stefan Adams and Chau Yuen. 'A Transfer Learning-based State of Charge Estimation for Lithium-Ion Battery at Varying Ambient Temperatures'. In: *arXiv preprint arXiv:2101.03704* (2021).

[168] T Raj. *Path Dependent Battery Degradation Dataset Part 1*. DOI: 10.5287/bodleian:v0ervBv6p. URL: https://ora.ox.ac.uk/objects/uuid:de62b5d2-6154-426d-bcbb-30253ddb7d1e.

[169] T Raj. *Path Dependent Battery Degradation Dataset Part 2*. 2021. DOI: 10.5287/bodleian:2zvyknyRg. URL: https://ora.ox.ac.uk/objects/uuid:be3d304e-51fd-4b37-a818-b6fa1ac2ba9d.

[170] T Raj. *Path Dependent Battery Degradation Dataset Part 3*. 2021. DOI: 10.5287/bodleian:j1a2eD7ow. URL: https://ora.ox.ac.uk/objects/uuid:78f66fa8-deb9-468a-86f3-63983a7391a9.

[171] Trishna Raj et al. 'Investigation of Path-Dependent Degradation in Lithium-Ion Batteries'. In: *Batteries & Supercaps* 3.12 (2020), pp. 1377–1385.

[172] Jürgen Remmlinger et al. 'State-of-health monitoring of lithium-ion batteries in electric vehicles by on-board internal resistance estimation'. In: *Journal of Power Sources* 196.12 (2011), pp. 5357–5363.

[173] Lei Ren et al. 'Remaining useful life prediction for lithium-ion battery: A deep learning approach'. In: *IEEE Access* 6 (2018), pp. 50587–50598.

[174] J M Reniers, G Mulder and D A Howey. *Oxford energy trading battery degradation dataset*. 2020. DOI: 10.5287/bodleian:gJPdDzvP4. URL: https://doi.org/10.5287/bodleian:gJPdDzvP4.

[175] Jorn M Reniers, Grietus Mulder and David A Howey. ‘Unlocking extra value from grid batteries using advanced models’. In: *Journal of Power Sources* 487 (2021), p. 229355.

[176] Jorn M Reniers et al. ‘Improving optimal control of grid-connected lithium-ion batteries through more accurate battery and degradation modelling’. In: *Journal of Power Sources* 379 (2018), pp. 91–102.

[177] Jorn M. Reniers. ‘Degradation-aware optimal control of grid-connected lithium-ion batteries’. PhD thesis. University of Oxford, 2019. URL: `https://ora.ox.ac.uk/objects/uuid:e0a33cb5-93f5-4e34-9b17-996a9d40755b`.

[178] Robert R Richardson, Michael A Osborne and David A Howey. ‘Battery health prediction under generalized conditions using a Gaussian process transition model’. In: *Journal of Energy Storage* 23 (2019), pp. 320–328.

[179] Robert R Richardson, Michael A Osborne and David A Howey. ‘Gaussian process regression for forecasting battery state of health’. In: *Journal of Power Sources* 357 (2017), pp. 209–219.

[180] Marco-Tulio F. Rodrigues et al. ‘Fast Charging of Li-Ion Cells: Part I. Using Li/Cu Reference Electrodes to Probe Individual Electrode Potentials’. In: *Journal of the Electrochemical Society* 166 (2019), A996–A1003.

[181] Martin Rogall et al. ‘DREMUS: A Data-Restricted Multi-Physics Simulation Model for Lithium-Ion Battery Storage’. In: *Journal of Energy Storage* 32 (2020), p. 102051.

[182] Ankit Rohatgi. *Webplotdigitizer: Version 4.4*. 2020. URL: `https://automeris.io/WebPlotDigitizer`.

[183] Darius Roman et al. ‘Machine learning pipeline for battery state of health estimation’. In: *Nature Machine Intelligence* (2021). DOI: `10.1038/s42256-021-00312-3`.

[184] Cynthia Rudin. ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[185] B. Saha and K. Goebel. ‘Battery Data Set’. In: *NASA Ames Prognostics Research Center,* (2007).

[186] Bhaskar Saha and Kai Goebel. ‘Model adaptation for prognostics in a particle filtering framework’. In: *International Journal of Prognostics and Health Management* 2 (2011), p. 61.

[187] Bhaskar Saha and Kai Goebel. ‘Modeling Li-ion battery capacity depletion in a particle filtering framework’. In: *Proceedings of the annual conference of the prognostics and health management society*. San Diego, CA. 2009, pp. 2909–2924.

[188] Bhaskar Saha et al. ‘Prognostics methods for battery health monitoring using a Bayesian framework’. In: *IEEE Transactions on instrumentation and measurement* 58.2 (Mar. 2009), pp. 291–296.

[189] Sandia National Lab. *Data for Degradation of Commercial Lithium-ion Cells as a Function of Chemistry and Cycling Conditions*. 2020. URL: https://www.batteryarchive.org/snl_study.html.

[190] Shriram Santhanagopalan and Ralph E. White. 'Quantifying Cell-to-Cell Variations in Lithium Ion Batteries'. In: *International Journal of Electrochemistry* 2012 (2012), pp. 1–10. DOI: 10.1155/2012/395838. URL: https://doi.org/10.1155/2012/395838.

[191] Ville Satopaa et al. 'Finding a "kneedle" in a haystack: Detecting knee points in system behavior'. In: *2011 31st international conference on distributed computing systems workshops*. IEEE. 2011, pp. 166–171.

[192] Abhinav Saxena et al. 'Designing data-driven battery prognostic approaches for variable loading profiles: Some lessons learned'. In: *European Conference of Prognostics and Health Management Society*. 2012, pp. 72–732.

[193] Saurabh Saxena, Christopher Hendricks and Michael Pecht. 'Cycle life testing and modeling of graphite/LiCoO2 cells under different state of charge ranges'. In: *Journal of Power Sources* 327 (2016), pp. 394–400.

[194] Stefan Schindler et al. 'Fast charging of lithium-ion cells: Identification of aging-minimal current profiles using a design of experiment approach and a mechanistic degradation analysis'. In: *Journal of Energy Storage* 19 (Oct. 2018), pp. 364–378. DOI: 10.1016/j.est.2018.08.002. URL: https://doi.org/10.1016/j.est.2018.08.002.

[195] Johannes Schmalstieg et al. 'A holistic aging model for Li(NiMnCo)O2 based 18650 lithium-ion batteries'. In: *Journal of Power Sources* 257 (July 2014), pp. 325–334. DOI: 10.1016/j.jpowsour.2014.02.012. URL: https://doi.org/10.1016/j.jpowsour.2014.02.012.

[196] Simon F. Schuster et al. 'Nonlinear aging characteristics of lithium-ion cells under different operational conditions'. In: *Journal of Energy Storage* 1 (2015), pp. 44–53.

[197] Kristen Severson et al. 'Data-driven prediction of battery cycle life before capacity degradation'. In: *Nature Energy* 4 (May 2019), pp. 1–9.

[198] Yunlong Shang et al. 'A compact resonant switched-capacitor heater for lithium-ion battery self-heating at low temperatures'. In: *IEEE Transactions on Power Electronics* 35.7 (2019), pp. 7134–7144.

[199] Sheng Shen et al. 'A Hybrid Machine Learning Model for Battery Cycle Life Prediction with Early Cycle Data'. In: *2020 IEEE Transportation Electrification Conference & Expo (ITEC)*. IEEE. 2020, pp. 181–184.

[200] AJ Smith et al. 'Long-term low-rate cycling of LiCoO2/graphite Li-ion cells at 55°C'. In: *Journal of The Electrochemical Society* 159.6 (2012), A705.

[201] L. Somerville et al. 'The effect of charging rate on the graphite electrode of commercial lithium-ion cells: A post-mortem study'. In: *Journal of Power Sources* 335 (2016), pp. 189–196.

[202] Lingjun Song et al. 'Intelligent state of health estimation for lithium-ion battery pack based on big data analysis'. In: *Journal of Energy Storage* 32 (2020), p. 101836.

[203] Yuchen Song et al. 'A hybrid statistical data-driven method for on-line joint state estimation of lithium-ion batteries'. In: *Applied Energy* 261 (2020), p. 114408.

[204] Shashank Sripad and Venkatasubramanian Viswanathan. 'Performance Metrics Required of Next-Generation Batteries to Make a Practical Electric Semi Truck'. In: *ACS Energy Letters* 2.7 (June 2017), pp. 1669–1673. DOI: 10.1021/acsenergylett.7b00432. URL: https://doi.org/10.1021/acsenergylett.7b00432.

[205] Georg Steinbuß et al. 'FOBSS: Monitoring Data from a Modular Battery System'. In: *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. 2019, pp. 456–459.

[206] Georg Steinbuß et al. *Frequent Observations from a Battery System with Subunits*. 37.01.03; LK 01. 2019. DOI: 10.5445/IR/1000094469. URL: https://dx.doi.org/10.5445/IR/1000094469.

[207] Calum Strange, Rasheed Ibraheem and Gonçalo dos Reis. 'Online Lifetime Prediction for Lithium-Ion Batteries with Cycle-by-Cycle Updates, Variance Reduction, and Model Ensembling'. In: *Energies* 16.7 (2023), p. 3273.

[208] Calum Strange and Gonçalo dos Reis. 'Prediction of future capacity and internal resistance of Li-ion cells from one cycle of input data'. In: *Energy and AI* 5 (Sept. 2021), p. 100097. DOI: 10.1016/j.egyai.2021.100097. URL: https://doi.org/10.1016/j.egyai.2021.100097.

[209] Calum Strange et al. 'Automatic method for the estimation of li-ion degradation test sample sizes required to understand cell-to-cell variability'. In: *Energy and AI* (2022), p. 100174.

[210] Calum Strange et al. 'Elbows of Internal Resistance Rise Curves in Li-Ion Cells'. In: *Energies* 14.4 (2021). ISSN: 1996-1073. DOI: 10.3390/en14041206. URL: https://www.mdpi.com/1996-1073/14/4/1206.

[211] Calum Strange et al. *Synthetic IR data for the Attia et al. (2020) battery dataset*. English. https://datashare.is.ed.ac.uk/handle/10283/3798. Accessed: 2020-11-30. 2020. DOI: 10.7488/ds/2957.

[212] D. A. Howey T. Raj. *Path Dependent Battery Degradation Dataset Part 2*. 2020. URL: https://ora.ox.ac.uk/objects/uuid:be3d304e-51fd-4b37-a818-b6fa1ac2ba9d.

[213] Liang Tang et al. 'An integrated health and contingency management case study on an autonomous ground robot'. In: *2011 9th IEEE international conference on control and automation (ICCA)*. IEEE. 2011, pp. 584–589.

[214] Xiaopeng Tang et al. 'Model migration neural network for predicting battery aging trajectories'. In: *IEEE Transactions on Transportation Electrification* (2020).

[215] A. T. Thorgeirsson et al. 'Probabilistic Prediction of Energy Demand and Driving Range for Electric Vehicles With Federated Learning'. In: *IEEE Open Journal of Vehicular Technology* 2 (2021), pp. 151–161. DOI: 10.1109/OJVT.2021.3065529.

[216] Toyota Research Institute. *Experimental data platform*. 2021. URL: https://data.matr.io/1/.

[217] Khiem Trad. 'Calendar ageing test results on commercial 18650 Li ion cell @ 25°C and 45°C'. In: (Mar. 2021). DOI: 10.4121/13804304.v1. URL: https://dx.doi.org/10.4121/13804304.v1.

[218] Khiem Trad. *D2.3 - Report containing aging test profiles and test results*. Tech. rep. EVERLASTING, 2020. URL: https://everlasting-project.eu/wp-content/uploads/2020/03/EVERLASTING_D2.3_final_20200228.pdf.

[219] Khiem Trad. *Lifecycle ageing tests on commercial 18650 Li ion cell @ $25°C$ and $45°C$*. Mar. 2021. DOI: 10.4121/13739296.v1. URL: http://dx.doi.org/10.4121/13739296.v1.

[220] Kuo-Hsin Tseng et al. 'Regression models using fully discharged voltage and internal resistance for state of health estimation of lithium-ion batteries'. In: *energies* 8.4 (2015), pp. 2889–2907.

[221] M Uitz et al. 'Aging of tesla's 18650 lithium-ion cells: Correlating solid-electrolyte-interphase evolution with fading in capacity and power'. In: *Journal of The Electrochemical Society* 164.14 (2017), A3503–A3510.

[222] *V. Sulzer's online spreadsheet of battery datasets*. URL: https://docs.google.com/spreadsheets/d/183uKKd0JTV46tGFsfvM-0etvHHSEL1L26Cetm6bJDDw.

[223] Pooja Vadhva et al. 'Electrochemical Impedance Spectroscopy for All-Solid-State Batteries: Theory, Methods and Future Outlook'. In: *ChemElectroChem* n/a.n/a (Apr. 2021). DOI: https://doi.org/10.1002/celc.202100108. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/celc.202100108.

[224] C. Vidal et al. 'Li-ion Battery State of Charge Estimation Using Long Short-Term Memory Recurrent Neural Network with Transfer Learning'. In: *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*. 2019, pp. 1–6. DOI: 10.1109/ITEC.2019.8790543.

[225] Carlos Vidal et al. 'Robust xEV Battery State-of-Charge Estimator Design Using a Feedforward Deep Neural Network'. In: *SAE Int. J. Adv. & Curr. Prac. in Mobility* 2 (Apr. 2020), pp. 2872–2880. DOI: 10.4271/2020-01-1181. URL: https://doi.org/10.4271/2020-01-1181.

[226] Wladislaw Waag, Christian Fleischer and Dirk Uwe Sauer. 'Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles'. In: *Journal of Power Sources* 258 (2014), pp. 321–339.

[227] Thomas Waldmann, Björn-Ingo Hogg and Margret Wohlfahrt-Mehrens. 'Li plating as unwanted side reaction in commercial Li-ion cells – A review'. In: *Journal of Power Sources* 384 (Apr. 2018), pp. 107–124. DOI: `10.1016/j.jpowsour.2018.02.063`. URL: `https://doi.org/10.1016/j.jpowsour.2018.02.063`.

[228] Thomas Waldmann et al. 'Temperature dependent ageing mechanisms in lithium-ion batteries – A Post-Mortem study'. In: *Journal of Power Sources* 262 (2014), pp. 129–135.

[229] Yujie Wang et al. 'Experimental data of lithium-ion battery and ultracapacitor under DST and UDDS profiles at room temperature'. In: *Data in Brief* 12 (2017), pp. 161–163. ISSN: 2352-3409. DOI: `https://doi.org/10.1016/j.dib.2017.01.019`. URL: `http://www.sciencedirect.com/science/article/pii/S2352340917300197`.

[230] Yujie Wang et al. 'Modeling and state-of-charge prediction of lithium-ion battery and ultracapacitor hybrids with a co-estimator'. In: *Energy* 121 (2017), pp. 739–750. ISSN: 0360-5442. DOI: `https://doi.org/10.1016/j.energy.2017.01.044`. URL: `https://www.sciencedirect.com/science/article/pii/S0360544217300452`.

[231] Logan Ward et al. 'Principles of the Battery Data Genome'. In: *arXiv preprint arXiv:2109.07278* (2021).

[232] Nick Williard et al. 'Comparative analysis of features for determining state of health in lithium-ion batteries'. In: *Int. J. Progn. Health Manag* 4.1 (2013), pp. 14–20.

[233] Billy Wu et al. 'Battery digital twins: Perspectives on the fusion of models, data and artificial intelligence for smart battery management systems'. In: *Energy and AI* (2020), p. 100016.

[234] Gao Xiangyang, Zhang Jun and Ning Ning. 'Transient behavior modeling and physical meaning analysis for Battery'. In: *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*. Vol. 2. IEEE. 2010, pp. V2–383.

[235] Yinjiao Xing et al. 'An ensemble model for predicting the remaining useful performance of lithium-ion batteries'. In: *Microelectronics Reliability* 53.6 (2013), pp. 811–820.

[236] Yinjiao Xing et al. 'State of charge estimation of lithium-ion batteries using the open-circuit voltage at various ambient temperatures'. In: *Applied Energy* 113 (2014), pp. 106–115.

[237] Fangfang Yang et al. 'Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model'. In: *Journal of Power Sources* 476 (2020), p. 228654.

[238] Ruixin Yang et al. ‘A fractional-order model-based battery external short circuit fault diagnosis approach for all-climate electric vehicles application’. In: *Journal of cleaner production* 187 (2018), pp. 950–959.

[239] Xiao-Guang Yang et al. ‘Fast charging of lithium-ion batteries at all temperatures’. In: *PNAS* 115 (2018), pp. 7266–7271.

[240] zenodo. *Homepage of Zenodo*. URL: `https://zenodo.org/`.

[241] Caiping Zhang et al. ‘Accelerated fading recognition for lithium-ion batteries with Nickel-Cobalt-Manganese cathode using quantile regression method’. In: *Applied Energy* 256 (2019), p. 113841.

[242] Jiucai Zhang and Xiaoli Zhang. ‘A Novel Internal Resistance Curve Based State of Health Method to Estimate Battery Capacity Fade and Resistance Rise’. In: *2020 IEEE Transportation Electrification Conference & Expo (ITEC)*. IEEE. 2020, pp. 575–578.

[243] Shuzhi Zhang. *Data for: A data-driven Coulomb counting method for state of charge calibration and estimation of lithium-ion battery, Version 1*. June 2018. URL: `https://data.mendeley.com/datasets/c5dxwn6w92/1`.

[244] Shuzhi Zhang et al. ‘A data-driven Coulomb counting method for state of charge calibration and estimation of lithium-ion battery’. In: *Sustainable Energy Technologies and Assessments* 40 (2020), p. 100752.

[245] Shuzhi Zhang et al. ‘A rapid online calculation method for state of health of lithium-ion battery based on coulomb counting method and differential voltage analysis’. In: *Journal of Power Sources* 479 (2020), p. 228740. ISSN: 0378-7753. DOI: `https://doi.org/10.1016/j.jpowsour.2020.228740`. URL: `https://doi.org/10.1016/j.jpowsour.2020.228740`.

[246] Yunwei Zhang et al. ‘Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning’. In: *Nature communications* 11.1 (2020), pp. 1–6.

[247] R. Zhao et al. ‘A Compact Methodology Via a Recurrent Neural Network for Accurate Equivalent Circuit Type Modeling of Lithium-Ion Batteries’. In: *IEEE Transactions on Industry Applications* 55.2 (2019), pp. 1922–1931. DOI: `10.1109/TIA.2018.2874588`.

[248] Fangdan Zheng et al. ‘Influence of different open circuit voltage tests on state of charge online estimation for lithium-ion batteries’. In: *Applied energy* 183 (2016), pp. 513–525.

[249] Jianbao Zhou et al. ‘Dynamic battery remaining useful life estimation: An on-line data-driven approach’. In: *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE. 2012, pp. 2196–2199.

[250] Z. Zhou et al. 'An efficient screening method for retired lithium-ion batteries based on support vector machine'. In: *Journal of Cleaner Production* 267 (2020). DOI: 10.1016/j.jclepro.2020.121882.